LIME: The Metadata Module for OntoLex

Manuel Fiorelli[1], Armando Stellato[1](✉), John P. McCrae[2],
Philipp Cimiano[2], and Maria Teresa Pazienza[1]

[1] ART Research Group, University of Rome "Tor Vergata", Rome, Italy
{fiorelli, stellato, pazienza}@info.uniroma2.it
[2] Cognitive Interaction Technology Center of Excellence,
University of Bielefeld, Bielefeld, Germany
{jmccrae, cimiano}@cit-ec.uni-bielefeld.de

Abstract. The OntoLex W3C Community Group has been working for more than three years on a shared lexicon model for ontologies, called *lemon*. The *lemon* model consists of a core model that is complemented by a number of modules accounting for specific aspects in the modeling of lexical information within ontologies. In many usage scenarios, the discovery and exploitation of linguistically grounded ontologies may benefit from summarizing information about their linguistic expressivity and lexical coverage by means of metadata. That situation is compounded by the fact that *lemon* allows the independent publication of ontologies, lexica and lexicalizations linking them. While the VoID vocabulary already addresses the need for general metadata about interlinked datasets, it is unable by itself to represent the more specific metadata relevant to *lemon*. To solve this problem, we developed a module of *lemon*, named LIME (Linguistic Metadata), which extends VoID with a vocabulary of metadata about the ontology-lexicon interface.

Keywords: Ontolex · Metadata · Ontologies · Natural language · Discovery

1 Introduction

Ontologies and widely shared vocabularies are the cornerstone of the Semantic Web as they provide the basis for interoperability as well as for reasoning, consistency detection, etc. Yet, the grounding of ontology and vocabulary elements in natural language is crucial to ensure communication with humans [1]. Enriching ontologies and Semantic Web vocabularies with information about how the vocabulary elements are expressed in natural language is crucial to support tasks such as ontology mediation [2] as well as in all tasks in which natural language needs to be interpreted with respect to a formal vocabulary or ontology (e.g. question answering [3, 4], ontology-based information extraction [5], ontology learning [6]) or in which natural language descriptions need to be generated from a given ontology or dataset [7–9].

A number of models have been proposed to enrich ontologies with information about how vocabulary elements are expressed in different natural languages, including the Linguistic Watermark framework [10, 11], LexOnto [12], LingInfo [13], LIR [14], LexInfo [1] and more recently *lemon* [15].

F. Gandon et al. (Eds.): ESWC 2015, LNCS 9088, pp. 321–336, 2015.
DOI: 10.1007/978-3-319-18818-8_20

The OntoLex W3C Community Group[1] has the goal of providing an agreed-upon standard by building on the aforementioned models, the designers of which are all involved in the community group. Additionally, linguists have acknowledged [16] the benefits that the adoption of the Semantic Web technologies could bring to the publication and integration of language resources. As such, the Open Linguistics Working Group[2] of the Open Knowledge Foundation is contributing to the development of a LOD (Linked Open Data) (sub)cloud of linguistic resources.[3]

These complementary efforts by Semantic Web practitioners and linguists are in fact converging, as the ontology lexicon model provides a principled way [17] to encode even notable resources such as the Princeton WordNet [18, 19] and other similar ones (which we will refer to hereafter as wordnets) for other languages.

The *lemon* model envisions an open ecosystem in which ontologies[4] and lexica for them co-exist, both of which are published as data on the Web. It is in line with a many-to-many relationship between: (i) ontologies and ontological vocabularies, (ii) lexicalization datasets and (iii) lexical resources. While an OWL T-Box consists essentially of classes and properties, a lexicon mainly consists of a collection of lexical entries. Lexicalizations in our sense are reifications of the relation between an ontology reference and the lexical entries by which these can be expressed within natural language. *lemon* foresees an ecosystem in which many independently published lexicalizations and lexica for a given ontology co-exist. Within such an ecosystem, it is crucial to support the discovery of lexica and lexicalizations for a given ontology according to a number of criteria. Relevant criteria in choosing a particular lexicalization or lexicon include the following:

- **Vocabulary Coverage:** How many vocabulary elements of a given ontology are covered by at least one lexicalization in the lexicon?
- **Language Coverage:** How many natural languages are covered in the lexicon?
- **Variation:** How many different lexicalizations are there per vocabulary element?
- **Linguistic Model:** Which model is used to express lexicalizations for vocabulary elements (rdfs:label, skos/skosxl:{pref,alt,hidden}Label, *lemon*, LexInfo, etc.?)

When data are immediately accessible, it may be the case that relevant metadata can be computed automatically by statistical profiling. However, its explicit representation through a dedicated vocabulary is still useful for many reasons. Firstly, it promotes architectural clarity, by separating metadata gathering and exploitation. Concerning the latter, available approaches include symbolic manipulation of structured metadata, as well its use in the construction of a feature space for the application of machine learning algorithms. The second advantage of explicit metadata is that metadata can be computed once and be reused multiple times, possibly avoiding computationally

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

[2] http://linguistics.okfn.org/.

[3] http://nlp2rdf.lod2.eu/OWLG/llod/llod.svg.

[4] It would be more appropriate to adopt the term "reference dataset" (including thus also SKOS thesauri and datasets in general), to express data containing the logical symbols for describing a certain domain. In line with the traditional name OntoLex (and thus the ontology-lexicon dualism), we will however often refer to them with the term ontology.

intensive queries over the actual data. In fact, the reuse of pre-computed metadata opens it up the possibility of aggregating metadata in Web accessible repositories that can answer queries expressed through the metadata vocabulary.

In this paper, we introduce LIME (**Li**nguistic **Me**tadata), the metadata vocabulary for the *lemon* model. The paper is structured as follows: in the next Sect. 2 we discuss related work, mainly related to the representation of metadata. Section 3 briefly introduces the Lexicon Model for Ontologies (*lemon*) reflecting the current agreements of the OntoLex community group. Section 4 introduces requirements on the metadata vocabulary, and Sect. 5 presents the actual vocabulary. In Sect. 6, we sketch an application scenario for the model in the context of ontology mediation or alignment. We conclude in Sect. 7.

2 Related Work

Semantic Web practitioners have accepted the necessity of metadata describing the interlinked datasets themselves (e.g. what is it about? [20]), rather than focusing only on the description of entities in the universe of discourse.

VoID (Vocabulary of Interlinked Datasets) [21] satisfied the need for a machine-understandable coarse-grained description of the LOD as a whole, by defining a vocabulary of metadata about datasets and their interconnections, as well as mechanisms to publish, locate and aggregate dataset descriptions. The VoID framework can be extended for different usages. VOAF (Vocabulary of a Friend)[5] is one such extension, supporting the description of OWL ontologies and RDFS schemas. VOAF distinguishes various types of dependencies between vocabularies, supports the categorization of vocabularies, and defines statistical metrics relevant to vocabularies (e.g. number of classes). VOAF can be complemented with modules providing additional metadata (e.g. the preferred prefix). Currently, the LOV (Linked Open Vocabularies)[6] service exploits VOAF metadata to support the navigation and discovery of vocabularies and to understand their relationships. LOV mashes up the data provided by LODStats [22] on the usage of vocabularies in the LOD.

DCAT (Data Catalog Vocabulary) [23] is a related vocabulary for the description of data catalogs on the Semantic Web, aiming at improving their discoverability and supporting federated queries across them. While DCAT is agnostic with respect to data models/formats, it is possible to combine it with other format-specific vocabularies, such as VoID in the case of RDF datasets.

In the field of HLT (Human Language Technology), structured metadata supports the reuse of Language Resources (LRs). The OLAC (Open Language Archives Community) [24] metadata model provides a template for the description of LRs, by extending the Dublin Core Metadata Element Set.[7] Supported metadata includes, among others, provenance metadata, resource typology and language identification.

[5] http://purl.org/vocommons/voaf.

[6] http://lov.okfn.org/.

[7] http://dublincore.org/documents/dces.

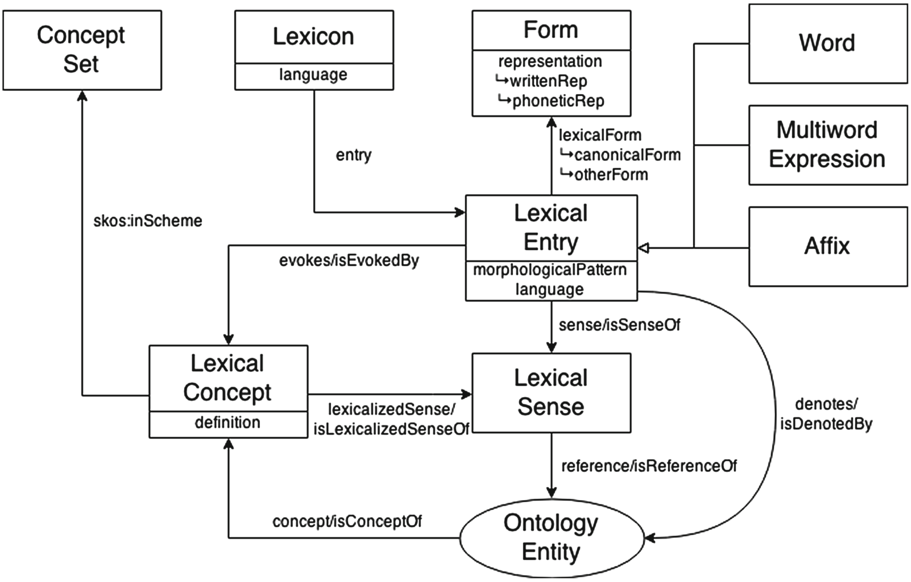

Fig. 1. The Lemon/OntoLex Model as presented in the Ontolex Final Model Specification, and available at: http://www.w3.org/community/ontolex/wiki/Final_Model_Specification. For some properties the inverse is denoted as ‘property/inverse property’; only the direction of the first property is indicated in the diagram.

OLAC is intended to specialize the general infrastructure provided by OAI (Open Archives Initiative) [25], which supports the federation of archives and the aggregation of the associated metadata.

While OLAC aims to define a distributed infrastructure for resource sharing, LRE Map [26] is a crowd-sourced catalog of LRs, initially fed by authors submitting papers to LREC Conferences. LRE Map defines numerous resource types and usage applications, whilst OLAC distinguishes only a handful of types. Similar in scope to OLAC, META-SHARE [27] has its own metadata schema. These works commit to a definition of LR that includes both software tools (e.g. part of speech taggers and parsers) and data (e.g. corpora, dictionaries and grammars) expressed in different formats. Because of their broad coverage, these works fail to provide specific metadata for the description of the relationship between ontologies and lexica, which is the core of OntoLex. Moreover, these works are not specifically tailored to the description of Semantic Web datasets, nor do they fit the metadata ecosystem that is being developed on the Semantic Web through initiatives such as VoID and DCAT.

Starting from previous works about metadata for linguistic resources [10], we filled this gap by proposing a standard (LIME) that extends VoID to provide descriptive statistics at the level of the lexicon-ontology interface, in particular for the *lemon* model developed by the OntoLex community group. The model we present here represents a refined version of the initial proposal [28] that was seeded to the community before *lemon* was finalized.

3 The *Lemon*/OntoLex Model

The *lemon* model (see Fig. 1) developed by the OntoLex community group is based on the original *lemon* model, which by now has been adopted by a number of lexica [29–32], and as such was taken as the basis of the OntoLex community group to develop an agreed-upon and widely accepted model. The *lemon* model is based onto the idea of a separation between the lexical and the ontological layer following Buitelaar [33] and Cimiano et al. [34], where the ontology describes the semantics of the domain and the lexicon describes the morphology, syntax and pragmatics of the words used to express the domain in a language. The model thus organizes the lexicon primarily by means of *lexical entries*, which are a word, affix or multiword expression with a single syntactic class (part-of-speech) to which a number of *forms* are attached, such as for example the plural, and each form has a number of *representations* (*string forms*), e.g. written or phonetic representation. Entries in a lexicon can be said to *denote* an entity in an ontology, however normally the link between the lexical entry and the ontology entity is realized by a *lexical sense* object where pragmatic information such as domain or register of the connection may be recorded.

In addition to describing the meaning of a word by reference to the ontology, a lexical entry may be associated with a *lexical concept*. Lexical concepts represent the semantic pole of linguistic units, and are the mentally instantiated abstractions which language users derive from conceptions [35]. Lexical concepts are intended primarily to represent such abstractions when present in existing lexical resources, e.g. synsets for wordnets. An example of a lexical entry lexicalizing the property knows in the FOAF (Friend of a Friend) vocabulary (http://xmlns.com/foaf/spec/) is as follows:

```
:acquainted_with a ontolex:LexicalEntry;
   lexinfo:partOfSpeech lexinfo:adjective;
   ontolex:canonicalForm :acquainted_form;
   synsem:synBehavior :acquainted_adjective_frame;
   ontolex:sense :acquainted_with_sense.

:acquainted_form a ontolex:Form;
   ontolex:writtenRep “acquainted”@en.

:acquainted_adjective_frame a lexinfo:AdjectivePPFrame;
   lexinfo:coplativeArg :acquainted_adjective_arg1;
   lexinfo:prepositionalObj :acquainted_adjective_arg2.

:acquainted_with_sense ontolex:reference foaf:friend;
   synsem:subjOfProp :acquainted_adjective_arg1;
   synsem:objOfProp :acquainted_adjective_arg2.

:acquainted_adjective_arg2 synsem:marker :with;
   synsem:optional “false”^^xsd:boolean .

:with a ontolex:LexicalEntry;
   ontolex:canonicalForm :with_form .

:with_form ontolex:writtenRep “with”@en .
```

The *lemon* model is structured into a core module (ontolex prefix in the example above) and four additional modules. Firstly, the *syntax and semantics* (synsem prefix) module describes the syntactic usage of a frame and furthermore how this syntax can be mapped into logical representations, as well as further conditions that may affect whether a word can be used for a concept in the ontology. This mapping is based on a proven mechanism for representing the meaning of ontological concepts with lexical elements [36]. The second module is concerned with *decomposition* of terms into their component elements, that is either the decomposition of multiword elements into individual words, or of synthetic words into individual lexemes. The next module is the *variation* module that describes how terminological and lexical variants and relations may be stated and in particular how we can represent translations of terms taking into account a meaning of a word in an ontology. The final module is the *metadata* module described in this paper.

4 Requirements for the Metadata Module

The design of LIME has been informed by the following requirements, which express information that is relevant to different use-cases and applications.

R1. *Compatibility with the lemon model.*
R2. *Compatibility with other lexicalization models*, such as RDFS, SKOS (Simple Knowledge Organization System), SKOS-XL (SKOS eXtension for Labels).
R3. *Distributed publication* of each component of the ontology-lexicon interface.
R4. *Encoding*. It must provide metadata describing how content is encoded.
R5. *Content summarization.* It must provide summaries about the dataset content.
R6. *Reuse of existing vocabularies*.

5 The Metadata Vocabulary

The LIME vocabulary (see Fig. 2) we present here, though inspired by the proposal in [28], is in fact very different because of the need for a better alignment with the overall scope of the working group and for accommodating the flexible publication scenario envisaged by *lemon*.

Following the conceptual model of the ontology-lexicon interface defined by *lemon* (see Requirement R1), we distinguish at the metadata level three entities:

1. the ontology (bearing semantic information),
2. the lexicon (bearing linguistic information),
3. the set of lexicalizations (intended as the mere correspondences between logical entities in the ontology and lexical entries in the lexicon).

From the perspective of a metadata vocabulary, LIME focuses on the representation of the relation between these three entities and summaries and descriptive statistics concerning these entities and their relations (see Requirement R5).

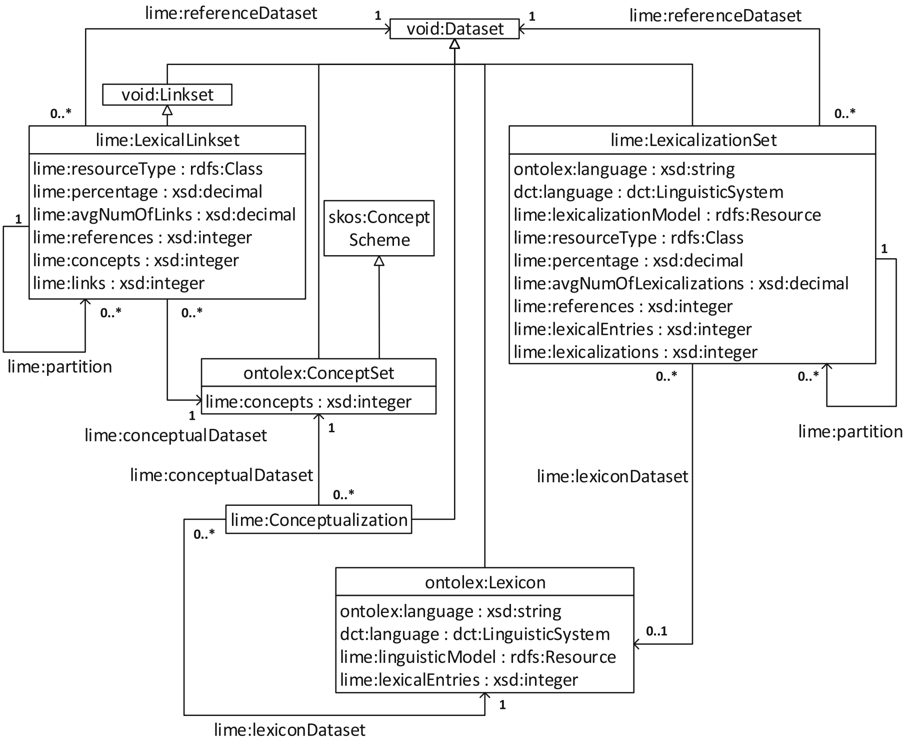
lime:referenceDataset
1
void:Dataset
1
lime:referenceDataset
void:Linkset
0..*
0..*
lime:LexicalLinkset
lime:resourceType : rdfs:Class
lime:percentage : xsd:decimal
lime:avgNumOfLinks : xsd:decimal
lime:references : xsd:integer
lime:concepts : xsd:integer
lime:links : xsd:integer
1
0..*
0..*
lime:partition
skos:Concept Scheme
lime:LexicalizationSet
ontolex:language : xsd:string
dct:language : dct:LinguisticSystem
lime:lexicalizationModel : rdfs:Resource
lime:resourceType : rdfs:Class
lime:percentage : xsd:decimal
lime:avgNumOfLexicalizations : xsd:decimal
lime:references : xsd:integer
lime:lexicalEntries : xsd:integer
lime:lexicalizations : xsd:integer
1
0..*
0..*
lime:partition
ontolex:ConceptSet
lime:concepts : xsd:integer
1
lime:conceptualDataset
1
lime:conceptualDataset
0..*
lime:Conceptualization
0..*
lime:lexiconDataset
ontolex:Lexicon
ontolex:language : xsd:string
dct:language : dct:LinguisticSystem
lime:linguisticModel : rdfs:Resource
lime:lexicalEntries : xsd:integer
0..1
1
lime:lexiconDataset

Fig. 2. The LIME Model

The three entities (ontology, lexicon and lexicalization set) are regarded as instances of `void:Dataset`. While the *lemon* model introduces a subclass of `void:Dataset` to represent lexica (`ontolex:Lexicon`), no such subclass exists for lexicalizations. LIME introduces such a subclass, `lime:LexicalizationSet`, to describe the relation between the lexicon and the ontology in question. A `lime:LexicalizationSet` object thus holds all the relevant metadata and descriptive statistics about the lexicalizations that relate ontology elements in the ontology to lexical entries (possibly found in a lexicon).

Moving away from our original assumption that lexicalizations are embedded within an ontology, we allow each entity to be published independently or combined with others into a single resource (see Requirement R3). By allowing this freedom, we support the following scenarios:

1. a lexicon is published as a stand-alone resource, independently of any specific ontology. We further distinguish the following two cases:
   (a) an ontology contains a set of lexicalizations by means of entries in the lexicon (thus ontology + lexicalization as a single data source)
   (b) an ontology exists independently of the lexicon, and a third party publishes a lexicalization of the ontology by adopting the above lexicon (thus all the three datasets are separate entities)

2. a lexicon is created for a specific ontology:
   (a) the lexicon and lexicalizations for an existing ontology are published together.
   (b) an ontology is published alongside with its lexicon (ontology, lexicon and the set of lexicalizations published together).

Obviously, since ontologies may be lexicalized for more languages, and as a general-purpose lexicon may be reused across different ontologies, multiple combinations of the above cases may happen for any single resource. Finally, linguistic enrichment of ontologies may occur by means of links with lexical concepts, rather than links with specific lexical entries, as suggested by Pazienza and Stellato [37]. The notion of Lexical Linkset accounts for this scenario, by specializing the notion of `void:Linkset` to make explicit its linguistic value.

5.1 Describing (Domain) Datasets

From the LIME viewpoint, any RDF dataset may be lexicalized in a natural language or aligned with a set of lexical concepts. The term dataset is meant hereafter to encompass ontologies, SKOS concept schemes and in general any set of RDF triples. In the ontology-lexicon dualism, the dataset corresponds to the ontology, in the sense that it provides formal symbols that need for grounding in a natural language.

At the metadata level, a dataset is then represented as an instance of the class `void:Dataset` or a more specific subclass, e.g. `voaf:Vocabulary` for vocabularies. LIME defines no specific term for the description of the dataset bearing the semantic references for the ontology-lexicon interface. Still, it suggests the use of appropriate metadata terms suggested by the VoID specification (see Requirement R6). For instance, in the following excerpt:

```
<http://xmlns.com/foaf/0.1/> a voaf:Vocabulary;
    foaf:homePage <http://xmlns.com/foaf/0.1/>;
    dct:title “The Friend of a Friend (FOAF) Vocabulary”@en;
    void:dataDump <http://xmlns.com/foaf/spec/index.rdf>;
    voaf:classNumber 13;
    voaf:propertyNumber 62 .
```

we declare an instance of `voaf:Vocabulary` describing the FOAF vocabulary. In the example, we show how to provide the name of the vocabulary, its home page (providing a unique key supporting data aggregation), a download file and the count of classes and properties. In the previous example, we followed LOV when reusing the URI of FOAF to provide additional metadata. This approach requires the publication of metadata via a SPARQL endpoint or some other API (Application Programming Interface). Alternatively, one can create a new URI for the metadata instance, so that it can be dereferenced. Meanwhile, the connection to the vocabulary is established via an `owl:sameAs` axiom, or some other uniquely identifying property.

5.2 Describing Lexica

A lexicon comprises a collection of lexical entries in a given natural language, and is generally independent from the semantic content of ontologies. The class `ontolex:Lexicon` represents lexica in both the core (data) and metadata levels of the OntoLex specification. This class extends `void:Dataset`, such that recommendations from the VoID specification apply.

Perhaps the most important fact about a lexicon is the language it refers to, an explicit marker for applicability of the resource in given scenarios. This information can be represented either as a literal (according to ISO 639 [38]) through property ontolex: language or as a resource (through the property `dct:language`), using any of the vocabularies assigning URIs to languages (e.g. http://www.lexvo.org/, http://www.lingvoj.org/, http://id.loc.gov/). The following example describes an English lexicon:

```
ex:myLexicon a ontolex:Lexicon;
   ontolex:language “en”;
   dct:language <http://lexvo.org/id/iso639-3/eng>;
   void:dataDump <http://example.org/lexicon/dump.rdf>;
   void:sparqlEndpoint <http://example.org/lexicon/sparql>;
   void:triples 10000 .
```

The description above contains terms from VoID (see Requirement R6), e.g. to provide a data dump and a SPARQL endpoint. An agent may choose between the available types of access based on various criteria: (i) the suitability of the local triple store for handling the advertised number of triples, (ii) the necessity of specialized processing not provided by the SPARQL endpoint, (iii) the willingness to avoid stressing the data provider with frequent/complex queries.

To support the actual exploitation of a lexicon, LIME supports metadata about the way a lexicon has been encoded (see Requirement R4). The reason is that *lemon* does not commit to a specific catalog of linguistic categories (e.g. part-of-speech), whereas it defers to the user the choice of a specific catalog. The adopted catalog may be indicated as a value of the property `lime:linguisticModel`. This property is defined as a subproperty of `void:vocabulary`, to better qualify the specific association between the lexicon and the ontology providing linguistic categories. For instance, we can say that `ex:myLexicon` uses LexInfo2 as repository of linguistic annotations:

```
ex:myLexicon a ontolex:Lexicon;
   lime:linguisticModel <http://www.lexinfo.net/ontology/2.0/lexinfo>
```

An important metric indicating the usefulness of a lexicon is the number of lexical entries it contains (see Requirement R5):

```
ex:myLexicon lime:lexicalEntries 13 .
```

5.3 Describing Lexicalization Sets

We use the term lexicalization for the reified relation between a lexical entry and the ontological meaning it denotes. A collection of such lexicalizations is modeled by the class lime:LexicalizationSet, which in turn subclasses void:Dataset. For example, the property foaf:knows can be lexicalized as “X is friend of”, “X knows Y”, “X is acquainted with X” etc., all corresponding to different lexicalizations.

A lime:LexicalizationSet is characterized (as an ontolex:Lexicon) by the natural language it refers, which can be indicated via the properties already used for the same purpose within ontolex:Lexicon. Moreover, a lime:LexicalizationSet may play an associative function, as it may relate a dataset with a lexicon providing lexical entries. The properties lime:referenceDataset and lime:lexiconDataset point to the dataset and the lexicon, respectively. The presence of explicit links with the dataset and lexicon will allow metadata indexes answering queries that seek, as an example, a lexicalization set in a natural language for a given dataset (see Requirement R3). This is an example of an English lexicalization set for FOAF utilizing an OntoLex lexicon:

```
ex:LexicalizationSet a lime:LexicalizationSet;
   ontolex:language “en”;
   dct:language <http://lexvo.org/id/iso639-3/eng>;
   lime:referenceDataset <http://xmlns.com/foaf/0.1/>;
   lime:lexiconDataset ex:myLexicon .
```

The mandatory property lime:referenceDataset tells which dataset the lexicalization is about. Similarly, the optional property lime:lexiconDataset holds a reference to the lexicon being used. This optionality allows supporting previous lexicalization models (see Requirement R2) that rely on plain literals (e.g. RDFS and SKOS) or introduce reified labels (e.g. SKOS-XL), but in any case have no separate notion of lexicon. It is thus necessary to introduce the mandatory property lime:lexicalizationModel, which holds the model used in a specific lexicalization set (see Requirement R4). We may say, for instance, that FOAF has an embedded lexicalization set expressed in RDFS:

```
<http://xmlns.com/foaf/0.1/> void:subset ex:embedLexSet .
ex:embedLexSet a lime:LexicalizationSet;
   ontolex:language “en”;
   lime: lexicalizationModel <http://www.w3.org/2000/01/rdf-schema#>
```

Knowing that a dataset is lexicalized in a given natural language does not guarantee that the available linguistic information is useful. In particular, the value of a lexicalization set may be assessed by means of metrics (see Requirement R5). For instance, in the following excerpt:

```
:myItalianLexicalizationOfFOAF a lime:LexicalizationSet;
   ontolex:language “it”;
   lime:referenceDataset <http://xmlns.com/foaf/0.1/>;
   lime:lexicalizationModel ontolex:;
   lime:lexiconDataset :italianWordnet;
   lime:partition [
      lime:resourceType owl:Class;
      lime:percentage 0.75;
```

```
    lime:avgNumOfLexicalizations 3.54;
    lime:references 13;
    lime:lexicalEntries 46;
    lime:lexicalizations 46
  ].
```

the property `lime:partition` (domain: `lime:LexicalizationSet` ⊔ `lime:LexicalLinkset`) points to a `lime:LexicalizationSet`, which is the subset of the lexicalization set dealing exclusively with instances of the class referenced by `lime:resourceType`. The properties `lime:references` and `lime:lexicalEntries` hold, respectively, the number of entities from the reference dataset and the number of lexical entries from the lexicon that participate in at least one lexicalization, while `lime:lexicalizations` holds the total number of lexicalizations. Additionally, `lime:avgNumOfLexilicazions` gives the average number of lexicalizations per resource, while `lime:percentage` indicates the ratio of resources having at least one lexicalization. There is a certain level of redundancy among these properties, so that it is at the discretion of the publisher to choose a number of properties. For instance, if metadata for the lexicalized ontology is not available, then it is mandatory to provide ratios (such in the above example), whereas clients can combine counts (if available from both the lexicalization and the reference datasets) in order to compute them.

5.4 Describing Lexical Concept Sets

The class `ontolex:ConceptSet` is a subclass of `void:Dataset` that defines a collection of `ontolex:LexicalConcepts`. It holds LIME-specific and other dataset-level metadata. Lexical concepts are instances of `skos:Concept` (as `ontolex:LexicalConcept` is a subclass of `skos:Concept`). In fact, following the pattern already adopted for the lexicon, we combined the concept scheme with the concept set, by making the latter a subclass of the former. It is possible to summarize the content of a concept set (see Requirement R5), by reporting (via the property `lime:concepts`) the total number of lexical concepts in a concept set. Beyond the need for such summarizing information, the rationale for the class `ontolex:ConceptSet` is to support the publication of lexical concepts as a separate dataset (see Requirement R3). This, in turn, allows the independent publication of the linguistic realization of those concepts in different natural languages, e.g. several wordnets sharing the synsets from the English WordNet. However, *lemon* and LIME are also compatible with the approach to multilingual wordnets, in which each wordnet has its own set of synsets, while an inter-language index establishes a mapping between them. In the following excerpt, we define a `void:Linkset` providing `skos:exactMatch` mappings between two `ontolex:ConceptSets` (defined elsewhere):

```
:ItalianWN_EnglishWN_index a void:Linkset;
  void:subjectsTarget ex:ItalianWN;
  void:objectsTarget ex:EnglishWN;
  void:linkPredicate skos:exactMatch .
```

5.5 Describing Conceptualizations

A lime:Conceptualization is a dataset relating a set of lexical concepts to a lexicon, indicated by the properties lime:conceptualDataset and lime:lexiconDataset, respectively. In the representation of wordnets, it plays a role like that of a lime:LexicalizationSet for the ontology lexicalization. A different class has been introduced, since the association between lexical concepts and words is different from the lexicalization of ontology concepts.

In addition to the explicit references to the lexicon and the lexical concept set, a conceptualization holds a number of resuming metadata (see Requirement R5). The properties lime:lexicalEntries and lime:concepts hold the number of lexical entries and lexical concepts that have been associated, respectively.

5.6 Describing Lexical Link Sets

An interesting use of wordnets is to enrich an ontology with links to lexical concepts, which may provide a less ambiguous inter-lingua (than natural language, which has inherent lexical ambiguity) for the task of ontology matching.

To represent a collection of these links, we introduced lime:LexicalLinkset, which extends void:Linkset with additional metadata tailored to this specific type of linking. The properties lime:referenceDataset and lime:conceptualDataset clearly distinguish between the different roles that the linked datasets play from the perspective of the *lemon* model, whereas properties from the VoID vocabulary only deal with lower-level features, e.g. to which dataset the subjects of the link belong to. Similarly to the case of lime:LexicalizationSet, the property lime:partition references a lime:LexicalLinkset dealing with a given resource type. Due to the lack of space, we will not provide specific examples for the relevant metrics. However, they are analogous to the ones already discussed for lexicalization sets, expect for the fact they now refer to links rather than lexicalizations.

6 A Use-Case: Ontology Matching

Ontology matching is the task of finding a set of correspondences between a pair of input ontologies. Although ensemble strategies – combining different kinds of matching techniques based on terminology, structure, extension and models of the compared resources – dominate evaluation campaigns, lexical comparison [2] is the basic step providing the initial “anchors” for further analysis performed through those techniques. While matchers can certainly find out how and in which languages labels are expressed by analyzing the data to determine the matching techniques to be applied, descriptive summaries of the linguistic characteristics of the ontologies in question would save computation time, making this information directly accessible.

We focus here on the activities that a coordinator needs to perform beforehand in order to define a successful mediation strategy. Linguistic metadata have been shown to be useful in support coordination activities in a semi-automatic process [39].

LIME metadata about the input ontologies allows the coordinator to estimate their level of linguistic compatibility, which in turn indicates how easily they can be matched. If the coordinator finds at least one pair of lexicalizations that sufficiently cover the ontologies, then it may use them to perform the match. When multiple lexicalizations exist, the coordinator may exclude those that do not sufficiently cover the input ontologies, or it could assign different weights to the scores computed with respect to each of them. Similarly, a coordinator may consider whether the input ontologies have been enriched with links to lexical concepts found in the same wordnet, which provide a less ambiguous inter-lingua than natural language (see Sect. 5.6).

The explicit linguistic metadata about the input ontologies allow the coordinator to reason upon them, and determine an appropriate matching strategy by applying some heuristics. The greatest benefit of an explicit metadata vocabulary is that it supports access to previously unknown information. Indeed, using LIME it would be possible to locate relevant data from remote repositories.

Such metadata aggregation would benefit from the protocols that VoID specifies to support the independent publication of dataset descriptions in a predictable way. The fact that LIME is an extension of VoID entails that the same protocols may support harvesting of LIME metadata. Moreover, the same services that aggregate and make available VoID descriptions in general should also support LIME metadata as well.

7 Conclusions and Future Work

We presented LIME, a vocabulary developed in the context of the Ontolex group, providing metadata terms specifically relevant to *lemon*. The publication of such metadata alongside the corresponding datasets intends to foster their discoverability, understandability and exploitability. LIME provides metadata terms related to the core module of *lemon*. Future work will probably include development of extensions dealing with other *lemon* modules. A question for future work is how to include aspects related to the quality of linguistic resources as metadata.

The URI of the LIME ontology is: http://www.w3.org/ns/lemon/lime and it is currently available at:

https://github.com/cimiano/ontolex/blob/master/Ontologies/lime.owl.

Acknowledgements. This research has been partially supported by the EU funded project SemaGrow (http://www.semagrow.eu/) under grant agreement no: 318497 and by the EU funded project LIDER (http://www.lider-project.eu).

References

1. Cimiano, P., Buitelaar, P., McCrae, J., Sintek, M.: LexInfo: a declarative model for the lexicon-ontology interface. Web Seman. Sci. Serv. Agents World Wide Web **9**(1), 29–51 (2011)
2. Pazienza, M.T., Sguera, S., Stellato, A.: Let’s talk about our “being”: a linguistic-based ontology framework for coordinating agents. Appl. Ontology Spec. Issue Formal Ontol. Commun. Agents **2**(3–4), 305–332 (2007)

3. Unger, C., Freitas, A., Cimiano, P.: An introduction to question answering over linked data. In: Koubarakis, M., Stamou, G., Stoilos, G., Horrocks, I., Kolaitis, P., Lausen, G., Weikum, G. (eds.) Reasoning Web. LNCS, vol. 8714, pp. 100–140. Springer, Heidelberg (2014)
4. Atzeni, P., Basili, R., Hansen, D.H., Missier, P., Paggio, P., Pazienza, M.T., Zanzotto, F.M.: Ontology-based question answering in a federation of university sites: the moses case study. In: Meziane, F., Métais, E. (eds.) NLDB 2004. LNCS, vol. 3136, pp. 413–420. Springer, Heidelberg (2004)
5. Basili, R., Vindigni, M., Zanzotto, F.M.: Integrating ontological and linguistic knowledge for conceptual information extraction. In: IEEE/WIC International Conference on Web Intelligence, Washington (2003)
6. Cimiano, P.: Ontology Learning and Population from Text Algorithms, Evaluation and Applications XXVIII. Springer, Heidelberg (2006)
7. Bouayad-Agha, N., Casamayor, G., Wanner, L.: Natural language generation in the context of the semantic web. Seman. Web **5**(6), 493–513 (2014)
8. Bontcheva, K., Wilks, Y.: Automatic report generation from ontologies: the MIAKT approach. In: Meziane, F., Métais, E. (eds.) NLDB 2004. LNCS, vol. 3136, pp. 324–335. Springer, Heidelberg (2004)
9. Galanis, D., Androutsopoulos, I.: Generating multilingual descriptions from linguistically annotated OWL ontologies: the NaturalOWL system. In: Proceedings of the Eleventh European Workshop on Natural Language Generation, Stroudsburg, pp. 143–146 (2007)
10. Pazienza, M.T., Stellato, A., Turbati, A.: Linguistic watermark 3.0: an RDF framework and a software library for bridging language and ontologies in the semantic web. In: 5th Italian Semantic Web Workshop on Semantic Web Applications and Perspectives (SWAP 2008), FAO-UN, Rome, Italy, 15–17 December 2008
11. Oltramari, A., Stellato, A.: Enriching ontologies with linguistic content: an evaluation framework. In: The Role of OntoLex Resources in Building the Infrastructure of Web 3.0: Vision and Practice (OntoLex 2008), Marrakech, Morocco, 31 May 2008
12. Cimiano, P., Haase, P., Herold, M., Mantel, M., Buitelaar, P.: LexOnto: a model for ontology lexicons for ontology-based NLP. In: Proceedings of the OntoLex 2007 Workshop (held in conjunction with ISWC 2007) (2007)
13. Buitelaar, P., Declerck, T., Frank, A., Racioppa, S., Kiesel, M., Sintek, M., Engel, R., Romanelli, M., Sonntag, D., Loos, B., Micelli, V., Porzel, R., Cimiano, P.: LingInfo: design and applications of a model for the integration of linguistic information in ontologies. In: OntoLex 2006, Genoa, Italy (2006)
14. Montiel-Ponsoda, E., Aguado-de-Cea, G., Gómez-Pérez, A., Peters, W.: Enriching ontologies with multilingual information. Nat. Lang. Eng. **17**, 283–309 (2011)
15. McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T.: Interchanging lexical resources on the Semantic Web. Lang. Resour. Eval. **46**(4), 701–719 (2012)
16. Chiarcos, C., McCrae, J., Cimiano, P., Fellbaum, C.: Towards open data for linguistics: linguistic linked data. In: Oltramari, A., Vossen, P., Qin, L., Hovy, E. (eds.) New Trends of Research in Ontologies and Lexical Resources, pp. 7–25. Springer, Heidelberg (2013). doi:10.1007/978-3-642-31782-8_2
17. McCrae, J., Fellbaum, C., Cimiano, P.: Publishing and linking wordnet using lemon and RDF. In: Proceedings of the 3rd Workshop on Linked Data in Linguistics, Reykjavik, Iceland (2014)
18. Miller, G.: WordNet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
19. Fellbaum, C.: WordNet: An Electronic Lexical Database. WordNet Pointers, MIT Press, Cambridge (1998)

20. Jain, P., Hitzler, P., Yeh, P., Verma, K., Sheth, A.: Linked data is merely more data. In: AAAI Spring Symposium: Linked Data Meets Artificial Intelligence (2010)
21. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets with the VoID vocabulary (W3C Interest Group Note). In: World Wide Web Consortium (W3C). http://www.w3.org/TR/void/. Accessed 3 March 2011
22. Auer, S., Demter, J., Martin, M., Lehmann, J.: LODStats – an extensible framework for high-performance dataset analytics. In: ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Acquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (eds.) EKAW 2012. LNCS, vol. 7603, pp. 353–362. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33876-2_31
23. W3C: data catalog vocabulary (DCAT). In: World Wide Web Consortium (W3C). http://www.w3.org/TR/vocab-dcat/. Accessed 16 Jan 2014
24. Bird, S., Simons, G.: Extending dublin core metadata to support the description and discovery of language resources. Comput. Humanit. **37**(4), 375–388 (2003)
25. Lagoze, C., Van de Sompel, H.: The open archives initiative: building a low-barrier interoperability framework. In: Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, New York, pp. 54–62 (2001)
26. Calzolari, N., Del Gratta, R., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., Soria, C.: The LRE map. Harmonising community descriptions of resources. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, pp. 1084–1089 (2012)
27. Piperidis, S.: The META-SHARE language resources sharing infrastructure: principles, challenges, solutions. In: Proceedings of the Eighth International Conference on Language, Istanbul, Turkey, pp. 36–42 (2012)
28. Fiorelli, M., Pazienza, M.T., Stellato, A.: LIME: towards a metadata module for ontolex. In: 2nd Workshop on Linked Data in Linguistics: Representing and Linking Lexicons, Terminologies and Other Language Data, Pisa, Italy (2013)
29. Borin, L., Dannélls, D., Forsberg, M., McCrae, J.: Representing swedish lexical resources in RDF with lemon. In: Proceedings of the ISWC 2014 Posters & Demonstrations Track a Track Within the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, pp. 329–332 (2014)
30. Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J., Cimiano, P., Navigli, R.: Representing multilingual data as linked data: the case of BabelNet 2.0. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, 26–31 May 2014, pp. 401–408 (2014)
31. Eckle-Kohler, J., McCrae, J., Chiarcos, C.: LemonUby - a large, interlinked syntactically-rich lexical resources for ontologies. Semantic Web Journal (2015, accepted)
32. Sérasset, G.: Dbnary: wiktionary as a LMF based multilingual RDF network. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, 23–25 May 2012, pp. 2466-2472 (2012)
33. Buitelaar, P.: Ontology-based semantic lexicons: mapping between terms and object descriptions. In: Huang, C.-R., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., Prevot, L. (eds.) Ontology and the Lexicon: A Natural Language Processing Perspective. Cambridge University Press, Cambridge (2010)
34. Cimiano, P., McCrae, J., Buitelaar, P., Montiel-Ponsoda, E.: On the role of senses in the ontology-lexicon. In: Oltramari, A., Vossen, P., Qin, L., Hovy, E. (eds.) New Trends of Research in Ontologies and Lexical Resources, pp. 43–62. Springer, Heidelberg (2013)
35. Evans, V.: Lexical concepts, cognitive models and meaning-construction. Cogn. Linguist. **17**(4), 491–534 (2006)
36. Cimiano, P., Unger, C., McCrae, J.: Ontology-based interpretation of natural language. Synth. Lect. Hum. Lang. Technol. **7**(2), 1–178 (2014)

37. Pazienza, M.T., Stellato, A.: An environment for semi-automatic annotation of ontological knowledge with linguistic content. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 442–456. Springer, Heidelberg (2006)
38. ISO, International organization for standardization: language codes - ISO 639. In: ISO, International Organization for Standardization. http://www.iso.org/iso/home/standards/language_codes.htm
39. Fiorelli, M., Pazienza, M.T., Stellato, A.: A meta-data driven platform for semi-automatic configuration of ontology mediators. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, May 2014