European Commission

WP2: PMKI core data model and extensions

Project Deliverable D2.2: PMKI data model (Ontology based on RDF(S)/OWL technologies)

PMKI–Public Multilingual Knowledge Infrastructure

Date: 25/01/2018
Doc. Version: 1.0
Template Version: 2.5

This template is based on PM² v2.5

For the latest version of this template please visit the PM² Wiki

Document Control Information

Settings	Value
Document Title:	WP2: PMKI core data model and extensions – D2.2: PMKI data model (Ontology based on RDF(S)/OWL technologies)
Project Title:	PMKI–Public Multilingual Knowledge Infrastructure
Document Author (s):	Brahim BATOUCHE
Project Owner:	Peter SCHMITZ
Delegated Project Owner:	Enrico FRANCESCONI
Project Manager:	Najeh HAJLAOUI
Doc. Version:	1.0
Sensitivity:	Public
Date:	25/01/2018

Document Approver(s) and Reviewer(s):

NOTE: All Approvers are required. Records of each approver must be maintained. All Reviewers in the list are considered required unless explicitly listed as Optional.

Name	Role	Action	Date
Najeh HAJLAOUI	Project Manager	*Approved/Reviewed*	04/04/2017
Enrico FRANCESCONI	Delegated Project Owner	*Approved / Reviewed*	
Peter SCHMITZ	Project Owner	*<Approve / Review>*	

Document history:

The Document Author is authorized to make the following types of changes to the document without requiring that the document be re-approved:

- Editorial, formatting, and spelling
- Clarification

To request a change to this document, contact the Document Author or Owner.

Changes to this document are summarized in the following table in reverse chronological order (latest version first).

Revision	Date	Created by	Short Description of Changes
1.0	07/02/2018	Najeh Hajlaoui	Consolidated version
0.4	30/10/2017	Armando Stellato	Draft for finalization Revised the OntoLex-Description and the examples Added description of SKOS Added section on standards for alignment Added section on further vocabularies Revised the conclusions
0.3	24/10/2017		
0.2	08/09/2017	Armando Stellato	Improvements and review
0.1	20/02/2017	Brahim Batouche	First draft using PM² template, goal, scope, structure

Configuration Management: Document Location

The latest version of this controlled document is stored in <link[1]>.

TABLE OF CONTENTS

[1] SharePoint accessible from OP environment.

1. Introduction

In the concluding deliverable of WP1 (D 1.2), we have studied different data models describing linguistic resources. We have concluded that OntoLex-Lemon is the best candidate to describe the types of resources covered by the PMKI project.

In this delivery, we would like to test the reliability and completeness of OntoLex-Lemon in describing some example of EuroVoc resource. The description of the resource concerns different enrichment point of views: topic concept and its hierarchy, linguistic variation, sense of term, multiword expression as phrases, and semantic description of a term. We need also to find out the possible limits of OntoLex-Lemon data model and if necessary complete them with extensions or hybridations.

The delivery is composed into 9 sections. SKOS is shortly recalled in section 2, whereas a descriptive summary of OntoLex-Lemon is given in Section 3. Section 4 describes deferments instances covering our requirements. A review of OntoLex-lemon is given in Section 5 while mentions of possible extensions are provided in section 6. Section 7 introduces other vocabularies to be adopted within PMKI while Section 8 discusses standards to be adopted for alignments. We conclude then the deliverable in Section 9.

2. SKOS Model

Simple Knowledge Organization System (SKOS, (World Wide Web Consortium (W3C), 2009)) is a W3C recommendation designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary. SKOS (an overview of the model in Figure 1) is part of the Semantic Web family of standards built upon RDF and RDFS, and its main objective is to enable easy publication and use of such vocabularies as linked data.

The model is already widely adopted in the Semantic Web community (it is actually considered one of the very core modeling vocabularies together with OWL and RDFS) and many resources in the Publication Office (in primis, EuroVoc) are modeled through it.

Specifically, EuroVoc adopts an extension of SKOS, SKOS-XL (World Wide Web Consortium (W3C), 2009) that allows for reification of the labels (as described in D1.2) so that additional metadata or even relations among labels can be added to the description of each label.

Due to its wide and mature adoption, SKOS (and SKOS-XL) are part of the data model adopted in PMKI.

SKOS Vocabulary Organized by Theme

Concepts	Labels & Notation	Documentation	Semantic Relations	Mapping Properties	Collections
Concept	prefLabel	note	broader	broadMatch	Collection
ConceptScheme	altLabel	changeNote	narrower	narrowMatch	orderedCollection
inScheme	hiddenLabel	definition	related	relatedMatch	member
hasTopConcept	notation	editorialNote	broaderTransitive	closeMatch	memberList
topConceptOf		example	narrowerTransitive	exactMatch	
		historyNote	semanticRelation	mappingRelation	
		scopeNote			

Figure 1: an overview of the SKOS Vocabulary with elements organized by them

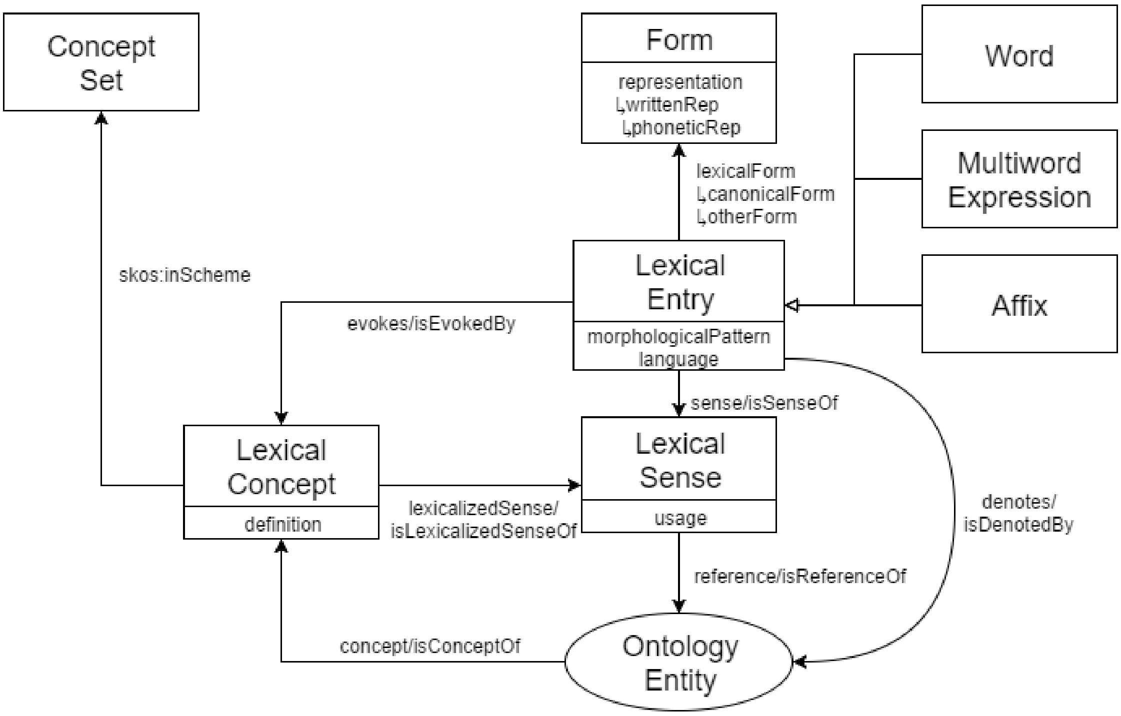

Figure 2: OntoLex data model

3. ONTOLEX-LEMON MODEL

The Ontology-Lexica community group develops models (in Figure 2**Error! Reference source not found.** the core vocabulary) for the representation of lexica and for their interfaces with ontologies, thesauri and RDF datasets in general. The models are thought and designed to be modular and extensible, thus giving a hub for connecting lexical and terminological description elements with ontologies. The model is under continuous evolution in order to cover more linguistic phenomena and account for different type of linguistic resources

OntoLex describes the meanings of a lexical entry (still in the context of a Lexicon) by coining senses for it, which in turn connect to instances of ontolex:LexicalConcept (which is in turn a subclass of skos:Concept). Lexical concepts represent the semantic pole of linguistic units and are the mentally instantiated abstractions which language users derive from conceptions. We consider the abstraction of lexical concept as skos:Concept an interesting feature because it allows to consider the topic and its hierarchy which is not covered by Lemon.

In interfacing a Lexicon to an existing ontology or thesaurus, it is possible to reference also elements of the ontology (classes, properties, instances etc..) through the ontolex:reference property, thus “lexicalizing” that element with the lexical entries from the lexicon.

4. ONTOLEX-LEMON INSTANCE

In order to illustrate the capability of OntoLex-Lemon we describe in this section the EuroVoc resource. We choose as example the resource http://eurovoc.europa.eu/100234 where its prefLabel is "*4416 organization of work and working conditions*"*@en*.

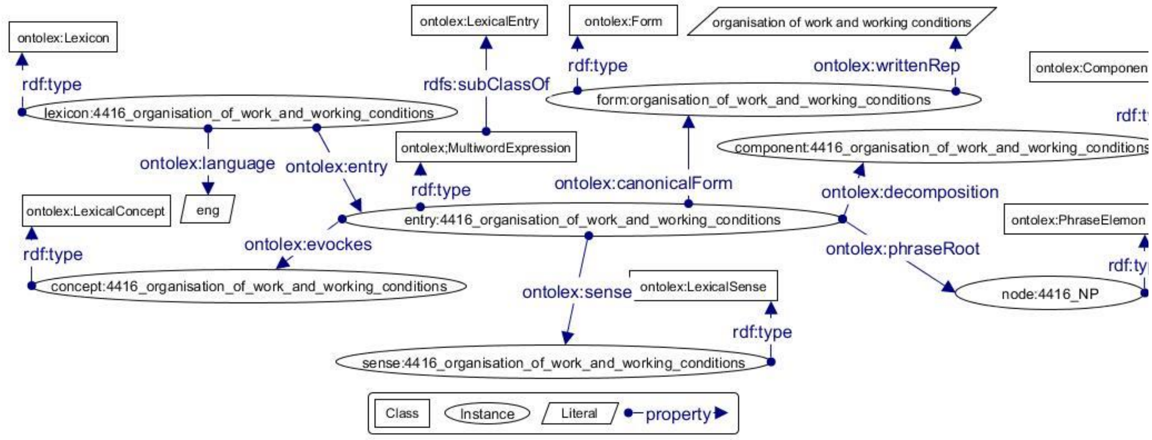

Figure 3: The entry "organisation of work and working conditions"

As mentioned at the previous delivery, our requirements are the consideration of: (1) multi word expression (2) linguistic variation (3) hierarchical classification of concepts (4) definition of terms (5) lexical sense (6) reference to a linguistic data model.

The following sub sections are structured according to these points (1-6). In section 4.1 we give an example illustrating the root description. Section 4.2 gives an example illustrating the point (1), section 4.3 illustrates the point (2) and section 4.4 illustrates points (3-6).

4.1. Lexicon and lexical entry

The lexicon has a lexical entry which it is defined with form, lexical sense and lexical concept. This later refers to the linguistic data model such as dbpedia.

- When the entry is multi-words expression, we define the component list and the node. The node of the sentence is the root of the sentence and the leave of node will be the component, where each component has proper lexical entry and form.

- When the entry is verb phrase we define the frame and the argument. Knowing that the thesauri rarely contain the verb phrase.

Figure 3 illustrates the entry "*4416 organization of work and working conditions*", in describing: the concept, the sense, the node, the component and the form.

4.2. Multi words expression

As illustrated in **Figure 4**, the multi word expression in the entry *"organization of work and working conditions*", is decomposed into a set of components (ontolex:Component), where each element (e.g. *organization, of, work, and, working, conditions*) can even be an entry in turn.

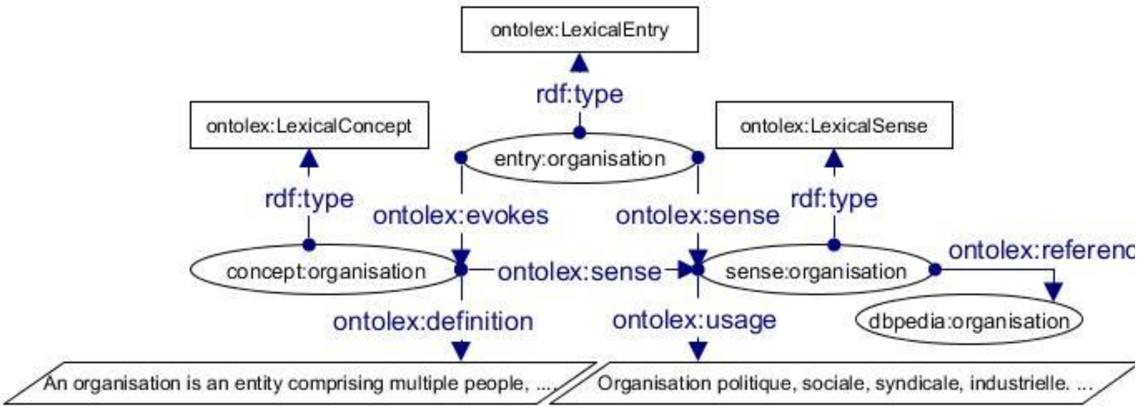

Figure 5: Description of the component

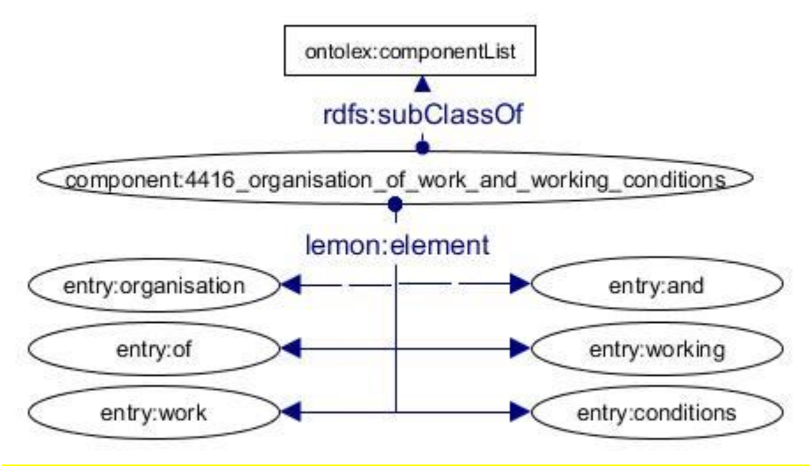

Figure 4: Multi words expression

The element entry has: form, sense, and can have a concept, where Figure 5**Error! Reference source not found.** illustrates the description of the element entry: *organisation*.

4.3. Linguistic variation

The linguistic variation concerns the noun phrase and verb phrase.

4.3.1. Noun phrase (NP)

As illustrated in Figure 6, the phrase *"organization of work and working conditions"* corresponds to a node, which is composed into three nodes:

- Noun phrases NP1=*"organization of work"* and NP2="*working conditions"*.
- Coordination conjunction CC=*"and"*, which related the noun phrases.

Each noun phrases will be initiated with a node which represents the root of the phrase. A node represents a tree which can be composed of other nodes (sub nodes) or directly composed of leaf. The leaf element has an entry, which means each element has the frame, a form, a sense and a reference and can have a concept.

To go to the end of the tree, we detail the noun Phrase NP1=*"organization of work and working condition"*, which corresponds to a node, and it is composed into three nodes:

NP1:*"organization of work"*, CC:*"and"*, and NP2: *"working condition"*. The node NP1*"organization of work*" is composed into three leafs: *organization*, *of*, and *work*.

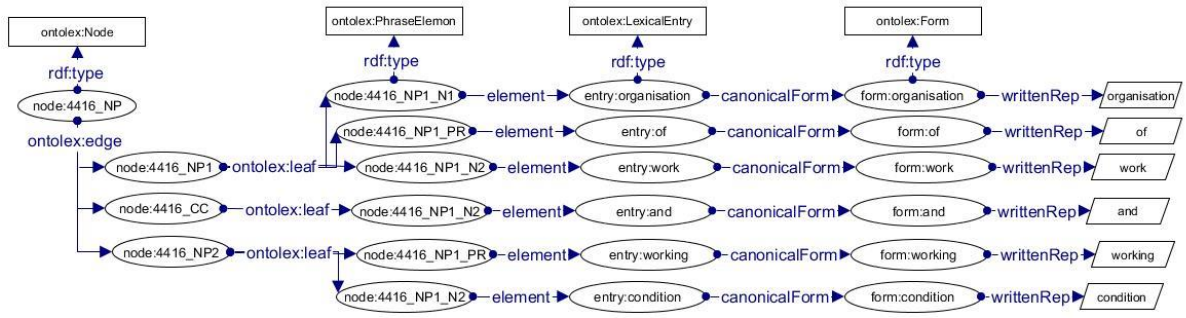

Figure 6: Linguistic variation of "*organization of work and working condition*"

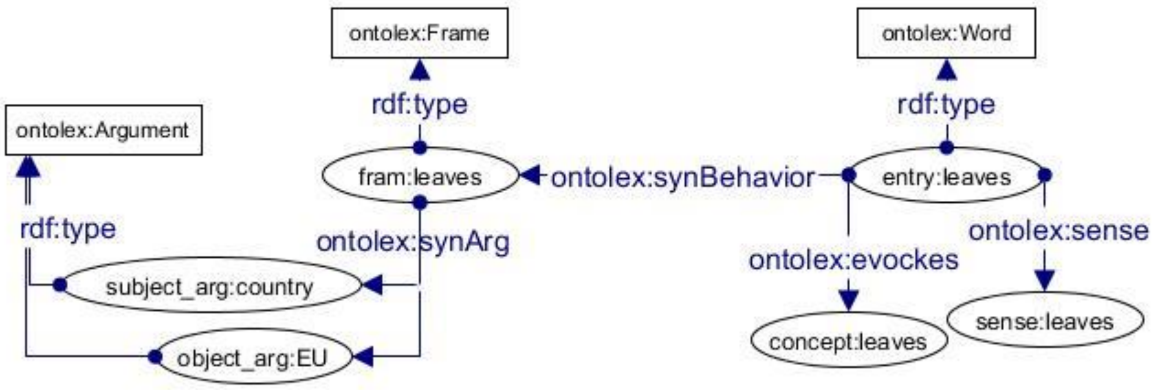

Figure 7: Linguistic variation of "country leaves EU"

4.3.2. Verb phrase (VP)

In general a thesaurus such as EuroVoc does not contain verbs because the resources are generally nouns or noun phrases. However the recent event of Brexit requires a new concept to define the events of country leaves EU. The Portuguese expression of this concept contains the verb *leaves*; i.e. "*country leaves EU*".

The verb phrase "*country leaves EU*" is composed by: verb, subject and object. The verb represents the frame *"leaves"*, where its arguments are: the subject *"country"* and the object *"EU"*.

4.4. Meaning: Concept and Sense

The meaning of the entry is composed into: sense and concept, where the sense is obligatory, but note the concept. The **Error! Reference source not found.** illustrates the sense and concept of our example and the **Error! Reference source not found.** illustrates them for the element entry "*organization*".

We would rather distinguish between "concept", which is domain-specific and pertains to an ontology, and "sense" which is more generic. A "concept" in an ontology limits the representation of word meaning to those distinctions that are actually relevant in the context of the given ontology and/or domain (Cimiano, McCrae, Buitelaar, & Montiel-Ponsoda, 2013),

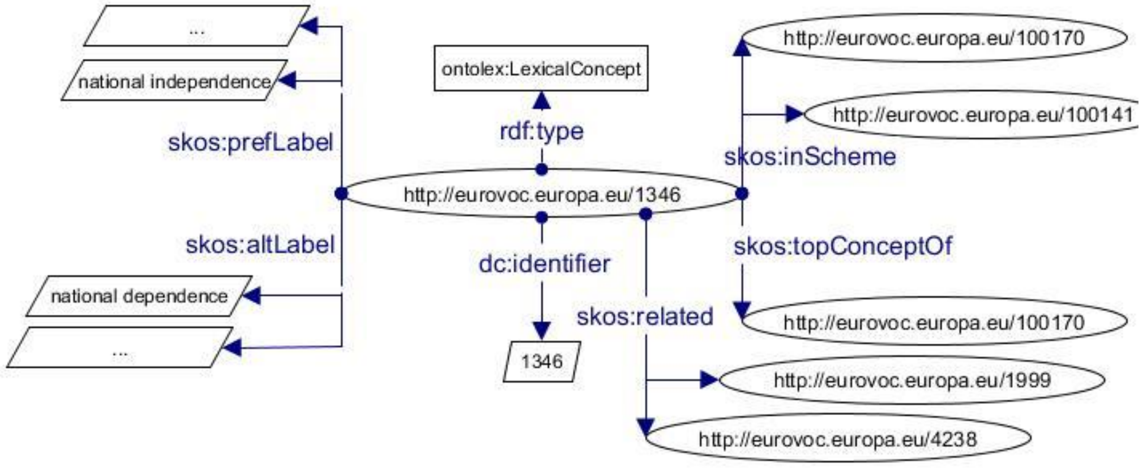

Figure 9: Concept labels and hierarchies

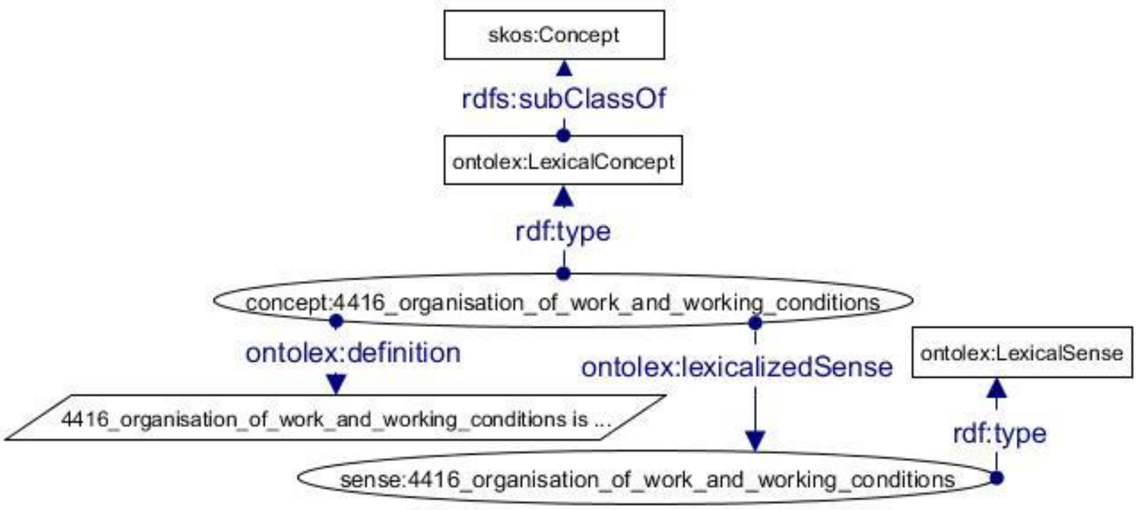

Figure 8: Sense and concept

while "sense" is a reification of the relation between a lexical entry and its possible conceptualization, that can even not be covered by the associated ontology. This allows also extending the coverage of the possible senses of a lexical entry without having to update the contextual ontology.

Moreover given a word, one can imagine as many senses as many context it appears in (Cimiano, McCrae, Buitelaar, & Montiel-Ponsoda, 2013), while an ontology is used to solve these possible ambiguities in a conceptualization which is domain-specific. For example in the domain of scientific publishing, the meaning of the lexical term "paper" as material is not relevant.

This can be capture as "sense" in the lexicon, while not in the scientific publishing domain ontology.

The reification of the "sense" property gives also the possibility to express additional characteristics to a send of a word.

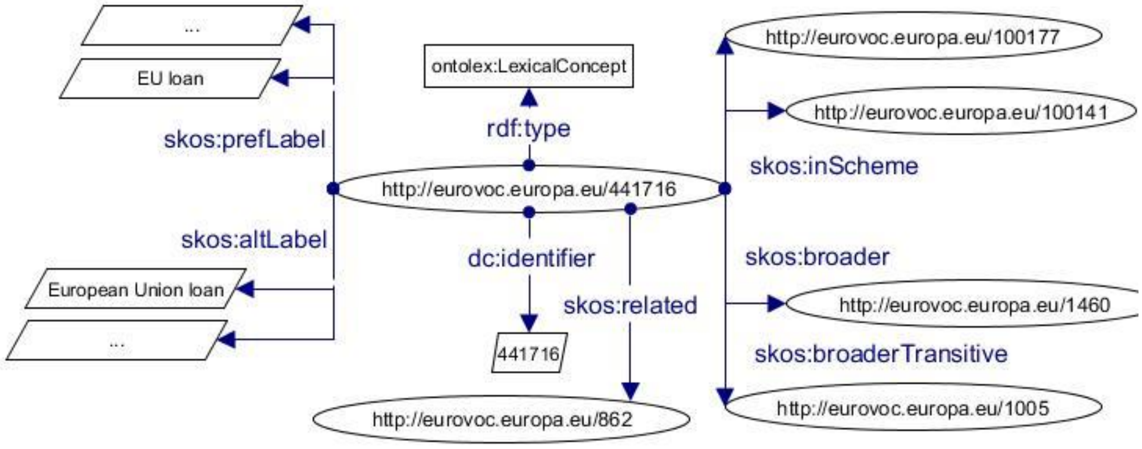

Figure 10: Concept broader/narrower

4.4.1. Concept

The lexical concept is on purpose a sub class of skos:Concept, so that it is possible to use inScheme (generic concept), topConceptOf (specific concept), prefLabel, altLabel, broader, narrower and all the SKOS properties characterizing a concept scheme. We consider the resource http://eurovoc.europa.eu/1346 which its prefLabel is "*national independence*"*@en* (Figure 9**Error! Reference source not found.**).

As illustrated in Figure 10**Error! Reference source not found.**, the concept can have narrower and broader concept, e.g. the concept http://eurovoc.europa.eu/1005: "*EU loan*"*@en*. In the case of different datasets, we can have matching between concepts, in using the properties skos:mappingRelation.

It is worth noticing that there has been some debate inside the OntoLex community group whether to consider concepts from any SKOS resource as lexical concepts. Lexical concepts were introduced with the intention of covering the semantic backbone of lexicons, thus concepts created to give meaning to existing lexical entries (semasiological approach). Concepts from existing thesauri should instead be treated as entities in any dataset[2] that needs to be lexicalized by OntoLex-Lemon. In practice, the different intention reflects in the property to be used: ontolex:evokes or ontolex:denotes (or their more complex patterns mediated by ontolex:Sense). We will inquiry the Ontology-Lexica community and converge on a best practice to be adopted in the case of pure thesauri being lexicalized by Lexical Entries in OntoLex-Lemon.

4.4.2. Sense

In order to illustrate the sense we give two similar terms where their sense is incompatible, but refers to the same dbpedia object and can refers also to the same lexical concept (Figure 11**Error! Reference source not found.**). The lexical sense describes the most specific level of the term (the end level) and as illustrated example we consider the following French terms: *riviere* and *fleuve*.

[2] The term "ontology" in the ontology-lexica dualism, is intended with a wider meaning, covering in fact all kind of entities present in any (kind of) RDF datasets which might need to be lexicalized

The lexical sense can have the properties such as broader, narrower, equivalent, incompatible, senseRelation, subsense, condition, context, definition, and example. The properties broader and narrower concern the concept and the sense.

The concept groups a set of label by language, where the set of labels has the same sense, and this latter is the sense of the concept.

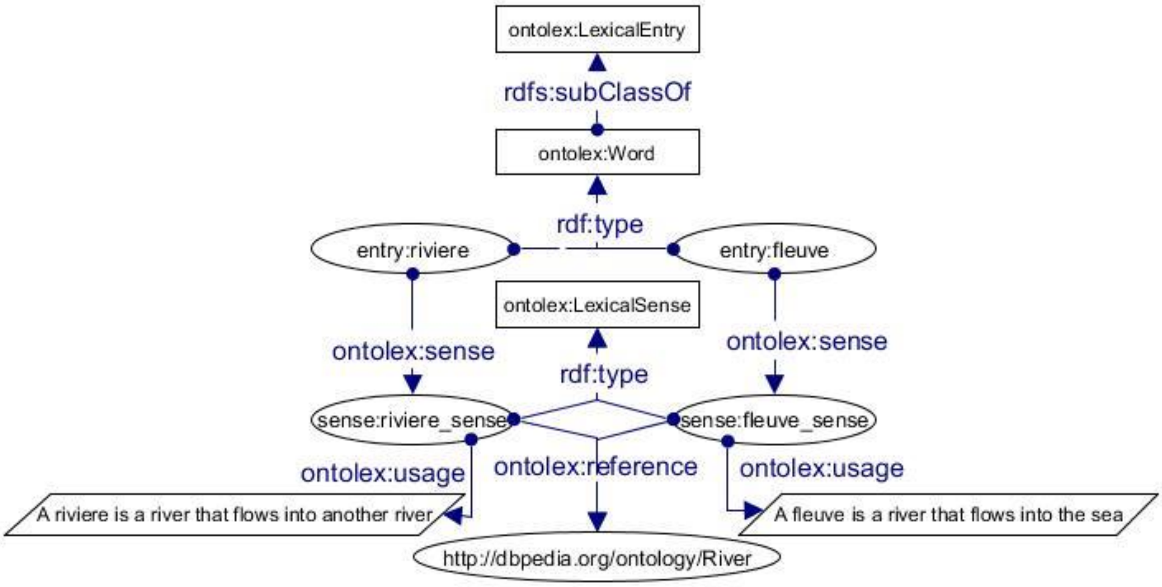

Figure 11: Sense of "river" and "fleuve"

5. OntoLex-Lemon reviews

The OntoLex-Lemon model answers our requirements, which are: (1) Description of Multi word expressions, (2) Description of the linguistic variation (3) Hierarchical classification of concepts (4) Definition of terms (5) Lexical Sense (6) Refer to a linguistic data model. These possibilities are independent of the conceptual model, and can thus be exploited to descriobe lexicalizations for any of the resources types considered in PMKI: Controlled vocabularies, Glossaries, Thesauri, Lexica (lexical-Semantic databases), Taxonomies and Semantic Networks.

The biggest contribution of OntoLex is the merging of two community models: librarian and linguistic communities, without losing any advantage of their data models: SKOS and Lemon. As counter-example, BabelNet merges these models loosing an advantage given by Lemon. The BabelNet data model combines Lemon and SKOS, by considering SKOS as linguistic ontology. This means the point of contact between skos and Lemon is the property lemon:reference. This combination is done to the detriment of the linguistic data model, and therefore the point (6) is not allowed in BabelNet. Basing on that, we can say that the OntoLex modelisation is better than BabelNet model.

As summary, OntoLex-Lemon model satisfies our requirements (1-6), where the given examples of instances support and show that.

6. Extensions and Evolutions of OntoLex

Other resources of interest that might fall into the interest of PMKI include encyclopedias, dictionaries and similar linguistic resources. These further set of resources are being considered in a second iteration of work in the Ontology-Lexica community group at the time of writing. We may consider to submit interesting resources as use cases to the group and actively participate to discussion on their representation in OntoLex-Lemon

7. Other Vocabularies in the PMKI model

The data model being used will include also elements from other commonly used vocabularies: FOAF, Dublin Core ontology will be used to manage the user and their added dataset. PROV-O can also play a role in connecting the descriptions of users with the action they carry on within the system. DCAT, VoID and LIME can provide the metadata for describing the resources, facilitating the semi-automatic tasks of lexicalization and alignment.

8. Alignment Tasks and Alignment Standards

As already analyzed in section 5 of D2.1, different resources may interact in different ways. Two conceptual resources may provide different perspectives (i.e. modeling solutions) over overlapping domains and under the same kind of description: they are in a sense competing models satisfying the same need. Where the selection/emergence of a single model is not the objective, a reconciliation can be performed by means of alignment. The semantics of the alignment depend on the nature of the involved resources (e.g. thesauri vs ontologies) and on the objective of the result. Usually, the objective of alignment thesauri, characterized by shallow semantics and rich lexicalization, is to expand the retrieval capabilities of systems based on their vocabularies. So, if a concept C_A from thesaurus A is aligned to a concept C_B from thesaurus B (whatever the mapping relation), the objective of this alignment is to be expand the set of documents retrieved by using concept C_A (because they have been semantically indexed with thesaurus A) including (all or part of, depending on the mapping relation) those documents indexed with concept C_B. The objective of an alignment between formal ontologies is conceptually similar, except that it involves the retrieval of instances by means of logical computation (as opposed to mere indexing in the case of thesauri). The alignment properties must thus ensure that the “enlarged” set of facts is still consistent with the original model.

In the case of lexical resources, the task can still be an alignment if performed between the conceptualizations of the resources, whereas it is an elaborated process of lexicalization when it involves the elements in a thesaurus or ontology and entries available in a lexical resource. The latter elements need not only to be associated to the former, but they may eventually be composed in order to adherently represent the element of interest from the conceptual resource.

Concerning alignments, various formats have emerged. We will consider the format adopted in the OAEI initiative (Dragisic, et al., 2014), which has been defined within the Alignment API (David, Euzenat, Scharffe, & Trojahn dos Santos, 2011) project, as it is the most widely accepted intermediate format for alignment, providing description of mappings which, being reified, can be enriched with further metadata. Metadata can include details such as the relevance of the mapping, the rationale, the reliability etc.. Compliancy with this standard will be a must for the machinery we will be developing in the concrete systems devised within PMKI. At the same time, it is important to project these mappings towards final mapping specifications from the most common vocabularies: OWL (i.e. owl:EquivalentClass, owl:EquivalentProperty, rdfs:subClassOf, owl:sameAs combined with the possibility to express logical re-elaborations of the basic elements) and SKOS (i.e. skos:exactMatch, skos:closeMatch, skos:broadMatch, skos:narrowMatch, skos:relatedMatch).

For the lexicalization task, the standard properties in OntoLex-Lemon (e.g. ontolex:denotes) will be adopted for linking the entries of the lexical resource to the elements of the target conceptual resource. However, the lexicalization will very often require to also create new compound lexical entries determined by the lexical composition of the core entries of the lexical resource, in order to better represent the target resource's conceptual entries.

9. Conclusion

This deliverable presents some use case described through the OntoLex-Lemon vocabulary in order to illustrate the capability of such data model. We confirm that the combination of SKOS and OntoLex-Lemon data models, enriched with use of popular vocabularies such as FOAF, PROV-O and Dublin Core answer all our needs, with no necessity to extend or hybridize the model. The given illustrated examples of lexicalizations based on OntoLex are basic and for more details see this link[3].

The OntoLex data model merges the models given by librarian and linguistic communities, which are respectively SKOS and OntoLex-Lemon. This merging of models is profitable for both the communities though better for the librarian one, because a lexical concept has mandatorily a lexical entry while the inverse is false. This means, large part of Lemon resources have not matched concept and then are not undescribed with SKOS. However, when the lexical entry has a matched concept, i.e. defines a frame of concept, the hierarchical classification of the concept and its different label will be described.

The immediate added value of OntoLex at the librarian community is the automatic indexation of the text, because the words/multi word expression are described semantically and then the concept are detected automatically through the description of OntoLex model. In other word, the alignment from linguistic resource to librarian resource is immediate.

Finally, metadata models such as DCAT, VoID and LIME can provide the required metadata for identifying and profiling the resources, in order to better support with automatisms processes such as lexicalization and alignment.

References

Cimiano, P., McCrae, J., Buitelaar, P., & Montiel-Ponsoda, E. (2013). On the Role of Senses in the Ontology-Lexicon. In A. Oltramari, P. Vossen, L. Qin, & E. Hovy (Eds.), *New Trends of Research in Ontologies and Lexical Resources* (pp. 43-62). Springer Berlin Heidelberg.

David, J., Euzenat, J., Scharffe, F., & Trojahn dos Santos, C. (2011). The Alignment API 4.0. *Semantic Web Journal*, *2*(1), 3-10.

Dragisic, Z., Eckert, K., Euzenat, J., Faria, D., Ferrara, A., Granada, R., et al. (2014). Results of the ontology alignment evaluation initiative 2014. *9th International Workshop on Ontology Matching, October 20, 2014. 1317*, p. 61-104. Riva del Garda, Trentino, Italy: CEUR-WS.org.

World Wide Web Consortium (W3C). (2009, August 18). *SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL).* (A. Miles, & S. Bechhofer, Eds.) Retrieved March 22, 2011, from World Wide Web Consortium (W3C): http://www.w3.org/TR/skos-reference/skos-xl.html

[3] https://www.w3.org/2016/05/ontolex/

World Wide Web Consortium (W3C). (2009, August 18). *SKOS Simple Knowledge Organization System Reference.* (A. Miles, & S. Bechhofer, Eds.) Retrieved March 22, 2011, from World Wide Web Consortium (W3C): http://www.w3.org/TR/skos-reference/