An Integrated Formal Representation for Terminological and Lexical Data included in Classification Schemes

Thierry Declerck[1,2], Kseniya Egorova[3], Eileen Schnur[1]
[1]DFKI GmbH, Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
[2]Austrian Academy of Sciences, Sonnenfelsgasse 19,1010 Vienna, Austria
[3]Saint Petersburg State University of Architecture and Civil Engineering,
2-nd Krasnoarmeiskaya St. 4, 190005 St. Petersburg, Russia
declerck@dfki.de, kseniya.a.egorova@gmail.com, eileen.schnur@dfki.de

Abstract

This paper presents our work dealing with a potential application in e-lexicography: the automatized creation of specialized multilingual dictionaries from structured data, which are available in the form of comparable multilingual classification schemes or taxonomies. As starting examples, we use comparable industry classification schemes, which frequently occur in the context of stock exchanges and business reports. Initially, we planned to follow an approach based on cross-taxonomies and cross-languages string mapping to automatically detect candidate multilingual dictionary entries for this specific domain. However, the need to first transform the comparable classification schemes into a shared formal representation language in order to be able to properly align their components before implementing the algorithms for the multilingual lexicon extraction soon became apparent. We opted for the SKOS-XL vocabulary for modelling the multilingual terminological part of the comparable taxonomies and for OntoLex-Lemon for modelling the multilingual lexical entries which can be extracted from the original data. In this paper, we present the suggested modelling architecture, which demonstrates how terminological elements and lexical items can be formally integrated and explicitly cross-linked in the context of the Linguistic Linked Open Data (LLOD).

Keywords: SKOS-XL, OntoLex-Lemon, Terminology, Lexicography

1. Introduction

The topic of extracting dictionaries from raw data was discussed in the context of the recently terminated European Network of e-Lexicography (ENeL) COST Action[1], and it was also the motto of the fifth biennial eLex conference[2] called "eLex 2017: Lexicography from scratch". The work presented in this paper takes its source within this context, meaning in detail that we aim to investigate the automated extraction of domain-specific multilingual lexicons from comparable classification schemes or taxonomies.

As a dataset for our investigation, we selected two comparable industry classification schemes that are used in various stock exchanges like Euronext, the New York Stock Exchange, the Toronto Stock Exchange, etc., or within business reports of companies. The decision to use these sources was also inspired by former ontology mapping work applied to this kind of data, as described in (Gromann and Declerck, 2014), building thus on related previous work. We planned to follow an approach based on cross-taxonomies and cross-languages string mappings to automatically detect candidate multilingual dictionary entries for this specific domain.

After an in-depth analysis of both data sources, which are available in Excel files, it soon became obvious that there is a necessity to proceed from a formally identical structure derived from the distinct sources. As a consequence of that, we decided to first transform the comparable classification schemes into a shared formal representation language in order to be able to properly align their components before implementing the algorithms for the domain specific multilingual lexicon extraction.

We opted for SKOS-XL[3] for modelling the multilingual terminological part of the comparable taxonomies and OntoLex-Lemon[4] for modelling the lexical items that can be extracted from the labels and the definitions included in the classification schemes.

As both vocabularies, SKOS-XL and OntoLex-Lemon, are using formal representation languages that supports the publication of terminological and lexical datasets on the Linguistic Linked Open Data (LLOD)[5] cloud, our approach can contribute to a significant increase of the linking of such terminological and lexical datasets in the LLOD framework.

In the next sections, we will first introduce the datasets, followed by the description of the formalisation in SKOS-XL of multilingual labels and definitions used in the two classification schemes. After that, we will show how the lexical items used in these labels and definitions can be modelled in OntoLex-Lemon, before displaying the suggested modelling architecture for integrating SKOS- XL and OntoLex-Lemon and finally illustrating how this integration can be published in the LLOD cloud.

2. The Data Sources

We are currently applying our approach on two comparable multilingual industry classification schemes: the In-

[1]See `http://www.elexicography.eu/` for more details.

[2]See `https://elex.link/elex2017/`.

[3]`https://www.w3.org/TR/skos-reference/skos-xl.html`.

[4]See `https://www.w3.org/2016/05/ontolex/` and (McCrae et al., 2017) for more details.

[5]See `http://linguistic-lod.org/llod-cloud` and (Chiarcos et al., 2012) for more details.

dustry Classification Benchmark (ICB)[6] and the Global Industry Classification Standard (GICS)[7]. Both classification schemes are using a four levels taxonomic structure, each level being indexed by a numeral combination associated with a short textual label, and in both classification schemes a definition text is added to the leaf element of the taxonomic structure.

ICB implements a taxonomic structure consisting of 10 industries, subdivided into 19 supersectors, which are further divided into 41 sectors including 114 subsectors, which are the leaf categories/labels that are equipped with a definition. The similar looking GICS consists of 11 sectors, subdivided in 24 industry groups, partitioned in 68 industries and 157 sub-industries, which are the leaf categories/labels to which the definitions are associated.

Figure 1 below shows an example of the taxonomic structure of ICB, in which its 4 levels are indicated by the increasing specification of numbers: $7000 > 7500 > 7530 > 7537$. Here, only the German and English labels are illustrated. Each leaf category/label is associated with a definition, which is shown by the example of index 7537, displayed in Figure 2.

ICB	**German Categories introducing the definition**	**English Categories introducing the definition**
Industry	7000 VERSORGER	7000 Utilities
Supersector	7500 Energieversorgung	7500 Utilities
Sector	7530 Elektrizität	7530 Electricity
Subsector	7537 Alternative Stromerzeugung	7537 Alternative Electricity

Figure 1: The four levels of the ICB classification with German and English labels.

In Figure 1, it can be observed that one and the same word of the source language (English) has been translated with different words in the target language, depending on its level in the taxonomy in which it occurs: "Electricity" is translated in German as being either "Elektrizität" (index 7530) or "Stromerzeugung" (index 7537) and "Utilities" as either "VERSORGER" (index 7000) or "Energieversorgung" (index 7500). We have no information about the reasons behind the existence of those different translations: if they are motivated by style considerations or by the position of the labels in the hierarchical structure remains unclear.

In Figure 2, two different types of statements can be noticed in the definitions. The first sentence, which could be considered an intensional description of the defined term/label, and the second sentence, which can be considered an extensional description, listing mainly subterms. It will be important for an automated extraction of (multilingual) terminologies and lexical elements to be able to distinguish those types of statements.

	7537 Alternative Electricity (EN) Alternative Stromerzeugung (DE)
English Definition	Companies generating and distributing electricity from a renewable source. Includes companies that produce solar, water, wind and geothermal electricity.
German Definition	Firmen, die Strom aus erneuerbaren Quellen erzeugen und vertreiben. Einschließlich Firmen, die Solar-, Wasser- und Windenergie sowie geothermische Energie erzeugen.

Figure 2: Definitions associated to the ICB leaf category/label with index 7537, in German and English.

Figures 3 and 4 display the similar structure of GICS, whereas the reader can observe the differences in labelling the classes and sub-classes on the one hand and the difference in the length of the provided definitions on the other.

In Figure 3, it can be seen that the designers of the taxonomy are using an indexing strategy that differs from the one for ICB, although both rely on numbers for this. It can also be noticed that in this concrete example, GICS uses the same German words to translate "Utilities" in the two categories in which the term occurs. The same remark applies to the two translations of the English noun "Electricity".

	German Categories introducing the definition	**English Categories introducing the definition**
Sector	55 Versorgungsbetriebe	55 Utilities
Industry Group	5510 Versorgungsbetriebe	5510 Utilities
Industry	551050 Unabhängige Energie- und Erneuerbare Elektrizitätshersteller	551050 Independent Power and Renewable Electricity Producers
Sub-Industry	55105020 Erneuerbare Elektrizität	55105020 Renewable Electricity

Figure 3: The four levels of the GICS classification with German and English labels.

In Figure 4, we can observe that the provided textual definitions are significantly longer than in the ICB case. Additionally, the provided definitions do not only offer a combination of intensional and extensional statements, but they also precise which terms should be excluded from the definition of the leaf category/label. This is why a specific algorithm should be applied to those definitions in order to be able to automatically extract terminologies and lexical items.

Those differences are motivating our proposal for a shared formal representation of the two multilingual classification schemes, so that their components can be more easily aligned and form the base for a more accurate lexical extraction. The vocabulary we selected for this modelling is SKOS-XL, which is introduced in section 3.

[6] http://www.icbenchmark.com. ICB covers 13 languages: Chinese, Danish, Estonian, French, Finnish, German, Icelandic, Italian, Japanese, Latvian, Lithuanian, Spanish and Swedish. For each language a different Excel file is available. English is the original language.

[7] https://www.msci.com/gics. GICS covers 10 languages: French, German, Italian, Japanese, Korean, Portuguese, Russian, Simplified Chinese, Traditional Chinese and Spanish. For each language a different Excel file is available. English is the original language.

	55105020 Renewable Electricity (EN) / Erneuerbare Elektrizität (DE)
English Definition	Companies that engage in the generation and distribution of electricity using renewable sources, including, but not limited to, companies that produce electricity using biomass, geothermal energy, solar energy, hydropower, and wind power. Excludes companies manufacturing capital equipment used to generate electricity using renewable sources, such as manufacturers of solar power systems, installers of photovoltaic cells, and companies involved in the provision of technology, components, and services mainly to this market.
German Definition	Unternehmen, die in der Herstellung und Verteilung von Elektrizität unter Verwendung von erneuerbaren Energien tätig sind. Eingeschlossen, aber nicht beschränkt auf, sind Unternehmen, die Elektrizität und/oder elektrische Leistung durch Energiequellen wie Biomasse, Biogase, Sonnen-, Wasser- und Windkraft herstellen. Ausgeschlossen sind Unternehmen, die in der Herstellung von Ausrüstungsgütern für Stromherstellung unter Verwendung erneuerbarer Energien tätig sind, wie Hersteller von Sonnenkraftanlagen, Installateure von photovoltaischen Zellen und Unternehmen, die ihre Technologie, Komponenten, und Dienste hauptsächlich diesem Markt anbieten.

Figure 4: Definitions associated to the GICS leaf category/label with index 55105020.

3. The SKOS-XL Modelling

We are quoting two sources for describing SKOS: "The Simple Knowledge Organization System (SKOS) is a common data model for sharing and linking knowledge organization systems via the Semantic Web."[8] and "SKOS is an RDF vocabulary for describing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, 'folksonomies', other types of controlled vocabulary, and also concept schemes embedded in glossaries and terminologies"[9]. A number of taxonomies, classification schemes and terminologies have successfully been ported to SKOS[10].

As the quoted texts are pointing out, SKOS is a RDF-based vocabulary, and it is making use of RDF(S)[11] "annotation properties" like `rdfs:label` or `rdfs:comment`. Those annotation properties have been introduced in the RDF(S) vocabulary in order to equip OWL[12] ontological elements, like classes, properties or instances with additional meta-data and also human readable descriptions of the modelled knowledge objects. SKOS introduces three additional annotation properties that can be considered as a specialisation of `rdfs:label` for addressing terminological purposes: `skos:prefLabel`, `skos:altLabel` and `skos:hiddenLabel`. The values of such annotation properties are literals, and have as such no ontological status and can thus not be designated by a URI and consequently can not be used as a subject in RDF triples or as an object in an `owl:ObjectProperty`. SKOS alone would thus not allow us to formally state relations between the terms represented by the labels in the two classification schemes we are dealing with. Fortunately, the W3C community has proposed a remedy to this situation, and defined a corresponding recommendation called SKOS-XL.

SKOS-XL stands for "Simple Knowledge Organization System eXtension for Labels", providing additional support for describing and linking label elements of knowledge systems[13]. SKOS-XL is thus in a sense elevating the values of the `skos:prefLabel`, `skos:altLabel` and `skos:hiddenLabel` properties to the same level as concepts defined in the knowledge sources, supporting thus the cross-linking of labels or their linking to other formal objects. This is exactly the point that makes this formal representation language interesting for our purpose: In SKOS-XL concepts and labels that describe them are the same type of object/entities to which an URI can be associated. Relations between SKOS-XL labels can thus be explicitly and formally defined. A `skos:Concept` can relate to a `skosxl:Label` object via a `skosxl:prefLabel`, a `skosxl:altLabel` or a `skosxl:hiddenLabel` property and users can define all types of relations between `skosxl:Label` objects. This way we can state explicit relations between labels within one classification scheme but also between two or more classification schemes, as can be seen in the following section.

3.1. SKOS and SKOS-XL Encoding of the Components of the GICS and ICB Schemes

In the following two sections we give examples of the encoding of both the taxonomic concepts and the labels of the original classification schemes[14].

3.1.1. The Encoding of the Concepts

We define in one conceptual space "Industry_Classification" a `skos:ConceptScheme` for each taxonomy, as exemplified below in the TTL[15] SKOS-XL code listing 1 for the GICS scheme.

SKOS-XL_Encoding 1: skos:ConceptScheme for GICS

```
gics:ConceptScheme_GICS
  rdf:type skos:ConceptScheme ;
  rdfs:comment "GICS stands for \"Global
      Industry Classification Standard\".
      This structure is effective after
      close of business (US, EST) Wednesday
      - August 31, 2016"@en ;
  rdfs:label "GICS"@en ;
  skos:hasTopConcept gics:Concept_10 ;
  skos:hasTopConcept gics:Concept_55 ;
  .....
.
```

All 11 top-level concepts of GICS and all 10 top-level concepts of ICB are encoded as `skos:Concept` being in a `skos:topConceptOf` relation to their corresponding `skos:ConceptScheme` and in a `skos:narrower` to the concepts placed lower in the original taxonomy, as shown in code listing 2 for the GICS class 55 ("Utilities"), where the reader can see how we introduce the SKOS-XL

[8] Quoted from `https://www.w3.org/TR/skos-reference/skos-xl.html`.

[9] Quoted from `https://www.w3.org/2009/08/skos-reference/skos.rdf`.

[10] See for example `https://www.w3.org/2001/sw/wiki/SKOS/Datasets`.

[11] RDF stands for "Resource Description Framework" and RDF Schema is adding a data model for for the basic RDF vocabulary. See also `https://www.w3.org/TR/rdf-schema/`.

[12] OWL stands for "Web Ontology Language", a Semantic Web representation language for modelling knowledge. See also `https://www.w3.org/OWL/`.

[13] See also `http://lov.okfn.org/dataset/lov/vocabs/skosxl`.

[14] Due to limitation of space, we focus here on GICS, but the encodings are the same for ICB.

[15] TTL stand for "Terse RDF Triple" or more commonly "Turtle", a syntax for serializing RDF triples. See also `https://www.w3.org/TR/turtle/`.

property for linking to two skosxl:Label elements[16], which are exemplified further down, in the SKOS-XL codes 5 (for English) and 6 (for German).

SKOS-XL_Encoding 2: skos:topConceptOf

```
gics:Concept_55
  rdf:type skos:Concept ;
  rdfs:comment "Id of a top-level concept of
       GICS"@en ;
  rdfs:comment "This concept is in the
      domain of \"Sector\""@en ;
  rdfs:label "Utilities"@en ;
  skos:narrower gics:Concept_5510 ;
  skos:topConceptOf gics:ConceptScheme_GICS
      ;
  skosxl:prefLabel gics:Label_55_de ;
  skosxl:prefLabel gics:Label_55_en ;
.
```

All other concepts are encoded as a skos:Concept being in a skos:inScheme relation to the corresponding skos:ConceptScheme, and organized in a skos:broader relation to the concept immediately higher in the original taxonomy. The SKOS-XL code listing 3 gives as an example the GICS ID 5510 ("Utilities").

SKOS-XL_Encoding 3: skos:inScheme concept

```
gics:Concept_5510
  rdf:type skos:Concept ;
  rdfs:comment "Id of a top-level concept of
       GICS"@en ;
  rdfs:comment "This concept is in the
      domain of \"Industry Group\""@en ;
  rdfs:label "Utilities"@en ;
  skos:broader gics:Concept_55 ;
  skos:inScheme gics:ConceptScheme_GICS ;
  skos:narrower gics:Concept_551050 ;
  skosxl:prefLabel gics:Label_5510_de ;
  skosxl:prefLabel gics:Label_5510_en ;
.
```

Finally, The SKOS-XL code in listing 4 is displaying the SKOS encoding for a leaf catagory/label of GICS (class 55105020), including (partially) the definition.

SKOS-XL_Encoding 4: GICS leaf category

```
gics:Concept_55105020
  rdf:type skos:Concept ;
  rdfs:comment "Id of a top-level concept of
       GICS"@en ;
  rdfs:comment "This concept is in the
      domain of \"Sub-Industry\""@en ;
  rdfs:label "Renewable Electricity"@en ;
  skos:broader gics:Concept_551050 ;
  skos:definition "Companies that engage in
      the generation and distribution of
      electricity using renewable sources,
      including, but not limited to, ... ."
      @en ;
  skos:inScheme gics:ConceptScheme_GICS ;
  skosxl:prefLabel gics:Label_55105020_de ;
  skosxl:prefLabel gics:Label_55105020_en ;
.
```

[16]With maximally one skosxl:prefLabel per language. We do not display here all the languages listed in GICS.

3.1.2. The Encoding of the Labels

As already stated, we propose a SKOS-XL encoding for the labels of the original taxonomies, in order to be able to formally express relations between those, within one classification system or between both taxonomies. SKOS-XL code listings 5 and 6 show the basic information associated with the English and German labels of the top level concepts. Those labels are now encoded as instances of an owl:Class and no longer as simple literals as this was the case in SKOS. The German label is marked as being a translation of the English label. In both cases we indicate with lex:identical that the used label is identical to the label of the immediately lower category.[17]

SKOS-XL_Encoding 5: skosxl:Label of a GICS top level concept

```
gics:Label_55_en
  rdf:type skosxl:Label ;
  lex:identical gics:Label_5510_en ;
  rdfs:comment "Labels of a GICS concept"@en
      ;
  rdfs:label "Utilities"@en ;
  skosxl:literalForm "Utilities"@en ;
.
```

SKOS-XL_Encoding 6: German label as a translation of GICS 55 label

```
gics:Label_55_de
  rdf:type skosxl:Label ;
  lex:identical gics:Label_5510_de ;
  lex:isTranslationOf gics:Label_55_en ;
  rdfs:comment "Labels of a GICS concept"@en
      ;
  rdfs:label "Versorgungsbetriebe"@de ;
  skosxl:literalForm "Versorgungsbetriebe"
      @de ;
.
```

The next two TTL code listings are displaying the encodings for the last two levels of the original hierarchy. For the code listing 8 we do not reproduce the full definition, which can be found in Figure 4.

SKOS-XL_Encoding 7: An intermediate German Label

```
gics:Label_551050_de
  rdf:type skosxl:Label ;
  lex:isTranslationOf gics:Label_551050_en ;
  lex:lessSpecific gics:Label_5510_de ;
  lex:moreSpecific gics:Label_55105020_de ;
  rdfs:comment "Labels of a GICS concept"@en
      ;
  rdfs:label "Unabhaengige Energie\- und
      Erneuerbare Elektrizitaetshersteller"
      @de ;
  skosxl:literalForm "Unabhaengige Energie-
      und Erneuerbare
      Elektrizitaetshersteller"@de ;
.
```

[17]In the SKOS-XL encodings of the labels we keep the rdfs:label property, as it is very often queried by Semantic Web applications. So that we have a kind of redundancy to the skosxl:literalForm property.

SKOS-XL_Encoding 8: A German leaf category label

```
gics:Label_55105020_de
  rdf:type skosxl:Label ;
  lex:lessSpecific gics:Label_551050_en ;
  rdfs:comment "Labels of a GICS concept"@en
      ;
  rdfs:label "Erneuerbare Elektrizitaet"@de
      ;
  skos:definition "Unternehmen, die in der
      Herstellung und Verteilung von
      Elektrizitaet unter Verwendung von
      erneuerbaren Energien taetig sind.
      Eingeschlossen, aber nicht ... ."@de ;
  skosxl:literalForm "Erneuerbare
      Elektrizitaet"@de ;
.
```

Once this mapping from the two original classification schemes into a unified SKOS-XL representation has been solved, we started to investigate how the lexical elements contained in the labels can be described in a comparable formalism. We selected for this OntoLex-Lemon (McCrae et al., 2017)[18], as this model already includes a link between lexical items encoded in a standardized RDF vocabulary and the SKOS vocabulary (see Figure 5). In this case, we just need to consider SKOS-XL instead of SKOS for modelling the relation between the conceptual world and the lexicon modelling proposed by OntoLex-Lemon. As examples for this modelling we take instances of the ICB classification scheme (see Figure 1). We display first the SKOS-XL encoding for the German term "Stromerzeugung", in code listing 9, and then in code listing 10 the corresponding ICB concept.

SKOS-XL_Encoding 9: The German label for ICB 7537

```
icb:Label_7537_de
  rdf:type skosxl:Label ;
  lex:isTranslationOf icb:Label_7537_en ;
  lex:lessSpecific icb:Label_7530_de ;
  rdfs:comment "Labels of a ICB concept"@en
      ;
  rdfs:label "Alternative Stromerzeugung"@de
      ;
  skos:definition """Firmen, die Strom aus
      erneuerbaren Quellen erzeugen und
vertreiben. Einschliesslich Firmen, die
    Solar-, Wasser- und
Windenergie sowie geothermische Energie
    erzeugen."""@de ;
  skos:related gics:Label_55105020_de ;
  skosxl:literalForm "Alternative
      Stromerzeugung"@de ;
.
```

In code listing 9 we give also an example on how we can now link two labels across distinct taxonomies, using for this the `skos:related` property, but any more specific property can be used. The reader can clearly see here the advantage of "elevating" labels to an ontological entity status.

[18]See also the corresponding W3C Community Report `https://www.w3.org/2016/05/ontolex/`.

SKOS-XL_Encoding 10: The encoding for the concept ICB 7537

```
icb:Concept_7537
  rdf:type skos:Concept ;
  rdfs:comment "Id of a concept of ICB"@en ;
  rdfs:comment "This concept is in the
      domain of \"Subsector\""@en ;
  rdfs:label "Alternative Electricity"@en ;
  skos:broader icb:Concept_7530 ;
  skos:definition "Companies generating and
      distributing electricity from a
      renewable source. Includes companies
      that produce solar, water, wind and
      geothermal electricity.{@en@" ;
  skos:inScheme icb:ConceptScheme_ICB ;
  skosxl:prefLabel icb:Label_7537_de ;
  skosxl:prefLabel icb:Label_7537_en ;
.
```

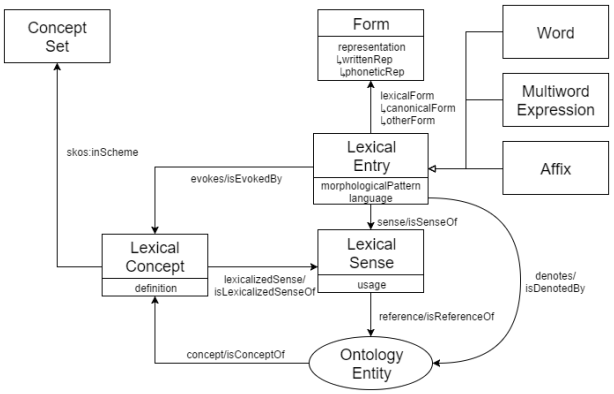

Figure 5: The core module of OntoLex-Lemon: Ontology Lexicon Interface. Graphic taken from `https://www.w3.org/2016/05/ontolex/`.

The following section describes our first steps in the OntoLex-Lemon modelling of lexical data included in ICB/GICS labels.

4. Modelling of the lexical Data

The core module of OntoLex-Lemon is displayed in Figure 5. There the reader can see how the lexical data is related to conceptual data encoded in SKOS. For our example concerning the German term "Stromerzeugung" (ICB ID 5737, see Figure 1), we just need to adapt the OntoLex-Lemon model, and integrate the SKOS-XL vocabulary instead of SKOS. This reflects also our view that terminological data should not be modelled as a lexical data, but rather within a representation framework conceived for terminologies, as this is the case for SKOS-XL. Code listing 11 is showing the suggestion for encoding the lexical item "Stromerzeugung" in OntoLex-Lemon.

SKOS-XL_Encoding 11: OntoLex-Lemon entry for "Stromerzeugung"

```
ind_class_lemon:lex_7537_2
  rdf:type ontolex:MultiWordExpression ;
  lexinfo:termElement <http://tutorial-
      topbraid.com/ind_class#Label_7537de> ;
```

```
  rdf:_1 ind_class_lemon:Component_1 ;
  rdf:_2 ind_class_lemon:Component_2 ;
  rdfs:label "Stromerzeugung"@de ;
  <http://www.w3.org/ns/lemon/decomp#
      constituent> ind_class_lemon:
      Component_1 ;
  <http://www.w3.org/ns/lemon/decomp#
      constituent> ind_class_lemon:
      Component_2 ;
  ontolex:denotes <http://de.dbpedia.org/
      page/Stromerzeugung> ;
  ontolex:evokes icb:Concept_7537 ;
.
```

The link between the lexical description and the terminology is established by using the property `ontolex:evokes`. As OntoLex-Lemon is based on the idea that lexical entries are getting their sense(s) by linking them to elements of ontologies, we are linking the entry to a DBpedia page, using the property `ontolex:denotes`. A nice feature is the fact that we can link the word "Stromerzeugung" to the SKOS-XL label 7537, stating that it is a part of it (`lexInfo:termElement`), and not only to the SKOS-XL concept 7537. We take advantage of the fact, that OntoLex-lemon is also supporting the modelling of compound words. We can decompose the word and link to its components ("Strom" and "erzeugung"). And from there to link the components to the related lexical entries "Strom" and "Erzeugung". Importantly, we can also link the component "Strom" directly to the right sense ("Electricity" versus "River")[19]. In doing this, we are fulfilling lexicographic requirements in the context of terminology and OntoLex-Lemon proved to be a very satisfactory modelling framework.

5. Linguistic Linked Data Cloud

The Linguistic Linked Open Data (LLOD) cloud[20] is an initiative to break the data silos of linguistic data and thus encourage NLP applications that can use data from multiple languages, modalities (e.g., lexicon, corpora, etc.) and develop novel algorithms. Looking at the current state of the LLOD, one can see that the data sets published in this cloud are classified along the lines of six categories:

- Corpora
- Terminologies, Thesauri and Knowledge Bases
- Lexicons and Dictionaries
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

Not all the date sets are equally linked to each other, and our approach can contribute in better linking the data sets in the fields of Terminologies, Thesauri and Knowledge Bases and those in the fields of Lexicons and Dictionaries.

[19]This approach is based on (Declerck and Lendvai, 2016)

[20]See `http://linguistic-lod.org/llod-cloud`.

6. Conclusion

We have implemented the integration of two different but closely related formal representation languages, SKOS-XL and OntoLex-Lemon, for encoding terminological and lexical data that are used in classification schemes as interrelated knowledge objects. This makes those data accessible in the Linked Open Data and also in the Linguistic Linked Open Data cloud[21]. This formalisation seemed to be a necessary pre-requisite for our original task, which consists in extracting multilingual domain-specific dictionaries from such classification systems. The next step in our work will consist in implementing the extraction algorithms based on the formal representation of the terms and the language data used in those terms.

7. Acknowledgements

The DFKI contribution to this work has been partially funded by the BMBF project "DeepLee - Tiefes Lernen für End-to-End-Anwendungen in der Sprachtechnologie" with number 01-W17001. The ACDH contribution is supported in part by the H2020 project "ELEXIS" with Grant Agreement number 731015. We would like to thank the participants of the ENeL WG3 meeting held in Budapest in February 2017 for comments on the very first steps of our investigation.

8. Bibliographical References

Chiarcos, C., Hellmann, S., and Nordhoff, S., (2012). *Linking Linguistic Resources: Examples from the Open Linguistics Working Group*, pages 201–216. Springer Berlin Heidelberg, Berlin, Heidelberg.

Declerck, T. and Lendvai, P. (2016). Towards a formal representation of components of german compounds. In Micha Elsner et al., editors, *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Humboldt University, ACL.

Gromann, D. and Declerck, T., (2014). *A Cross-Lingual Correcting and Completive Method for Multilingual Ontology Labels*, pages 227–242. Springer.

McCrae, J. P., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In *Proceedings of eLex 2017*.

[21]`http://linguistic-lod.org/llod-cloud`.