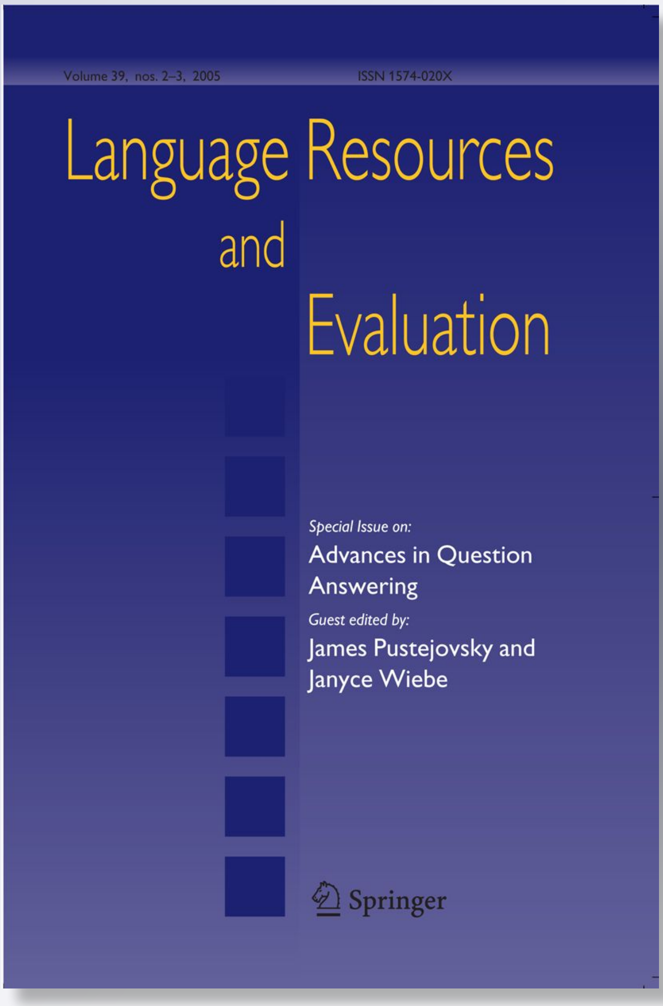
Interchanging lexical resources on the Semantic Web

John McCrae, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, et al.

Language Resources and Evaluation

ISSN 1574-020X

Lang Resources & Evaluation
DOI 10.1007/s10579-012-9182-3



Volume 39, nos. 2–3, 2005 ISSN 1574-020X

Language Resources
and
Evaluation

Special Issue on:
Advances in Question Answering
Guest edited by:
James Pustejovsky and
Janyce Wiebe

Springer

Springer

Lang Resources & Evaluation
DOI 10.1007/s10579-012-9182-3

ORIGINAL PAPER

Interchanging lexical resources on the Semantic Web

John McCrae · Guadalupe Aguado-de-Cea · Paul Buitelaar · Philipp Cimiano · Thierry Declerck · Asunción Gómez-Pérez · Jorge Gracia · Laura Hollink · Elena Montiel-Ponsoda · Dennis Spohr · Tobias Wunner

Abstract Lexica and terminology databases play a vital role in many NLP applications, but currently most such resources are published in application-specific formats, or with custom access interfaces, leading to the problem that much of this data is in "data silos" and hence difficult to access. The Semantic Web and in particular the Linked Data initiative provide effective solutions to this problem, as well as possibilities for data reuse by inter-lexicon linking, and incorporation of data categories by dereferencable URIs. The Semantic Web focuses on the use of ontologies to describe semantics on the Web, but currently there is no standard for providing complex lexical information for such ontologies and for describing the relationship between the lexicon and the ontology. We present our model, lemon, which aims to address these gaps

J. McCrae (✉) · P. Cimiano · D. Spohr
CITEC, University of Bielefeld, Universitätsstraße, Bielefeld, Germany
e-mail: jmccrae@techfak.uni-bielefeld.de

P. Cimiano
e-mail: cimiano@techfak.uni-bielefeld.de

D. Spohr
e-mail: dspohr@techfak.uni-bielefeld.de

G. Aguado-de-Cea · A. Gómez-Pérez · J. Gracia · E. Montiel-Ponsoda
Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, Boadilla del Monte, Spain

G. Aguado-de-Cea
e-mail: lupe@fi.upm.es

A. Gómez-Pérez
e-mail: asun@fi.upm.es

J. Gracia
e-mail: jgracia@fi.upm.es

E. Montiel-Ponsoda
e-mail: emontiel@fi.upm.es

P. Buitelaar · T. Wunner
DERI, National University of Ireland, Galway, Galway DERI IDA Business Park, Galway, Ireland

Published online: 08 May 2012

while building on existing work, in particular the Lexical Markup Framework, the ISOcat Data Category Registry, SKOS (Simple Knowledge Organization System) and the LexInfo and LIR ontology-lexicon models.

Keywords Lexica · Terminology · Semantic Web · Linked data · Ontologies

1 Introduction

Lexica and terminology databases form an essential part of many modern NLP systems and frequently consist of large amounts of highly detailed and well curated entries. Examples of such resources are the lexical semantic network WordNet (Fellbaum 1998) and subcategorisation lexica such as COMLEX (Grishman et al. 1994). However, there is currently a great diversity of formats for representing such lexical resources, thus making it difficult to share and interlink them. Current work on the Semantic Web, in particular that of the Linking Open Data project (Bizer et al. 2009), has focused on the challenge of using Web representation formalisms, RDF in particular, to connect such "data silos" and allows for interlinking different datasets. This linking supports the reuse of entries from a lexicon within another one and allows third parties to extend an existing lexicon. However, so far there is no principled model to connect ontological knowledge with lexical knowledge, that would enable the creation of an interface between an ontology and an appropriate lexicon. Such an ontology-lexicon interface represents an essential component in the scenario of the Semantic Web, since it will enable an appropriate exploitation of the available knowledge by end-user applications, which are frequently language-based. Thus, it seems natural that any attempt to exchange lexica on the Semantic Web should build on Semantic Web representation formalisms, i.e. RDF and OWL. While there exist many terminology resources, they rarely have sufficient semantic information to enable these resources to be used for more complex reasoning. Similarly, while there exist many large semantic resources, such as DBPedia (Auer et al. 2007), and in particular models of domain semantics such as the Gene Ontology (Ashburner et al. 2000), they are rarely connected to complex linguistic and lexical information. Our goal is thus to provide a formalisms that 'connects

P. Buitelaar
e-mail: paul.buitelaar@deri.ie

T. Wunner
e-mail: tobias.wunner@deri.ie

T. Declerck
DFKI, Stuhlsatzenhausweg 3, Saarbrücken, Germany
e-mail: declerck@dfki.de

L. Hollink
Technical University of Delft, Mekelweg 4, Delft, Netherlands
e-mail: l.hollink@tudelft.nl

these worlds', i.e. the world of lexical resources and the world of ontologies and semantic data as available on the Semantic Web.

Towards this goal we have developed a model we call *lemon* (**Le**xicon **M**odel for **On**tologies), which is designed to represent lexical information about words and terms relative to an ontology on the Web. *lemon* is what we term an *ontology-lexicon*, in that, following Buitelaar (2010), the ontology-lexicon expresses how the elements of the ontology, i.e. classes, properties, and individuals, are realized linguistically. In providing this model, we follow a principle that we call *semantics by reference* in the sense that the (lexical) meaning of the entries in the lexicon is assumed to be expressed exclusively in the ontology and the lexicon merely points to the appropriate concepts. This is in contrast to other lexical resources which include lexico-semantic relations such as hypernymy or synonymy as part of the lexicon. *lemon* is not intended to be a collection of resources but rather a basic model supporting the exchange of ontology-lexica on the Semantic Web. We focus primarily on domain terminology, as ontologies generally refer to specific domains, however, *lemon* is not domain-specific and could be used for any task. The *lemon* model builds on previous research by the authors in the design of lexica for interfacing with ontologies, in particular that of the LexInfo (Buitelaar et al. 2009) and LIR (Montiel-Pondsoda et al. 2010) models, as well as existing work on lexicon (meta-)models, in particular the Lexical Markup Framework (ISO-24613:2008) (Francopoulo et al. 2006). In addition, we build on work that is currently being performed in using the Web to link resources to the ISOcat meta-data registry (Kemps-Snijders et al. 2008) as well as in the OLiA project (Chiarcos 2010). The *lemon* model attempts to be a highly scalable format in the sense that its modelling of lexical and linguistic information related to concepts in ontologies has to scale from very simple to quite complex lexical entries. In many ways, *lemon* is closely related to the work of the SKOS project (Miles and Bechhofer 2009), which attempts to model simple knowledge organisation systems such as thesauri, classification schemes and taxonomies on the Semantic Web. However, the model we propose differs from SKOS in that it is an independent and external model, intended to be published with arbitrary ontology-based conceptualisations, or any other type of knowledge organisation systems, in order to provide a richer description of the knowledge captured in those resources in one or several natural languages. In fact, the model is agnostic to the specific linguistic data categories used, allowing to reuse any data category (e.g. part-of-speech) together with its values. *lemon* is able to incorporate such externally defined data categories by including their URIs as a unique specification of a property, giving additional information such as the ownership which becomes accessible when dereferencing the corresponding URI.

The remainder of the paper is structured as follows: In Sect. 2 we give a brief overview of the main standardisation initiatives for linguistic and terminological description that have inspired our work. We also provide a brief description of those models intended to interface ontologies we draw upon, and of standards for interchanging linguistic and lexical information on the Web. Then, in Sect. 3, we present the *lemon* model and provide several examples of linking possibilities provided by the model that contribute to the reuse of and interoperability with existing standards. Further, in Sect. 4 we report on available tools that support the creation of specific lexicon instances. Finally, in Sect. 5 we summarise the main benefits of the model and conclude the paper.

2 Foundations and related work

In this section we will briefly describe some of the extant models for the representation of lexica and work on establishing the correspondence between syntactic and semantic resources.

2.1 WordNet and FrameNet

WordNet (Fellbaum 1998) is arguably the most significant lexical database for English, in which word senses are organised in sets of semantic equivalents (so-called "synsets"). Over the years, the WordNet model has been applied to many other languages besides English. As part of the EuroWordNet project (Vossen 1998), many of these multilingual WordNets have been linked by means of an interlingual index, a set of core meanings assumed to exist in all languages. More recently, WordNet has been adapted to RDF and published on the Semantic Web as linked data (Van Assem et al. 2006). However, as WordNet aims to be a general lexicon for English, it does not contain many domain-specific terms, although Vossen et al. (1999) have outlined how such terms could be added to the interlingual index. In addition, WordNet assumes a rather informal interpretation of its lexical-semantic relations (e.g. synonymy, antonymy, hypernymy and meronymy), and does not allow for sophisticated linguistic information. In fact, WordNet only provides information on four parts of speech: nouns, verbs, adjectives and adverbs. In general, its data model is not easy to carry over to lexica with significantly different purposes.

FrameNet (Baker et al. 1998) is a hierarchically structured collection of prototypical situations (called "semantic frames") that are *evoked* by lexical units. Each frame has a set of slots or roles (called "frame elements"), describing participants involved in a particular situation. In contrast to WordNet, semantic relations are not expressed between senses, but between frames and frame elements. Scheffczyk et al. (2006) have shown how these can be represented in OWL and linked to ontologies like SUMO. Similar to the case of WordNet, however, FrameNet is not intended to be a general lexicon model, and does not provide a vocabulary for deeper linguistic description nor a clear methodology how such could be integrated. In fact, FrameNet is mainly concerned with defining a repertoire of cognitively inspired linguistic frames and not with providing a general model for the ontology-lexicon interface.

2.2 LMF

The Lexical Markup Framework (LMF) is an ISO standard for representing lexica in XML/UML, which has had a strong influence on the design of *lemon* in terms of how lexical information is represented. It provides a framework for representing lexical objects, including morphological, syntactic, and semantic aspects of these. It was conceived for the purpose of providing a common model for the representation of electronic lexical resources in order to permit data exchange and thus foster interoperability. The description of an entry is very detailed and relies on previous standards for linguistic category description, namely ISO 12620 Data Categories or ISOcat (see Sect. 2.5), thus making the data highly reusable. In this sense, the LMF standard has been conceived as a meta-model for representing the whole lexicon of a language, in which all possible senses of a word are accounted for. Instead, *lemon*'s purpose is to enrich the conceptualisation represented by a given ontology by means of a lexico-terminological layer.

A simple example of an LMF entry in RDF[1] is given below in Turtle syntax (Beckett and Berners-Lee 2008). The lexicon consists of a single entry representing a common noun with lemma “tax”:

```
@prefix lmf: <http://www.tagmatica.fr/lmf#>.

:lexicon1 a lmf:Lexicon .

:entry1 a lmf:LexicalEntry ;
  lmf:isPartOf :lexicon1 ;
  lmf:isAdorned [ lmf:att “partOfSpeech” ;
                  lmf:val “commonNoun” ] .
:lemma1 a lmf:Lemma ;
  lmf:isPartOf :entry1 ;
  lmf:isAdorned [ lmf:att “writtenForm” ;
                  lmf:val “tax” ] .
```

The RDF/OWL version of LMF, however, uses only the properties `isAssociated`, `isPartOf` and `isAdorned`, and the size of the RDF models generated is very large due to the number of unnecessary elements introduced by the conversion to RDF. Moreover, lexical properties like “writtenForm” and data categories like “partOfSpeech” or “commonNoun” are hidden inside literal values. As a consequence, the format does not exploit the full potential of RDF and it is thus very difficult to query and work with lexica represented using this schema. In contrast, *lemon* takes an RDF-native approach in using a different name for each property. In *lemon*, the above example can be represented as follows[2]:

[1] Note that LMF is a meta-model and hence other serializations could be consistent with the model. However for the purposes of this paper, we refer to the RDF and XML serializations described at http://www.lexicalmarkupframework.org/.

[2] Note that “lemma” approximately corresponds to “canonical form” in *lemon* and we specify the `xml:lang` special property on each string in *lemon*.

```
@prefix lemon: <http://www.monnet-project.eu/lemon#> .

:lexicon lemon:entry :tax .

:tax lemon:canonicalForm [ lemon:writtenRep “tax” @en ] ;
      :partOfSpeech :commonNoun .
```

2.3 SKOS

SKOS (Simple Knowledge Organisation System) was developed as a system to provide a way to formalise many knowledge organisation systems and share them on the Web: (from Miles and Bechhofer 2009)

> Different families of knowledge organisation systems, including thesauri, classification schemes, subject heading systems, and taxonomies are widely recognised and applied in both modern and traditional information systems. In practise it can be hard to draw an absolute distinction between thesauri and classification schemes or taxonomies, although some properties can be used to broadly characterise these different families. The important point for SKOS is that, in addition to their unique features, these families have much in common with one another, and can often be used in similar ways. *However, there is currently no widely deployed standard for representing these knowledge organisation systems as data and exchanging them between computer systems*. (emphasis added)

We focus on SKOS here as it is based on RDF and is the format most widely used on the Semantic Web for representing terminologies, thesauri and taxonomies. However, it is also important to note that there exist other models for representing terminologies such as OTR (Reymont et al. 2007) and TBX (ISO 30042), which are based on Linked Data and XML standards, respectively. In many ways *lemon* aims to achieve similar goals to SKOS in making lexica available on the Semantic Web.

Within the use cases that motivated the design of SKOS (Isaac et al. 2009), it was identified that it is important for many of these knowledge organisation systems to have labels that relate to one another, indicating for instance that one label is an acronym of another label. For this reason, an extension called SKOS-XL (SKOS eXtension for Labels) was introduced in which the label property is ‘reified’, so that further properties of labels can be specified. In this context, a concept like tuberculosis with a label “tuberculosis” can be additionally associated to an alternative label “TB”, thus indicating that one is an acronym of the other.

```
@prefix skosxl: <http://www.w3.org/2008/05/skos-xl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix isocat: <http://www.isocat.org/datcat/> .
:tuberculosis skosxl:prefLabel :tuberculosis_label ;
         skosxl:altLabel :tuberculosis_shortform .
:tuberculosis_label skosxl:literalForm “tuberculosis” @en.
:tuberculosis_shortform skosxl:literalForm “TB” @en .
# DC-66=acronymFor
:tuberculosis_shortform isocat:DC-66 :tuberculosis_label .
```

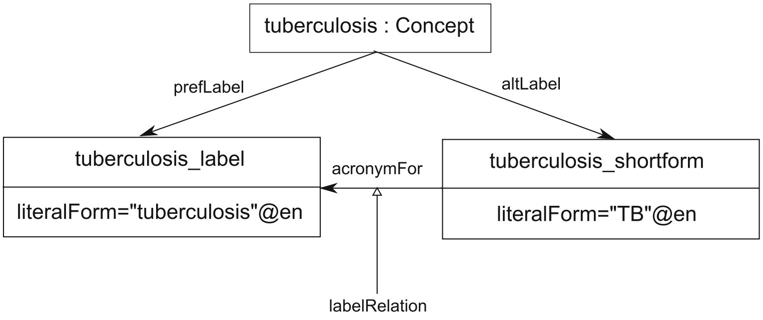


Fig. 1 Example of adding lexical information with SKOS-XL

The main modelling decisions of *lemon* are based on the SKOS model. However, we extend the idea of reifying labels by introducing a well defined "textual-conceptual" path, which is defined simply as *the number of nodes between the string literal value and the concept* (in the above example, illustrated in Fig. 1 `:tuberculosis`). In SKOS-XL, an extra node is introduced between the text and the concept, reifying labels and thus allowing to state relations between labels. However, there is no clear linguistic motivation for the introduction of this node. One key aspect of *lemon* is that it introduces a longer but, from a linguistic point of view, more principled chain, in particular differentiating between syntactic and terminological variation and clearly separating pragmatic and syntactic constraints.

2.4 LexInfo and LIR

As *lemon* follows the principle of "semantics by reference", no complex semantic information needs to be stated in the lexicon. Consequently, we build on our previous models to represent the interface between lexica and ontologies and in particular how syntactic information in the lexicon can be linked to semantic information in the ontology. In the LexInfo project (Buitelaar et al. 2009) we identified the key requirements of a lexicon-ontology model as follows:

1. **Separation of the lexicon and ontology**. There must be a clear separation of the lexical layer and the ontological layer, and lexica must be interchangeable (i.e., multiple lexica can describe the same ontology).
2. **Structured linguistic information**. It should be possible to represent linguistic descriptions, e.g., part of speech.
3. **Syntactic behaviour**. The model should represent the syntactic behaviour of its entries, e.g., valency of verbs.
4. **Morphological decomposition**. The model should allow for the representation of the decomposition of terms.
5. **Arbitrary and multiple ontologies**. The lexicon model should be general enough to work with any ontology and conceptualization. Further, it should support the reuse of lexica across ontologies.

The LIR (Linguistic Information Repository) model (Montiel-Pondsoda et al. 2010) supports the 1st, 2nd and 5th requirements. LIR has a strong focus on multilingualism and thus provides mechanisms to establish links among lexical and terminological elements within and across different languages. Thus, LIR and LexInfo can be viewed as complementary models which have fixed data categories. To overcome this rigidity, our goal was to design a model that was agnostic with respect to which data categories are used by a specific instantiation of the model, i.e. a concrete lexicon. Both LIR and LexInfo are problematic in this sense as they make too strong commitments to certain linguistic data categories, while at the same time being rather "incomplete" in the values defined for these categories. We thus designed *lemon* as an open, flexible and linguistically agnostic model that could be applied to many different purposes by avoiding to make unnecessary assumptions by introducing specific data categories.

2.5 ISOcat, OLiA, GOLD

There have a been a number of attempts to enable the exchange of computer lexica and, as described by Romary (2010), there is increasing convergence among different formats. One of the key challenges identified in developing an exchange lexicon is whether to define a specific model or rather provide general guidelines. In particular Romary notes that "the choice to provide an actual format potentially facilitates immediate interoperability across applications, but bears the risk of not being flexible enough if some phenomena occur that have not been anticipated in the standard." One of the key solutions to this issue is the idea of *data categories* that aim to provide the following (again from Romary 2010):

- A generic entry point and unique identifier for sharing concepts
- Fine grained information about a linguistic concept that may only be relevant to certain languages or resources

The goals of data category projects can be seen to parallel those of the Semantic Web, in particular as identified by Shadbolt et al. (2006). The first goal of data categories parallels the usage of dereferencable URIs to identify resources on the Semantic Web, while the second goal parallels the creation of large scale RDF(S) taxonomies and OWL ontologies for describing particular domains.

There has been some work on harmonizing and integrating different date category ontologies. The GOLD ontology (Farrar et al. 2003) for instance combines many of the most common lexical categories into a single large ontology. A similar but more recent project is ISOcat, which is connected to the work on standardization of LMF and relies on a format called DCIF (ISO 12620). However, each data category are also published as RDF allowing for some interface with existing Semantic Web standards. Finally, OLiA (Chiarcos 2010) is an ontology that builds on existing taxonomies of linguistic annotation and provides a core reference model that covers similar ground to GOLD and ISOcat. In particular, OLiA has annotation linking models that are used to describe alignments between the OLiA reference model and other annotation schemes (for example Penn Treebank tags). The OLiA

Table 1 Size of existing resources for data categories of linguistic properties and values

	Values	Properties
GOLD	506	83
ISOcat	1,506	1,538
OLiA	595	45

Note that in ISOcat values are called "simple" data categories and properties are referred to as "complex" data categories. For OLiA and GOLD, properties and values are modelled as OWL classes and OWL individuals, respectively

project is also working on publishing links between the OLiA reference model and the GOLD and ISOcat models. The relative sizes of each resource are given in Table 1.[3]

3 The lemon model

In the light of current existing linguistic resource standards, we propose *lemon* as a model for exchanging lexicon resources on the Web with the following goals:

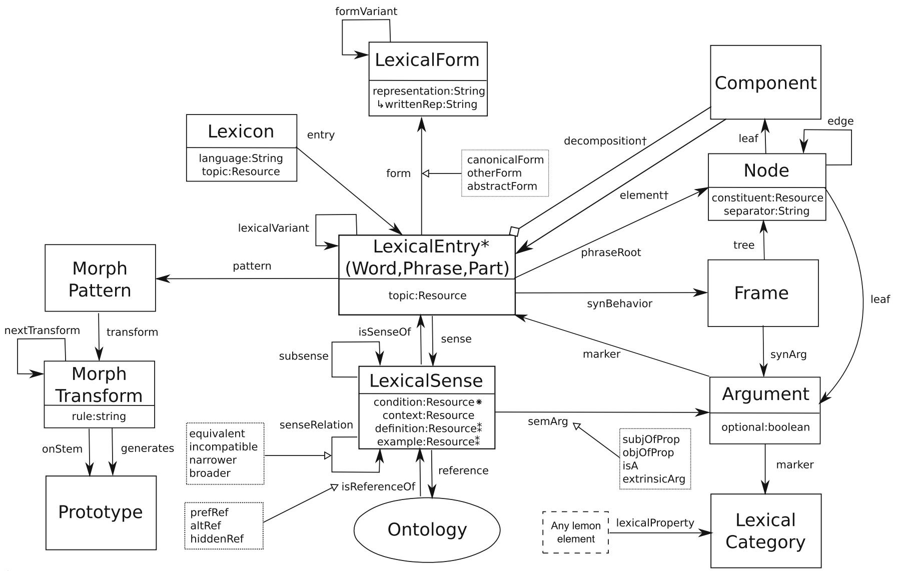
- LMF-like structure facilitating the conversion to existing formats (TBX, TEI, TIGER etc.).
- RDF-native form to enable leveraging existing Semantic Web technologies (SPARQL, OWL, RIF etc.).
- Separation of the lexicon vs. ontology layers with the result that the semantic information and lexical information are separated, but interlinked. This modularity enables straightforward exchange, addition, extension and substitution of lexica.
- The semantic inventory (ontology) is external to the lexicon model. Thus the model does not prescribe a representation of the meaning of entries and is open to any semantic distinction the user of the lexicon requires.
- Linking to data categories in order to allow for arbitrarily complex linguistic description.
- A small model using the *principle of least power*—the less expressive the language, the more reusable the data (Shadbolt et al. 2006).

The *lemon* model, as illustrated in Fig. 2, is available in RDF with extra OWL constraints at http://www.monnet-project.eu/lemon.

3.1 The core

The core of *lemon* covers the basic elements required to define lexical entries and associate them to their lexical forms as well as to concepts in the ontology

[3] The ISOcat results are based on public data categories, of which there are a total of 3,036 of these 9 lack a specified type and 17 are typed as both simple and complex. Results retrieved 17th January 2012.



* LexicalEntry has three subclasses: Word, Phrase, Part
‡ definition and example are stated as nodes with a value
⁎ condition has subproperties propertyDomain and propertyRange
† decomposition and element may also be used with Frames and Arguments resp.

Fig. 2 The *lemon* model

representing their meaning. This is done primarily by defining the following elements:

- **Lexicon:** The object representing the lexicon as a whole. This must be marked with a language, with the consequence that all objects in the lexicon are assumed to belong to this language.
- **Lexical Entry:** An entry in a lexicon is a container for one or several forms and one or several meanings of a lexeme. All forms of an entry must be realised with the same part of speech, and while an entry may have multiple meanings, homonyms (as they have different etymologies) are treated as separate lexical entries.
- **Lexical Form:** An inflectional form of an entry. The entry must have one canonical form and may have any number of other forms. It may also have abstract forms, which are intended to model stems and other partial morphological units.
- **Representation:** A given lexical form may have several representations in different orthographies, for example a phonetic representation in addition to a standard written representation.
- **Lexical Sense:** A sense links the lexical entry to the **reference** used to describe its meaning, i.e. a concept, property or individual in the ontology.
- **Component:** A lexical entry may also be broken up into a number of components.

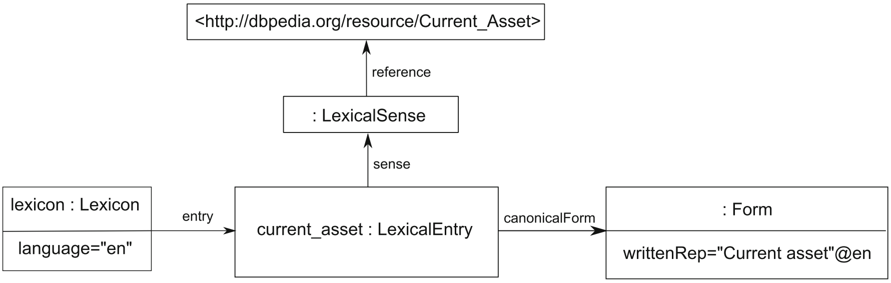

<http://dbpedia.org/resource/Current_Asset>
reference
: LexicalSense
sense
lexicon : Lexicon
language="en"
entry
current_asset : LexicalEntry
canonicalForm
: Form
writtenRep="Current asset"@en

Fig. 3 A simple example of a *lemon* lexicon with a single entry “Current asset”

In this way we give a clearer textual-conceptual path than is possible with SKOS. The following example gives a simple lexicon with a single lexical entry:

```
@prefix lemon: <http://www.monnet-project.eu/lemon#> .
@prefix dbpedia: <http://www.dbpedia.org/resource/> .
:lexicon lemon:entry :current_asset ;
   lemon:language “en” .
:current_asset
   lemon:canonicalForm[ lemon:writtenRep “Current asset” @en] ;
   lemon:sense[ lemon:ref dbpedia:Current_asset] .
```

In this example (illustrated in Fig. 3), we have an English lexicon with a single entry, with canonical form “Current asset”, and a sense that refers to the entry in the Linked Data resource DBPedia (Auer et al. 2007) from which further semantic information about the entry can be obtained.

3.2 Linking to data categories

While the core is useful for representing many aspects of lexical information, it is frequently necessary to include more information about morphology, syntax, terminological distinctions, versioning, authorship information etc. It would be very difficult to include all such categories in a way that would satisfy all users of the model. As a solution to this, we follow current Semantic Web practices, supporting the reuse of existing data categories by referencing their URIs. By this, users of the *lemon* model have absolute flexibility with respect to the choice of specific data categories. Consequently, the approach also scales in the sense that arbitrarily complex linguistic information can be included in the *lemon* model by referencing sources such as ISOcat, OLiA, GOLD for linguistic information, and vocabularies such as Dublin core[4] for authorship information. It is important to note that this does

[4] See http://dublincore.org/.

not solve the interoperability problem between category ontologies. However, this is also not the goal of *lemon*, but our position is that the reuse of unique identifiers and the ability to dereference these identifiers would support the alignment of categories across lexical resources. For example, we will show an entry for the Dutch feminine noun “vergunning” (“permit”), with plural form “vergunningen”.[5]

```
@prefix lemon: <http://www.monnet-project.eu/lemon#> .
@prefix isocat: <http://www.isocat.org/datcat/> .
@prefix dublincore: <http://purl.org/dc/elements/1.1/> .

:vergunning
     lemon:canonicalForm [ lemon:writtenRep ‟vergunning″ @nl;
                                # number=singular
                              isocat:DC-1298 isocat:DC-1387 ] ;
     lemon:altForm       [ lemon:writtenRep ‟vergunningen″ @nl ;
                                # number=plural
                              isocat:DC-1298 isocat:DC-1354 ] ;
     isocat:DC-1345 isocat:DC-1333 ; # partOfSpeech=noun
     isocat:DC-1297 isocat:DC-1880 ; # gender=feminine
     dublincore:contributor ‟John McCrae″ .

isocat:DC-1298 rdfs:subPropertyOf lemon:property .
isocat:DC-1345 rdfs:subPropertyOf lemon:property .
isocat:DC-1297 rdfs:subPropertyOf lemon:property .
```

Here we use ISOcat URIs to reference each of the properties, so that extra information about the data category can be obtained by dereferencing this link. The relation between these external properties and the *lemon* model is established by declaring them as sub-properties of *lemon*'s `property`, so that the role of the property in the *lemon* model is properly defined. The use of URIs implies that the specification of the linguistic category becomes unambiguous. Furthermore, the source and provenance as well as ownership and responsibility for the data category can be clearly identified. In addition, we use the Dublin Core vocabulary to provide non-linguistic annotations, e.g. to indicate the author of the lexical entry. The use of RDF for data categories may allow to express ontological relationships and constraints on a lexicon (see McCrae et al. 2011 for a preliminary discussion).

[5] We reference ISOcat by the use of the data category number, and put a readable comment to each property. In the diagrams, we put only the readable description.

3.3 Linking between lexica

One of the most interesting aspects of using RDF and Semantic Web standards is that there are possibilities of data reuse not available to static resources. For example the medical term “hospital-acquired pneumonia”, is composed of the words “hospital”, “acquired” and “pneumonia”, and we can provide appropriate morpho-syntactic and terminological information for each of these entries. However, it is inefficient for every single lexicon to repeat non-domain-specific words like “acquired”. Thus, we shall expand on our previous example to show how RDF can aid in data reuse:

```
@base <http://www.example.org/biomedical_lexicon> .
@prefix common: <http://www.example.org/common_lexicon#> .
@prefix lemon: <http://www.monnet-project.eu/lemon#> .
@prefix isocat: <http://www.isocat.org/datcat/> .

:hospital_acquired_pneumonia
     lemon:canonicalForm
       [ lemon:writtenRep “hospital-acquired pneumonia” @en ] ;
     lemon:decomposition (
       [ lemon:element common:hospital ]
       [ lemon:element common:acquire ;
          isocat:DC-1427 isocat:DC-1341 ; # mood=participle
          isocat:DC-1286 isocat:DC-1347 ] # tense=past
       [ lemon:element :pneumonia ]
     ) .

:pneumonia
   lemon:canonicalForm [ lemon:writtenRep “pneumonia” @en ] .
```

In this example (illustrated in Fig. 4), we see that “hospital-acquired pneumonia” is defined as being composed of an ordered list of components each of which refers to a lexical entry.[6] Two of these entries have URIs in the “common lexicon” (identified by, for example, http://www.example.org/common_lexicon#hospital) and one in the same lexicon, the “biomedical lexicon” (identified by the URI http://www.example.org/biomedical_lexicon). As such, any extra information that is stated in the common lexicon about the entries is then automatically available for users of the domain lexicon. Because these lexical entries are included by use of their URIs, they can be imported from any lexicon published on the Semantic Web, not just those controlled by the same author. This has the advantage that if the

[6] We note that more precise modelling of the phrase structure of the term is possible using the *lemon* model. This is described further in the “lemon cookbook” available at http://lexinfo.net/lemon-cookbook.pdf.

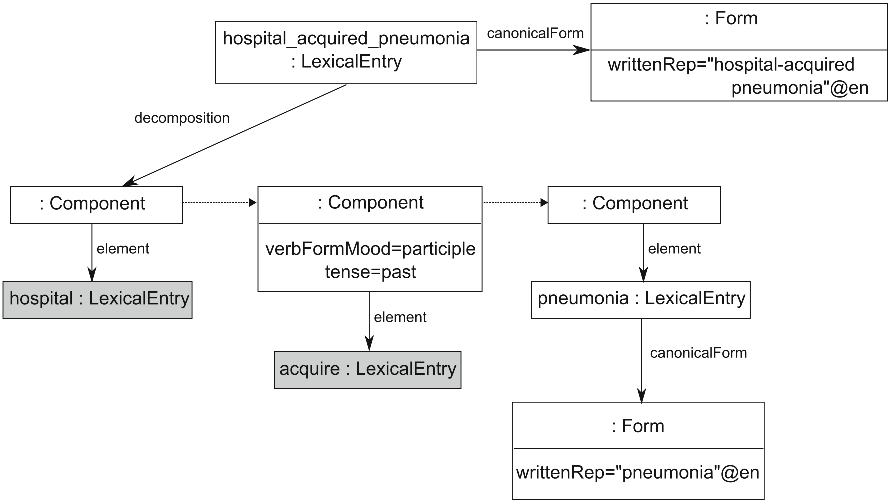


Fig. 4 Linking between lexica. The entries in white are part of the biomedical lexicon and the greyed entries are part of the general lexicon. Note that "acquired" is modelled as the past participle of the verb "acquire"

lexical entries are updated, the lexicon importing it will also automatically update these changes, a clear benefit of referencing in contrast to static import or duplication.

3.4 Lexicon-ontology mapping

The *lemon* model does not intend to be a semantic model, but instead it allows semantics to be represented by referencing extant semantic resources, in particular ontologies. The *lemon* model approaches this by means of its "(lexical) sense" object, which differs significantly from the concept of a word sense found in existing models, which has been criticised by many authors (Kilgariff 1997). Technically, a sense is unique for every pair of lexical entry and reference, i.e., the sense refers to a single ontology entity and a single lexical entry. Thus, each word has a different sense for each distinct reference. In fact, a sense may have multiple reference URI values, but this implies that the reference URIs represent ontologically equivalent entities.[7] The sense object in *lemon* plays three roles: first, the set of all senses defines a many-to-many mapping between lexical entries and ontological entities. This models the fact that lexical entries can have different meanings with respect to a given ontology and the fact that ontology elements can be verbalised linguistically in various ways. Second, the sense object represents the (lexical) meaning of the lexical entry when interpreted as the given ontological concept. Third, the sense also represents an ontological specialisation of the

[7] i.e., s `lemon:reference` x_1, s `lemon:reference` $x_2 \vdash x_1$ `owl:sameAs` x_2, if both x_1 and x_2 are individuals.

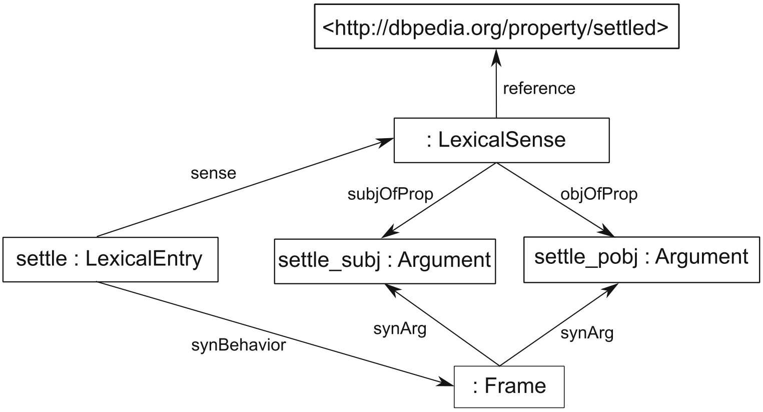


Fig. 5 Linking a verb's subcategorisation to an ontology property

referenced ontology entity which accounts for the specific lexico-semantic connotations that the lexical entry introduces.

This relationship does not state that the meaning of the entity and the lexicalisation are equivalent. It rather indicates that there are linguistic contexts in which the lexical entry is used with this meaning and conversely this entity may sometimes be lexicalised with the given lexical entry. Therefore, it follows that the sense object belongs neither truly to the lexicon nor the ontology but instead acts as a bridge between the two and represents an *underspecified* relationship between the actual uses of a given lexical entry and the ontology entity it refers to. The contexts in which it is legitimate to interpret the lexical entry as representing the meaning of a given concept can be further constrained by attaching additional contextual and pragmatic conditions at the sense object. For example, we might express that the verb "fressen" (in German) is typically used for animals as agents, while "essen" (in German) is used for human agents.

The *lemon* model represents subcategorisation with a frame object that can have a number of syntactic arguments indicated with the `synArg` property, which may be sub-typed to indicate specific roles played by syntactic arguments. The link to the ontology is then represented by linking the sense to each of these arguments with `subjOfProp`, `objOfProp` and `isA` (used for classes which we model as unary predicates). An example of such a mapping for the subcategorisation "X was settled in Y" is as follows, where "X" is the subject entity and "Y" the object entity.[8]

[8] Here we use our *lemon*-aligned version of LexInfo, as ISOcat does not currently have many data categories for subcategorisation. Note that it is not strictly necessary to define these properties as subproperties of *lemon*, as they are already published as such.

```
@prefix lemon: <http://www.monnet-project.eu/lemon#> .
@prefix dbpedia: <http://dbpedia.org/property/> .
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .
:settle lemon:syntacticBehavior [
      lexinfo:subject :settle_subj ;
      lexinfo:prepositionalObject :settle_pobj ] ;
    lemon:sense [
      lemon:reference dbpedia:settled ;
      lemon:subjOfProp :settle_subj ;
      lemon:objOfProp :settle_pobj ] .
```

This example (illustrated in Fig. 5) shows how we define a subcategorisation frame for a verb, in this case by indicating its arguments with ISOcat data categories that are specified as sub-properties of `synArg`. These arguments are then also linked to the sense, and indicated as the subject and object of the property referred to by this sense. In this way we can precisely describe the correspondence between a lexical entry and an ontology property or class.

These examples cover only a small part of the model, a full technical manual is available at http://lexinfo.net/lemon-cookbook.pdf, which also covers other features of the *lemon* model including:

- Mapping with ternary (e.g., "donative") and other multiple-argument subcategorisations
- Relations between lexical entries
- Representing syntax trees
- Combining syntax trees with subcategorisations
- Specifying sense contexts and conditions
- Assigning subject fields to lexica
- Asserting global lexicon constraints
- Providing compact representations of inflection and agglutination

4 Using *lemon*

In general, the most important step for instantiating a *lemon* lexicon is identifying the sets of data categories that we wish to (re-)use in a specific instantiation in *lemon*. In contrast to other formats, there is no need to create a data category selection file to state which set of data categories are used in a given file. Instead, as each data category is uniquely identified by a URI, they can simply be used without prior identification. As such, in order to use *lemon* to represent a lexicon, the following steps should be carried out:

1. Identify which properties/relations you wish to use to define specific linguistic concepts.

2. Look up appropriate data categories from some source (e.g., ISOcat) and include them in the lexicon by stating them as subproperties of the appropriate *lemon* property.
3. If there are properties that are not covered by any standardised source, you may define them yourself. The URIs of the properties should be dereferencable, i.e., an RDF description of it should be available at the address.
4. Align the data source with the *lemon* core model. For example, it is commonly necessary to identify how canonical and alternative labels are identified in the source.
5. Publish the lexicon as an RDF/XML document. The URIs for each entry should be resolvable at the given URI.

We have also created a number of tools to support the creation and use of *lemon* models. Firstly, as *lemon* is developed from LMF, we have implemented methods supporting the conversion from and to the LMF format.[9] We are also working on import/export facilities to a number of other formats including XLIFF and TBX. We have also developed a web interface that allows people to upload and modify *lemon* models.[10] This service can also create *lemon* models automatically from OWL ontology files. It works by extracting the labels for each concept from the ontology through an annotation such as `rdfs:label` or `skos:prefLabel`. Otherwise the system uses the URI of the entity to attempt to obtain a label for the concept, for example by de-camel-casing the fragment. Then, the system applies a tokeniser and then a part-of-speech tagger and uses this to create the core structure of the *lemon* entry. Finally, as in the work of Cimiano et al. (2011), we apply syntactic analysis to infer the subcategorisation frame of the term and the phrase structure, if desired.

Another important aspect of *lemon* is that it is based on established Semantic Web technologies and hence a number of tools already exist to enable the sharing and integration of models on the Web. For example, several Semantic Web search engines (SWSE) maintain an index of all RDF data published on the Semantic Web. As such, if someone chooses to publish their *lemon* lexicon on the Web, it can be submitted to a SWSE. It is then easy for other users to find lexica and share them, as SWSEs allow for particular properties to be queried. For example, querying for all triples using the property `lemon:writtenRep` and value `"cat"` would retrieve all *lemon* lexica that use the word “cat.” In addition there is a Java API for handling *lemon* documents and converting them to other formats.[11]

5 Conclusion

In this paper, we have proposed the *lemon* model as a format allowing to bring together two “worlds”: the world of ontological knowledge, which builds on Web-based knowledge representation formalisms such as OWL and RDF(S) and the

[9] Available as part of the *lemon* Java API.

[10] Available at http://monnetproject.deri.ie/lemonsource.

[11] https://github.com/jmccrae/lemon.api/

world of lexical and linguistic resources, which builds on various standards to represent lexica (e.g. LMF) and terminological resources (e.g. TBX). The *lemon* model has been designed to describe ontology lexica, which specify how the concepts in a given ontology are lexicalised, thus providing NLP applications with the required background knowledge to interpret natural language with respect to an ontology. The *lemon* RDF model allows for ontology-lexica to be shared and interlinked on the Web and integrated with ontologies in the Web Ontology Language (OWL). This allows for greater reuse of existing data than is possible using current lexicon formats and the integration with ontologies allows for deeper semantic relationship than a lexicon alone would provide. The *lemon* model is based on several existing resources, in particular LMF, SKOS, LIR, and LexInfo and as such maintains a high degree of compatibility with these models. However, its focus on compactness and expressivity allows for a large amount of linguistic information to be represented, while keeping the model fairly small. It maintains a high degree of flexibility and extensibility by the use of data categories, allowing the model to act as a lexicon meta-model as well as a format in its own right. We have also discussed tools that facilitate easy usage of the *lemon* model and interaction with existing standards in both lexicography and the Semantic Web. As such we hope that this will lead to a consensus model for the exchange of lexica on the Semantic Web. We are currently working towards building a community that can continue to develop and apply the model.[12]

Acknowledgments The *lemon* model and associated tools have been developed in the context of the Monnet project, which is funded by the European Union FP7 program under grant number 248458, the CITEC excellence initiative funded by the DFG (Deutsche Forschungsgemeinschaft), the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2), the Spanish Ministry of Science and Innovation within the Juan de la Cierva program and the Spanish Project BabelData (TIN2010-17550). We would like to thank the following people for their contributions and advice: Axel Polleres (DERI), Antoine Zimmermann (DERI), Dimitra Anastasiou (CNGL), Susan Marie Thomas (SAP), Christina Unger (CITEC), Sue Ellen Wright (Kent State University), Menzo Windhouwer (Universiteit van Amsterdam).

References

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., et al. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics, 25*(1), 25.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In: *Proceedings of the 6th international Semantic Web conference (ISWC)* (pp. 722–735).

Baker, C., Fillmore, C., & Lowe, J. (1998). The Berkeley FrameNet project. In: *Proceedings of the 36th annual meeting of the association for computational linguistics (ACL)* (pp. 86–90).

Beckett, D., & Berners-Lee, T. (2008). Turtle–Terse RDF triple language. http://www.w3.org/TeamSubmission/turtle/. Accessed October 19, 2010.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data—the story so far. *International Journal on Semantic Web and Information Systems, 5*, 1–22.

Buitelaar, P. (2010). Ontology-based Semantic Lexicons: Mapping between terms and object descriptions. In: C.R. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari & L. Prevot (Eds.), *Ontology and the Lexicon* (pp. 212–223). Cambridge: Cambridge University Press.

[12] http://www.w3.org/community/ontolex/.

Buitelaar, P., Cimiano, P., Haase, P., & Sintek, M. (2009). Towards linguistically grounded ontologies. In: *Proceedings of the European Semantic Web conference (ESWC)* (pp. 111–125).

Chiarcos, C. (2010). Grounding an ontology of linguistic annotations in the data category registry. In: *Proceedings of the international conference on language resource and evaluation (LREC)* (pp. 37–40).

Cimiano, P., Buitelaar, P., McCrae, J., & Sintek, M. (2011). LexInfo: A declarative model for the lexicon-ontology interface. *Journal of Web Semantics, 9*(1), 29–51.

Farrar, S., & Langendoen, D. (2003). Markup and the GOLD ontology. In: *Proceedings of workshop on digitizing and annotating text and field recordings* (pp. 845–862).

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT press.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., et al. (2006). Lexical markup framework (LMF). In: *Proceedings of the international conference on language resource and evaluation (LREC)* (pp. 233–236).

Grishman, R., Macleod, C., & Meyers, A. (1994). COMLEX syntax: Building a computational lexicon. In: *Proceedings of the 15th international conference on computational linguistics (COLING)* (pp. 268–272).

Isaac, A., Phipps, J., & Rubin, D. (2009). SKOS use cases and requirements. http://www.w3.org/TR/2009/NOTE-skos-ucr-20090818/, Accessed 19 October 2010.

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., & Wright, S. (2008). ISOcat: Corralling data categories in the wild. In: *Proceedings of the international conference on language resource and evaluation (LREC)* (pp. 887–891).

Kifer, M. (2008). Rule interchange format: The framework. In: *Proceedings of the 2nd international conference on web reasoning and rule systems* (pp. 1–11).

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities, 31*(2), 91–113.

Kim, J., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics, 19*(1), 180–182.

McCrae, J., Spohr, D., & Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In: *Proceedings of the 8th extended Semantic Web conference (ESWC-11)* (pp. 245–259).

Miles, A., & Bechhofer, S. (2009). SKOS simple knowledge organization system reference. http://www.w3.org/TR/skos-reference/. Accessed October 19, 2010.

Montiel-Ponsoda, E., Aguado de Cea, G., Gómez Pérez, A., & Peters, W. (2010). Enriching ontologies with multilingual information. In B. Boguraev, J. Tait, & M. Palmer (Eds.), *Natural language engineering* (pp 1–27). Cambridge: Cambridge University Press.

Reymonet, A., Thomas, J., & Aussenac-Gilles, N. (2007). Modelling ontological and terminological resources in OWL-DL. In: *Proceedings of the 6th international Semantic Web conference (ISWC)* (pp. 415–425).

Romary, L. (2010). Standardization of the formal representation of lexical information for NLP. In: *Dictionaries: An international encyclopedia of lexicography*. Mouton de Gruyter. http://arxiv.org/abs/0911.5116v1.

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The semantic web revisited. *IEEE Intelligent Systems, 21*(3), 96–101.

Scheffczyk, J., Pease, A., & Ellsworth, M. (2006). Linking FrameNet to the suggested upper merged ontology. In: *Formal ontology in information systems (FOIS-2006)* (pp. 289–300).

Van Assem, M., Gangemi, A., & Schreiber, G. (2006). Conversion of WordNet to a standard RDF/OWL representation. In: *Proceedings of the fifth international conference on language resources and evaluation (LREC)* (pp. 237–242).

Vossen, P. (1998). EuroWordNet: A multilingual database with lexical semantic networks. *Computational Linguistics, 25*(4), 628–630.

Vossen, P., Bloksma, L., Peters, W., Kunze, C., Wagner, A., Pala, K., et al. (1999). *Extending the Inter-Lingual-Index with new concepts*. Deliverable 2D010, EuroWordNet, LE2-4003.