Standardizing Lexical-Semantic Resources – Fleshing out the abstract standard LMF

Judith Eckle-Kohler[‡]
‡ Ubiquitous Knowledge Processing
Lab (UKP-TUDA)
Department of Computer Science
Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Iryna Gurevych[†‡]
† Ubiquitous Knowledge Processing
Lab (UKP-DIPF)
German Institute for Educational
Research and Educational Information
www.ukp.tu-darmstadt.de

Abstract

This paper describes the application of the Lexical Markup Framework (LMF) for standardizing lexical-semantic resources in the context of NLP. More specifically, we highlight the question how lexical-semantic resources can be made semantically interoperable by means of LMF and ISOCat. The LMF model UBY-LMF, an instantiation of LMF specifically for NLP, serves as an example to illustrate the path towards semantic interoperability of lexical resources.

1 Introduction

Lexical-semantic resources (LSR) are used in major NLP tasks, such as word sense disambiguation, semantic role labeling and information extraction. In recent years, the aspects of reusing and merging LSRs have gained significance, mainly due to the fact that LSRs are expensive to build. Standardization of LSRs plays an important role in this context, because it facilitates integration and merging of LSRs and makes reuse of LSRs easy. NLP systems that are built according to standards can simply plug in standardized LSRs and are thus able to easily switch between different standardized LSRs. In other words, standardizing LSRs makes them interoperable.

Two aspects of interoperability are to be distinguished in NLP: syntactic interoperability and semantic interoperability (Ide and Pustejovsky, 2011). While NLP systems can perform the same kind of processing with *syntactically interoperable* LSRs, there is no guarantee, that the results can be interpreted the same way. Two syntactically interoperable LSRs might use the same term to denote different meanings. *Semantically interoperable* LSRs, on the other hand, use terms that share a common definition of their meaning. Consequently, NLP systems that switch between semantically interoperable LSRs can perform the same kind of processing and the results produced can still be interpreted the same way.

In this paper, we focus on the question how to achieve semantic interoperability by means of the ISO 24613:2008 LMF (Francopoulo et al., 2006) and ISOCat.[1] The comprehensive LMF lexicon model UBY-LMF (Eckle-Kohler et al., 2012) serves as an example to show how the abstract LMF standard is to be fleshed out and instantiated in order to make LSRs semantically interoperable for NLP purposes.

UBY-LMF covers very heterogeneous LSRs in two languages, English and German, and has been used to standardize a range of LSRs resulting in the large-scale LSR UBY (Gurevych et al., 2012), see http://www.ukp.tu-darmstadt.de/uby/. UBY currently contains ten resources in two languages: English WordNet (Fellbaum, 1998), Wiktionary[2], Wikipedia[3], FrameNet (Baker et al., 1998), and VerbNet (Kipper et al., 2008); German Wiktionary, Wikipedia, GermaNet (Kunze and Lemnitzer, 2002) and IMSlex-Subcat (Eckle-Kohler, 1999) and the English and German entries of OmegaWiki[4].

[1] http://www.isocat.org/

[2] http://www.wiktionary.org/

[3] http://www.wikipedia.org/

[4] http://www.omegawiki.org/

2 LMF and semantic interoperability

First, we give an overview of the LMF standard and briefly describe how to use it. We put a special focus on the question how LSRs can be made semantically interoperable by means of LMF.

LMF – an abstract standard LMF defines a *meta-model* of lexical resources, covering both NLP lexicons and machine readable dictionaries. The standard specifies this meta-model in the Unified Modeling Language (UML) by providing a set of UML diagrams. UML packages are used to organize the meta-model and each diagram given in the standard corresponds to an UML package. LMF defines a mandatory core package and a number of extension packages for different types of resources, e.g., morphological resources or wordnets. The core package models a lexicon in the traditional headword-based fashion, i.e., organized by lexical entries. Each lexical entry is defined as the pairing of one to many forms and zero to many senses.

Instantiating LMF The abstract meta-model given by the LMF standard is not immediately usable as a format for encoding (i.e., converting) an existing LSR (Tokunaga et al., 2009). It has to be instantiated first, i.e., a full-fledged lexicon model has to be developed by choosing LMF classes and by specifying suitable attributes for these LMF classes.

According to the standard, developing a lexicon model involves

1. selecting classes from the UML packages,
2. defining attributes for these classes and
3. linking the attributes and other linguistic terms introduced (e.g., attribute values) to standardized descriptions of their meaning.

Selecting a combination of LMF classes from the LMF core package and from the extension packages establishes the structure of a lexicon model. While the LMF core package models a lexicon in terms of lexical entries, the LMF extensions provide classes for different types of lexicon organization, e.g., covering the synset-based organization of wordnets or the semantic frame-based organization of FrameNet.

Fixing the structure of a lexicon model by choosing a set of classes contributes to syntactic interoperability of LSRs, as it fixes the high-level organization of lexical knowledge in an LSR, e.g., whether synonymy is encoded by grouping senses into synsets (using the `Synset` class) or by specifying sense relations (using the `SenseRelation` class), which connect synonymous senses.

Defining attributes for the LMF classes and specifying the attribute values is far more challenging than choosing from a given set of classes, because the standard gives only a few examples of attributes and leaves the specification of attributes to the user in order to allow maximum flexibility.

Finally, the attributes and values have to be linked to a description of their meaning in an ISO 12620:2009 compliant Data Category Registry (DCR), see (Broeder et al., 2010). ISOCat is the implementation of the ISO 12620:2009 DCR providing descriptions of terms used in language resources.

These descriptions in ISOCat are standardized, i.e., they comply with a predefined format and provide some mandatory information types, including a unique administrative identifier (e.g., partOfSpeech) and a unique and persistent identifier (PID, e.g., http://www.isocat.org/datcat/DC-396) which can be used to link to the descriptions. The standardized descriptions of terms are called *Data Categories* (DCs).

Semantic Interoperability Connecting the linguistic terms used for attributes and their values in a lexicon model with their meaning defined externally in ISOCat contributes to *semantic interoperability* of LSRs (see also Windhouwer and Wright (2012)). The definitions of DCs in ISOCat constitute an interlingua that can be used to map idiosyncratically used linguistic terms to a set of reference definitions (Chiarcos, 2010). Different LSRs that share a common definition of their linguistic vocabulary are said to be *semantically interoperable* (Ide and Pustejovsky, 2010).

Consider as an example the `LexicalEntry` class of two different lexicon models A and B. Lexicon model A could have an attribute `partOfSpeech` (POS), while lexicon model B could have an attribute `pos`. Linking both attributes to the meaning “A category assigned to a word based on its grammatical and semantic properties." given in ISOCat (http://www.isocat.org/datcat/DC-396) makes the two lexicon models semantically interoperable with respect to the POS attribute.

A human can look up the meaning of a term occurring in a lexicon model by following the link to the ISOCat DC and consulting its description in ISOCat. Linking the attributes and their values to ISOCat DCs results in a so-called Data Category Selection.

It is important to stress that the notion of "semantic interoperability" in the context of LMF has a limited scope: it only refers to the meaning of the linguistic vocabulary used in an LMF lexicon model – not to the meaning of the lexemes listed in a LSR.

3 UBY-LMF – Instantiating LMF

Considering the fact that only a fleshed-out LMF lexicon model, i.e., an instantiation of the LMF standard, can be used for actually standardizing LSRs, it is obvious that LMF-compliant LSRs are not necessarily interoperable, neither syntactically nor semantically.

Therefore it is important to develop a single, comprehensive instantiation of LMF, which can immediately be used for standardizing LSRs. The LMF lexicon model UBY-LMF strives to be such a comprehensive instantiation of LMF to be used in NLP.

3.1 UBY-LMF characteristics

UBY-LMF covers a wide range of lexical information types, since it has been designed as a uniform format for standardizing heterogeneous types of LSRs, including both expert-constructed resources – wordnets, FrameNet, VerbNet – and collaboratively constructed resources – Wikipedia, Wiktionary, OmegaWiki.

In UBY-LMF, there is one `Lexicon` per integrated resource, i.e., one `Lexicon` for FrameNet, for WordNet and so on. This way, the `Lexicon` instances can be aligned at the sense level by linking pairs of senses or synsets using instances of the `SenseAxis` class.

The full model consists of 39 classes and 129 attributes. Please refer to (Eckle-Kohler et al., 2012) and (Gurevych et al., 2012) for detailed information on UBY-LMF and the corresponding large-scale LSR UBY.

UBY-LMF is represented by a DTD which can be used to automatically convert any given resource into the corresponding XML format. Converters for ten LSRs to UBY-LMF format are publicly available on Google Code, see http://code.google.com/p/uby/.

3.2 UBY-LMF attributes

In UBY-LMF, the definition of attributes for the LMF classes was guided by two requirements that we identified as important in the context of NLP: (i), comprehensiveness, and (ii) extensibility (Eckle-Kohler et al., 2012).

Comprehensiveness implies that the model should be able to represent all the lexical information present in a wide range of LSRs, because NLP applications usually require different types of lexical knowledge and it is difficult to decide in advance which type of lexical information will be useful for a particular NLP application.

Extensibility is also crucial in the NLP domain, because UBY-LMF should be applicable across languages (Gurevych et al., 2012), (Eckle-Kohler and Gurevych, 2012) and as well be able to adopt automatically extracted lexical-semantic knowledge.

4 UBY-LMF and semantic interoperability

In section 2, we have pointed out that linking attributes and values used in an LMF lexicon model to DCs in ISOCat is crucial for semantic interoperability.

Now we will take a closer look at the semantic interoperability of UBY-LMF compliant resources by describing in detail the grounding of UBY-LMF attributes and values in the ISOCat repository. First, we introduce ISOCat, then, we describe the process of selecting and creating an ISOCat Data Category Selection for UBY-LMF, and finally, we look at some limitations of the current version of UBY-LMF.

4.1 Overview of ISOCat

The Data Category Registry ISOCat is a collaboratively constructed repository where everybody can register and create DCs. Users can assign their DCs to so-called Thematic Domains, such as Morphosyntax, Syntax or Lexical Resources. The ISOCat Web interface provides a form that guides the user through the process of creating a DC, also indicating which kind of information has to be provided mandatorily. The well-formedness of a newly created DC is automatically checked and displayed by a flag – a green marking indicating well-formedness.

Users can also group DCs, including self-created ones, into a Data Category Selection. Typically, Data Category Selections are created for projects or resources, e.g., there are Data Category Selections for RELISH[5] or CLARIN[6] or for resources, such as the STTS tagset[7] or UBY.[8] Data Category Selections can be made publicly available, in order to allow for linking to particular DCs.

It is possible to submit subsets of well-formed DCs to standardization. While the standardization of ISOCat DCs has not yet started, standardized DCs might be important for resources where sustainability is an issue, because standardized DCs can be considered as stable. Non-standardized DCs, on the other hand, could in principle be changed at any time by their owners which might also involve changes in their meaning.

Two types of DCs distinguished in ISOCat are relevant for LMF lexicon models (Windhouwer and Wright, 2012): first, complex DCs which have a typed value domain and typically correspond to attributes in an LMF lexicon model. According to the size of the value domain, DCs are classified further into open DCs (they can take an arbitrary number of values), closed DCs (their values can be enumerated) and constrained DCs (the number of values is too big in order to be enumerated, but yet constrained). Second, there are simple DCs which describe values of a closed DC.

The attributes and values defined in UBY-LMF refer to 175 ISOCat DCs; most of them are simple DCs. The corresponding Data Category Selection is publicly available in ISOCat, see http://www.isocat.org/rest/dcs/484.

Figure 1 shows a screenshot of the public Data Category Selection for UBY as of 2012.

4.2 UBY-LMF: Selecting DCs from ISOCat

Selecting DCs from ISOCat which are suitable descriptions of terms used in a lexicon model such as UBY-LMF is a task that requires the identification of fine-grained and subtle differences in meaning and hence, is a task where humans are superior to machines in terms of quality. For UBY-LMF, the intended meanings of the lexicon model terms and the textual descriptions of the DCs were manually compared in order to first identify candidate DCs with equivalent or similar meaning and then to select one of the candidate DCs as reference for a specific term.

Searching for DCs in ISOCat Identifying candidate DCs in ISOCat means accessing the ISOCat Web interface (http://www.isocat.org/interface/index.html) and searching for the lexicon model term or variants thereof. Searching for candidate DCs in ISOCat is time-consuming and not particularly user-friendly, because there are many DCs with similar names and equivalent or near-equivalent meaning. Currently ISOCat does not display relations between such terms (e.g., the equivalence relation).

This is offered to a limited extent by RelCat, a companion registry to ISOCat (Windhouwer, 2012). However, RelCat is still at an early stage of development and currently provides only relations between selected Data Category Selections, such as GOLD[9] and RELISH. Therefore, we did not use it for the selection of DCs for UBY-LMF.

There are basically two ways of looking for DCs in ISOCat: first, by entering a search term and second by browsing Thematic Domains, such as Syntax, or by browsing Data Category Selections published by particular groups or projects, such as CLARIN or RELISH.

Choosing among several candidate DCs Typically, an ISOCat search query for a term (using the option *exact match*) yields a list of DCs that have slightly different names, but are very similar in meaning. Consider as an example the linguistic

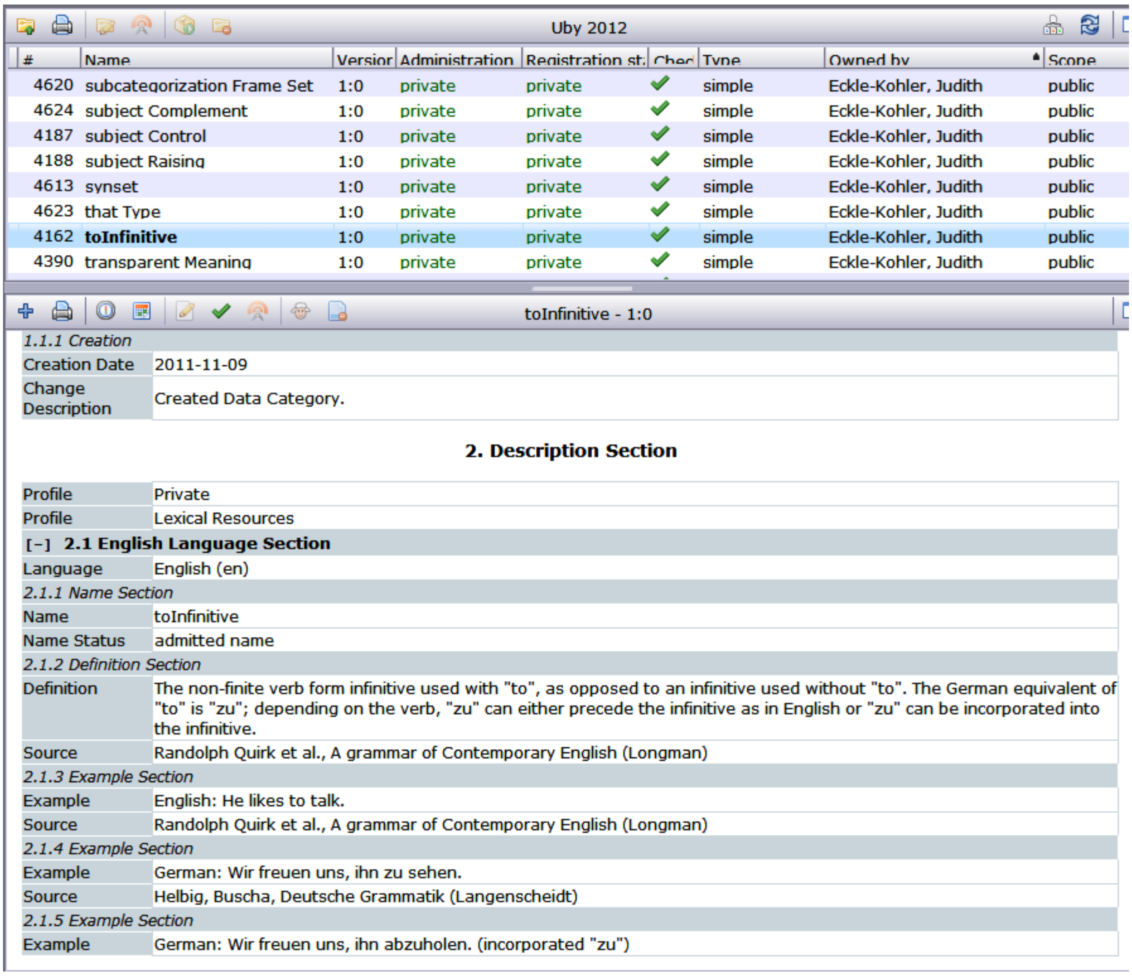
[5] http://tla.mpi.nl/relish/

[6] http://www.clarin.eu

[7] http://www.isocat.org/rest/dcs/376

[8] http://www.isocat.org/rest/dcs/484

[9] http://linguistics-ontology.org/



Uby 2012

#	Name	Versior	Administration	Registration st.	Chec	Type	Owned by	Scope
4620	subcategorization Frame Set	1:0	private	private	✓	simple	Eckle-Kohler, Judith	public
4624	subject Complement	1:0	private	private	✓	simple	Eckle-Kohler, Judith	public
4187	subject Control	1:0	private	private	✓	simple	Eckle-Kohler, Judith	public
4188	subject Raising	1:0	private	private	✓	simple	Eckle-Kohler, Judith	public
4613	synset	1:0	private	private	✓	simple	Eckle-Kohler, Judith	public
4623	that Type	1:0	private	private	✓	simple	Eckle-Kohler, Judith	public
4162	**toInfinitive**	1:0	private	private	✓	simple	Eckle-Kohler, Judith	public
4390	transparent Meaning	1:0	private	private	✓	simple	Eckle-Kohler, Judith	public

toInfinitive - 1:0

1.1.1 Creation	
Creation Date	2011-11-09
Change Description	Created Data Category.

2. Description Section

Profile	Private
Profile	Lexical Resources
[−] 2.1 English Language Section	
Language	English (en)
2.1.1 Name Section	
Name	toInfinitive
Name Status	admitted name
2.1.2 Definition Section	
Definition	The non-finite verb form infinitive used with "to", as opposed to an infinitive used without "to". The German equivalent of "to" is "zu"; depending on the verb, "zu" can either precede the infinitive as in English or "zu" can be incorporated into the infinitive.
Source	Randolph Quirk et al., A grammar of Contemporary English (Longman)
2.1.3 Example Section	
Example	English: He likes to talk.
Source	Randolph Quirk et al., A grammar of Contemporary English (Longman)
2.1.4 Example Section	
Example	German: Wir freuen uns, ihn zu sehen.
Source	Helbig, Buscha, Deutsche Grammatik (Langenscheidt)
2.1.5 Example Section	
Example	German: Wir freuen uns, ihn abzuholen. (incorporated "zu")

Figure 1: Screenshot of the public Data Category Selection “Uby 2012“ in ISOCat, see http://www.isocat.org/interface/index.html.

term *determiner* which is a possible value of the attribute `partOfSpeech` in UBY-LMF. There are 25 DCs in ISOCat that exactly match the term *determiner*.

Another example is the term *direct object* that is sometimes used for specifying the accusative noun phrase argument of transitive verbs in German. In ISOCat, there are two different specifications of this term, one explicitly stating that this accusative noun phrase argument can become the clause subject in passivization (http://www.isocat.org/datcat/DC-1274), the other not mentioning passivization at all (http://www.isocat.org/datcat/DC-2263).

We tried to select a DC with a meaning as close as possible to the intended meaning in UBY-LMF. When selecting a particular DC from a list of candidates, we followed two simple strategies:

- We sorted the search results by Owner, i.e., by the person who created the DC, and preferred DCs which are owned by experts, either in the standardization community or in the linguistics community.

- We preferred those DCs that had passed the ISOCat test for well-formedness.

Sometimes, selecting an existing DC from ISOCat involved making compromises. For instance, the attribute value `taxonomic` of the attribute `relType` of the `SenseRelation` class links to the ISOCat DC taxonomy (http://www.isocat.org/rest/dc/4039) which has a narrower meaning than the meaning intended in UBY-LMF: While `taxonomic` in UBY-LMF denotes a type of sense relation that defines a taxonomy in a general sense, covering all kinds of lexemes, (see Cruse (1986)), the DC 4039, taxonomy, is restricted to controlled vocabulary terms organized in a taxonomy.

Note that many attributes or attribute values in UBY-LMF refer to ISOCat DCs with a different (so-called admitted) name. This is explicitly supported by ISOCat, because each DC definition may optionally contain Data Element Name Sections in order to record other names for the DC as used in different sources, such as a given database, format or application. In this manner, the number of parallel DCs in ISOCat with equivalent meaning can be limited. We left the specification of Data Element Name Sections in the newly created DCs (holding a deviating UBY name) to future work.

Semantic Drift As ISOCat is a collaboratively constructed and thus dynamically evolving repository, DCs that are not standardized yet could eventually change their meaning over time, e.g., they might be further fleshed out and specified in more detail.

Since the public release of UBY in March 2012, we have already encountered such a case of semantic drift for the DC verbFormMood (see http://www.isocat.org/rest/dc/1427) which we selected as a description of the attribute `verbForm` attached to the `SyntacticCategory` class in the Syntax part of UBY-LMF.

The DC verbFormMood has been changed after March 2012 (according to the change log, this DC was changed on 2012-06-09) and is no longer similar enough to the intended meaning. Therefore, we will create a new DC which specifies `verbForm` as a property of verb phrase complements in the context of subcategorization.

4.3 UBY-LMF: Creating new ISOCat DCs

The majority of attributes and values in UBY-LMF refer to already existing ISOCat DCs. Yet, we had to create 38 new DCs in ISOCat, as particular definitions were missing. We decided to create new DCs in those domains where much standardization work has already been done or large sources of linguistic expert knowledge are available. In particular, these are the domains lexical syntax (related to subcategorization), derivational morphology, and frame semantic information.

We used the following expert-based sources as references for newly created DCs:

- the EAGLES synopsis on morphosyntactic phenomena[10] (Calzolari and Monachini, 1996), as well as the EAGLES recommendations on subcategorization[11] have been used to identify DCs relevant for lexical syntax
- traditional grammars such as the English grammar by Quirk (1980)
- a Web-based encyclopedia of linguistic terms, collaboratively built by linguists, see http://www.glottopedia.de
- the detailed documentation of the FrameNet resource (Ruppenhofer et al., 2010)

Fifteen of the newly created DCs are required for representing subcategorization frames in UBY-LMF. Most of these DCs are simple DCs that are linked to attribute values in UBY-LMF. The corresponding DCs were either missing in ISOCat or there were only DCs with a related, but not sufficiently similar meaning.

For instance, we created a DC for to-infinitive complements as in *He tried to address all questions.*. While both *infinitive* and *infinitiveParticle* are present in ISOCat, all available definitions of *infinitive* do not explicitly exclude the presence of a particle, neither http://www.isocat.org/datcat/DC-1312 nor http://www.isocat.org/datcat/DC-2753. What is required, however, for the specification of verbal complements are individual DCs for bare infinitives used without particle and for infinitives used with particle (i.e., *to* in English).

Ten newly created DCs are necessary to represent FrameNet in LMF and primarily describe specific properties of frame-semantic frames, e.g., coreness, incorporated semantic argument or transparent meaning.

4.4 UBY-LMF: Limitations of semantic interoperability

The semantic interoperability of a UBY-LMF compliant resource is naturally limited to those domains within UBY-LMF where attributes and

[10] http://www.ilc.cnr.it/EAGLES96/morphsyn/

[11] http://www.ilc.cnr.it/EAGLES96/synlex/

their values are fleshed out at a fine-grained level and refer to ISOCat DCs. Consequently, attributes of UBY-LMF classes that are string-valued currently limit the semantic interoperability of UBY-LMF. There are three main areas in UBY-LMF where attributes or their values are currently string-valued due to a lack of standardization: first, the names of semantic relations, second, the names of semantic roles, and third, the types of arbitrary semantic classification schemes, as well as the labels used in these schemes.

Semantic Relations Names of semantic relations are values of the attribute `relName` of the `SenseRelation` and `SynsetRelation` class. It is not clear, if the set of possible names for semantic relations is restricted, considering the names of semantic relations found in Wiktionary, OmegaWiki, WordNet and FrameNet, For instance, in OmegaWiki, relations such as "works in a", "partners with", "is practiced by a", "orbits around" (for moons moving around a planet) are listed. This does not fit in the set of classical lexical-semantic relations described by Cruse (1986).

Semantic Roles Names of semantic roles are values of the attribute `semanticRole` of the `SemanticArgument` class. Although DCs for semantic roles have been proposed by Schiffrin and Bunt (2007), we have not entered them to ISOCat, because there is still ongoing work in ISO committees on the standardization of semantic roles.

Semantic classification schemes Finally, names of semantic classification schemes and the labels used in these schemes are values of the attributes `type` and `label` of the `SemanticLabel` class. On the one hand, the `type` attribute covers such divergent classification schemes as Wikipedia categories, WordNet semantic fields (i.e., the lexicographer file names), VerbNet classes, selectional preference schemes or register classification which is present in Wiktionary.

On the other hand, the string-values of the `label` attribute are even more diverse. For instance, the selectional preference schemes in FrameNet and VerbNet employ different label sets for selectional preferences. While VerbNet makes use of a combination of high-level concepts from the EuroWordNet hierarchy (Kipper-Schuler, 2005), e.g., $+$*animate* & $+$*organization*, FrameNet selectional preferences (called ontological types in FrameNet) correspond to synset nodes of WordNet (Ruppenhofer et al., 2010), such as *Message, Container, Speed.*

5 Discussion

In this section, we first compare UBY-LMF as an LMF instantiation with the TEI guidelines for Dictionaries. For a detailed discussion of related work, please refer to (Eckle-Kohler et al., 2012). Then, we give an outlook of future work on providing alternative formats for UBY-LMF which make the linking to ISOCat accessible to larger communities.

5.1 The TEI guidelines for Dictionaries

The Text Encoding Initiative (TEI) provides guidelines which define standardized representation formats for various kinds of digitized texts in the Digital Humanities. In particular, the TEI provides guidelines for Dictionaries which are in principle applicable to all kinds of lexical resources, but primarily to those that are intended for human use.

Romary (2010a) has suggested to develop the TEI guidelines for Dictionaries into a full LMF instantiation. In their current form, the TEI guidelines cover already core classes of LMF which are required for representing dictionaries. Turning the TEI guidelines for Dictionaries into an LMF instantiation would require first to select the subset of these guidelines that can be mapped to LMF and then to extend the guidelines, e.g., in order to also cover lexical syntax in more detail (Romary, 2010b).

We did not follow this proposal for UBY-LMF, however, because we aimed at representing very heterogeneous resources, which are important for NLP systems, by a single, uniform lexicon model. This particular goal could be achieved mainly due to the high degree of flexibility offered by the LMF standard. Using the TEI guidelines as a starting point seemed to bring about too many constraints, because many aspects of the lexicon model would have been fixed already.

5.2 Linking to ISOCat – Outlook

The actual representation format for the linking of UBY-LMF terms and ISOCat DCs should be useful both for humans and for machines. For humans, it is helpful to be able to look up the meaning of the terms used in UBY by reading the descriptions of these terms in ISOCat. Machines, on the other hand, can determine the degree of semantic interoperability of two lexical resources by comparing the ISOCat PIDs each lexical resource links to.

At the time of writing this paper, UBY-LMF is the only LMF lexicon model with a publicly accessible Data Category Selection in ISOCat. We hope that DataCategory Selections used in other LMF models will be added to ISOCat soon, because referring to ISOCat DCs is the only way to automatically determine the degree of semantic interoperability of any two LMF-compliant LSRs.

Currently, UBY-LMF provides human-readable links to ISOCat DCs at the schema level (as suggested by Windhouwer and Wright (2012)), i.e., at the level of the UBY-LMF DTD. Pairs of attributes or attribute values and ISOCat DCs are listed in XML comments. As part of future work, we plan to provide additional versions of the UBY-LMF model with machine-readable links to ISOCat, in particular an XSD version as well as an RDFS version for the Semantic Web community.

In addition, we plan to create a metadata instance for UBY-LMF based on the component metadata infrastructure which has been adopted by large communities developing infrastructures of language resources, such as CLARIN and META-SHARE (Broeder et al. (2012), Labropoulou and Desipri (2012), Gavrilidou et al. (2011)). Such a metadata instance specifies the characteristics of UBY-LMF as a language resource, also including the linking of UBY-LMF terms to ISOCat.

6 Conclusion

We have shown how LMF contributes to semantic interoperability of LSRs which is crucial for NLP systems using LSRs. The two key aspects are first, using a single instantiation of LMF for multiple resources, such as UBY-LMF, and second, establishing a linking between an LMF lexicon model and ISOCat DCs.

Many DCs required for LSRs are still missing in ISOCat, mainly due to ongoing standardization in various areas. Yet, for UBY-LMF, we have fleshed out a Data Category Selection of considerable size and detail in such areas as morphosyntax and subcategorization. We made the UBY-LMF Data Category Selection available in ISOCat as a starting point for further work on standardizing lexical resources according to LMF.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank Silvana Hartmann, Michael Matuschek and Christian M. Meyer for their contributions to this work.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 86–90, Montreal, Canada.

Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, and Claus Zinn. 2010. A Data Category Registry- and Component-based Metadata Framework. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 43–47, Valletta, Malta.

Daan Broeder, Dieter van Uytvanck, Maria Gavrilidou, Thorsten Trippel, and Menzo Windhouwer. 2012. Standardizing a component metadata infrastructure. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1387 – 1390, Istanbul, Turkey.

Nicoletta Calzolari and Monica Monachini. 1996. EAGLES Proposal for Morphosyntactic Standards: in view of a ready-to-use package. In G. Perissinotto, editor, *Research in Humanities Computing*, volume 5, pages 48–64. Oxford University Press, Oxford, UK.

Christian Chiarcos. 2010. Towards robust multi-tool tagging. an OWL/DL-based approach. In *Proceed-*

ings of the 48th Annual Meeting of the Association for Computational Linguistic, pages 659–670. Association for Computational Linguistic.

D.A. Cruse. 1986. *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Judith Eckle-Kohler and Iryna Gurevych. 2012. Subcat-LMF: Fleshing out a standardized format for subcategorization frame interoperability. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 550–560, Avignon, France.

Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer. 2012. UBY-LMF - A Uniform Format for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 275–282, Istanbul, Turkey.

Judith Eckle-Kohler. 1999. *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora*. Logos-Verlag, Berlin, Germany. PhDThesis.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 233–236, Genoa, Italy.

Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis, Monica Monachini, Francesca Frontini, Gil Francopoulo, Victoria Arranz, and Valérie Mapelli. 2011. A Metadata Schema for the Description of Language Resources (LRs). In *Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, pages 84–92, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. Uby - A Large-Scale Unified Lexical-Semantic Resource. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590, Avignon, France.

Nancy Ide and James Pustejovsky. 2010. What Does Interoperability Mean, anyway? Toward an Operational Definition of Interoperability. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, Hong Kong.

Nancy Ide and James Pustejovsky. 2011. Issues and Strategies for Interoperability among NLP Software Systems and Resources, Tutorial at the 5th International Joint Conference on Natural Language Processing (IJCNLP).

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A Large-scale Classification of English Verbs. *Language Resources and Evaluation*, 42:21–40.

Karin Kipper-Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon. PhD. Thesis. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.

Claudia Kunze and Lothar Lemnitzer. 2002. GermaNet representation, visualization, application. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, pages 1485–1491, Las Palmas, Canary Islands, Spain.

Penny Labropoulou and Elina Desipri. 2012. Documentation and User Manual of the META-SHARE Metadata Model, http://www.meta-net.eu.

Randolph Quirk. 1980. *A Grammar of contemporary English*. Longman.

Laurent Romary. 2010a. Standardization of the formal representation of lexical information for NLP. In *Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*. Mouton de Gruyter.

Laurent Romary. 2010b. Using the TEI framework as a possible serialization for LMF, presentation at the RELISH workshop, August.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. FrameNet II: Extended Theory and Practice.

Amanda Schiffrin and Harry Bunt. 2007. Documented compilation of semantic data categories, LIRICS Deliverable D4.3, http://lirics.loria.fr/documents.html.

Takenobu Tokunaga, Dain Kaplan, Nicoletta Calzolari, Monica Monachini, Claudia Soria, Virach Sornlertlamvanich, Thatsanee Charoenporn, Yingju Xia, Chu-Ren Huang, Shu-Kai Hsieh, and Kiyoaki Shirai. 2009. Query Expansion using LMF-Compliant Lexical Resources. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 145–152, Suntec, Singapore.

Menzo Windhouwer and Sue Ellen Wright. 2012. Linking to linguistic data categories in ISOcat. pages 99–107. Springer, Heidelberg.

Menzo Windhouwer. 2012. RELcat: a Relation Registry for ISOcat data categories. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3661 – 3664, Istanbul, Turkey.