Linking a domain thesaurus to WordNet and conversion to WordNet-LMF

Antonio Toral[*] **Monica Monachini**[*] **Claudia Soria**[*] **Montse Cuadros**[†]
German Rigau[◇] **Wauter Bosma**[∧] **Piek Vossen**[∧]
[*]Istituto di Linguistica Computazionale, CNR. Pisa, Italy
[†]TALP Research Center, UPC. Barcelona, Spain
[◇]Informatika Fakultatea, EHU. Donostia, Spain
[∧]Faculteit der Letteren, VUA. Amsterdam, The Netherlands

Abstract

We present a methodology to link domain thesauri to general-domain lexica. This is applied in the framework of the KYOTO project to link the Species2000 thesaurus to the synsets of the English WordNet. Moreover, we study the formalisation of this thesaurus according to the ISO LMF standard and its dialect WordNet-LMF. This conversion will allow Species2000 to communicate with the other resources available in the KYOTO architecture.

1 Introduction

The goal of the KYOTO project[1] is the construction of a system for facilitating the exchange of information across cultures, domains and languages. This system is expected to allow people in communities to define the meaning of their words and terms in a shared Wiki platform so that it becomes anchored across languages and cultures but also so that a computer can use this knowledge to detect knowledge and facts in text. Whereas the current Wikipedia uses free text to share knowledge, KYOTO will represent this knowledge so that a computer can understand it. The system is being developed for the domain of environment. For example, the notion of environmental footprint will become defined in the same way in all these languages but also in such a way that the computer knows what information is necessary to calculate a footprint. With these definitions it will be possible to find information on footprints in documents, websites and reports so that users can directly ask the computer for actual information in their environment.

This endeavour presupposes the sharing of lexical and knowledge bases, both general and domain-related, under the form of lexical repositories and ontologies. The lexical resources that will be integrated in KYOTO are wordnets for the English, Dutch, Italian, Basque, Spanish, Chinese and Japanese languages. Special-domain wordnets and ontology will be developed: they are to be seen as a plug-in extension of the generic wordnet and ontology. These extensions contribute to the development of the Global Wordnet Grid[2].

As in KYOTO the integration of resources is viewed as a need, the use of formats that facilitates interoperability is essential. Interoperability allows an easier integration among general domain lexicons sharing the same structure (i.e other wordnets) and domain lexicons, but, more importantly, eases the integration of resources with different theoretical and implementation approaches, such as the ones being used within the project: Web 2.0 sources (DbPedia), species taxonomies (Species2000) and ontologies (DOLCE, SUMO, SIMPLE). There is no means to speak about interoperability if not paired with standards: they are bound to be the communicative channel by means of which diverse data, resources, formats, and models can interact on a common ground, in a controlled way.

In this paper we present a methodology to map the entries of domain thesauri to general-domain lexica. Specifically, we link the Species2000[3] thesaurus to the synsets of the English Princeton WordNet version 3.0 (PWN) (Fellbaum, 1998). Moreover, we study the formalisation of this thesaurus according to a standard. This conversion will allow Species2000 to communicate with the other resources available in the KYOTO architecture; Species2000 has been linked to PWN, whilst both the ontology and the wordnets for the rest of the languages of the project are also connected to the latter.

[1] `http://www.kyoto-project.eu`

[2] `http://www.globalwordnet.org`

[3] `http://www.sp2000.org`

The rest of the paper is organised as follows. The next section discusses the standard LMF and its dialect WN-LMF. After that, we report on the automatic mapping of the Species2000 thesaurus to PWN and the conversion of the resulting linked thesaurus to WN-LMF. Finally, we draw some general conclusions and sketch future work directions.

2 Standardisation

2.1 LMF

The Lexical Markup Framework (LMF) (ISO 24613, 2008) is an ISO standard for the representation of lexical resources. LMF has been chosen as representation format because it gathers experiences and harmonisation efforts started by the interested community in the '90s. This format for lexical resource representation has now reached a high level of sophistication, theoretical consensus, and official international standard status. Being ratified as an ISO standard LMF was specifically designed to accommodate as many models of lexical representations as possible. Purposefully, it is designed as a meta-model, i.e. a high-level specification for lexical resources defining the structural constraints of a lexicon.

Before being issued as an official ISO standard, LMF has passed a range of officially needed stages and has been extensively discussed and commented in a wide community comprising both academia and industry. LMF is thus mature enough to be taken as "the" choice when coming to selecting a standardised format for the representation and encoding of computational lexicons. Time is ripe now to start assessing LMF, providing the community with real examples of use.

2.2 WN-LMF

Wordnet-LMF (WN-LMF) is an LMF dialect tailored to the encoding of lexical resources adhering to the PWN model of lexical knowledge representation. No real attempt has been made so far to fully apply LMF to wordnet-like lexicons: WN-LMF is an example of the practical use of LMF in a real-world application (Soria et al., 2009). The KYOTO project represents an ideal test case for this format: going beyond the level of toy examples it allows to make a crash test, as the various resources need to be fully integrated. This will put us in the position to both have a preview on any problems we might encounter and make the LMF standard easy to adopt.

WN-LMF fully complies with the standard LMF as for its general framework. It builds on the representational devices made available by LMF and tailors them to the specific content requirements of the PWN model of lexical knowledge representation. The LMF library provides the hierarchy of lexical objects with structural relations among them. The Data Category (DC) library provides the elementary descriptors to be used in combination with the structural elements, necessary to represent lexical information (Francopoulo et al., 2006). Figure 1 shows a general diagram of WN-LMF.

2.2.1 WN-LMF overall design

The main conceptual components of PWN-like lexicons that need to be represented in LMF are the following:

- Synsets, variants and synset relations, including information about synset identifiers and sense-keys;
- Domain attribution, linking to ontologies, administrative information;
- Interlingual information, i.e. mapping of synsets in a given language to Interlingual Index (ILI).

The LMF semantic package naturally lends itself to the representation of wordnet-like resources, since it already contains lexical objects devised for the representation of synsets, their associated gloss and examples, variants, and synset relations.

Expression of PWN-related types of information (such as synset relations, external sources linked to wordnets) falls into the realm of LMF DCs, which are by definition either selectable from the predefined standard registry or custom-defined. The WN-LMF format, accordingly, has defined a DC Selection, necessary to fully represent the various wordnets to be integrated in KYOTO. Examples of custom DCs are values for describing synset relations, inter-lingual relations, for identifying external resources and their associated nodes, etc. For the sake of better parsing efficiency, in WN-LMF, DCs are represented by means of XML attributes and values instead of nested lexical objects. Consider the following sample of LMF code:

```
<Lemma>
  <feat att="partOfSpeech" val="n"/>
```

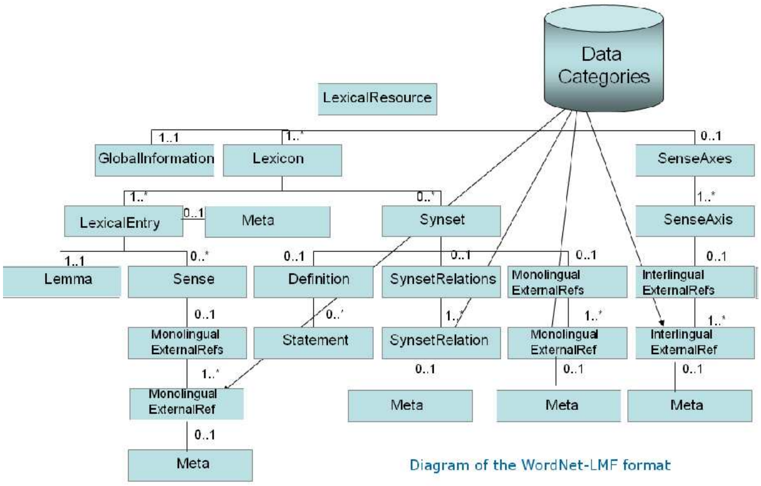

Figure 1: WN-LMF diagram

```
  <feat att="writtenForm" val="abbey"/>
</Lemma>
```

and its equivalent in WN-LMF:

```
<Lemma partOfSpeech="n" writtenForm="abbey"/>
```

By explicitly naming the attributes, we also make a stronger claim about the features and properties of the structure of a wordnet. This will enforce better compatibility and interoperability across the many wordnets for different languages that are available. In this respect, the WN-LMF DTD implementation has to be seen as a dialectal variant of the LMF DTD. Motivation behind this choice is to reach efficiency, while keeping adherence to standards.

2.2.2 The WN-LMF core component

The WN-LMF core package component provides the structural skeleton to represent the basic hierarchies of the lexicon.

KYOTO wordnets are represented as a grid of lexicons: *LexicalResource* is the container for all of them. A specific set of lexical objects is devoted to record general information about the lexical resource.

The lexical resource, besides the monolingual lexicons, contains the interlingual correspondences which are grouped in a section *SenseAxes* which is separated from the lexicons proper and contains inter-lexicon correspondences only.

Lexicon contains a monolingual resource, instantiated as a set of *LexicalEntry* instances. This element is a container for representing a lexeme in a lexicon. A *LexicalEntry* element contains the basic building blocks: lemma and senses. *Lemma* represents a word form chosen by convention to designate the lexical entry, whereas *Sense* represents one meaning of a lexical entry. For wordnet representation, this triplet is used to represent the variant(s), or literal(s) of a synset.

MonolingualExternalRef represents linking between a *Sense* or *Synset* and another resource, be it a knowledge organisation system, a database, or another lexical resource. Mapping among different versions of the same resource, reference to external information, such as mapping onto entries of another lexical database and or referencing additional sources can be dealt with by the *MonolingualExternalRef* object.

When linked to a *Sense* element, it can be used to express a mapping between the sense and its correspondent in another lexical resource (such as in Cornetto (Maks et al., 2008)[4]). In the particular case of the representation of PWN, *MonolingualExternalRef* serves as a representational device to express the Sense Key. When linked to the *Synset* element, then *MonolingualExternalRef* allows to encode reference to the domain and/or one or more links to an ontological system.

2.2.3 The WN-LMF semantic component

The Semantic component is in charge of describing information about a wordnet synset by means of the *Synset* element. A *Synset* clusters senses of different *LexicalEntry* instances within the same part of speech. The element *Definition* allows to

[4] Cornetto is a Dutch lexical database that has combines a database with lexical units with a wordnet database. Lexical units are synonyms in synsets and for each lexical unit morpho-syntactic, pragmatic and combinatorial information is provided.

```
<LexicalResource>
 <GlobalInformation label="Wordnet entries using
  Kyoto-LMF"/>
 <Lexicon languageCoding="ISO 639-3"
  label="English Wordnet 3.0" language="eng"
  owner="Princeton" version="3.0">
  <LexicalEntry id="footprint">
   <Lemma writtenForm="footprint"
    partOfSpeech="n"/>
   <Sense id="footprint_1"
    synset="eng-30-06645039-n">
    <MonolingualExternalRefs>
     <MonolingualExternalRef
      externalSystem="Wordnet3.0"
      externalReference="footprint&#37;1:10:00::"/>
    </MonolingualExternalRefs>
   </Sense>
  </LexicalEntry>
  <LexicalEntry id="footmark">
   <Lemma writtenForm="footmark" partOfSpeech="n"/>
   <Sense id="footmark_1"
    synset="eng-30-06645039-n">
    <MonolingualExternalRefs>
     <MonolingualExternalRef
      externalSystem="Wordnet3.0"
      externalReference="footmark&#37;1:10:00::"/>
    </MonolingualExternalRefs>
   </Sense>
  </LexicalEntry>
  [...]
 <Lexicon>
<LexicalResource>
```

Figure 2: Example of the core component

represent the gloss associated with each synset. Relations between synsets are codified by means of *SynsetRelation* elements (represented by means of XML attributes), one per relation.

A set of harmonised KYOTO DCs has been defined. This is to be used in conjunction with the *SynsetRelation* elements for representing the various relations holding between synsets. This DC library, for the sake of coherence, is being maintained as a centralised repository. This option has been followed in order to enforce better compatibility and interoperability across the many monolingual wordnets.

MonolingualExternalRef, which is used to represent linking between the lexical resource and another resource, when linked to the *Synset* element, allows to encode reference to the domain and/or one or more links to an ontological system.

2.2.4 The WN-LMF multilingual component

The Multilingual notation component is used in KYOTO for expressing interlingual correspondences. This component is designed as an independent package in order not to overload the representation of monolingual lexicons. The model is based on the notion of "Axes" that link synsets pertaining to different languages. For the purposes of creating a grid of wordnets linked via Interlingual Index, the *SenseAxis* device is specifically suited to implement approaches based on an interlingual pivot. Any *SenseAxis* element groups together monolingual synsets that correspond one to another by means of a particular type of relation.

```
<Synset id="eng-30-06645039-n" baseConcept="1">
 <Definition gloss="mark of a foot or shoe on a
  surface">
  <Statement example="the police made casts of the
   footprints in the soft earth outside the window"/>
 </Definition>
 <SynsetRelations>
  <!-- (mark, print) -->
  <SynsetRelation target="eng-30-06798750-n"
   relType="has_hyperonym">
   <Meta author="AH" date="2008-07-01"
    source="Wordnet3.0" status="yes"
    confidenceScore="1.0"/>
  </SynsetRelation>
  <!-- (footprint, evidence) -->
  <SynsetRelation target="eng-30-06645266-n"
   relType="has_hyponym">
   <Meta author="AH2" date="2008-07-01"
    source="eng-Wordnet3.0" status="yes"
    confidenceScore="1.0" />
  </SynsetRelation>
 </SynsetRelations>
 <MonolingualExternalRefs>
  <MonolingualExternalRef externalSystem="Wordnet1.6"
   externalReference="eng-16-01234567-n"/>
  <MonolingualExternalRef externalSystem="SUMO"
   externalReference="superficialPart" relType="at"/>
 </MonolingualExternalRefs>
</Synset>
```

Figure 3: Example of the semantic component

The *SenseAxis* element is a means for grouping together synsets belonging to different monolingual wordnets that correspond one to another and share the same equivalence relation (e.g. a synonymy or near_synonymy relation) to a pivot synset, which by convention is an English one. This is a compact way of encoding correspondences among wordnets, avoiding to have several LanguageX-English single correspondences.

InterlingualExternalRef is used in WN-LMF to express a linking between a *SenseAxis* instance and an external system such as an ontology, and represents the means to anchor a multilingual group of synsets to an ontological node. Its intended use, thus, is to provide a representational device to link a group of synsets from different wordnets to the same ontological concept.

```
<SenseAxis id="sa_001" relType" val="eq_synonym">
 <Target ID="ita-16-00001251-n"/>
 <Target ID="spa-16-09688541-n"/>
 <Target ID="eng-30-13480848-n"/>
 <InterlingualExternalRef
  ExternalSystem="SUMO"
  ExternalReference="Combustion"
  relType="at"/>
 </InterlingualExternalRef>
</SenseAxis>
```

Figure 4: Example of the multilingual component

3 Automatically mapping Species2000 to WN synsets

3.1 The structure of Species2000

A domain specific thesaurus such as Species 2000, provides an important vocabulary that can be used to model the knowledge in the environment domain. It contains around two million species structured according to a biological taxonomy shown in figure 5.

Each concept has at least a Latin name and often many alternative labels in different languages. An example of a Latin hierarchy is shown in figure 6. Implicitly, each level of the hierarchy corresponds to a particular level of the biological classification.

To be able to exploit the data, we converted the Species2000 format to SKOS format and published it in Virtuoso. The taxonomic relations have been converted to skos:broader relations. To extend the language labels, we looked for the Latin name in DBPedia and collected all language labels for a matching record. The results are shown in Table 1.

Language	**Species 2000**	**DBPedia extension**
English	69,045	834,821
Spanish	1,731	358,499
Italian	17,552	215,511
Dutch	5,397	185,437
Chinese	58,774	83,756
Japanese	4,625	139,754
Total	157,124	1,817,778

Table 1: Language labels for Species 2000 concepts after alignment with DBPedia

The number of language labels increased from 157,124 to 1,817,778 labels. Note that a single concept can have many different synonymous labels. However, there are still many language gaps. That is, there are many Species 2000 concepts that only have a Latin name. Figure 7 shows an example of the SKOS entry corresponding to the subspecies ITS-207724, whose scientific Latin name is “Eleutherodactylus augusti”. This subspecies is also known as “Barking Frog” in English and “Rana-ladrarora común” in Spanish. The rest of alternative labels for English, French, Dutch, Spanish and Portuguese have been acquired using the multilingual correspondences of DBpedia.

If sufficient nodes in the vocabulary are represented by labels in a language, the hierarchy can be used to create a mapping across the database and the wordnet in a language. For mapping the SKOS Species 2000 database to PWN version 3.0, we thus can use the original Latin names occurring in the Species 2000 hierarchies and the corresponding 834,821 English labels. In fact, many species are named by its Latin name in PWN 3.0.

3.2 Integrating semantic structures

Ontology alignment has been recognised as a major issue in the semantic web community (van Hage, 2008). On the Semantic Web, data is structured by means of ontologies which describe the semantics of the data (Maedche and Staab, 2001). In this scenario, data is represented by many different ontologies. However, information processing across ontologies is not possible without knowing corresponding mappings between them. Manually finding such mappings is tedious, not systematic, and clearly not possible with large-scale ontologies representing large collections of content data. With the proliferation of applications sharing information represented in multiple ontologies, the development of automatic methods for robust and accurate ontology matching will be crucial to their success.

Due to the importance of the problem, many works have addressed ontology mapping using a variety of matching heuristics, e.g. (McGuinness et al., 2000), (Noy and Musen, 2001), (Rodriguez and Egenhofer, 2003). Recently, the Relaxation Labelling algorithm and structural constraints have been integrated successfully in a multi-strategy process for mapping ontologies (Daudé et al., 2000), (Doan et al., 2002).

There is also a meta-approach to ontology integration. The Linking Open Data Project (Bizer et al., 2008), launched by the W3C, aims to interlink existing ontologies. It encourages people to make RDFS/OWL data sets available on-line as Web services. On top of these Web services, it establishes links between equivalent concepts in different data sets.

3.3 Integrating Species2000 and PWN

In order to perform the integration, we designed a novel approach to align Species2000 concepts to the PWN synsets. First, we manually align to the PWN synsets the Kingdoms concepts appearing in the Species2000. Then, Then, we perform an automatic alignment on each of the taxonomic

```
Kingdom -> Class  -> Order -> Family -> Genus -> Species -> Infra species
```

Figure 5: Biological taxonomy in Species2000

```
Kingdom: Animalia ->
         Class: Chordata ->
                Order: Amphibia ->
                       Family: Anura ->
                               Genus: Leptodactylidae ->
                                      Species: Eleutherodactylus ->
                                               Infra species: Eleutherodactylus augusti
```

Figure 6: Example of the biological classification of an Species2000 concept

branches occurring in the Species2000 ontology. For example:

- Animalia : Chordata
- Animalia : Chordata : Amphibia
- Animalia : Chordata : Amphibia : Anura
- Animalia : Chordata : Amphibia : Anura : Leptodactylidae
- Animalia : Chordata : Amphibia : Anura : Leptodactylidae : Eleutherodactylus
- ...

During the process, we also keep record of the alignment of a particular Species2000 concept occurring in a partial branch allowing to maintain an appropriate consistency of the alignment.

The alignment process has been carried out by using a robust and accurate knowledge-based Word Sense Disambiguation algorithm. We used a version of the Structural Semantic Interconnections algorithm (SSI) called SSI-Dijkstra (Cuadros and Rigau, 2008), (Laparra and Rigau, 2009). SSI is a knowledge-based iterative approach to Word Sense Disambiguation (Navigli and Velardi, 2005). Previously, the SSI-Dijkstra algorithm has been used for constructing KnowNets (Cuadros and Rigau, 2008) and for the integration of PWN and FrameNet (Laparra and Rigau, 2009).

The original SSI algorithm is very simple and consists of an initialisation step and a set of iterative steps. Given W, an ordered list of words to be disambiguated, the SSI algorithm performs as follows. During the initialisation step, all monosemous words are included into the set I of already interpreted words, and the polysemous words are included in P (all of them pending to be disambiguated). At each step, the set I is used to disambiguate one word of P, selecting the word sense which is closer to the set I of already disambiguated words. Once a sense is disambiguated, the word sense is removed from P and included into I. The algorithm finishes when no more pending words remain in P.

SSI-Dijkstra uses the Dijkstra algorithm to obtain the shortest path distance between a node and the rest of nodes of the whole graph. The Dijkstra algorithm is a greedy algorithm that computes the shortest path distance between one node an the rest of the nodes of a graph. The BoostGraph[5] library can be used to compute very efficiently the shortest distance between any two given nodes on very large graphs. We also use already available knowledge resources to build very large connected graphs. In fact, we use two graph to perform the alignment. The first graph uses only hyponym/hypernym relations with 97,666 edges and the second uses the set of direct relations between synsets gathered from PWN and the relations extracted from the sense annotated PWN glosses, totalising 595,339 edges. That is, the first one with only PWN hyponymy/hypernymy relations and a second one with all PWN and gloss relations.

Note that initially the list I of interpreted words should include the senses of the monosemous words in W, or a fixed set of word senses. In our case, we already have the top Kingdom concepts of each taxonomic branch from Species2000 manually aligned to its appropriate synset.

Consider the example above:

```
Animalia : Chordata : Amphibia : Anura :
  Leptodactylidae : Eleutherodactylus
```

In this case, only “animalia” (aligned manually to animal#n#1) and “amphibia” appear in PWN. However, in English “eleutherodactylus” is also “barking_frog” which appears in PWN. Thus, we can establish the following alignment:

- Interpretation: barking_frog n 01643507-n “of southwest United States and Mexico; call is like a dog’s bark”

[5] http://www.boost.org/doc/libs/1_35_0/libs/graph/doc/index.html

```
<skos:Concept
rdf:about="http://kyoto-project.eu/col2009ac/Animalia/Chordata/Amphibia/Anura/Leptodactylidae/Eleutherodac-
          tylus/ITS-207724">
        <skos:prefLabel xml:lang="la">Eleutherodactylus augusti</skos:prefLabel>
        <skos:prefLabel xml:lang="en">Barking Frog</skos:prefLabel>
        <skos:prefLabel xml:lang="es">Rana-ladradora común</skos:prefLabel>
        <skos:altLabel xml:lang="en">Eleutherodactylus</skos:altLabel>
        <skos:altLabel xml:lang="fr">Eleutherodactylus</skos:altLabel>
        <skos:altLabel xml:lang="nl">Eleutherodactylus</skos:altLabel>
        <skos:altLabel xml:lang="es">Eleutherodactylus</skos:altLabel>
        <skos:altLabel xml:lang="pt">Eleutherodactylus coqui</skos:altLabel>
        <skos:broader
rdf:resource="http://kyoto-project.eu/col2009ac/Animalia/Chordata/Amphibia/Anura/Leptodactylidae/Eleuthero-
          dactylus"/>
</skos:Concept>
```

Figure 7: Example of SKOS concept enriched with language labels from Dbpedia

- Interpretation: amphibia n 01625747-n "the class of vertebrates that live on land but breed in water; frogs; toads; newts; salamanders; caecilians"
- Interpretation: animal n 00015388-n "a living organism characterised by voluntary movement"

The mapping also provides the proximity scores of the two graphs used and the Lexicographer file from PWN, in this case 05 ANIMAL[6]. We use the two scores provided by the SSI-Dijkstra algorithm and the Lexicographer files to filter out inappropriate matchings. We select only those alignments appearing in the ANIMAL lexicographer file and with the scores above average. Finally, a total number of 150,486 Species2000 concepts have been aligned to a PWN synset, while discarding 330,167 potential connections. The total number of concepts in Species2000 is 3,006,105. Thus, we are connecting to PWN just a small amount of concepts. The rest can be considered as new domain concepts in PWN.

3.4 Evaluation

In order to perform an initial evaluation of the alignment process, we selected randomly a small set of one-hundred filtered alignments. An independent evaluator (not an expert in the field) established the correctness of the mapping according to the following categories: C (correct), B (matches the broader term), BB (matches even higher up in the hierarchy) and X (incorrect).

We ignored the infraspecies level, concentrating on the species level. Surprisingly, the results show up no incorrect cases (X). However, almost all cases are B (48) or BB (52), and only one case is C. However, it seems that the filtering process performed correctly. For instance, the following branch was not included as a result of the mapping:

```
Animalia : Mollusca : Gastropoda : Baso mmatophora :
  Planorbidae : Armiger
score  WN hierarchy = 0.272727272727273
score WN+gloss = 0.0769230769230769
synset = eng-30-09808591-n
lexicographer file = PERSON
```

Possibly, by adjusting the filtering parameters we would obtain different coverage/accuracy figures.

3.5 Error analysis

We can partly explain this behaviour by looking at the following example trying to establish the connection at the "genus" level of drosophila.

```
Animalia : Arthropoda : Insecta : Diptera :
  Drosophilidae : Drosophila
score WN hierarchy = 0.5
score WN+gloss = 0.19047619047619
synset = eng-30-02197413-n
lexicographer file = ANIMAL
```

The "genus Drosophila" also occurs in PWN as synset eng-30-02197545-n. Thus, we are matching too high in the hierarchy. We are probably missing potential candidates since we are not taking into account the information of the level description of the Species2000 hierarchy. Thus, in order to improve the matching process, the general look-up strategy could be extended with domain specific heuristics (e.g. use the genus, order, family clues). Such look-up modules need to be made for each domain and used optionally in the software.

Furthermore, if the concept is not found in PWN, we use the previous aligned concept in Species2000 hierarchy. This is always a more abstract concept. In that case we should also change the skos mapping to skos:broaderMatch.

[6] http://wordnet.princeton.edu/man/lexnames.5WN.html

4 Formalising Species2000 into WN-LMF

This section reports on the automatic procedure developed to derive a WN-LMF compliant domain lexicon from the mapping between the Species2000 thesaurus and the PWN lexicon (see section 3). The resulting lexicon contains elements belonging both to WN-LMF's core (LexicalResource, Lexicon, LexicalEntry, Lemma, Sense) and semantic (Synset, SynsetRelation) components.

The procedure followed for each instance of the mapping depends on the type of the mapping:

- Monosemous mappings (output-M), the Species entry is mapped to a monosemous word in PWN. In this case we add an equivalence relation (eq-plugin) to the corresponding synset.
- Polysemous mappings (output-P), the Species entry is mapped to a polysemous word in PWN and the correct sense automatically disambiguated. In this case, we proceed as for monosemous mappings, we add an eq-plugin relation to the synset.
- New mappings (output-N), the Species entry is not found in PWN, so the mapping is void. The entry is encoded in WN-LMF but not connected to PWN

Apart from this, for all the entries and regardless of their mapping type, we add a hyperonym relation (has_hyperonym) to the direct broader term (e.g. from mollusca to animalia). This information is fetched from the hyperonymy chain as represented in the URI of the mapping input (e.g. `http://kyoto-project.eu/col2009ac/Animalia/Mollusca`).

As the domain lexicon created is connected to the general-domain wordnet, it is required to perform some checking in order to avoid identifier conflicts. A conflict would arise if an identifier of the domain lexicon is already present in the general-domain lexicon. In order to prevent this situation, all the identifiers in the domain lexicon have an extra ”d” character (e.g. LE_d_Mollusca_n).

As an example of this conversion procedure, from this sample of the mapping file:

```
output-M: LA [...]/col2009ac/Animalia : Animalia->
  01313093-n 0 05 animal
output-M: LA [...]/col2009ac/Animalia/Mollusca :
  Mollusca-> 01940488-n 0.25 05 animal
output-N: LA [...]/col2009ac/Animalia/Mollusca/
  Gastropoda/Basommatophora : Basommatophora-> ? 0
In: animalia|n|1 mollusca|n|1 gastropoda|n|1
  armiger|n
output-P: LA [...]/col2009ac/Animalia/Mollusca/
  Gastropoda/Basommatophora/Planorbidae/Armiger :
Armiger-> 09808591-n 0.0769230769230769 18 person
```

This is the output produced:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE LexicalResource SYSTEM "kyoto_wn.dtd">
<LexicalResource>
 <GlobalInformation>
  <feat att="label" val="WN-LMF domain lexica from
   the Species2000 thesaurus"/>
 </GlobalInformation>
 <Lexicon label="English lexicon from Species2000
  thesaurus" language="eng"
  languageCoding="ISO 639-3" owner=""
  version="20090716">
  <LexicalEntry id="LE_d_Animalia_n">
   <Lemma partOfSpeech="n" writtenForm="Animalia"/>
   <Sense id="S_d_Animalia_1"
    synset="eng-d-00000001-n"/>
  </LexicalEntry>
  <LexicalEntry id="LE_d_Mollusca_n">
   <Lemma partOfSpeech="n" writtenForm="Mollusca"/>
   <Sense id="S_d_Mollusca_1"
    synset="eng-d-00000002-n"/>
  </LexicalEntry>
  <LexicalEntry id="LE_d_Basommatophora_?">
   <Lemma partOfSpeech="?"
    writtenForm="Basommatophora"/>
   <Sense id="S_d_Basommatophora_1"
    synset="eng-d-00000003-n"/>
  </LexicalEntry>
  <LexicalEntry id="LE_d_Armiger_n">
   <Lemma partOfSpeech="n" writtenForm="Armiger"/>
   <Sense id="S_d_Armiger_1"
    synset="eng-d-00000004-n"/>
  </LexicalEntry>
  <Synset id="eng-d-00000001-n">
   <SynsetRelation relType="eq-plug-in"
    target="01313093-n">
    <Meta confidenceScore="0"/>
   </SynsetRelation>
  </Synset>
  <Synset id="eng-d-00000002-n">
   <SynsetRelation relType="eq-plug-in"
    target="01940488-n">
    <Meta confidenceScore="0.25"/>
   </SynsetRelation>
   <SynsetRelation relType="has_hyperonym"
    target="eng-d-00000001-n"/>
  </Synset>
  <Synset id="eng-d-00000003-n">
   <SynsetRelation relType="hyp-plug-in"
    target="?">
    <Meta confidenceScore="0"/>
   </SynsetRelation>
   <SynsetRelation relType="has_hyperonym"
    target=""/>
  </Synset>
  <Synset id="eng-d-00000004-n">
   <SynsetRelation relType="eq-plug-in"
    target="09808591-n">
    <Meta confidenceScore="0.0769230769230769"/>
   </SynsetRelation>
   <SynsetRelation relType="has_hyperonym"
    target=""/>
  </Synset>
 </Lexicon>
</LexicalResource>
```

5 Conclusions

This paper has presented a methodology to link domain thesauri to general-domain lexica and has studied the formalisation of the resulting mapping. The method uses the SSI-Dijkstra algorithm and exploits the semantic relations found in lexical resources. A case study carried out in the framework of the KYOTO project to link the Species 2000 thesaurus to the English WordNet has been described, including evaluation and error analysis. The resulting resource is then converted to WN-LMF, a dialect of the ISO standard LMF, so that it can communicate with the other resources available in the KYOTO architecture. Regarding future work, we plan to extend the general look-up strategy with domain specific heuristics to improve matching.

Acknowledgements

This work has been partially funded by the EU Commission under the project KYOTO (ICT-2007-211423).

References

Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. 2008. Linked data on the web. In *WWW*, pages 1265–1266.

M. Cuadros and G. Rigau. 2008. Knownet: Building a large net of knowledge from the web. In *Proceedings of COLING*.

J. Daudé, L. Padró, and G. Rigau. 2000. Mapping WordNets Using Structural Information. In *Proceedings of 38th annual meeting of the Association for Computational Linguistics (ACL'2000)*, Hong Kong.

AnHai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. 2002. Learning to map between ontologies on the semantic web. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 662–673, New York, NY, USA. ACM.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

Gil Francopoulo, Monica Monachini, Thierry Declerck, and Laurent Romary. 2006. The relevance of standards for research infrastructure. In *LREC 2006, Workshop Towards Research Infrastructures for Language Resources*. European Language Resources Association (ELRA).

ISO 24613. 2008. Languages Resources Management – Lexical Markup Framework (LMF), rev.15 ISOTC37SC4 FDIS. [Online; accessed 25-March-2008].

E. Laparra and G. Rigau. 2009. Integrating wordnet and framenet using a knowledge-based word sense disambiguation algorithm. In *Proceedings of the International Conference, Recent Advances on Natural Language Processing RANLP'09*, Borovets, Bulgaria.

Alexander Maedche and Steffen Staab. 2001. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79.

Isa Maks, Piek Vossen, Roxane Segers, and Hennie van der Vliet. 2008. Adjectives in the dutch semantic lexical database cornetto. In *Proceedings of LREC 2008*, Marrakech, Morocco. European Language Resources Association (ELRA).

Deborah L. McGuinness, Richard Fikes, James Rice, and Steve Wilder. 2000. The chimaera ontology environment. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 1123–1124. AAAI Press / The MIT Press.

R. Navigli and P. Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(7):1063–1074.

Natalya Noy and Mark Musen. 2001. Anchor-prompt: Using non-local context for semantic matching. In *In Proc. IJCAI 2001 workshop on ontology and information sharing*, pages 63–70.

M. A. Rodriguez and M. J. Egenhofer. 2003. Determining semantic similarity among entity classes from different ontologies. *Knowledge and Data Engineering, IEEE Transactions on*, 15(2):442–456.

Claudia Soria, Monica Monachini, and Piek Vossen. 2009. Wordnet-lmf: fleshing out a standardized format for wordnet interoperability. In *IWIC '09: Proceeding of the 2009 international workshop on Intercultural collaboration*, pages 139–146, New York, NY, USA. ACM.

W. R. van Hage. 2008. *Evaluating Ontology-Alignment Techniques*. Ph.D. thesis, Vrije Universiteit Amsterdam.