Towards modeling Arabic lexicons compliant LMF in OWL-DL

Abstract. Elaborating reusable lexical databases and especially making interoperability operational are crucial tasks effecting both Natural Language Processing (NLP) and Semantic Web. With this respect, we consider that modeling Lexical Markup Framework (LMF) in Web Ontology Language Description Logics (OWL-DL) can be a beneficial attempt to reach these aims. This proposal will have large repute since it concerns the reference standard LMF for modeling lexical structures. In this paper, we study the requirement for this suggestion. We first make a quick presentation of the LMF framework. Next, we define the three ontology definition sublanguages that may be easily used by specific users: OWL Lite, OWL-DL and OWL Full. After comparing of the three, we have chosen to work with OWL-DL. We then define the ontology language OWL and describe the steps needed to model LMF in OWL. Finally, we apply this model to develop an instance for an Arabic lexicon.

Keywords: Lexical Markup Framework LMF, Web Ontology Language Description Logics OWL-DL, Interoperability.

1 Introduction

Consistent lexical resources represent a crucial requirement for several NLP tasks. This necessity arises by the rising need of automatic tools to deal with Information retrieval, Information Filtering, Information Extraction, Question-Answering, etc. However, these tasks suffer from the lack of reusable linguistic, and in particular lexical, resources. These deficiencies vary quite a great deal from one language to another. Arabic is one of the languages which suffers most from this shortcoming. The common problem of the majority of the languages is that the lack of resources limits any progress in the computational linguistic sciences for these languages [1].

From another angle, the need of standardized lexicons is even harder to achieve because standardization requires significant time resources. First, human resources are needed to ensure compatibility with the chosen standards making the task of putting together a conformant lexical structure more complex. LMF is one of the standards in language technology that intends to cover all languages in the world. Providing compliance to such a standard thus makes our work comparable with similar endeavours worldwide.

In this paper, we thus propose an initiative enabling us to model the LMF standard in the OWL-DL ontology language, with the aim to facilitate the elaboration of reusable lexical data bases and make interoperability operational in future works. As a matter of fact, there are few standards dedicated to digital lexica in comparison to available standards for language resources at large. International Standardization Organization (ISO) lexicons have critical effect in NLP. Indeed, this standardization

identifies an informative common coverage for all lexicons. The developed coverage is fundamental for introducing tools allowing the exchange and the share of lexical resources. So that interoperability can be easily introduced. This notion means that information and communication systems will be able to exchange data and enable sharing knowledge [2, 3]. Nowadays, having interoperable framework is so required then before. It will be a mixture of standards and guidelines such as Text Encoding Initiative (TEI) [4]. So, standards will be consistently correlated and guidelines will explain application of standards specification. However, between these two axis (standards and guidelines), a transformation prototype should be present. This prototype should have a lot of characteristics which will be explained later. However now, we can prove that OWL-DL can be an important factor in this prototype [5].

In this paper, we will present first of all a scope for LMF in order to make sure that this standard will be able to be mapped to OWL-DL [6]. Secondly, we will present OWL with its three sublanguages: OWL Lite, OWL-DL and OWL Full and we will prove our choice for the OWL-DL. The next section will be the most important one while it interests our transformation prototype from LMF to OWL-DL. Finally, we will instantiate this model to develop an instance for an Arabic lexicon.

2 LMF overview

After successful scientific activities and teamworks in developing lexicons, NLP and Machine Readable Dictionaries (MRDs) communities decided to start ISO tasks in 2003. Several theoretical divergences in structures languages make these activities hardly achieved in 2008. In fact, a group of 60 researchers and during these 5 years of developing the requirements produced finally the LMF standard [3]. LMF is an ISO standard covering monolingual and multilingual lexica. LMF specification follows UML modeling principles defined by Object Management Group (OMG). It is composed of a core model and extensions packages. The modeling principles of LMF take up the general principles developed in ISO committee TC 37 and allow a lexical database designer to combine any component of the LMF meta-model with data-categories [7] in order to create an appropriate model. These data categories function as UML attribute-value pairs in the diagrams. The core model covers the backbone of a lexical entry. It specifies the basic concepts of vocabulary, word, form and sense. LMF core model is a hierarchical structure consisting on several components. Lexical Entry is one of the components that represents the basic resource in the lexicon. In fact, this unit represents the lexeme and contains associated form and sense.

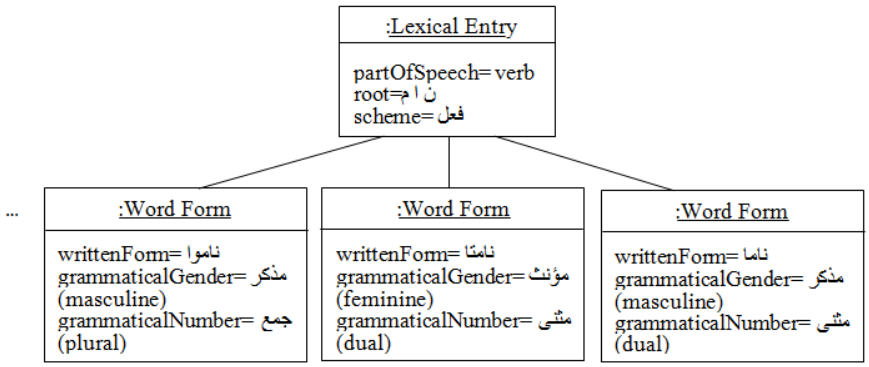

Fig. 1. Inflected forms of the verb "نام" "nAm"[1] (to sleep)

The example in Fig. 1 illustrates one prototype among the entire inflected forms of the verb "nAm" (to sleep). This example is an instance from the Arabic LMF core model.

Extensions are used according to the requirements of the users. Thus, lexicons developers have to choose packages that are useful for their needs. However, an extension package can not be drawn regardless of the core package [3].

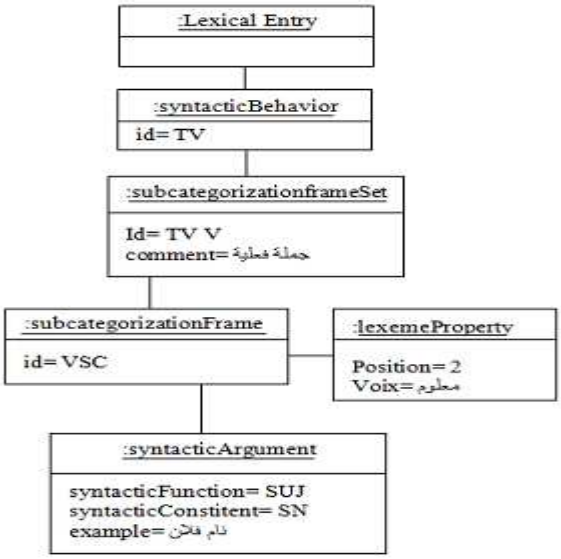

Fig. 2. Syntactic extension of the verb "نام" "nAm"(to sleep)

[1] http://www.qamus.org/transliteration.htm

The example illustrated in Fig. 2 shows the syntactic extension concerning the verb "nAm". This example is a part of the Arabic LMF extensions packages.

There have been some works dealing with LMF lexicons for Arabic language. In [8], authors have studied the importance of the reuse of syntactic Arabic lexicons. They have consequently encoded a lexicon compliant LMF after the examination of HPSG and LTAG lexicons specificity [9]. In the same context, we can mention the work concerning Arabic lexicons projection from HPSG to LMF [10].

As we have already explained previously, LMF and lexical standard in general can make the notion of interoperability more operational if we use a transformation prototype of LMF in OWL-DL. In the next section, we will provide a presentation of this language in order to describe the main lines for the prototype.

3 OWL overview

In general, ontology is a philosophy concept that allows studying the existing [11]. Yet, in the computer sciences, it must be defined as a structured and formal set of concepts offering meaning to informations. Particularly, OWL, recommended by W3C and strongly inspired from DAML+OIL, is a language which represents ontologies. These ontologies are quite useful on the Semantic Web [12]. In fact, data occurring there could be easily published and shared. From a technical point of view, OWL includes comparison tools of properties and classes that match properly with LMF such as identity, cardinality, inheritance. In fact, OWL presents more capacity for content web interpretation than RDF and RDFS due to the largest vocabulary and the right formal semantic [6], [12]. While modeling in OWL, we will notes that OWL supplies three increasingly expressive sublanguages that may be easily used by specific users: OWL Lite, OWL-DL and OWL Full [5]:

OWL Lite is the sublanguage which is used by those who need simple classification hierarchy and constraint features. For instance, if we study cardinality constraints supported by OWL Lite, we note that only the values of 0 or 1 are allowed. Indeed, the most advantage of this kind of OWL is that it looks simpler than OWL-DL and OWL FULL which seem to be more expressive. However, fast mapping path for large taxonomies lets it yet repute.

OWL-DL is the sublanguage whose users look for highest number of expressiveness. In spite of this expressiveness, completeness and decidability are assured as well [13]. It means that all inferences are computed and in limited time. Technically, OWL-DL involves absolutely all paradigms of OWL. These concepts have some restrictions. For example, a class can not be considered as an individual or property. This phenomenon is called type separation. As well, a property can not be considered as an individual or a class. This restriction is allowed in OWL FULL and consequently makes it non decidable [14]. OWL-DL is named from its underlying logical formalism, description logics, which offers adequate inference capabilities while preserving expressive power. And for this purpose, we want to model LMF with this expressive sublanguage.

Finally, OWL Full is destined for whose they want the greatest expressiveness without carrying about how much completeness and decidability are guaranteed. This liberty makes one use the syntactic of RDF with large freedom. From a technical corner, a class can be discussed at once as a collection of individuals and as an individual in its own right. In a word, we can say that maximum of expressiveness limits decidability and makes reasonable mechanisms more and more complex.

Works dealing with the generation of OWL resources in Arabic language are so few. We can mention the approach proposed for the generation of domain ontology from LMF standardized dictionaries. An additional alternative of automatic domain ontology enrichment based on the semantic component of LMF has been suggested [15]. In the following section, we will explain details for the developed prototype for the transformation of LMF to OWL-DL model since it will facilitate having interoperable framework in the future. In fact, this prototype will play the role of the pivot between standards and guidelines. Currently, projects require such a construction; otherwise they will be out of business. A recent report by TAUS declares that: "The lack of interoperability costs the translation industry a fortune" [16]. Fortune is compensated for adjusting data formats. Thus, setting up an interoperability framework will gain us much more time and fortune. The step before building this framework is to seek for a pivot language. This language will be described as a dynamic environment where standards will be consistently related and guidelines evidently explain the specifications application to several types of resource.

4 Modeling LMF in OWL-DL

The built of interoperable framework is our future target. However, the construction of a similar framework requires an environment making possible interoperability between applications exchanging non formal and non structured informations through the web. So, it helps exchanging data and simplifies documents description. These characteristics are available in OWL [17]. Then, we have chosen OWL-DL because of its expressiveness and decidability as we have described in the previous section. Thus, modeling LMF in OWL-DL is a crucial task in the process of making an interoperable managing lexicons framework. In fact, we have to check the possibility of mapping the whole LMF concepts to OWL-DL ones. This task is hard to carry out since the components in LMF model are so nested and complex. In this part, we are going to describe the prototype of transformation LMF into OWL-DL. This prototype is divided in seven parts:

4.1 Building OWL-DL Entities

In order to simplify some entries in OWL modeling, we have to use first of all required entities. Here, we have defined the following entities:

```
<!DOCTYPE rdf:RDF [
<!ENTITY xsd "http://www.w3.org/2001/XMLSchema#">
<!ENTITY owl "http://www.w3.org/2002/07/owl#">
```

```
<!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#">
<!ENTITY lmf "http://www.lexicalmarkupframework.org#">
]>
```

The xsd, owl, rdf and rdfs entities are related to the OWL-DL language. Yet, "lmf" entity is related to the LMF model.

4.2 Used Namespaces

In order to make ontologies more comprehensible and non-ambiguous, OWL offers the possibility of a new component definition: namespaces. This component is an indication for specified vocabularies used in the ontology.

```
<rdf:RDF
    xmlns:owl="&owl;"
    xmlns:rdf="&rdf;"
    xmlns:rdfs="&rdfs;"
    xmlns:xsd="&xsd;"
    xmlns:lmf="&lmf;" >
</rdf:RDF>
```

The above namespaces can be useful for the terms related to the LMF standard. Thus, the first component of the ontology is the definition of a declaration set of XML namespace contained in an opening tag <rdf: RDF>. These statements are used to interpret identifiers and make the following presentation of the ontology much more readable.

4.3 LMF Header and Classes

A set of assertion should be described after the definition of namespaces. These assertions adorn the modeling file by comments, labels, version control and inclusion of other ontologies.

```
<owl:ontology
rdf:about="http://www.lexicalmarkupframework.org">
 <rdfs:comment>Ontology LMF in OWL</rdfs:comment>
 <rdfs:label>LMF Ontology </rdfs:label>
</owl:ontology>
```

Classes in OWL are considered as basic components. All these classes are in fact members of “Thing class”. Concerning our LMF core model, we have defined eight classes as follows:

```
<owl:Class rdf:ID="Lexical Resource"/>
<owl:Class rdf:ID="Global Information"/>
<owl:Class rdf:ID="Lexicon"/>
```

```
<owl:Class rdf:ID="Lexical Entry"/>
<owl:Class rdf:ID="Form"/>
<owl:Class rdf:ID="Sense"/>
<owl:Class rdf:ID="Definition"/>
<owl:Class rdf:ID="Statement"/>
```

The classes described above concern only the LMF core model. Classes in other extensions packages have to be mapped also to OWL-DL.

4.4 LMF SubClasses

Generally, all ontologies contain a list of restrictions. Subclasses are one of those restrictions. In LMF core model, Form Representation are restrictions of the class Representation:

```
<owl:Class rdf:ID="Representation"/>
<owl:Class rdf:ID="Form Representation">
  <rdfs:subClassOf rdf:resource="#Representation"/>
</owl:Class>
```

Only two subclasses are defined above. The LMF model includes several subclasses that should be converted to OWL-DL.

4.5 LMF Properties

Some general and specific information are interpreted as attributes in LMF core model. For example, Global Information is an administrative class involving general attributes, such as /language coding/ or /script coding/ which are suitable for the whole lexical resource:

```
<owl:DatatypeProperty rdf:ID="language coding ">
 <rdfs:domain rdf:resource="#Global Information"/>
 <rdfs:range rdf:resource="#&xsd; String"/>
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="script coding ">
 <rdfs:domain rdf:resource="#Global Information "/>
 <rdfs:range rdf:resource="#&xsd; String"/>
</owl:DatatypeProperty>
```

All attributes figured in the LMF model have to be converted to OWL-DL.

4.6 LMF Relations

LMF relations define a list of domain and co-domain restrictions. For instance, “has lexicon” is an “ObjectProperty” restriction:

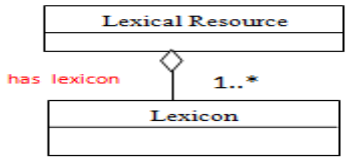

Fig. 3. LMF Relations

The Fig. 3 shows the name relation added for the LMF model. This name will allows us making the transformation to OWL-DL. Modeling the restriction "has lexicon" in OWL, we obtain:

```
<owl:ObjectProperty rdf:ID="hasLexicon">
 <rdfs:domain rdf:resource="#LexicalResource"/>
 <rdfs:range rdf:resource="#Lexicon"/>
</owl:ObjectProperty>
```

We have to add relations names in the LMF core model with the aim of modeling relations restrictions in OWL-DL.

4.7 LMF Cardinalities

Cardinalities are transformed to restrictions in OWL-DL. Thus, they are defined as follows:

```
<owl:Class rdf:ID="LexicalResource">
  <owl:Restriction>
    <owl:onProperty rdf:resource="#hasLexicon"/>
     <owl:minCardinality rdf:datatype="&xsd; nonNegativeInteger">
1</owl:minCardinality>
  </owl:Restriction>
</owl:Class>
```

These cardinalities make the designed model richer in term of restrictions. Thus, individuals created later will have number constraints.

5 Instantiation for Arabic lexicon

The instantiation part is done according to the OWL-DL built scheme. So, in this section, we are going to choose morphological extension in LMF extensions packages. This choice is based upon the importance of this part for the most lexicons in NLP. Morphological extension is treated by two different ways in LMF. The first represents explicitly inflected forms. The second uses the paradigms of flexions to

generate different forms derived from the Lexical Entry. We represent a part from the entire inflectional description of the verb "جلس" "jls" (to sit). This description is in fact an instantiation of the built prototype which is explained in the previous section:

```
<LexicalEntry rdf:ID="جلس فعل">
    <PartOfSpeech>verb</PartOfSpeech>
    <Root>ج ل س</Root>
    <Scheme>فعل</Scheme>
    <WordForm rdf:resource="">
        <writtenForm>تجلسا</writtenForm>
        <grammaticalTense>inaccomplished</grammaticalTense>
        <grammaticalMood>apocated</grammaticalMood>
        <person>2</person>
        <grammaticalGender>feminine</grammaticalGender>
        <grammaticalNumber>dual</grammaticalNumber>
    </WordForm>
</LexicalEntry>
```

The example shows a prototype of one possible inflected form from a set of 56 Word Form that an Arabic verb could take.

6 Discussions

The proposal of modeling Arabic lexicons compliant LMF in OWL-DL is based upon several paradigms: (i.e. header, classes, subclasses, properties). Applying these concepts, we have built a new ontology designed on OWL-DL. However, constructing such a comprehensive ontology is hindered by the complexity of the LMF model. Once the entire LMF ontology (core model and extension packages) is already built, we have to populate it with individuals. With this prototype, making lexicons, in any language, is so easier to build since we have just to instantiate the OWL-DL prototype with the appropriate individuals. However, such modeling includes shortcomings: mapping prototype can lead to the loss of certain informations such as aggregation.

7 Conclusion

We have studied the structure and representation of the LMF mode in order to design an OWL-DL ontology that would be able to match its components maximally. The next step will be to use this model as a tool to check the actual coverage of existing LMF serialisations such as the one anticipated in [18] on the basis of the TEI framework. The underlying vision is to create an interoperable framework describing a dynamic environment among standards and guidelines. Such environments should be both internally coherent and facilitate the continuous update of modeling standards and their serializations when use cases and associate tool development provide new representational needs.

References

1. Farghhly, A., Shaalan, K.: Arabic Natural Language Processing: Challenges and Solutions. In New York, NY, USA, ACM Transactions on Asian Language Information Processing. Vol. 8, No. 4, Article 14, 22 pages(2009)
2. Romary, L., Salmon-Alt, S., Francopoulo, G.: Standards going concrete: from LMF to Morphalou. Workshop Enhancing and Using Electronic Dictionaries - http://hal.inria.fr/inria-00100195 (2004)
3. Francopoulo, G.: Lexical Markup Framework. US, Great Britain and the United States: ISTE Ltd and John Wiley & Sons, Inc. (2013)
4. Sperberg-McQeen, C.M., Burnard, L.: TEI P5: Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative Consortium Charlottesville, Virginia: the TEI Consortium (2014)
5. Smith, M., Welty, C., McGuinness, D.: eds. OWL Web Ontology Language Guide, W3C Recommendation, http://www.w3.org/TR/2004/REC-owl-guide-20040210/ (2004)
6. Heflin, J.: An Introduction to the OWL Web Ontology Language. Lehigh University. National Science Foundation (NSF) (2007)
7. Ide, N., Romary, L.: A Registry of Standard Data Categories for Linguistic Annotation. In Proc. 4th International Conference on Language Resources and Evaluation - LREC'04 135–138 - http://hal.inria.fr/inria-00099858 (2004)
8. Loukil, N., Haddar, K., BenHamadou, A.: Normalisation de la représentation des lexiques syntaxiques arabes pour les formalismes d'unification. 26th conference on Lexis and Grammar, Bonifacio (2007)
9. Pollard, C., Sag, I.: Head-Driven Phrase Structure Grammars. Chigaco University Press (1994)
10. Haddar, K., Fehri, H., Romary, L.: A prototype for projecting HPSG syntactic lexica towards LMF. JLCL Band27(1), pp. 21–46 (2012)
11. Gruber, T.: A Translation Approach to Portable Ontology Specifications. In Stanford University Stanford, California, Knowledge Acquisition Elsevier, pp.199–220 (1993)
12. Noy, N., Hafner, C.: The State of the Art in Ontology Design. AI Magazine. 18(3), pp. 53–74 (1997)
13. Horrocks, I., Patel-Schneider, P.: Reducing OWL Entailment to Description Logic Satisfiability. In Fensel, D., Sycara, K., Mylopoulos, J. (eds.) International Semantic Web Conference (ISWC), pp. 17–29. Springer (2003)
14. Wang, T., Parsia, B., Hendler, J. A Survey of the Web Ontology Landscape. International Semantic Web Conference (ISWC) Springer, pp 682-694. (2006)
15. Baccar, F., Gargouri, B., Ben Hamadou, A.: Towards Generation of Domain Ontology from LMF Standardized Dictionaries. SEKE. 515-520 (2010)
16. TAUS, Report on a TAUS research about translation interoperability, 25 February 2011.
17. Heflin, J., Pan, Z.: A Model Theoretic Semantics for Ontology Versioning. In Third International Semantic Web Conference (ISWC). Springer, pp 62-76. (2004)
18. Romary, L. TEI and LMF crosswalks, to appear in Gradmann S. (Ed.) Digital Humanities: Wissenschaft vom Verstehen, Humboldt Universität zu Berlin - http://hal.inria.fr/hal-00762664 (pre-print, 2013)