LIME: Towards a Metadata Module for OntoLex

Manuel Fiorelli
University of Tor Vergata
Via del Politecnico 1, 00133
Rome, Italy
fiorelli@info.uniroma2.it

Maria Teresa Pazienza
University of Tor Vergata
Via del Politecnico 1, 00133
Rome, Italy
pazienza@info.uniroma2.it

Armando Stellato
University of Tor Vergata
Via del Politecnico 1, 00133
Rome, Italy
stellato@info.uniroma2.it

Abstract

The OntoLex W3C Community Group has been working for more than a year on realizing a proposal for a standard ontology lexicon model. As the core-specification of the model is almost complete, the group started development of additional modules for specific tasks and use cases. We think that in many usage scenarios (e.g. linguistic enrichment, localization and alignment of ontologies) the discovery and exploitation of linguistically grounded datasets may benefit from summarizing information about their linguistic expressivity. While the VoID vocabulary covers the need for general metadata about linked datasets, this more specific information demands a dedicated extension. In this paper, we fill this gap by introducing LIME (Linguistic Metadata), a new vocabulary aiming at completing the OntoLex standard with specifications for linguistic metadata.

1 Introduction

Linguistic grounding of formalized knowledge is a long-standing principle in ontological modelling, at least traceable back to the “clarity criterion” (Gruber, 1995). Recently, natural language characterization of ontologies has proved useful both in the semantic reconciliation of heterogonous ontologies, and in many tasks interfacing natural language and ontologies, such as ontology verbalization, natural language ontology querying, ontology-based information extraction, ontology learning, validation and evolution.

Therefore, many research works aimed at defining common models and best-practices for linguistically grounding the Semantic Web, or even theorised a Linguistic Linked Open Data (Chiarcos, et al., 2012) cloud. The OntoLex W3C Community Group[1] is currently working on a principled ontology lexicon model that combines and improves previous proposals. Similarly, the Open Linguistics Working Group[2] of the Open Knowledge Foundation is pushing forward the publication of linguistic resources according to the Linked Open Data principles, thus developing a LOD (sub-)cloud of linguistic resources[3].

While focusing on representing linguistic information, existing proposals mostly overlook the characterization of ontologies, datasets and linguistic resources for what concerns their linguistic expressivity. This information should be provided in the form of metadata about linked data resources, providing summarizing information on how a dataset is linguistically represented, which formalism have been adopted, which languages have been used for representing its formal content and so on.

Such metadata would enable resolution strategies to be tuned to the specificities of a given task (e.g. is this a cross-language ontology alignment task?), and to retrieve suitable resources for supporting this resolution (e.g. is this a bi-lingual dictionary between the pair of languages used in a specific cross-language task?).

In this paper, we try to address the lack of a standardized vocabulary for linguistic metadata by proposing LIME, which is an abbreviation for **Li**nguistic **Me**tadata, which aims to become a module of the future OntoLex specification.

The rest of the paper is organized as follows. In section 2, we describe previous works on linguistic enrichment of ontologies/datasets and introduce the general usefulness of metadata in

[1] http://www.w3.org/community/ontolex/
[2] http://linguistics.okfn.org/
[3] http://nlp2rdf.lod2.eu/OWLG/llod/llod.svg

the Linked Data paradigm. In section 3, we introduce some application scenarios that would benefit from a dedicated vocabulary of linguistic metadata. In section 4, we describe the design of the vocabulary and some usages examples. Finally, in section 5, the conclusions.

2 Background and Related work

Currently, Knowledge Modelling languages for the Semantic Web do not support the representation of linguistic information to a large extent.

In RDF (Carroll & Klyne, 2004), natural language expressions are simply treated as language-tagged literals. RDFS (Guha & Brickley, 2004) provides standard properties for attaching these literals to conceptual resources as human-friendly names (`rdfs:label`) or longer narrative descriptions (`rdfs:comment`). SKOS (Bechhofer & Miles, 2009) introduces a finer-grain characterization of labels by means of a few sub-properties of `rdfs:label` accounting for differences at the terminological-correspondence level (Pastor-Sanchez, et al., 2009). SKOS-XL (W3C, 2009) models natural language expressions as individuals of a dedicated class (`skosxl:Label`). Providers of large KOSs (Hodge, 2000), such as AGROVOC (Caracciolo, et al., 2013) and EUROVOC (Paredes, et al., 2008), are widely adopting this modelling style, since they need to treat natural language expressions as "first-class citizens", at least for attaching editorial metadata to them. For instance, in the AGROVOC thesaurus, natural language labels are associated with a wide range of metadata, including creation/modification date and publication status, which are required for publication as well as for supporting the thesaurus collaborative development workflow (Caracciolo, et al., 2012).

Further works proposed even richer models for linguistically grounded ontologies/dataset. LingInfo (Buitelaar, et al., 2006) allows the description of the morphological and syntactic decomposition of natural language labels. On the other hand, LexOnto (Cimiano, et al., 2007) focuses on the mapping of linguistic predicate-argument structures to the join of semantic (binary) properties. Buitelaar et al. (2009) combined these two complementary models into a unified model, called LexInfo, highly based on the RDF porting of the LMF (Francopoulo, et al., 2006), thus benefitting from a principle conceptual model and higher compatibility with existing resources. These works informed the Lemon Model (Mccrae, et al., 2012) , which focuses on modularity and extensibility.

A complementary aspect consists in characterizing linguistic resources as a whole (Pazienza & Stellato, 2006b) with proper metadata.

A classification of linguistic resources (later backed by a suite of ontologies in (Pazienza, et al., 2008)), called Linguistic Watermark, was defined by us to support the development of a software library for accessing heterogeneous linguistic resources under a common API. A reflection mechanism in the library allows system and tools to access seamlessly different linguistic resources, understanding their nature, what these have to offer and exploiting their content in several application contexts.

The publication of linguistic resources (e.g. dictionaries, thesauri, corpora) as Linked Open Data is attracting the attention of Semantic Web practitioners. While using NLP tools to create semantic annotations with respect to formal ontologies, Kiryakov et al. (2004) advocated the representation in RDF of the linguistic resources that empower these tools, thus entailing a technological and a methodological reuse. When reconciling heterogeneous ontologies, linguistic resources may prove useful as well, since they provide a common grounding across different semantic theories, as they reflect the organic development of a language within a community.

The difficulties related to the triplification of linguistic resources is exemplified by the number of works that informed the development of the W3C RDF/OWL representation of WordNet (Van Assem, et al., 2006). WordNet, and similar resources, are not ontologies (Hirst, 2004), therefore any systematic translation into an ontology necessarily violates the formal semantics of the modelling language and ontological adequacy principles (Guarino & Welty, 2004). Gangemi et al. (2003a) restructured WordNet through the upper-ontology DOLCE (Gangemi, et al., 2002). OntoWordNet (Gangemi, et al., 2003b) is a notable output of this research line aiming at equipping WordNet with a formal semantics.

Another approach consists in a two-step process: produce an ontology modelling the core concepts found in the resource, then, instantiate that conceptual model with information found in a specific resource. The definition of a shared upper-model for linguistic resources is in fact another requirement of the forthcoming OntoLex model.

Concerning the importance of metadata in Linked Open Data, the necessity of summarizing information about a dataset as a whole has been considered and assessed. Jain et al. (2010) insisted on the lack of conceptual characterization of a dataset (e.g. what is it about?). Similar concerns motivated the development of VoID (Alexander, et al., 2011), a vocabulary for describing linked datasets.

In the field of Human Language Technology it has been promoted the reuse of Language Resources (LRs) through structured metadata. OLAC (Bird & Simons, 2003) extends the Dublin Core Metadata Element Set[4] for defining a simple template for the description of LRs that includes, among others, provenance metadata, resource typology and language identification. While OLAC aims at defining a distributed infrastructure for resource sharing, LRE Map (Calzolari, et al., 2012) is a crowd-sourced catalogue of LRs, initially fed by authors submitting papers to LREC Conferences. LRE Map defines numerous resource types and usage applications, whilst OLAC distinguishes a handful of types. Similar in scope to OLAC, META-SHARE (Piperidis, 2012) has its own metadata schema. In META-SHARE the taxonomy of LRs is not developed in a top-down manner, rather it originates from the adoption of metadata combination as a criterion for classifying LRs (Gavrilidou, et al., 2012).

These works have a wider scope than ours, as their definition of LRs include both software tools (e.g. postaggers and parsers) and data (e.g. corpus, dictionaries and grammars), managing heterogeneous formats. In contrast, we focus only on linguistic resources and linguistically enriched datasets, both expressed in RDF. Like META-SHARE we emphasize the importance of properties for the selection and interpretation of resources. Although Dublin Core can be used in conjunction with our model, we believe that some aspects, namely the provenance tracking, deserve dedicated models. Furthermore, our interest in quantitatively describing the extent to which a dataset has been lexicalized does not seem to be in the scope of these works.

It is worth of notice that these works are not grounded in the Semantic Web, as they do not use RDF for metadata representation nor their metadata are modelled using Semantic Web modelling languages. In fact, these works stress validation and mandatory nature of some metadata, something that is still being discussed within the Semantic Web community[5]. Despite being interesting, the broader definition of LR is out of the scope of most works about the representation of linguistic information as Linked Data, such as OntoLex.

3 Motivating Applications

Our previous research work with Linguistic Watermark revealed that many applications may benefit not only from a common linguistic model, but also from a shared (linguistic) metadata vocabulary for characterizing and summarizing the nature of linguistic resources.

In the following sections, we describe some use cases that would benefit from a metadata module, complementing the ontology lexicon model provided by the core OntoLex specification.

The requirement recurring in all scenarios is “discovery of (linguistic) resources”, which is also the main requirement that motivated VoID. While providing a sound framework for coarse-grain description of datasets, VoID alone does not match this requirement, since it lacks vocabulary terms for language related metadata. These metadata should support both the description of linguistic resources, and the description of how ontologies and datasets have been enriched with their content.

3.1 Linguistic enrichment of ontologies

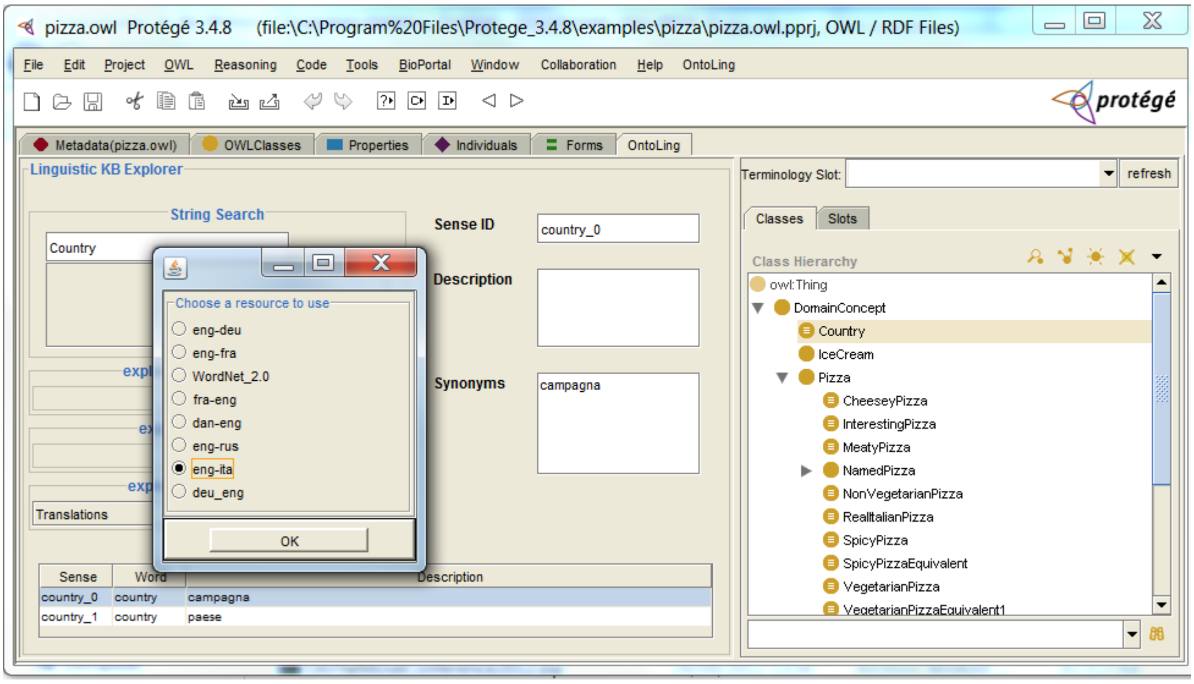
Algorithms and systems for automatically enriching ontologies with content from linguistic resources (Pazienza & Stellato, 2006a; Pazienza & Stellato, 2006c) may be written in terms of a common linguistic model, instead of being tightly coupled to specific resources.

In Figure 1, we see a screenshot of OntoLing (Pazienza & Stellato, 2005), a Protégé (Gennari, et al., 2003) plugin for the linguistic enrichment of ontologies. OntoLing uses metadata to uniformly load heterogeneous linguistic resources, by dynamically configuring its own UI to appropriately show their content and use it to enrich ontologies.

Discovery of linguistic resources can also be supported by linguistic metadata, provided in a way (e.g. in a VoID description) that can be recognized and indexed by Linked Data search engines. Agents may thus issue queries to these search engines to discover relevant linguistic resources in the LOD. The key point here is imme-

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

[5] https://www.w3.org/2012/12/rdf-val/



pizza.owl Protégé 3.4.8 (file:\C:\Program%20Files\Protege_3.4.8\examples\pizza\pizza.owl.pprj, OWL / RDF Files)
File Edit Project OWL Reasoning Code Tools BioPortal Window Collaboration Help OntoLing
protégé
Metadata(pizza.owl) OWLClasses Properties Individuals Forms OntoLing
Linguistic KB Explorer
String Search
Terminology Slot:
refresh
Classes Slots
Class Hierarchy
owl:Thing
DomainConcept
Country
IceCream
Pizza
CheeseyPizza
InterestingPizza
MeatyPizza
NamedPizza
NonVegetarianPizza
RealItalianPizza
SpicyPizza
SpicyPizzaEquivalent
VegetarianPizza
VegetarianPizzaEquivalent1
Translations
OK

String Search: Country

Sense ID: country_0

Description:

Synonyms: campagna

Choose a resource to use

- ○ eng-deu
- ○ eng-fra
- ○ WordNet_2.0
- ○ fra-eng
- ○ dan-eng
- ○ eng-rus
- ● eng-ita
- ○ deu_eng

Sense	Word	Description
country_0	country	campagna
country_1	country	paese

Figure 1. Loading different linguistic resources in OntoLing

diacy: in fact in a closed scenario an agent might profile the resources it controls by itself, while in open settings agents must necessarily depend on pre-compiled metadata to discover resources of interest.

3.2 Ontology Localization

Ontology Localization is about “the adaptation of an ontology to a particular language and culture” (Suárez-Figueroa & Gómez-Pérez, 2008). This definition was generalized by Cimiano et al. (2010), to account for variations in the cultural and socio-political context in a broader sense. They discussed thoroughly the interdependencies between the lexical and the conceptual layers, thus showing how an alteration of the former might require a modification of the latter, as well.

Nonetheless, bilingual dictionaries are valuable resources in an ontology localization process, as they provide translations of existing labels into the target natural language.

In this scenario, a localization agent might depend on linguistic metadata to determine its requirements, and, as discussed in previous section, query a LOD search engine for a list of matching resources. Semantically structured linguistic resources (such as the original WordNet for English, and the various wordnets created for many languages, such as EuroWordNet (Vossen, 1998) and Balkanet (Stamou, et al., 2002)) may help in understanding the conceptual heterogeneities which are bound to the different sociocultural contexts underlying each language.

3.3 Ontology Alignment

The Ontology Alignment task can benefit from a common metadata model.

Pazienza et al. (2007) extended the FIPA Ontology Service Specification with linguistically-aware methodologies for communication, describing a wide-scope framework for multi-agent systems design, semantic integration and coordination. In that perspective, Ontology Mediators should be able to understand which linguistic resources may be of support for a mediation activity between two ontologies/datasets. Such an understanding may happen at different levels, by making explicit the (natural) languages in which a given dataset is published, or the model being adopted for linguistically enriching the dataset. Even very specific facts, such as knowing that a certain popular resource (such as WordNet) has been used to support the lexicalization of a given dataset, may support the mediation activity: making the adoption of linguistic resources more explicit may be helpful in providing a common interlingua for aligning datasets sharing the same kind of linguistic development.

While Ontology Matching aims at supporting the automatic generation of alignments, a review of the state-of-the-art seems to support that in real scenarios the scarce availability of metadata hampers the achievement of this goal.

Shvaiko and Euzenat (2013) define the state-of-the-art in the field, by analysing the results of recent evaluation campaigns organized annually by the Ontology Alignment Evaluation Initiative[6] (OAEI). They stress the fact that no system outperforms the others in all matching scenarios, and that further advancement of the field requires the exploration of new paths. Among others, they cite the use of background knowledge and the design of meta-matchers able to construct the best strategy for solving a specific ontology alignment problem. We believe that for both purposes a metadata vocabulary may be useful, if not necessary, to describe a matching scenario, to plan a resolution strategy, and to support the discovery of relevant resources in the LOD cloud.

In the OAEI 2012 campaign (Shvaiko, et al., 2012), the Library track[7] provides evidences of the shortcomings in state-of-the-art matching systems. The track deals with two real-world thesauri encoded in SKOS: STW[8] (Neubert, 2009) for economics and TheSoz[9] (Zapilko, et al., 2013) for social sciences. Given the popularity of this genre of resources within large organizations and the growing adoption of SKOS, this track gives an important insight about the real-world performances of matching technologies. The results indicate clearly that current technologies (at least those participating in this international evaluation) have in fact some problems with these real-world matching scenarios. By first, most of the systems under evaluation were unable to deal with SKOS, therefore the organizers had to translate both thesauri into OWL. Unfortunately, this conversion can both introduce modelling errors, due to the stricter semantics of OWL, and cause loss of information, because the distinction between preferred and alternative labels is lost after the conversion. It turned out that the baseline matching all labels (both preferred and alternative ones) behaves more or less as the best system participating in the evaluation. This surprising result indicates that current matching strategies, developed for ontologies, are in fact quite inadequate for matching thesauri, which clearly deserve a special treatment. In this scenario, as evidenced by the contest results, terminology-based methods perform particularly well, and the importance of (linguistic) resources adopted in the alignment process seems to prevail over the adopted algorithms. Moreover, even for well-assessed multi-language resources, it should be noted that the quality of labels might vary drastically. For instance, both the thesauri used in the library track have been primarily developed in German, with translations made available in English. Therefore, it is unsurprising that German labels resulted sufficient alone for producing a good alignment, whereas English ones did not.

4 Vocabulary Design

With the work on the core OntoLex specification going on, and after recognizing a clear need for a linguistic metadata vocabulary, we have revised our previous work on the Linguistic Watermark suite of vocabularies, aiming at the definition of a suitable metadata module for OntoLex. We called this module: LIME, which is an abbreviation for **Li**nguistic **Me**tadata[10]. As most metadata apply equally to ontologies representing conceptual knowledge, and datasets representing ground facts, in the forthcoming discussion we will use the term dataset to broadly refer to both.

In line with previous works on the general description of datasets, LIME has been defined as an extension of VoID. Accordingly, LIME metadata should be put in a VoID description of linguistically grounded resources.

By following the same approach adopted in Linguistic Watermark, we start by distinguishing metadata related to linguistic resources from metadata describing the linguistic expressivity of a dataset.

4.1 Linguistic Resources Metadata

There are a number of very simple facts that are relevant for assessing the usefulness of a linguistic resource in a task, which are practically missing from currently available metadata standards.

By first, the main discriminator for judging the usefulness of a linguistic resource in a given scenario is the set of (natural) *language(s)* it covers. Each of these languages should appear as a distinct value of the property `lime:language`. These values must conform to the specification of language tags in RDF. As natural language

[6] http://oaei.ontologymatching.org/

[7] http://web.informatik.uni-mannheim.de/oaei-library/2012/

[8] http://zbw.eu/stw/versions/latest/about

[9] http://lod.gesis.org/thesoz/

[10] This name resembles Lemon, one of the various lexicon models which have informed the development of the OntoLex specification

```
Class: lime:LinguisticResource
  SubClassOf: void:Dataset
Class: lime:Dictionary
  SubClassOf: lime:LinguisticResource
Class: lime:SenseAwareDictionary
  SubClassOf: lime:Dictionary
Class: lime:ConceputalizedResource
  SubClassOf: lime:SenseAwareDictionary
Class: lime:MonolingualDictionary
  EquivalentClass: lime:Dictionary and lime:language exactly 1
Class: lime:BilingualDictionary
  EquivalentClass: lime:Dictionary and lime:language exactly 2
Class: lime:UnidirectionalBilingualDictionary
  SubClassOf:   lime:BilingualDictionary
  SubClassOf:   lime:sourceLanguage exactly 1
  SubClassOf:   lime:targetLanguage exactly 1
  DisjointWith: lime:BidirectionalBilingualDictionary
Class: lime:BidirectionalBilingualDictionary
  SubClassOf:   lime:BilingualDictionary
  DisjointWith: limeUnidirectionalBilingualDictionary
Class: lime:ConsistentBidirectionalBilingualDictionary
  SubClassOf:   lime:BidirectionalBilingualDictionary
DataProperty: lime:language
  Range: xsd:string
DataProperty: lime:sourceLanguage
  SubPropertyOf: lime:language
DataProperty: lime:targetLanguage
  SubPropertyOf: lime:language
```

Figure 2. An excerpt of the LIME vocabulary definition expressed in Manchester Syntax

expressions are usually hold by language tagged literals, this design avoids the need for a suitable mapping for relating metadata to data. This property does not hold when relying on other identification mechanisms, including the use of URIs[11].

Currently, no standard RDF vocabulary provides summarizing information about the coverage of natural language expressions in a dataset. In particular, Linguistic Resources should also be classifiable (see Figure 2) as *monolingual*, *bilingual* (as of translation resources), or *multilingual*.

Bilingual dictionaries are a kind of lexical resource providing direct translations between terms. These resources are modelled as individuals of `lime:BilingualDictionary`, which extends the class `lime:Dictionary`. These translations may or may not be divided according to the senses of the input terms (e.g. consider a popular free bilingual dictionary such Freelang[12] for the first case, and most of the FreeDict[13] dictionaries for the latter). To account for this difference, we have introduced the class `lime:SenseAwareDictionary`.

The translations may be available in one direction only (`lime:UnidirectionalBilingualDictionary`), or allow to go from each of the two languages to the other one (`lime:BidirectionalBilingualDictionary`). These two classes are declared disjoint. Concerning directional resources, we have defined two properties `lime:sourceLanguage` and `lime:targetLanguage`, which reflect the direction of the translation. Symmetry may be guaranteed or not (e.g. some dictionaries may not guarantee that an inverse translation of a translated term always brings back to the original term).

Resources with a strong conceptual backbone (`lime:ConceptualizedResource`) may provide consistent multilingual denotation of their entries. In this sense, any multilingual SKOS concept scheme with a strong linguistic grounding could be classified as a multilingual linguistic

[11] Look at http://www.lexvo.org/ for an example

[12] http://www.freelang.net/

[13] http://freedict.org

resource. The metadata model should be as agnostic with respect to the resource theory as possible, while still being able to tell whether a conceptualization of any kind exists. The metadata should describe to which extent the conceptualization is structured. For instance, the property lime:hasTaxonomy tells whether lexical concepts are organized into a taxonomy or not. In our model conceptualized resources are subclass of sense-aware dictionaries, as each attachment of a natural language expression to a concept corresponds to a distinguished sense of that expression. Other properties should trivially tell whether certain information is available or not, so that systems may know what to rely on. An example could be knowing that a given dictionary provides glosses (lime:hasGlosses) or usage examples (lime:hasUsageExamples). Furthermore, we assume that glosses and examples are attached either to senses in sense-aware resources, or to words otherwise.

4.2 OntoLinguistic Metadata

Whether a given dataset adopts vocabularies for an elaborated linguistic description (such as SKOS-XL or the under-development OntoLex) or just relies on simple labelling primitives, it is important to describe these facts through proper metadata. Thus, while the previous metadata relate to the description of linguistic resources (expressed as linked data), the onto-linguistic metadata provide quantitative and qualitative information about the linguistic expressivity of any linked dataset.

As for linguistic resources, the very first fact that should be declared about a dataset consists in the languages (lime:language) in which it is expressed. In the context of an alignment process, this enables immediate verification of the linguistic-compatibility between datasets. Obviously, the sole fact that lexicalizations exist for a given language is not enough for telling whether that language is sufficiently covering and representing the conceptual content of the resource.

In particular, for each language, the metadata should provide the percentage of RDF resources, per type (classes, individuals, properties, SKOS concepts) described by at least a lexicalization in that language. Additional information, such as the average number of lexicalizations per resource, may provide more insights on the "weight" of a language in describing the resource.

The following RDF snippet illustrates the use of LIME for asserting that English lexicalizations cover 75% (lime:percentage) of the SKOS concepts in the dataset :dat, and that there are, on average, 3.5 English lexical entries per concept.

```
:dat lime:languageCoverage [
  lime:lang "en";
  lime:resourceCoverage [
    lime:class skos:Concept;
    lime:percentage 0.75;
    lime:avgNumOfEntries 3.5
  ]
].
```

We use OWL 2 to restrict the range of lime:percentage to the interval [0.0, 1.0].

```
lime:percentage a
    owl:DatatypeProperty;
  rdfs:range [
    rdf:type rdfs:Datatype ;
    owl:onDatatype xsd:float ;
    owl:withRestrictions (
      [xsd:minInclusive 0.0]
      [xsd:maxInclusive 1.0]
    )
  ].
```

The range of lime:avgNumOfEntries is similarly restricted to non-negative floats.

```
lime:avgNumOfEntries a
    owl:DatatypeProperty;
  rdfs:range [
    rdf:type rdfs:Datatype ;
    owl:onDatatype xsd:float ;
    owl:withRestrictions (
      [xsd:minInclusive 0.0]
    )
  ].
```

The inclusion of zero in both ranges allows the representation of the lack of lexicalizations in a given natural language.

The grounding of two datasets to a common natural language allows them to be compared on the basis of the implicit knowledge about the use of that language by the community of its speakers. However, if mappings to popular (conceptualized) linguistic resources are represented explicitly, then these resources may be exploited as a kind of semantic hub between any two datasets sharing the same linguistic development. Being these resources a sort of less-ambiguous interlingua, the metadata about their usage are in fact very similar to the ones we have mentioned for natural languages. Below we reframe the previous example by considering the enrichment of a dataset with links to synsets from WordNet.

```
:dat
 lime:lexicalResourceCoverage [
  lime:lexresource
     ewn:WordNet;
  lime:resourceCoverage [
    lime:class skos:Concept;
    lime:lexConceptClass
                    wn:Synset;
    lime:percentage 0.75;
    lime:avgNumOfEntries 3.5
  ]
].
```

The property lime:lexConceptClass informs the LIME consumer of the specific class of the linguistic resource which is subclassing the generic OntoLex class ontolex:LexicalConcept.

The presence of any linguistic description does not guarantee that an agent might exploit it. Indeed, the agent must know whether linguistic information is available in the form of traditional rdfs:labels, SKOS labels, SKOS-XL reified labels, or OntoLex attachments. Most datasets are likely to use multiple linguistic models simultaneously, each one for different needs (e.g. the distinction between preferred and alternative labels may be or not of interest). These models are hold by the property lime:linguisticModel, which extends the property void:vocabulary, as the former expresses a more specific association with the vocabulary. When a dataset adopts multiple linguistic models, we assume that they express the same information about the metadata terms that apply to them. For instance, when both SKOS and RDFS are used (the latter being possibly materialized from the former), they must express the same labels, though RDFS loses the SKOS-specific finer grain distinctions.

Finally, the metadata vocabulary should account for the widely adopted practice of using evocative names as local name of the resources URIs. Local names are often not natural language expressions per se, since they are constrained by limitations of the URI syntax or by some naming convention. Luckily, the relation between local names and natural language expressions is generally very simple. Moreover, it is often expressed through a limited set of common patterns (e.g. camel-case, underscore separated words). These simple relations might be modelled through simple transducers, perhaps finite state ones. LIME provides default transducers for some of this popular naming schemes.

Local names are the weakest mechanism for linguistic enrichment, as synonymy and multilingualism are hardly supported. Actually, local names mostly serve as an aid for knowledge developers, who can get a sense of the data they are working on, without the need of considering complex lexicalization models. Therefore, some metadata should express whether (cleaned) local names are subsumed or not by lexicalizations provided in other manners.

5 Conclusion

In this paper, we presented LIME, a vocabulary for **Li**nguistic **Me**tadata, which aims to become a standard module of the OntoLex model.

Relevant metadata include statistics about natural language lexicalisations and mappings to linguistic resources. By following the same approach used in VoID, we defined dedicated terms, instead of relying on a fully-fledged (but maybe harder to parse) statistical vocabulary. However, as Data Cube (Cyganiak & Reynolds, 2013) establishes for the representation of (statistical) multi-dimensional data, we should consider providing mappings to it, or even adopting it.

While at present the coverage of a linguistic resource is interpreted only with respect to explicit mappings to its conceptual content, we could consider as well to define a merely lexical coverage. This information correlates with the linguistic compatibility of two datasets, as well can guide their linguistic enrichment to increase such compatibility, when it appears to be low.

An extension of LIME could attempt to go beyond simple coverage statistics, and try to capture the quality of linguistic information in deeper ways. By first, we should agree on a definition of quality, perhaps as some confidence measure. Then, we should decide the granularity of the metadata, i.e. whether to quantify the overall confidence of the linguistic description, or to qualify each linguistic attachment individually.

While developing LIME, we discussed about the very nature of linguistic resources, and how they relate to terminological thesauri or even just lexicalized conceptualizations. Actually, answering these questions is fundamental for the advancement of the field of ontology lexicalization.

Acknowledgements

This research has been partially supported by the EU project SemaGrow (Grant agreement no: 318497).

References

Alexander, K., Cyganiak, R., Hausenblas, M., & Zhao, J. (2011, March 3). *Describing Linked Datasets with the VoID Vocabulary (W3C Interest*

Group Note). Retrieved May 16, 2012, from World Wide Web Consortium (W3C): http://www.w3.org/TR/void/

Bechhofer, S., & Miles, A. (2009, aug). *SKOS Simple Knowledge Organization System Reference.* W3C Recommendation, W3C.

Bird, S., & Simons, G. (2003). Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources. *Computers and the Humanities, 37*(4), 375-388.

Buitelaar, P., Cimiano, P., Haase, P., & Sintek, M. (2009). Towards Linguistically Grounded Ontologies. *In Proceedings of the 6th Annual European Semantic Web Conference (ESWC2009),* (pp. 111-125).

Buitelaar, P., Declerck, T., Frank, A., Racioppa, S., Kiesel, M., Sintek, M., . . . Cimiano, P. (2006). LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies. *OntoLex06.* Genoa, Italy.

Calzolari, N., Del Gratta, R., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., & Soria, C. (2012). The LRE Map. Harmonising Community Descriptions of Resources. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)* (pp. 1084-1089). ELRA.

Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., & Keizer, J. (2013). The AGROVOC Linked Dataset. (P. Hitzler, & K. Janowicz, Eds.) *Semantic Web Journal, 4*(3), 341–348. doi:10.3233/SW-130106

Caracciolo, C., Stellato, A., Rajbahndari, S., Morshed, A., Johannsen, G., Keizer, J., & Jacques, Y. (2012, August Tuesday, 14). Thesaurus Maintenance, Alignment and Publication as Linked Data. *International Journal of Metadata, Semantics and Ontologies (IJMSO), 7*(1), 65-75.

Carroll, J. J., & Klyne, G. (2004, feb). *Resource Description Framework (RDF): Concepts and Abstract Syntax.* W3C Recommendation, W3C.

Chiarcos, C., Nordhoff, S., & Hellmann, S. (Eds.). (2012). *Linked Data in Linguistics.* Springer.

Cimiano, P., Haase, P., Herold, M., Mantel, M., & Buitelaar, P. (2007). LexOnto: A Model for Ontology Lexicons for Ontology-based NLP. *In Proceedings of the OntoLex07 Workshop (held in conjunction with ISWC'07).*

Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., & Gómez-Pérez, A. (2010, April). A note on ontology localization. *Applied Ontology, 5*(2), 127-137.

Cyganiak, R., & Reynolds, D. (2013). *The RDF Data Cube Vocabulary.* W3C.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., & Soria, C. (2006). Lexical Markup Framework (LMF). *LREC2006.* Genoa, Italy.

Gangemi, A., Guarino, N., Masolo, C., & Oltramari, A. (2003). Sweetening WORDNET with DOLCE. *AI Magazine, 24*(3), 13-24.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., & Schneider, L. (2002). Sweetening ontologies with DOLCE. In *Knowledge engineering and knowledge management: Ontologies and the semantic Web* (pp. 166-181). Springer.

Gangemi, A., Navigli, R., & Velardi, P. (2003). The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet. In *On the move to meaningful internet systems 2003: CoopIS, DOA, and ODBASE* (pp. 820-838). Springer.

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., . . . Mapelli, V. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. *Proceedings of the Eighth International Conference on Language* (pp. 1090-1097). ELRA.

Gennari, J., Musen, M., Fergerson, R., Grosso, W., Crubézy, M., Eriksson, H., . . . Tu, S. (2003). The evolution of Protégé-2000: An environment for knowledge-based systems development,. *International Journal of Human-Computer Studies, 58*(1), 89–123.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies, 43*(5-6), 907-928.

Guarino, N., & Welty, C. (2004). An Overview of OntoClean. In S. Staab, & R. Studer (Eds.), *The Handbook on Ontologies* (pp. 151-172). Berlin: Springer-Verlag.

Guha, R. V., & Brickley, D. (2004, feb). *RDF Vocabulary Description Language 1.0: RDF Schema.* W3C Recommendation, W3C.

Hirst, G. (2004). Ontology and the Lexicon. In S. Staab, & R. Studer (Eds.), *Handbook on Ontologies* (pp. 209-230). Springer.

Hodge, G. (2000, April). *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files.* Washington, DC: Council on Library and Information Resources.

Jain, P., Hitzler, P., Yeh, P. Z., Verma, K., & Sheth, A. P. (2010). Linked Data Is Merely More Data. *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence.* AAAI Press.

Kiryakov, A., Popov, B., Terziev, I., Manov, D., & Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web, 2*(1), 49-79.

Mccrae, J., Aguado-De-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., . . . Wunner, T. (2012, dec). Interchanging lexical resources on the Semantic Web. *Lang. Resour. Eval., 46*(4), 701-719.

Neubert, J. (2009). Bringing the "Thesaurus for Economics" on to the Web of Linked Data. In C. Bizer, T. Heath, T. Berners-Lee, & K. Idehen (Ed.), *Proceedings of the Linked Data on the Web Workshop (LDOW2009). 538.* Madrid, Spain: CEUR-WS.org.

Paredes, L. P., Rodrıguez, J. M., & Azcona, E. R. (2008). Promoting Government Controlled Vocabularies for the Semantic Web: the EUROVOC Thesaurus and the CPV Product Classification System. *Semantic Interoperability in the European Digital Library*, (p. 111).

Pastor-Sanchez, J.-A., Martínez Mendez, F. J., & Rodríguez-Muñoz, J. V. (2009). Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. *Information Research, 14*(4), 10.

Pazienza, M. T., & Stellato, A. (2005). The Protégé Ontoling Plugin - Linguistic Enrichment of Ontologies in the Semantic Web. *In poster proceedings of the 4th International Semantic Web Conference (ISWC-2005).* Galway, Ireland.

Pazienza, M. T., Stellato, A., & Turbati, A. (2008). Linguistic Watermark 3.0: an RDF framework and a software library for bridging language and ontologies in the Semantic Web. *Semantic Web Applications and Perspectives, 5th Italian Semantic Web Workshop (SWAP2008).* FAO-UN, Rome, Italy.

Pazienza, M., & Stellato, A. (2006). An Environment for Semi-automatic Annotation of Ontological Knowledge with Linguistic Content. In Y. Sure, & J. Domingue (A cura di), *The Semantic Web: Research and Applications, 3rd European Semantic Web Conference, ESWC 2006, Budva, Montenegro, June 11-14, 2006, Proceedings. Lecture Notes in Computer Science. 4011*, p. 442-456. Springer.

Pazienza, M., & Stellato, A. (2006). Exploiting Linguistic Resources for building linguistically motivated ontologies in the Semantic Web. *Second Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006).* Genoa, Italy.

Pazienza, M., & Stellato, A. (2006). Linguistic Enrichment of Ontologies: a methodological framework. *Second Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006).* Genoa, Italy.

Pazienza, M., Sguera, S., & Stellato, A. (2007, December 26). Let's talk about our “being”: A linguistic-based ontology framework for coordinating agents. (R. Ferrario, & L. Prévot, Eds.) *Applied Ontology, special issue on Formal Ontologies for Communicating Agents, 2*(3-4), 305-332.

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. *Proceedings of the Eighth International Conference on Language* (pp. 36-42). ELRA.

Shvaiko, P., & Euzenat, J. (2013). Ontology Matching: State of the Art and Future Challenges. *IEEE Transactions on Knowledge and Data Engineering, 25*(1), 158-176.

Shvaiko, P., Euzenat, J., Kementsietsidis, A., Mao, M., Noy, N., & Stuckenschmidt, H. (Eds.). (2012). Proceedings of the 7th International Workshop on Ontology Matching, Boston, MA, USA, November 11, 2012. *OM. 946.* CEUR-WS.org.

Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufiş, D., . . . Grigoriadou, M. (2002). BALKANET: A Multilingual Semantic Network for the Balkan Languages. *International Wordnet Conference*, (pp. 12-14). Mysore, India.

Suárez-Figueroa, M. C., & Gómez-Pérez, A. (2008). First Attempt towards a Standard Glossary of Ontology Engineering Terminology. In B. N. Madsen, & H. E. Thomsen (Eds.), *Managing Ontologies and Lexical Resources. TKE 2008 8th International Conference on Terminology and KE.* Copenhagen: Institut for Internationale Sprogstudier og Vidensteknologi (ISV).

Van Assem, M., Gangemi, A., & Schreiber, G. (2006). Conversion of WordNet to a standard RDF/OWL representation. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy.*

Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer Academic Publishers.

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

Zapilko, B., Schaible, J., Mayr, P., & Mathiak, B. (2013). TheSoz: A SKOS Representation of the Thesaurus for the Social Sciences. (P. Hitzler, & K. Janowicz, Eds.) *Semantic Web Journal*, *4*(3), 257–263.