320

Knowl. Org. 40(2013)No.5
E. W. De Luca. Extending the Linked Data Cloud with Multilingual Lexical Linked Data

# Extending the Linked Data Cloud
# with Multilingual Lexical Linked Data

Ernesto William De Luca

University of Applied Sciences Potsdam,
Friedrich-Ebert-Str. 4, 14467 Potsdam, Germany,
<deluca@fh-potsdam.de>

Ernesto William De Luca is Professor of Information Science at the University of Applied Sciences Potsdam. His research areas include semantic web technologies, recommender systems, and information retrieval. He has authored more than 80 papers at national and international conferences, books, and journals in these fields. He has organized a large number of workshops and served as programme committee member at top level conferences.

**ABSTRACT:** A lot of information that is already available on the Web, or retrieved from local information systems and social networks, is structured in data silos that are not semantically related. Semantic technologies make it apparent that the use of typed links that directly express their relations are an advantage for every application that can reuse the incorporated knowledge about the data. For this reason, data integration, through reengineering (e.g., triplify) or querying (e.g., D2R), is an important task in order to make information available for everyone. Thus, in order to build a semantic map of the data, we need knowledge about data items itself and the relation between heterogeneous data items. Here we present our work of providing Lexical Linked Data (LLD) through a meta-model that contains all the resources and gives the possibility to retrieve and navigate them from different perspectives. After giving the definition of Lexical Linked Data, we describe the existing datasets we collected and the new datasets we included. Here we describe their format and show some use cases where we link lexical data, and show how to reuse and inference semantic data derived from lexical data. Different lexical resources (MultiWordNet, EuroWordNet, MEMODATA Lexicon, the Hamburg Methaphor Database) are connected to each other towards an Integrated Vocabulary for LLD that we evaluate and present.

## 1.0 Introduction

With the advent of Linked Open Data (LOD, http://linkeddata.org/), more and more different resources are interconnected and shared on the Web. The idea of Linked Open Data is to connect and share data, information, or knowledge following semantic web principles like using URIs and RDF descriptions. While most linked data concentrate on linking facts, like music, movies, and geo- or demographic information, we believe that one important task is to connect language resources in order to support the process of language engineering. We also believe that natural language processing plays an important role in order to achieve this goal. language engineering involves the development and application of software systems that perform tasks concerning the processing of human natural language (Cunningham 1999). Different tools have been designed, constructed, and are used for tasks like translation, language teaching, information extraction, and indexing.

Other, more intangible "language engineering tools" are language resources. Language resources are essential components of language engineering, containing a wide range of linguistic information with different degrees of complexity. These linguistic resources are sets of language data and descriptions in machine readable form, used for building, improving, or evaluating natural language and speech systems or algorithms. (Cole et al. 1997) give a brief overview about the various types of language resources, i.e., written and spoken language corpora, lexicons, and terminological databases.

### 1.1 Lexical resources

In the following, we concentrate on lexical resources that provide linguistic information about words. This information can be represented in very diverse data structures, from simple lists to complex repositories with many types of linguistic information and relations attached to each

Knowl. Org. 40(2013)No.5
E. W. De Luca. Extending the Linked Data Cloud with Multilingual Lexical Linked Data

321

entry, resulting in network-like structures. Lexical re-sources are used in natural language processing, for example, to obtain descriptions and usage examples of different word senses. Different word senses refer to different concepts, and concepts can be distinguished from each other not only by their definitions or "glosses," but also by their specific relations to other concepts. Such disambiguating relations are intuitively used by humans. However, if we want to automate the process of distinguishing between word senses (word sense disambiguation), we have to use resources that provide appropriate knowledge, i.e., sufficient information about the usage context of a word. One of the most important resources available for this purpose is WordNet (Fellbaum 1998) and its multilingual variants, including MultiWordNet (Pianta et al. 2002), and EuroWordNet (Vossen 1999).

## 1.2 Lexical linked data

Because the Web is evolving from a global information space of linked documents to one where both documents and data are linked, we agree that a set of best practices for publishing and connecting structured data on the web, known as linked data, is needed. The Linked Open Data (LOD) project (Bizer et al. 2009) is bootstrapping the Web of Data by converting it into RDF and publishing existing available "open datasets." In addition, LOD datasets often contain natural language texts, which are important to link and explore data not only in a broad LOD cloud vision, but also in localized applications within large organizations that make use of linked data (Baldassarre et al. 2010; Nuzzolese et al. 2011).

The combination of natural language processing and semantic web techniques has become important in order to exploit lexical resources directly represented as linked data. One of the major examples is the WordNet RDF dataset (van Assem et al. 2006), which provides concepts (called synsets), each representing the sense of a set of synonymous words (Gangemi et al. 2003). It has a low level of concept linking, because synsets are linked mostly by means of taxonomic relations, while LOD data are mostly linked by means of domain relations, such as parts of things, ways of participating in events or socially interacting, topics of documents, temporal and spatial references, etc. (Nuzzolese et al. 2011). An example of inter-linking lexical resources like Framenet (http://framenet.icsi.berkeley.edu/) (Baker et al. 1998) to the LOD Cloud is given in Gangemi and Presutti (2010). They create a LOD dataset that provides new possibilities to the lexical grounding of semantic knowledge (Gangemi and Presutti 2010), and boosts the "lexical linked data" section of LOD, by linking, for example, FrameNet to other LOD datasets such as WordNet RDF (van Assem et al. 2006).

## 1.3 Knowledge organization and linked data

In recent years, the future of knowledge organization on the web has been presented and discussed with the special focus on linked data at ISKO UK 2010, among other venues. This was the most successful ISKO UK event to date. Several speakers referred to Tim Berners-Lee's thoughts on linked data principles, which he first wrote about in July 2006. The idea is simple: you can add value to your information by linking it to that of others (http://www.iskouk.org/events/linked data sep2010.htm). Already at ISKO UK 2009, David Crystal (2009) gave a keynote on "Semantic targeting: past, present and future," the evolution of the linguistic approach to content analysis which he has been developing over the past 20 years. It began with the knowledge management taxonomy used for the Cambridge family of general encyclopedias and followed its transformation into an Internet taxonomy, with applications in automatic document classification, search engine assistance, e-commerce, online advertising, and Internet security. The recent developments focus on advertising, a field which has seen ideas develop from simple keyword analysis to contextual advertising and now to semantic targeting. These notions also included the ways of handling site sensitivity, sentiment, intention, and cultural localization. Other work has been done in this area in a multilingual context, comparing two indexing vocabularies for image retrieval (Ménard 2010) or tagging Medline abstracts semantically for enhancing information access (Ibekwe-SanJuan 2009). The exploitation of data in the cloud has been also addressed by another presentation given by Paul Miller in his talk "Exploiting data in the cloud," where the increase of the quantity of data has been discussed and the need of standardization (e.g., through linked data) for a better structuring process has been recognized. Knowledge organization systems are needed in order to better access information and knowledge using the new technologies like RDF representations (Scott and Smethurst 2009), social tagging (Matthews et al. 2010), relational databases (Sarinder et al. 2010), or faceted classification approaches (Ménard 2010).

## 2.0 RDF/OWLWordNet and RDF/OWLEuroWordNet

The WordNet Princeton (Fellbaum 1998) has already been converted into an OWL format as described in W3C (2004) using the OWL-DL sublanguage. This representation in RDF/OWL is based on the WordNet data model shown in Figure 1.

When we compare the original Princeton WordNet synset (having only word senses) with the OWL representation, we can see that the RDF/OWL schema (in its full
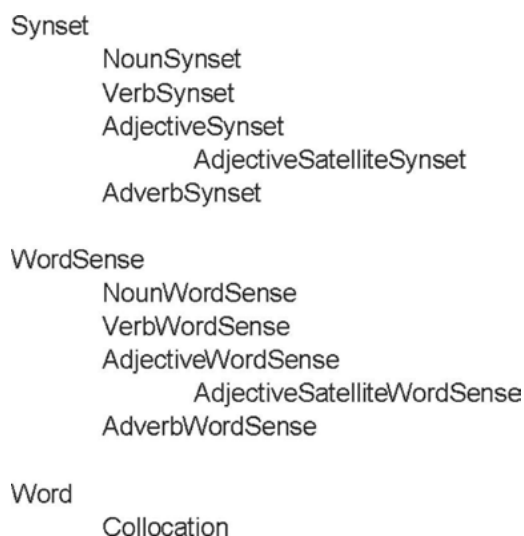
322

Knowl. Org. 40(2013)No.5
E. W. De Luca. Extending the Linked Data Cloud with Multilingual Lexical Linked Data

```
Synset
      NounSynset
      VerbSynset
      AdjectiveSynset
            AdjectiveSatelliteSynset
      AdverbSynset

WordSense
      NounWordSense
      VerbWordSense
      AdjectiveWordSense
            AdjectiveSatelliteWordSense
      AdverbWordSense

Word
      Collocation
```

*Figure 1.* OWL Hierarchy of Princeton WordNet.

version) has three main classes: Synset, Word, and Word-Sense. The basic version contains only the Synset class. The two classes Synset and WordSense further have four subclasses, which are based on the distinction of lexical groups; these are NounSynset, VerbSynset, AdjectiveSynset (which has another subclass, AdjectiveSatelliteSynset), and AdverbSynset. The Word class holds the subclass Collocation, which denotes terms that are composed of two or more words. In order to disambiguate the meanings of each instance of a synset, WordSense and Word have a unique URI that can be used for retrieving words and word senses independently from the synsets. This property was not available in the original version of WordNet. The URIs provide some information about the entity meaning and are built with patterns similar to:

wn20instances: + synset- + lexical form- + type- + sense number.

For example, if we want to retrieve the fourth word sense of the word "bank," we would get a URI like: "http://www.w3.org/2006/03/wn/wn20/instances/synsetbank-noun-4."

The properties of the RDF schema are divided into three kinds of relations:

1. those that relate two synsets to each other (e.g., hyponymOf)
2. those that relate two word senses to each other (e.g., antonymOf)
3. and a set of properties that give informations on entities (e.g., XML Schema datatypes like xsd:string as it is used in synsetId).

In order to avoid redundancy, only relations in one transitive direction (e.g., hyponymOf and not hypernymOf) are listed, the others can be retrieved with the owl:inverseOf property implemented in the RDF Schema (see the Figures 4 and 5). Altogether there are 27 relations implemented in the RDF/OWL representation of WordNet. The instances of all classes and properties are separated in several data files, one for the synsets, one for the WordSenses and Words, and one for each relation. Although the RDF Schema is used to describe most class and property definitions, there are several OWL statements integrated in the schema to provide better semantic descriptions, like checking the correctness of the data or defining inverse relations. For these statements software have to support the OWL DL standard in order to store and query the data.

### 3.0 RDF/OWL EuroWordNet

Because of the different problems related to WordNet and its variants, we decided to convert it into an RDF/OWL representation (see below), in order to enable the development of more flexible revision methods. In EuroWordNet, one synset contains all related word senses, synonyms and relations to other synsets and to the Inter-Lingual-Index. This information had to be prepared for inclusion in the appropriate RDF Schema and reorganized for a new data representation. The decision of converting EuroWordNet was also based on the need of extending it (because not all meanings are covered) with other resources. Furthermore, since most domain-specific ontologies are represented in OWL and a WordNet monolingual RDF/OWL representation has already been implemented, a EuroWordNet conversion would add multilingual capabilities to these resources. Therefore, we converted EuroWordNet into an RDF/OWL representation based on the work presented in van Assem et al. (2004). Since EuroWordNet has several relations and a structure that are different from the Princeton WordNet, several steps were required to adapt the data to the RDF/OWL Schema of WordNet and to extend this RDF Schema with the new relations. We first analyzed the requirements for EuroWordNet and adapted the WordNet RDF Schema to a multilingual representation of EuroWordNet. Then, we converted the EuroWordNet relations into OWL properties and extended the ontology with two domain ontologies (De Luca et al. 2007).

In the following, we describe the steps of this conversion and the problems that arose in more detail. An example on how to add ontologies is later given in Section 4.1 based on the OWL pizza and travel ontologies. The RDF/OWL-EuroWordNet Representation has also been extended with linguistic data included in the Hamburg Metaphor Database (HMD) in Section 4.

Knowl. Org. 40(2013)No.5

323

E. W. De Luca. Extending the Linked Data Cloud with Multilingual Lexical Linked Data

*3.1 Conversion of EuroWordNet in RDF/OWL*

The steps required to convert EuroWordNet in RDF/OWL can be subdivided into:

– Analysis of the requirements for EuroWordNet
– Adaptation of The WordNet RDF-Schema to EuroWordNet
– Multilinguality
– OWL Property Conversion
– OWL Domain Extension

van Assem et al. (2004) distinguish Word and WordSense in their datamodel for two reasons. First of all, several relations are defined for word senses and synsets, and WordNet uses this distinction in its database. Secondly, for the sake of ontological clarity, they assume that synsets include word senses, in order to partition the logical space of the lexicon (words as forms or meanings, and synsets as clusters of word senses by abstracting their distributional context). Agreeing with their model, we adapted their schema to convert EuroWordNet, applying this assumptions also for a multilingual task. An example of an OWL-EuroWordNet synset is given in Figure 2. Here the word sense "bank" is shown within its synset (and synsetId), WordSense, Word, and synonyms (containsWordSense).

Analyzing the structures of EuroWordNet and of the OWL representation of WordNet, we could recognize that some relations are supported in both versions (see Figure 4). Since EuroWordNet contains relations and properties that are not supported in the WordNet OWL representation, we had to adapt the RDF-OWL Schema to our needs in order to cover these gaps. Therefore, we cre-

ated and stored new RDF structures, containing these new relations and thus extended the WordNet OWL implementation (see Figure 5). Some other properties that are covered in the WordNet OWL representation (e.g., the property tagCount used in the WordSense OWL declaration) are not available in EuroWordNet, so that we could not consider them.

Because we also tried to avoid redundancy as discussed in van Assem et al. (2004), we decided to delete the relations in EuroWordNet, which have an inverse form. We compared them updating their inverse relation property, where necessary (see the Figures 4 and 5). This means, if an instance of an inverse relation of a distinct instance is present, the instance can be deleted. If no inverse instance is available, the instance is added as inverse relation instance. An example is the hyperonym-hyponym relation resulting in a simple hyponymOf relation. Here the hyponym relation was available in both representations, but its name had to be changed from has hyperonym EuroWordNet format to the hyponymOf OWLWordNet format (see Figure 3). Another point to consider was that EuroWordNet also contains different relations belonging to the same "upper relation description" (e.g., ROLE AGENT, ROLE INSTRUMENT, ROLE LOCATION, etc. belonging to ROLE), because of their similar functionality. We decided in this case to merge them all into the same "Upper Relation" RDF file. A similar decision was done in the Princeton Conversion within the pertainsTo relation. The complete mapping between EuroWordNet, OWL-WordNet and OWL-EuroWordNet relations is given in Figure 4. The relations that are only available in EuroWordnet and have been included as new RDF/OWL-EuroWordNet are shown in Figure 5.

```
<ewn20schema:NounSynset rdf:about="&ewn20instances;synset-bank-noun-1"
    rdfs:label="bank">
    <ewn20schema:synsetId>102690337</ewn20schema:synsetId>
</ewn20schema:NounSynset>
<ewn20schema:Word rdf:about="&ewn20instances;word-bank"
    ewn20schema:lexicalForm="bank"/>
<ewn20schema:NounWordSense rdf:about="&ewn20instances;wordsense-bank-noun-1"
    rdfs:label="bank">
    <ewn20schema:word rdf:resource="&ewn20instances;word-bank"/>
</ewn20schema:NounWordSense>
<rdf:Description rdf:about="&ewn20instances;synset-bank-noun-1">
    <ewn20schema:containsWordSense rdf:resource="&ewn20instances;wordsense-bank-noun-1"/>
    <ewn20schema:containsWordSense rdf:resource="&ewn20instances;wordsense-bank_building-noun-1"/>
</rdf:Description>
```

*Figure 2.* RDF/OWL-EuroWordnet synset Example.

```
<rdf:Description rdf:about="&ewn20instances;synset-bank-noun-1">
    <ewn20schema:hyponymOf rdf:resource="&ewn20instances;synset-deposit-noun-1"/>
</rdf:Description>
```

*Figure 3.* RDF/OWL-EuroWordnet hyponymOf Example for the word "bank".

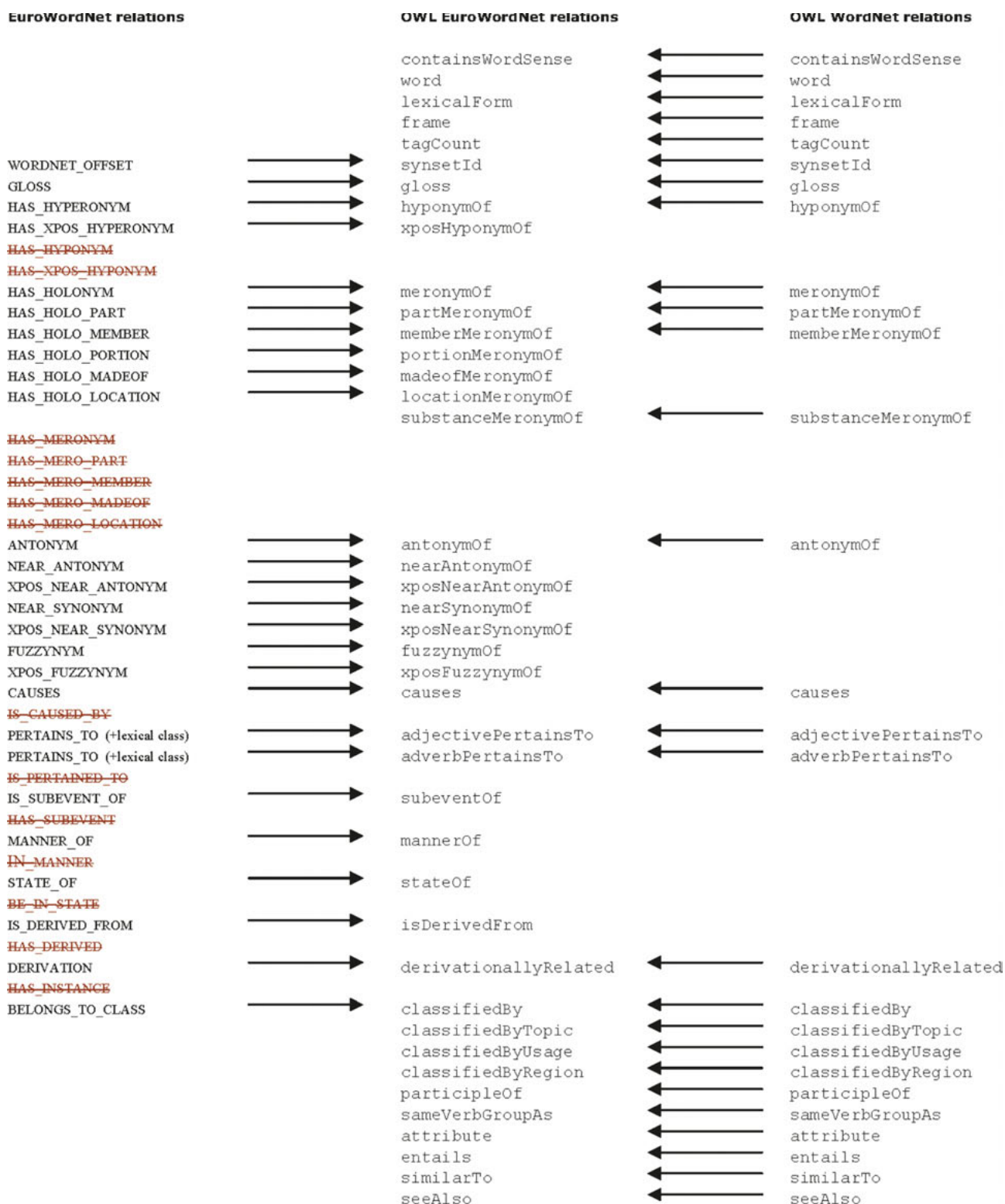| EuroWordNet relations | | OWL EuroWordNet relations | | OWL WordNet relations |
|---|---|---|---|---|
| | | containsWordSense | ← | containsWordSense |
| | | word | ← | word |
| | | lexicalForm | ← | lexicalForm |
| | | frame | ← | frame |
| | | tagCount | ← | tagCount |
| WORDNET_OFFSET | → | synsetId | ← | synsetId |
| GLOSS | → | gloss | ← | gloss |
| HAS_HYPERONYM | → | hyponymOf | ← | hyponymOf |
| HAS_XPOS_HYPERONYM | → | xposHyponymOf | | |
| ~~HAS_HYPONYM~~ | | | | |
| ~~HAS_XPOS_HYPONYM~~ | | | | |
| HAS_HOLONYM | → | meronymOf | ← | meronymOf |
| HAS_HOLO_PART | → | partMeronymOf | ← | partMeronymOf |
| HAS_HOLO_MEMBER | → | memberMeronymOf | ← | memberMeronymOf |
| HAS_HOLO_PORTION | → | portionMeronymOf | | |
| HAS_HOLO_MADEOF | → | madeofMeronymOf | | |
| HAS_HOLO_LOCATION | → | locationMeronymOf | | |
| | | substanceMeronymOf | ← | substanceMeronymOf |
| ~~HAS_MERONYM~~ | | | | |
| ~~HAS_MERO_PART~~ | | | | |
| ~~HAS_MERO_MEMBER~~ | | | | |
| ~~HAS_MERO_MADEOF~~ | | | | |
| ~~HAS_MERO_LOCATION~~ | | | | |
| ANTONYM | → | antonymOf | ← | antonymOf |
| NEAR_ANTONYM | → | nearAntonymOf | | |
| XPOS_NEAR_ANTONYM | → | xposNearAntonymOf | | |
| NEAR_SYNONYM | → | nearSynonymOf | | |
| XPOS_NEAR_SYNONYM | → | xposNearSynonymOf | | |
| FUZZYNYM | → | fuzzynymOf | | |
| XPOS_FUZZYNYM | → | xposFuzzynymOf | | |
| CAUSES | → | causes | ← | causes |
| ~~IS_CAUSED_BY~~ | | | | |
| PERTAINS_TO (+lexical class) | → | adjectivePertainsTo | ← | adjectivePertainsTo |
| PERTAINS_TO (+lexical class) | → | adverbPertainsTo | ← | adverbPertainsTo |
| ~~IS_PERTAINED_TO~~ | | | | |
| IS_SUBEVENT_OF | → | subeventOf | | |
| ~~HAS_SUBEVENT~~ | | | | |
| MANNER_OF | → | mannerOf | | |
| ~~IN_MANNER~~ | | | | |
| STATE_OF | → | stateOf | | |
| ~~BE_IN_STATE~~ | | | | |
| IS_DERIVED_FROM | → | isDerivedFrom | | |
| ~~HAS_DERIVED~~ | | | | |
| DERIVATION | → | derivationallyRelated | ← | derivationallyRelated |
| ~~HAS_INSTANCE~~ | | | | |
| BELONGS_TO_CLASS | → | classifiedBy | ← | classifiedBy |
| | | classifiedByTopic | ← | classifiedByTopic |
| | | classifiedByUsage | ← | classifiedByUsage |
| | | classifiedByRegion | ← | classifiedByRegion |
| | | participleOf | ← | participleOf |
| | | sameVerbGroupAs | ← | sameVerbGroupAs |
| | | attribute | ← | attribute |
| | | entails | ← | entails |
| | | similarTo | ← | similarTo |
| | | seeAlso | ← | seeAlso |

*Figure 4.* EuroWordnet, RDF/OWL-EuroWordNet and RDF/OWL-WordNet relations. The crossed out relations have been removed, since they are inverse relations to the existing ones.
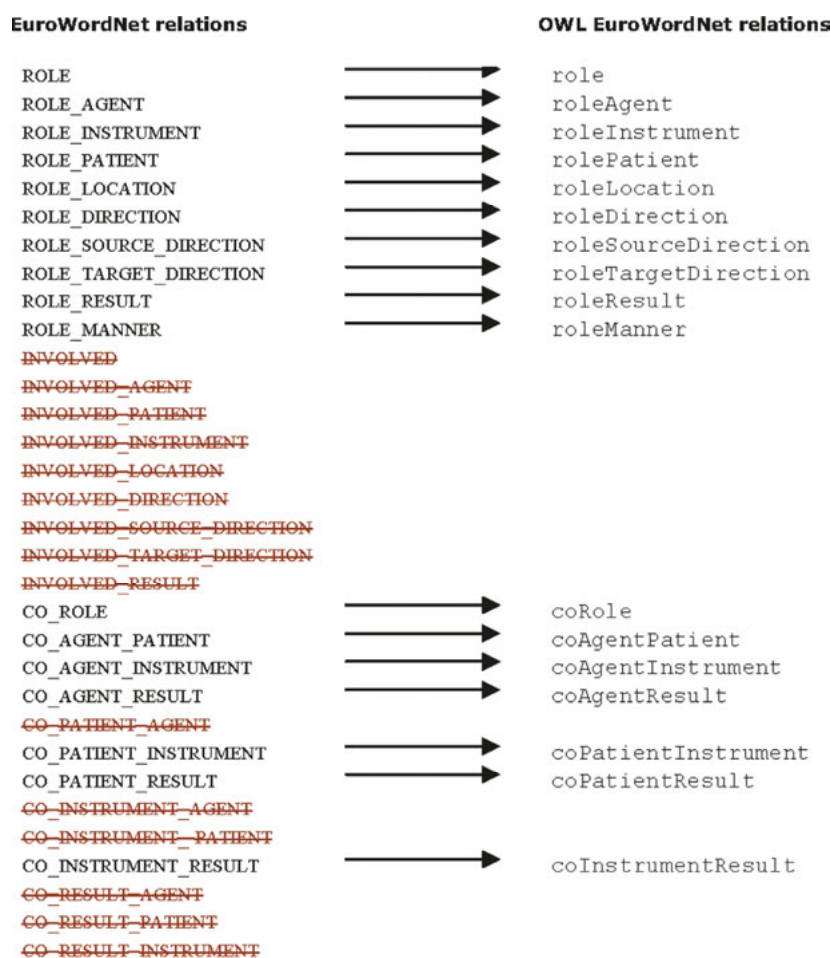
Knowl. Org. 40(2013)No.5
E. W. De Luca. Extending the Linked Data Cloud with Multilingual Lexical Linked Data

325

**EuroWordNet relations**

**OWL EuroWordNet relations**

| EuroWordNet relations | | OWL EuroWordNet relations |
|---|---|---|
| ROLE | → | role |
| ROLE_AGENT | → | roleAgent |
| ROLE_INSTRUMENT | → | roleInstrument |
| ROLE_PATIENT | → | rolePatient |
| ROLE_LOCATION | → | roleLocation |
| ROLE_DIRECTION | → | roleDirection |
| ROLE_SOURCE_DIRECTION | → | roleSourceDirection |
| ROLE_TARGET_DIRECTION | → | roleTargetDirection |
| ROLE_RESULT | → | roleResult |
| ROLE_MANNER | → | roleManner |
| ~~INVOLVED~~ | | |
| ~~INVOLVED_AGENT~~ | | |
| ~~INVOLVED_PATIENT~~ | | |
| ~~INVOLVED_INSTRUMENT~~ | | |
| ~~INVOLVED_LOCATION~~ | | |
| ~~INVOLVED_DIRECTION~~ | | |
| ~~INVOLVED_SOURCE_DIRECTION~~ | | |
| ~~INVOLVED_TARGET_DIRECTION~~ | | |
| ~~INVOLVED_RESULT~~ | | |
| CO_ROLE | → | coRole |
| CO_AGENT_PATIENT | → | coAgentPatient |
| CO_AGENT_INSTRUMENT | → | coAgentInstrument |
| CO_AGENT_RESULT | → | coAgentResult |
| ~~CO_PATIENT_AGENT~~ | | |
| CO_PATIENT_INSTRUMENT | → | coPatientInstrument |
| CO_PATIENT_RESULT | → | coPatientResult |
| ~~CO_INSTRUMENT_AGENT~~ | | |
| ~~CO_INSTRUMENT_PATIENT~~ | | |
| CO_INSTRUMENT_RESULT | → | coInstrumentResult |
| ~~CO_RESULT_AGENT~~ | | |
| ~~CO_RESULT_PATIENT~~ | | |
| ~~CO_RESULT_INSTRUMENT~~ | | |

*Figure 5.* EuroWordnet and RDF/OWL-EuroWordNet relations missing in RDF/OWLWordNet. The crossed out relations have been removed, since they are inverse relations to the existing ones.

Since EuroWordNet is a multilingual resource (and not a monolingual like WordNet), we had to create for every language a unique set of files (containing the language-dependent synsets and relations). Every file of this set was additionally tagged with the name of the corresponding language (e.g., for English: eurowordnet-english-synset.rdf, eurowordnet-english-wordsensesandwords.rdf, euroword-netenglish-hyponymOf.rdf, etc.). We used the available EuroWordNet Inter-Lingual-Index that contains synsets, having the same indentifier (synsetId) for all word meanings in all languages and an illustrative gloss. Because of the redundancy problem already described above, we decided to maintain only the gloss information included in the Inter-Lingual-Index deleting the word senses and synsetIds (already included in the english conversion). Therefore, the gloss entries were extracted and stored in a separate RDF file. Depending on this decision, the synsets of all languages are connected to another through the same identifier (synsetId) describing the same concept in different languages, instead of the Inter-Lingual-Index entries.

## 4.0 Interlinking RDF/OWL-EuroWordNet with other Lexical Resources

### 4.1 Interlinking RDF/OWL-EuroWordNet with the pizza.owl and travel.owl domain-ontologies

After a first conversion of EuroWordNet in an OWL representation, we decided to try to interlink it with two OWL ontologies (pizza.owl and travel.owl). For integrating them in our EuroWordnet OWL representation, we analyzed their hierarchy of classes that were also built in OWL DL. Every term is declared as a class (owl:Class), and every underlying term as a subclass (rdfs:subClassOf). There are additional restrictions, e.g., owl:disjointWith or owl:someValuesFrom statements. But there are no additional properties (except the OWL DL statements defined).

In order to extend EuroWordNet with these ontologies, we used a two steps approach:

326

Knowl. Org. 40(2013)No.5
E. W. De Luca. Extending the Linked Data Cloud with Multilingual Lexical Linked Data

1. Conversion from the OWL format to the EuroWord-Net OWL format
2. Integration of the converted data in the EuroWordNet OWL hierarchy

First of all, we converted all classes (owl:class) into RDF/OWL synset classes (e.g., ewn20Schema:NounSynset), so that they were easier to add into the OWLEuroWordNet hierarchy. Using some of the merging methods (De Luca and Nürnberger 2006a), we started a query for finding the synset we wanted to use as hyperonym of the domain ontology to be added (in this example "pizza"). We first explored the hierarchy with the LexiRes RDF/OWL tool. Then we tried to disambiguate the word senses of the searched word ("pizza") in order to find the correct synset to be extended. After having found the correct synset, we manually merged the complete converted domain ontology under the appropriate hyperonym (synset). In this case, we could enlarge the EuroWordNet coverage with domain-specific terms. The extension can be easily recognized (and deleted if needed), because we added to the synsetId the name of the domain ontology followed by the number of word senses (see Figure 6). The pizza.owl ontology, for example, also has a language description for every class (e.g., xml:lang="en"), so that we could add it to the correct set of language files, if available. The same procedure was applied to the travel.owl ontology. A more detailed description of this extension work is given in De Luca et al. (2007).

### 4.2 Interlinking RDF/OWL-EuroWordNet with the Hamburg Metaphor Database

The RDF/OWL-EuroWordNet representation has also been interlinked with data included in the Hamburg Metaphor Database (HMD), a relational database of French and German corpus attestations containing metaphorical expressions (Lönneker-Rodman 2008). In the HMD, each metaphor is manually analyzed and annotated at several levels; among other lexical features, HMD provides references to EuroWordNet synsets. In addition, conceptual information is indicated in terms of domain labels from the Berkeley Master Metaphor List (Lakoff et al. 1991).

To provide an RDF/OWL representation of HMD data, we started by defining a new relation between the different synsets, the conceptual relation extMetaphorOf ("extension by metaphor of ... "). This conceptual relation holds between a synset with a metaphorical meaning and a synset with a literal meaning of at least one of the contained word senses. The relation as such is defined by an RDF schema. We then populated the extMetaphorOf-relation by deriving 107 instances from the HMD data for French. This was done by converting the data concerning attested metaphorical mappings between EWN synsets from the HMD relational database into RDF. The 107 instances of the extMetaphorOf-relation thus represent cases where both the literal and the metaphorical synset were already contained in the original version of EuroWordNet. As with each relation in RDF/OWL EuroWordNet, the resulting information is stored in a separate RDF-file (extMetaphorOf.rdf) and can be distributed as such. A detailed description about the integration of the Hamburg Metaphor Database into the RDF/OWL-EuroWordNet format can be found in De Luca and Lönneker-Rodman (2008).

### 4.3 Interlinking RDF/OWL-EuroWordNet with the Basic Multilingual Lexicon MEMODATA (BMD)

Because of the well knownWordNet problem of conceptual coverage (De Luca and Nürnberger 2006b), we decided to interlink RDF/OWL-EuroWordNet with the Ba-

```
<ewn20schema:NounSynset rdf:about="&ewn20instances;synset-vegetarian_pizza-noun-1"
    rdfs:label="vegetarian pizza">
    <ewn20schema:synsetId>5068318-pizza.owl-1</ewn20schema:synsetId>
</ewn20schema:NounSynset>
<ewn20schema:Collocation rdf:about="&ewn20instances;word-vegetarian_pizza"
    ewn20schema:lexicalForm="vegetarian pizza"/>
<ewn20schema:NounWordSense rdf:about="&ewn20instances;wordsense-vegetarian_pizza-noun-1"
    rdfs:label="vegetarian pizza">
    <ewn20schema:word rdf:resource="&ewn20instances;word-vegetarian_pizza"/>
</ewn20schema:NounWordSense>
<rdf:Description rdf:about="&ewn20instances;synset-vegetarian_pizza-noun-1">
    <ewn20schema:containsWordSense rdf:resource="&ewn20instances;wordsense-vegetarian_pizza-noun-1"/>
</rdf:Description>
<rdf:Description rdf:about="&ewn20instances;synset-vegetarian_pizza-noun-1">
    <ewn20schema:hyponymOf rdf:resource="&ewn20instances;synset-pizza-noun-1"/>
</rdf:Description>
```

*Figure 6*. OWL-EuroWordnet (Merged) "vegetarian pizza" synset Example.

Knowl. Org. 40(2013)No.5
E. W. De Luca. Extending the Linked Data Cloud with Multilingual Lexical Linked Data

327

sic Multilingual Lexicon MEMODATA (BMD) (http://catalog.elra.info/product info.php?products id=100). This resource includes different multilingual concepts that can be used for enriching the already available RDF/OWL-EuroWordNet structure. First of all, before converting the BMD, we analyzed the different structures of the resources in order to find out similarities and differences. The BMD contains words associated by the meaning in 5 languages: English, French, German, Italian, Spanish. The lexical categories included are: nouns (5 *18 000), verbs (5 * 8 000), adjectives (5 * 6 000), and adverbs (5 * 1 500). Sixteen parts of speech (POS) are distinguished and grammatical information is also contained. This resource is divided into five files, one for each language. Every file has different lines, each of them representing one word. Each line (e.g., 19223;E;Guyanese;s masc plur) includes

1. id number - this number links the word to the respective word represented in other languages
2. language code – it represents the language of the word (E=Englisch, F=French, G=German, I=Italian, S=Spanish)
3. word - word or a word group
4. POS - part of speech or grammatical information (as, e.g., "s masc plur" for male plural nouns).

separated by semicolon. In addition to the five language files, there is a meta file including the glosses of the words (but only in French), each sorted at line level including the id number, the language code A and one (or more) descriptions, all separated by semicolon. The descriptions are categorized by tags and can be used to disambiguate word senses in combination with the complete database, *The Integral Dictionary* (http://www.springerlink.com/content/4wc0lb70m57km7eu/).

In order to interlink the BMD to RDF/OWL-Euro WordNet, we analyzed the RDF/OWL-EuroWordNet and BMD classes and relations and decided to merge them with the same procedure we already applied in (De Luca et al. 2007). We converted the BMD classes and adapted them to the RDF/OWL WordNet classes (Laske and De Luca, 2010). The sixteen parts of speech (POS) of the BMD have been introduced as additional classes to WordSense and Synset. The properties that connect the main classes (containsWordSense and word), the lexical-Form —the string representation of a word—and the synsetId have been extended: each synset contained in the BMD having several POS, have been split into more synsets with only one POS, like in WordNet. In order to be sure that each new synsets has a unique synsetId, we altered the synsetIds from a number into a complex string with the pattern:

%original synsetID in the BMD%-BMD-%part of speech%

In this way, two synsets belonging to the same source synsetID, will differ in our extension by their part of speech (POS). Some new relations that exist in the BMD had to be added to the RDF/OWL EuroWordNet representation. For the grammatical POS information included in some BMD entries, we added a grammaticalForm as a relation fromWord to xsd:string. Futhermore, the URIs of theWords had to be changed from wn20instances: + word- + %label% to bmd20instances: + word- + %label% + (%grammatical form%). This avoided conflicts that a Word could have more than one grammatical form (e.g., transitive and intransive form for a verb).

Figure 7 shows the adaptation of the RDF/OWL BMD classes that give a more precise description of the Synsets (concepts) included in this resource. Here we extend the WordSense and SynSet classes with the additional subclasses included in the RDF/OWL BMD. Figure 8 presents the comparison between the RDF/OWL-Euro WordNet and the RDF/OWL BMD relations. The crossed out relations have been removed, since they are inverse relations to the existing ones. The gloss relation has been extended (doing it more specific) due to the different glosses included in the BMD. The 'grammatical-Form' and the 'derivationallyRelated' relation of a word have been added to the already existing relations. Finally, the interlinkage of BMD with RDF/OWL EuroWordNet is useful for multilingual information retrieval or for expanding the coverage of the EuroWordNet synsets, as, e.g., the BMD glosses with different gloss types and the grammatical information available are used (Laske and De Luca 2010).

## 7.0 Conclusions

The web of data, with his canonical datasets (DBpedia, geographical and biological data, social network data, bibliographical, musical, and multimedia data, etc.), and the data emerging from the use of RDFa, Microformats, etc., has eventually provided an empirical basis to the semantic web, and indirectly to knowledge engineering. In recent years, on one hand, much attention has been paid to the representation of lexical meaning and the development of lexical-semantic resources on the other. Wealth of data has implications. Firstly, a lot of data that are produced have problems in their structure; secondly, large datasets are difficult to describe in ways that enable their consumption: what is typically described by those data? How are data characteristically organized? Vocabularies do not help much, since they provide a set of predicates and axioms,
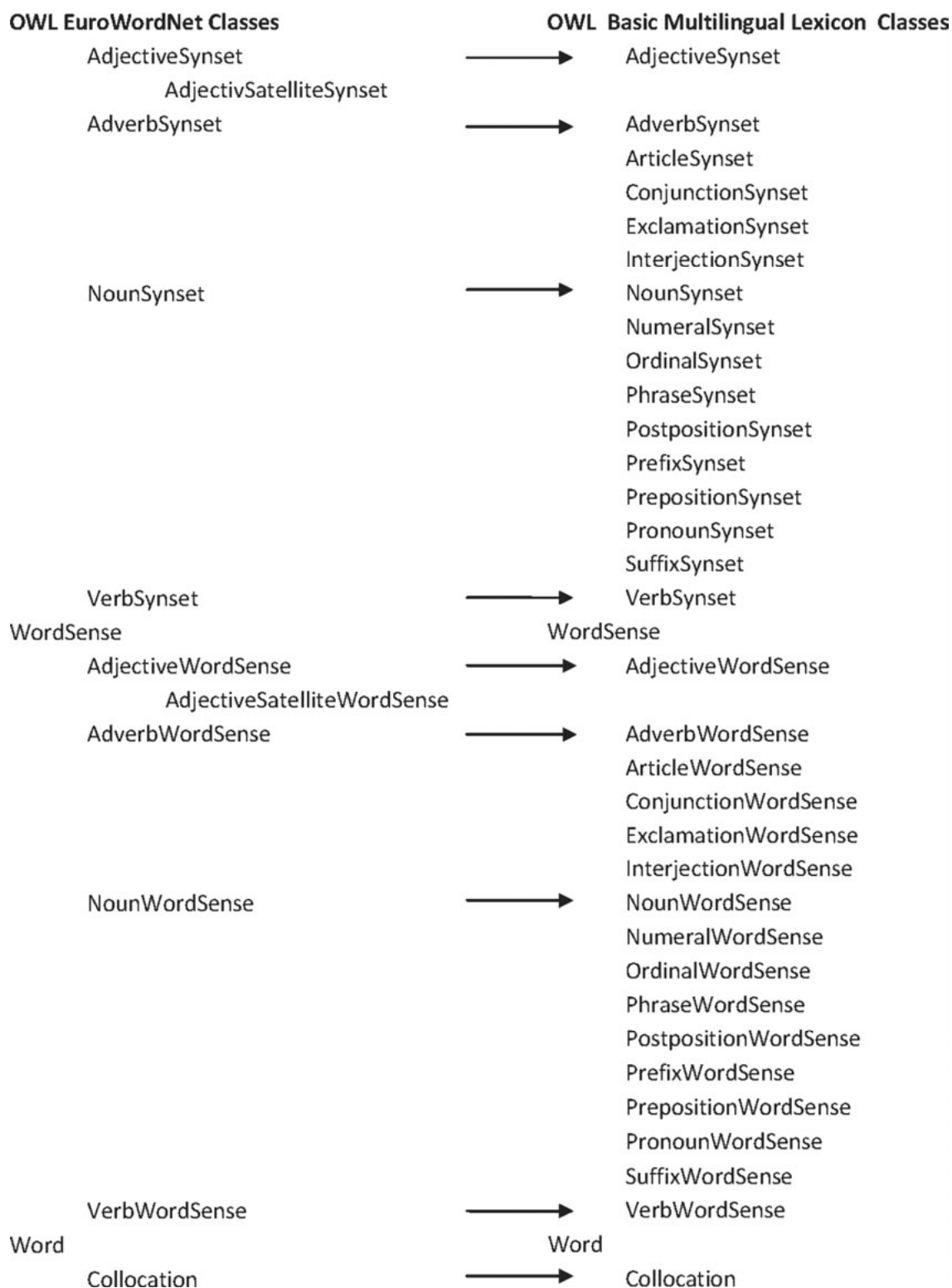
**OWL EuroWordNet Classes**                    **OWL Basic Multilingual Lexicon Classes**

AdjectiveSynset                    ⟶          AdjectiveSynset
    AdjectivSatelliteSynset
AdverbSynset                       ⟶          AdverbSynset
                                              ArticleSynset
                                              ConjunctionSynset
                                              ExclamationSynset
                                              InterjectionSynset
NounSynset                         ⟶          NounSynset
                                              NumeralSynset
                                              OrdinalSynset
                                              PhraseSynset
                                              PostpositionSynset
                                              PrefixSynset
                                              PrepositionSynset
                                              PronounSynset
                                              SuffixSynset
VerbSynset                         ⟶          VerbSynset
WordSense                                     WordSense
AdjectiveWordSense                 ⟶          AdjectiveWordSense
    AdjectiveSatelliteWordSense
AdverbWordSense                    ⟶          AdverbWordSense
                                              ArticleWordSense
                                              ConjunctionWordSense
                                              ExclamationWordSense
                                              InterjectionWordSense
NounWordSense                      ⟶          NounWordSense
                                              NumeralWordSense
                                              OrdinalWordSense
                                              PhraseWordSense
                                              PostpositionWordSense
                                              PrefixWordSense
                                              PrepositionWordSense
                                              PronounWordSense
                                              SuffixWordSense
VerbWordSense                      ⟶          VerbWordSense
Word                                          Word
Collocation                        ⟶          Collocation

*Figure 7*. Comparison of the RDF/OWL-EuroWordNet and the new RDF/OWL BMD classes.

**OWL EuroWordNet relations**　　　　　　　　　　**OWL Basic Multilingual Dictionary relations**

| containsWordSense | → | containsWordSense |
| word | → | word |
| lexicalForm | → | lexicalForm |
| | | grammaticalForm |
| synsetID | → | synsetID |
| tagCount | | |
| gloss | | |

descriptionGloss
domainGloss
physicalGloss
encyclopedicalGloss
topicGloss
glossHyperonymGloss
coHyponymOf
characterisedBy
grammaticalGloss

frame
hyponymOf
entails
similarTo
memberMeronymOf
substanceMeronymOf
partMeronymOf
classifiedByTopic
classifiedByUsage
classifiedByRegion
causes
sameVerbGroupAs
attribute
adjectivePertainsTo
adverbPertainsTo
derivationallyRelated　　　　　　→　　　　　　derivationallyRelated
antonymOf
seeAlso
participleOf
classifiedBy
meronymOf

*Figure 8*. Comparison of the RDF/OWL-EuroWordNet and RDF/OWL BMD relations.

330

Knowl. Org. 40(2013)No.5
E. W. De Luca. Extending the Linked Data Cloud with Multilingual Lexical Linked Data

which is not tailored to the size and shape of data in the large; size and shape can only be empirically discovered.

We addressed these issues and showed how hybridization research can be done by porting different lexical resource to the LOD cloud (see Figure 9). Our contribution here is twofold: 1) the production and publishing of Lexical LOD datasets; and, 2) the description of a method to produce a common lexical linked data knowledge repository. In this article, we listed the already available datasets and how they can be linked to another through a meta-model we developed.

## References

Baker, Collin F., Fillmore, Charles J. and Lowe, John B. 1998. The Berkeley FrameNet project. In COLING '98 Proceedings of the 17th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, pp. 86-90.

Baldassarre, Claudio, Daga, Enrico, Gangemi, Aldo, Gliozzo, Alfio Massimiliano, Salvati, Alberto and Troiani, Gianluca. 2010. Semantic scout: Making sense of or-

ganizational knowledge. In Cimiano, Philipp and Pinto, Helena Sofia, eds., *EKAW*. Lecture notes in computer science 6317. Springer, pp. 272-86.

Bizer, Christian, Heath, Tom and Berners-Lee, Tim. 2009. Linked data - The story so far. *International journal on semantic web and information systems* 5: 1-22.

Cole, Ronald A., Mariani, Joseph, Uszkoreit, Hans, Zaenen, Annie and Zue, Victor. 1997. *Survey of the state of the art in human language technology*. Pittsburgh, PA: Carnegie Mellon University.

Crystal, David. 2010. Semantic targeting: past, present, and future. *Aslib proceedings* 62. pp. 355-65.

Cunningham, Hamish. 1999. A definition and short history of language engineering. *Natural Language Engineering* 5: 1-16.

De Luca, Ernesto William. 2010. A corpus for evaluating semantic multilingual web retrieval systems: The sense folder corpus. In Calzolari, Nicoletta, Choukri, Khalid, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, Piperidis, Stelios, Rosner, Mike and Tapias, Daniel, eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Malta, Valletta, May 19-*
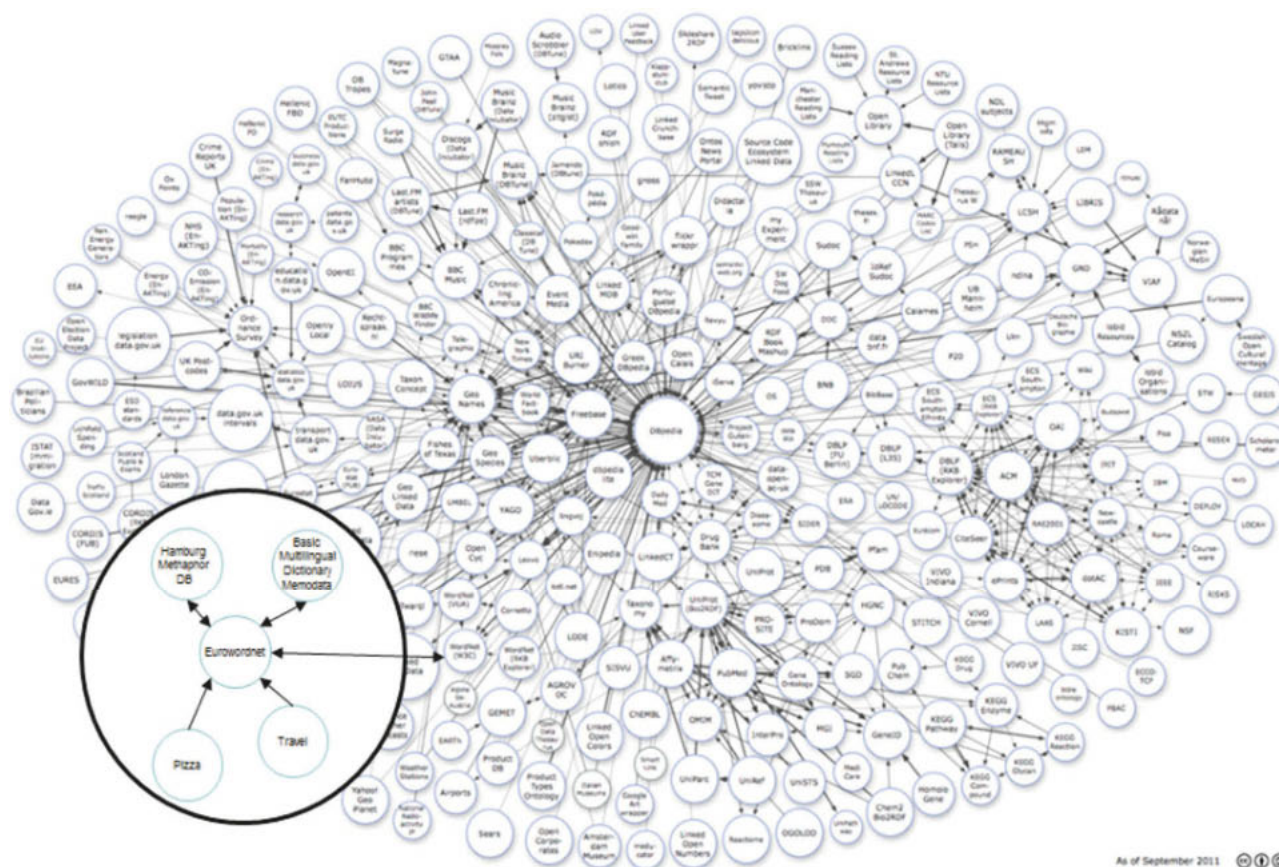


*Figure 9.* Extending the Linked Data Cloud with Lexical Linked Data information.

*21.* Paris: European Language Resources Association (ELRA), pp. 3475-80.

De Luca, Ernesto William and Lönneker-Rodman, Birte. 2008. Integrating metaphor information into RDF/ OWL EuroWordNet. In Calzolari, Nicoletta, Choukri, Khalid, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, Piperidis, Stelios and Tapias, Daniel, eds*., Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, May 28-30.* Paris: European Language Resources Association (ELRA), pp. 422-8.

De Luca, Ernesto William and Nürnberger, Andreas. 2006a. Rebuilding lexical resources for information retrieval using sense folder detection and merging methods. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy.* Paris: European Language Resources Association (ELRA), pp. 99-102.

De Luca, Ernesto William and Nürnberger, Andreas 2006b. Using clustering methods to improve ontology-based query term disambiguation. *International journal of intelligent systems* 21: 693-709.

De Luca, Ernesto William, Eul, Martin and Nürnberger, Andreas. 2007. Converting EuroWordNet in OWL and extending it with domain ontologies. In: *Proceedings of the Workshop on Lexical-Semantic and Ontological Resources. In conjunction with the GLDV-Frühjahrstagung (GLDV 2007). Tübingen, 2007.*

Fellbaum, Christiane. 1998. *WordNet, an electronic lexical database.* Cambridge, MA: MIT Press.

Gangemi, Aldo and Presutti, Valentina. 2010. Towards a pattern science for the Semantic Web. *Semantic Web journal* 1: 61-8.

Gangemi, Aldo, Navigli, Roberto and Velardi, Paola. 2003. The OntoWordNet project: Extension and axiomatization of conceptual relations in WordNet. In Meersman, Robert, Tari, Zahir and Schmidt, Douglas C., eds., *Proceedings On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7,* Berlin: Springer-Verlag, pp. 820-38.

Ibekwe-SanJuan, Fidelia. 2009. Semantic metadata annotation: tagging Medline abstracts for enhanced information access. *Aslib proceedings* 62: 476-88.

Lakoff, George, Espenson, Jane and Schwartz, Alan. 1991. *Master metaphor list. Second draft copy.* Berkeley: University of California.

Laske, Anja and De Luca, Ernesto William. 2010. *Conversion of the basic multilingual dictionary into the RDF/ OWL format. Technical Report.* Magdeburg, Germany: Otto-von-Guericke University of Magdeburg.

Lönneker-Rodman, Birte. 2008. The hamburg metaphor database project: issues in resource creation. *Language resources and evaluation* 42: 293-318.

Matthews, Brian, Jones, Catherine, Puzon, Bartlomiej, Moon, Jim, Tudhope, Douglas, Golub, Koraljka and Nielsen, Marianne Lykke. 2010. An evaluation of enhancing social tagging with a knowledge organization system. *Aslib proceedings* 62: 447-65.

Ménard, Elaine. 2010. Ordinary image retrieval in a multilingual context: a comparison of two indexing vocabularies. *Aslib proceedings* 62: 476-88.

Nuzzolese, Andrea Giovanni, Gangemi, Aldo and Presutti, Valentina. 2011. Gathering lexical linked data and knowledge patterns from FrameNet. *K-CAP '11 Proceedings of the Sixth International Conference on Knowledge Capture in cooperation with the AAAI, 25 - 29 June Banff, Alberta, Canada.* New York: ACM, pp. 41-8.

Pianta, Emanuele, Bentivogli, Luisa and Girardi, Christian. 2002. MultiWordNet: Developing an aligned multilingual database. *Proceedings of the 1st International WordNet Conference, January 21-25, Mysore, India,* pp. 293-302. Available http://multiwordnet.fbk.eu/paper/MWN-India-published.pdf

Sarinder, K.K.S., Lim, L.H.S., Merican, A.F. and Dimyati, K. 2010. Biodiversity information retrieval across networked data sets. *Aslib proceeding* 62: 514-22.

Scott, Tom and Smethurst, Michael. 2009. Building coherence at bbc.co.uk. In Content Architecture, exploiting and managing diverse resources. Presented at *Content Architecture: Exploiting and Managing Diverse Resources ISKO UK Conference, London 22-23 June.*

van Assem, Mark, Gangemi, Aldo and Schreiber, Guus. 2006. *RDF/OWL Representation of WordNet: W3C Working Draft 19 June 2006.* Available http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/.

Vossen, Piek. 1999. *EuroWordNet General Document, Version 3, Final.* Available www.illc.uva.nl/EuroWordNet/docs/GeneralDocPS.zip.

W3C. 2004. *Wordnet in RDFS and OWL: Editor's Draft.* Available http://www.w3.org/2001/sw/BestPractices/WNET/wordnet-sw-20040713.html