LMM: an OWL-DL MetaModel to Represent Heterogeneous Lexical Knowledge

Davide Picca, Alfio Massimiliano Gliozzo*, Aldo Gangemi*

University of Lausanne, *ISTC-CNR
CH 1015-Lausanne-Switzerland, *Via Nomentana 56-00161-Roma-Italy
davide.picca@unil.ch, {*alfio.gliozzo,aldo.gangemi}@istc.cnr.it

Abstract

In this paper we present a Linguistic Meta-Model (LMM) allowing a semiotic-cognitive representation of knowledge. LMM is freely available and integrates the schemata of linguistic knowledge resources, such as WordNet and FrameNet, as well as foundational ontologies, such as DOLCE and its extensions. In addition, LMM is able to deal with multilinguality and to represent individuals and facts in an open domain perspective.

1. Introduction

Bridging lexical resources and ontologies is becoming a prominent research topic in the Semantic Web community as well as in Computational Linguistics and Information Retrieval (Freitag, 1998), interest manifested by several workshops as *Ontolex,Ontology Learning and Population* and research projects as DBpedia (Auer et al., 2007) and YAGO (Suchanek et al., 2007). The main reason for this interest is that new generation Web content (e.g. social tagging produced by web communities, known as Web 2.0, ontologies from the Semantic Web, collaboratively developed text and semantic networks) could be combined and boosted by an adequate linguistic interpretation of terms and predicates expressed in a language with a formal semantics. In addition, the computational linguistic community has developed large scale and open domain repositories of lexical knowledge such as WordNet and FrameNet. Nowadays, such repositories are big enough to cover almost any area of human interest and activity, providing a first shallow linguistic interpretation of the basic linguistic concepts.

Since the privileged medium for internet communication is still natural language text in any of its forms (i.e. web pages, chats, blogs, VoIP), the development of a truly Semantic Web (i.e. the development of formalisms for knowledge representation, technologies for knowledge acquisition and automatic systems that are able to exploit such knowledge in order to provide intelligent services to the user over the Web) passes trough the definition of a semiotic model, which is able to represent natural language expressions, their meaning, and a formal semantics to give an interpretation to individuals and facts.

To this aim, we developed LMM, a Linguistic Meta-Model that provides a semiotic-cognitive representation of linguistic knowledge and grounds it in a formal semantics. LMM integrates linguistic knowledge sources, such as WordNet (Fellbaum, 1998) and Framenet (Baker et al., 1998), as well as foundational ontologies, such as DOLCE (Gangemi et al., 2002) and its extensions, notably the Descriptions and Situations framework (Gangemi, 2008)[1]. LMM expands all social-cognitive aspects as they are defined in DOLCE and in the Descriptions and Situations framework, in order to adapt them to a semiotic perspective. This ploy offers a new linguistic enforcement to the foundational layer.

LMM is described in OWL - DL, achieving the desirable goal of interoperability with existing Semantic Web applications.

LMM is characterized by the following features:

- LMM is compatible with the more consolidated and accepted semiotic theories.
- LMM is fully aligned with the existing foundational ontologies and lexical resources, and in particular DOLCE-Ultralite (the OWL version including basic DOLCE and the Descriptions and Situations framework), WordNet and FrameNet.
- LMM is capable to represent information coming from standard Information Extraction technology from text technology, such as Named Entity recognition, Relation Extraction and Frame Detection.
- LMM allows to represent expressions from different languages and their relations.
- LMM is compatible with existing Semantic Web technology, such as OWL and reasoners.
- LMM is able to represent multilingual knowledge provided by existing large scale, collaboratively developed knowledge bases, such as YAGO and DBpedia.

We delivered LMM as a public Semantic Web resource, that can be downloaded from $http://www.loa-cnr.it/codeps/owl/LMM\_Alignments.owl$. In the rest of the paper, we will firstly describe the semiotic notions underlying the development of LMM (see section 2.), then we describe its basic components and relations (see Section 3.). In Section 4. we show how LMM has been aligned to WordNet, FrameNet and DOLCE. Finally, section 5. concludes the paper, highlighting interesting perspectives for future applicability of LMM to socially produced Knowledge Bases, such as Yago, DBpedia and Dmoz.

[1] http://wiki.loa-cnr.it/index.php/LoaWiki:Ontologies is a wiki that collects information about those ontologies in the OWL format, and where they can be downloaded from.

2. A Semiotic Model

The most important feature of LMM is its ability to support the representation of different knowledge sources developed according to different underlying semiotic theories. This is possible because most knowledge representation schemata, either formal or informal, can be put into the context of so-called *semiotic triangle* (Peirce, 1958).

The semiotic triangle, represented in Figure 1 (the original version), and in Figure 2 (the OWL ontology version we have designed), is used to discuss the differences between objects, concepts and symbols. It has been originally proposed by (Peirce, 1958).

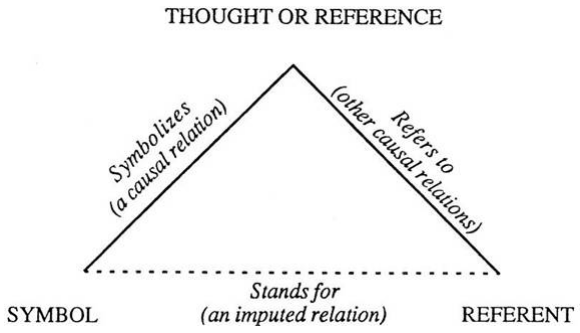

Figure 1: The structure of *Semiotic Triangle*

The interpretation of the semiotics insight behind the semiotic triangle is quite complex, and it's outside the scope of this paper to provide the required philosophical background. In this section we will simply provide some intuitive explanation motivating our decisions in developing the model.

Referents. Intuitively, the reference level (or referent, in the picture) is populated by any possible individual in the logical world, and by the fact where they occur together in some relation. Everything could be a reference object, including expression of the language itself (e.g. the word dog has three letters).

Thoughts. Providing a formal definition characterizing the class meaning (thought in our picture) is certainly a very complex task. In fact different notions of meaning could involve:

- Meaning of a term as a paraphrase (or 'gloss', or 'definition'), this is the meaning conceived by lexicographers.
- Meaning as concept schemes like thesauri and lexicons, which assume that the meaning of a term is a 'concept', encoded as a 'lemma', 'synset', or 'descriptor'.
- Meaning as a concept encoded in a cognitive system. This is the aspect captured mainly by psychologists and cognitive scientists.
- Meaning as a social object spread across the members of a community that use that object. This is the aspect captured mainly by social scientists and semioticians.

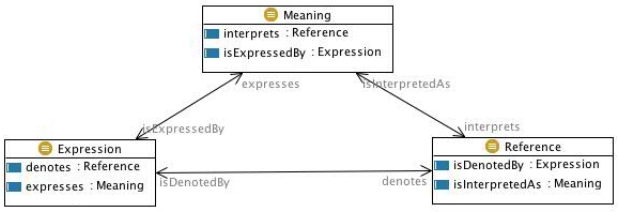

Figure 2: The semiotic triangle in LMM1

- Meaning as a logical component, equivalent to the set of individuals that the term can be applied to; for example, the meaning of 'Ali' is e.g. an individual person called Ali, the meaning of 'Airplane' is e.g. the set of airplanes, etc.
- Meaning taken by structuralist linguistics and frame semantics is the relational context in which an information object can be applied; for example, a meaning of 'Airplane' is situated e.g. in the context ('frame') of passenger airline flights.

In developing LMM, we tried to take into account all these aspects of meaning, as illustrated by Figure 4. In our work, we simply accept that instances of the class `Meaning` can be any kind of meta objects, such as classes of instances (i.e. logical concepts), classes of events (i.e. frames), vocabulary definitions, topics, etc.

Symbols. Language expressions (populating the sign corner of the triangle) can be related to either instances of the class meaning or instances of the class reference (in the first case they will be *interpreted*, in the second they *denote*). To this aim, the semiotic triable introduces an additional layer containing symbols, which are expressions formulated in a given semiotic system, such as a natural language or an iconic code. The fundamental property for an object to be a symbol is that it should be expressed in the context of a speech act, involving an agent acting in a particular frame and topic with some communicative intention (e.g. asking for information about buying product and services). In this case, the symbol will be interpreted following the relations established by the sides of the semiotic triangle. Otherwise, expressions are only potentially related to their references and meanings, but actually they remains ambiguous until such a speech act is fully interpreted. Polysemy of expressions is represented by connecting the same sign to more than one meaning/reference. Following the opposite path, the semiotic triangle provide a way to interpret the real world contexts, also providing a bridge between knowledge and language.

3. A Formal Lexical MetaModel

Following the intuition described in Section 2., we developed the core component of LMM, an OWL-DL ontology represented by Figure 2 called LMM1.

LMM1 is composed by three classes: `Reference`, `Meaning` and `Expression`, formalizing the distinctions of the semiotic triangle introduced in the previous section.

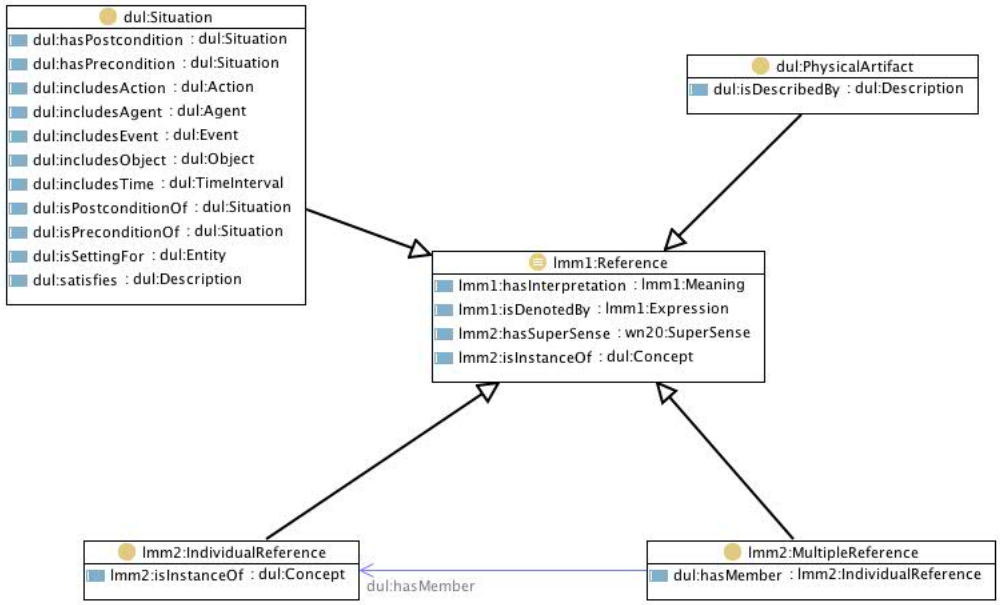

Figure 3: The class `Reference`

3.1. Reference

The `reference` level, represented by Figure 3, is populated by any possible individual in the logical world, being it either a concrete object or any other social object whose existence is stipulated by a community. Individuals are related by the fact that they co-occur into events.
Instances of the class `Reference` are all those *entities* belonging to the universe of discourse, including e.g. *physical objects*, *events*, etc., and they have a explicit reference "in the world".
In particular LMM distinguish between `physical objects`, `individual references`, `multiple references` and `situations`.

PhysicalArtifact allows to talk of artifacts in a very general sense, i.e. including recycled objects, objects with an intentional functional change, natural objects that are given a certain function, even though they are not modified or structurally designed, etc. Immaterial (non-physical) artifacts (e.g. texts, ideas, cultural movements, corporations, communities, etc.) can be modelled as social objects, which are all 'artifactual' in the weak sense assumed here. This concept is derived from DOLCE.

IndividualReference can have members that are individual references. They are typically Named Entities, and this class makes it easy the automatic population task by extraction tools based on Named Entity Recognition (NER).

MultiReference can have instances that are 'collective' individuals, whose members have a superclass in `MultiReference`. For example, JohnDoe (an IndividualReference) isMemberOf ACME_Employees (a MultipleReference), that dul:isCoveredBy (a DOLCE-Ultralite relation) the Employee Concept.

Situation is the circumstantial context where entities and events occur. This is a very important class, because it can belong either to the class `Reference` or the class `Meaning`. For example, the sentence 'John hit the ball' can refer to a Situation where someone called John actually hit a ball. On the other hand, the meaning of that sentence arises from the observation of someone, which uses her interpretive capabilities to single it out. A Situation is therefore still a 'constructed' entity, therefore it can also be a Meaning.

It is worthwhile to remark here that the class `Reference` can be populated by exploiting existing tools for Named Entity Recognition, therefore instantiating individuals of the class `IndividualReference`, or alternatively by alignment with ontological resources such as DBpedia, containing references to individuals and facts.

3.2. Meaning

Concepts are represented as instances of `meaning` objects (see Figure 4). Concepts are related between each other in two different ways. Subsumption relations organize concepts into hierarchies of subclasses (e.g. the dog is an animal). These relations are reflected at the extensional level by the fact that the set of instances denoted by the subclasses are contained into the set of instances of their superclasses. Conceptual relations are in turn represented by descriptions (whose definition is mutuated from the Description and Situations framework), expressing the possibility for events to occur. Descriptions can be considered reified relations, therefore they are actually instances of the meaning class themselves.

Frames are examples of descriptions (e.g. the frame of drinking express the possibility for a person to drink some beverage at a certain location and time). Concrete instantiations of frames (which are situations in the sense outlined above) are reflected by events at the reference level (e.g. the fact that Alfio Gliozzo is drinking a diet coke in the airplane today is an occurrence of the frame of drinking). Frames can be related between each other (e.g. subframe of , and so on). The meaning level is also populated by objects called `topics`, whose property is to define collections of concepts, frames and events characterized by the fact that they are contained in the conceptual area covered by the topic. Following a spatial metaphor introduced by Chris Welty (Welty and Jenkins, 1999) , the relations between topics are broader than (indicating that the conceptual area covered by a supertopic contain those of its subtopic, e.g. science and chemistry) and similar to (indicating that two different regions covered by a topic somehow overlaps). In contrast to concepts, topics are collections of concepts, whereas concepts denote collections of individuals.
The Figure 4 shows the sub-classes of the class `Meaning`.

Description can be thought also as a 'descriptive context' that uses or defines concepts in order to create a view on a 'relational context' (cf. `Situation`) out of a set of data or observations. For example, a Plan is a Description of some actions to be executed by agents in a certain way, with certain parameters. It is linked to the class `Concept` by means of the main relation `defines`.

Collection has as main task to give a unique coherent term to the class `Description` by means of the main relation `IsunifiedBy`. It can be thought as any container for entities that share one or more common properties. E.g. "stone objects", "the nurses", "the Louvre Egyptian collection", all the elections for the Italian President of the Republic.

Situation is the realization of a certain description. This is a very important class because the class `Situation` is *in acto* what the class `Description` is *in potentia.* Thank to this characteristic, the class `Situation` serves as a bridge between the class `Meaning` and the class `Reference` passing through the class `Description`.

Concept can be used in other descriptions by means of the main relation `isConceptUsedIn`. Concept is intertwined with SKOS () Concept, but SKOS notion also covers our notion of Topic.

All those classes are explicitly inherited from the Descriptions and Situations framework as represented in DOLCE-Ultralite, and they aim at catching the basic semiotic aspects involved in semantic technologies.

3.3. Expression

Finally, the two layers of meaning and reference are connected to the language by means of the `expression` layer (see Figure 6). Expressions are social objects produced by agents in the context of communicative acts. They are natural language terms, symbols in formal languages, icons, and whatever can be used as a vehicle for communication. Expressions denoting concepts, frames and topics (such as person, drink and sport) are interpreted by means of their connections to the meaning layer, while expressions denoting instances (e.g. Alfio Gliozzo) are directly connected to their corresponding individuals. It is important to remark that expressions, regardless of the context where they have been expressed, are usually strongly polysemic, reflecting the fact that they can be related to several concepts and/or instances. On the other hand, when they are inside a particular frame and topic, they become less ambiguous, allowing the model to work as a knowledge base.
As shown, instances of the classes `Expression` and `Reference` can be either directly connected through the relation `denotes` or indirectly connected by means of an intermediate conceptual level, which is assumed as being shared by speakers from a given community. The first case is typical for named entities, since in that case we can assume (simplifying the more complex interpretation function) that names refer directly to entities in the external world in virtue of some pre-existing ostensible act. For example, the proper noun "Leonardo da Vinci" `denotes` the person Leonardo da Vinci.
The LMM2 module extends LMM in order to talk about specific linguistic constructs and references, which are represented as a kind of *information objects*. An information object is a piece of information, such as a musical composition, a text, a word, a picture, independently from how it is concretely realized (concrete realizations are called information realizations). It uses the ontology IOL.owl, an ontology that contains several classes and relations for types of information objects and information realizations.
The classes of linguistic constructs in LMM2 closely mirror the distinctions between references. In particular, the class `lmm2:Name`, denotes either named entities (cf. `lmm2:NamedEntity`) or collective references (cf. `lmm2:ExtensionalReference`). A name is a proper noun that denotes an IndividualReference, be it singular or plural, e.g. 'John Zorn', 'Daimler Benz', 'FaceBook' (as a community).
The class `lmm2:ConceptExpression`, denotes multireferences (cf. `lmm2:MultipleReference`). A concept expression is a a term that expresses a Meaning, and denotes a MultipleReference, e.g. 'Dog', 'Black box'. Concept expression can be simple (e.g. single words) or polyrhematic (e.g. multiwords).
The class `lmm2:ContextualExpression`, denotes either contextual references (cf. `lmm2:ContextualReference`, or collective references (cf. `lmm2:ExtensionalReference`). A contextual expression is a a term that denotes a reference via anaphora or deixis, e.g. 'the dog over there', 'all my family', 'the current ACME employees', 'the lion described above'.

4. Alignments

The primary goal of LMM is to allow a simultaneous representation of multiple linguistic knowledge sources, al-

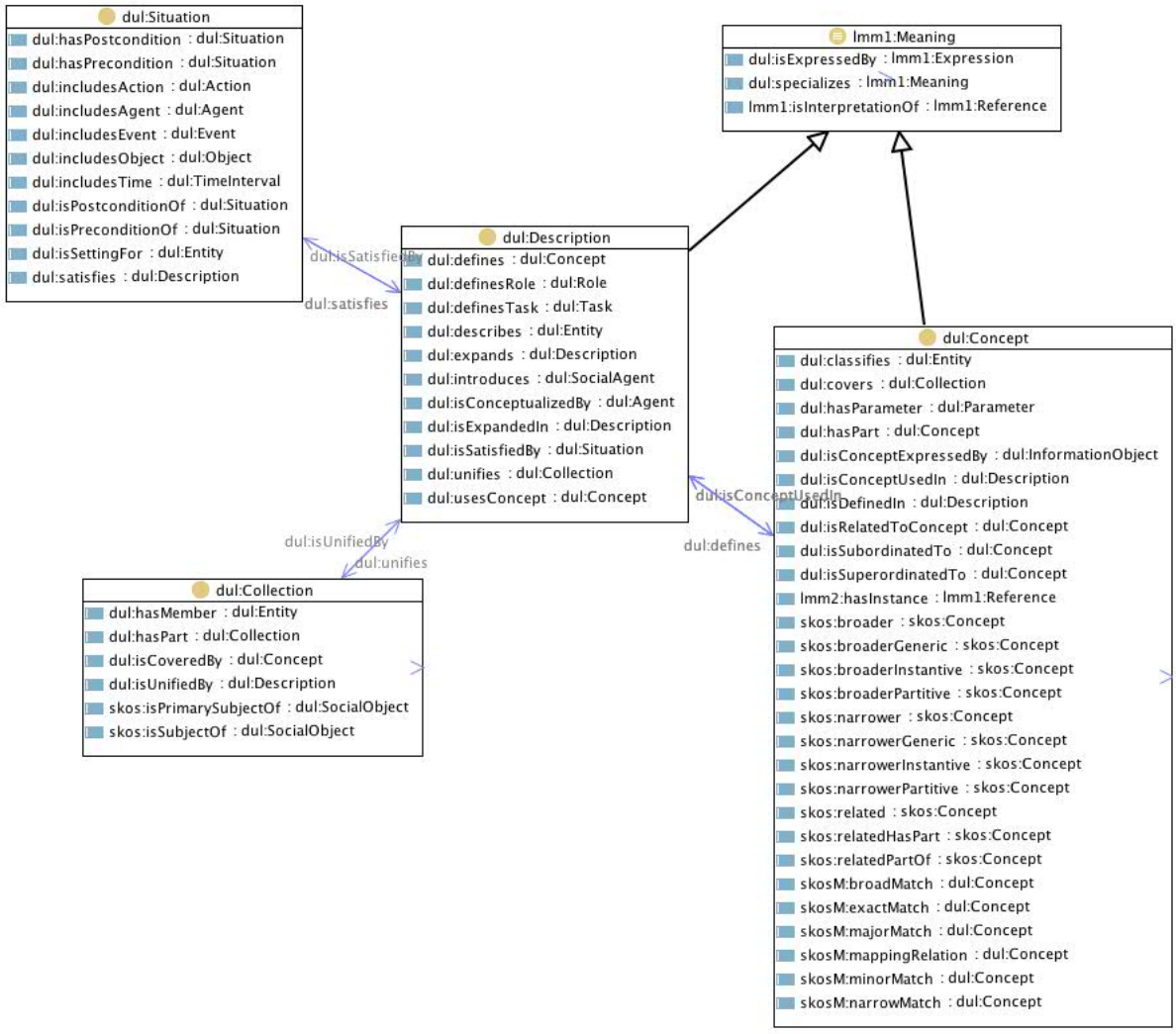

Figure 4: The class `Meaning`

lowing interoperability and an improved comprehension and exploitation of knowledge. To this purpose, a third level (LMM_Alignments) imports LMM_L1 and LMM_L2, which import DOLCE-Ultralite (including DOLCE and Descriptions and Situations) and other plugins that contain patterns to model communication acts, special kinds of information objects, plans, systems, etc.

As a first task, we have aligned WordNet, FrameNet, as well as some Web2.0 schemas and seamntic web ontologies. The alignments are included in the standard distribution of LMM.

The coverage currently includes mappings from WordNet (2.0) (Fellbaum, 1998) (in the W3C OWL version of the schema), the WordNet SuperSenses (i.e. the top level concepts of WordNet used by NLP researchers and lexicographers), the WordNet domains as designed by (Gliozzo, 2005), OntoWordNet (Gangemi et al., 2003), SKOS (Miles et al., 2005) (the W3C schema for thesauri), FrameNet (1.2) (Baker et al., 1998) (in the OntoFrameNet reengineering, (Gangemi et al., to appear)), OwlOdm (an OWL1.0 metamodel), etc.

LMM_Alignments is also the level that permits to link LMM to YAGO (Suchanek et al., 2007) and DBpedia (Auer et al., 2007).

The alignments consisted e.g. in specializing the class `Concept` with different notions from the aligned resources: in WordNet with `synset`, in FrameNet with the class `FrameElement` and in SKOS with the class `Concept`. As an example, Figure 5 illustrate the class `Concept` and its connections to other elements in the ontology.

Thanks to the generality of the semiotic theory represented in LMM_L1, any possible linguistic resource can be easily aligned. the alignment of Concept-related notions allow to frame WordNet synsets lexical information in a context provided by FrameNet frame elements, as well as to catch the idea of a formal relational context from DOLCE Ultralite. Similarly with relations: for example, the hyponymy/hyperonymy between from WordNet synsets is aligned under the `dul:specializes` transitive relation, so the formal context of LMM becomes available to make integrated queries and navigation in the different knowledge bases.

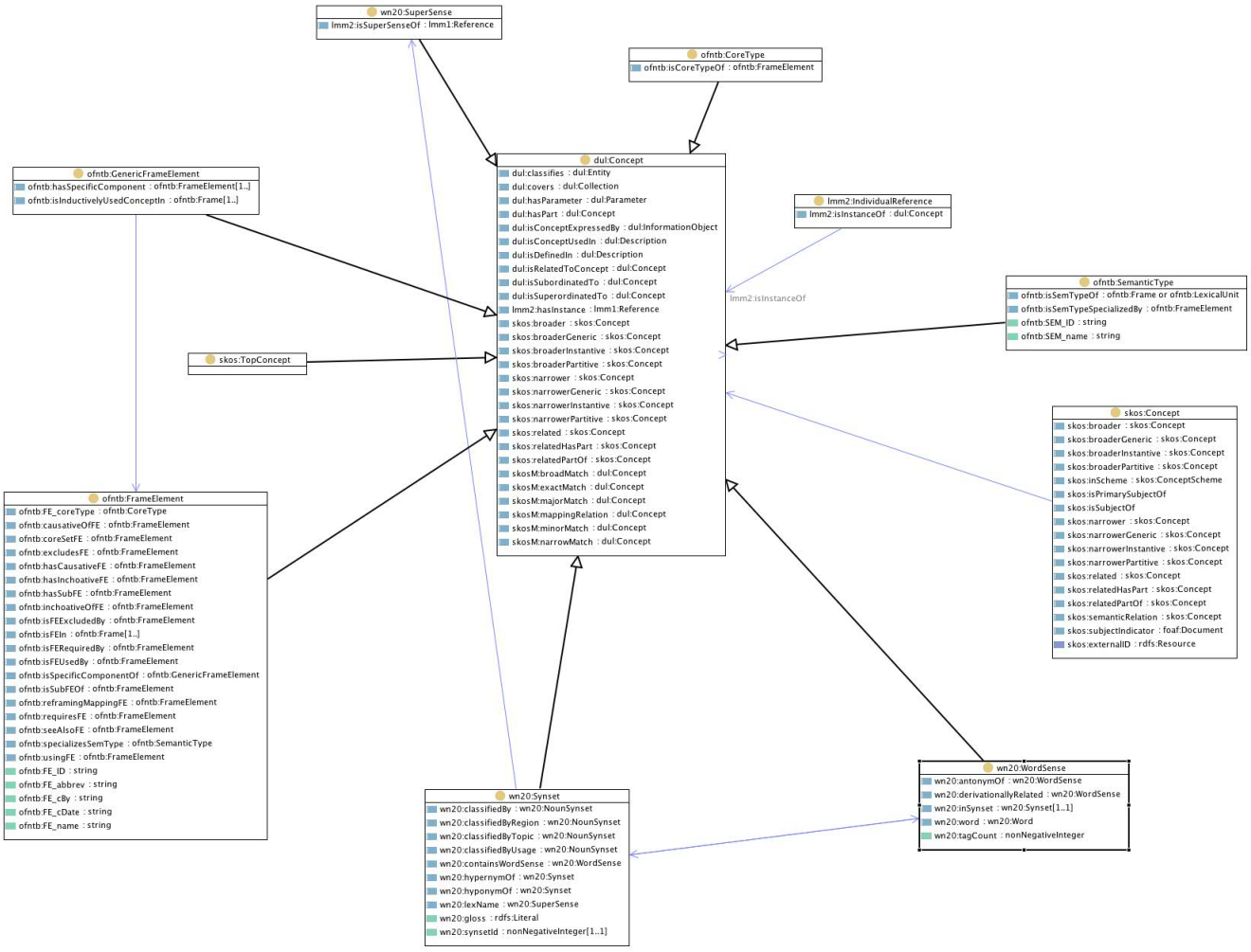

Figure 5: The structure of `Concept`

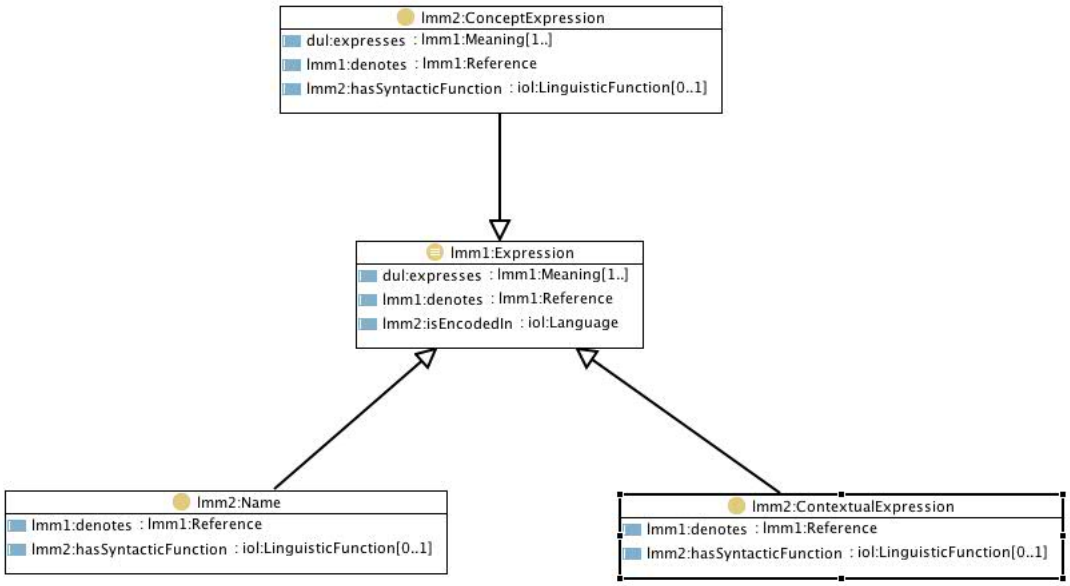

Figure 6: The class `Expression`

5. Conclusion and Future Work

In this paper we have presented LMM, an innovative Formal Linguistic Meta-Model that is designed as a semiotic-cognitive representation of lexical knoweldge. LMM is freely available and integrates *dictionary-like* and *encyclopedia-like* knowledge sources into a unique robust structure that also allows to represent multilinguality. As a matter of fact, as shown in the previous section, the class `Concept` can be characterized either extensionally or intensionally. In particular, we claimed that the intensional characterizations can be created by establishing semantic relations among concepts, which follows the structuralist paradigm (de Saussure, 1922), and realizing the Eco's notion of *dictionary* (Eco, 1976) . For example, taxonomies in WordNet are represented by adopting the relation `isHyperonymOf` as a subrelation of the `dul:specializes` relation, but holding between instances of the class `Synset`, a subclass of `Concept` on its turn.

On the other hand, due to the flexibility of the interpretation schema provided by our formalization of the semiotic triangle, concepts can be also characterized by extensional interpretations. In fact, we have planned to create an automatic connection to DBpedia and YAGO. This is allowed through the alignment of their schemata, as well as of the class `skos:Concept`. In fact both YAGO and DBpedia use SKOS for stocking information. Thanks to the LMM alignment, we are able to refer to the encyclopedic knowledge of YAGO and DBpedia. It is worth remarking that this novel feature permits to formalize Eco's notion of *encyclopedia* (Eco, 1976). In such a way knoweldge engineering is done without being affected by obstacles arising from the different realizations within specific natural languages, and it can be considered an ideal framework to design a multilinguality model.

6. References

S Auer, C Bizer, G Kobilarov, and J Lehmann. 2007. Dbpedia: A nucleus for a web of open data. *In proceedings of 6th Int'l Semantic Web Conference (iswc2007)*, Jan.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*.

F. de Saussure. 1922. *Cours de linguistique générale*. Payot, Paris.

U Eco. 1976. A theory of semiotics. *Indiana University Press*, Jan.

C Fellbaum. 1998. Wordnet: An electronic lexical database. *MIT Press*, page 423, Jan.

D. Freitag. 1998. *Machine Learning for Information Extraction in Informal Domains*. Ph.D. thesis, Carnegie Mellon University.

A Gangemi, N Guarino, C Masolo, and A Oltramari. 2002. Sweetening ontologies with dolce. *Proceedings of EKAW*, Jan.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The ontowordnet project: extension and axiomatization of conceptual relations in wordnet. In *Proceedings of ODBASE03*.

Aldo Gangemi, C. Catenacci., and M. Nissim:, to appear. *What's in a Schema*. Cambridge UP.

A. Gangemi. 2008. Norms and plans as unification criteria for social collectives. *Journal of Autonomous Agents and Multi-Agent Systems*, 16(3).

A. Gliozzo. 2005. *Semantic Domains in Computational Linguistics*. Ph.D. thesis, University of Trento.

Nicola Guarino and Pierdaniele Giaretta. 1995. Ontologies and knowledge bases: Towards a terminological clarification. *Towards Very Large Knowledge Bases*.

A Miles, B Matthews, M Wilson, and D Brickley. 2005. Skos core: Simple knowledge organisation for the web. *in Proceedings of International Conference on Dublin Core and Metadata Applications*, Jan.

Charles Sanders Peirce. 1958. *Collected Papers of Charles Sanders Peirce*. MIT Press, Cambridge, Mass.

F Suchanek, G Kasneci, and G Weikum. 2007. Yago: A large ontology from wikipedia and wordnet. *Technical Report*.

Christopher A. Welty and Jessica Jenkins. 1999. Formal ontology for subject. *Data Knowl. Eng.*, 31(2):155–181.