Terminology Harmonization in Industry Classification Standards

Dagmar Gromann[1] and Thierry Declerck[2]

[1] Vienna University of Economics and Business,
Nordbergstrasse 15, 1090 Vienna, Austria
dgromann@wu.ac.at

[2] DFKI GmbH, Language Technology Department,
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
declerck@dfki.de

Abstract. Terminology as a research area has shifted from portraying terms as lexical units to a concept-oriented approach. Accordingly, the process of terminology harmonization has to cope with the concept orientation of term entries. One approach to harmonization is the integration of several terminologies into one centralized terminology repository, which is either formalized as a conceptual system or points to such systems. In contrast, we propose an approach adopting the linked data strategy by linking resources that preserve the initial terminologies with the corresponding lexical items and the related ontology concepts. As ontologies traditionally link concepts but not the natural language designation of concepts, we propose a model that utilizes terminologies for terminological and ontology lexicons for morpho-syntactic information. We illustrate our suggested approach, applying it to closely related but competing industry classification standards.

Keywords: Terminology, lexicon, ontology, harmonization, industry classification

1 Introduction

Industry classification standards allow for a thorough analysis of the industrial landscape. Investors and asset managers rely on the transparency these standards offer by means of global comparisons by industry. But despite of very similar categories, (competing) systems of industry classification often employ different terminology. Harmonization of these systems experiences issues not only on the terminological level, also on the hierarchical level various degrees of granularity can be observed. For instance, the Industry Classification Benchmark (ICB)[3] defines and refers to *Banks*, whereas the Global Industry Classification Standard (GICS)[4] differentiates between *Diversified Banks*, *Regional Banks*, and *Thrifts & Mortgage Finance*. A strategy for harmonization could consist in subsuming

[3] http://www.icbenchmark.com/

[4] http://www.standardandpoors.com/indices/gics/en/us

these categories under one concept or modifying the existing classifications in order to make them interoperable.

Alternatively, our approach suggests a strategy based on the linked data [10] framework in that harmonization is achieved by interlinking terminologies, including their associated lexicons and related ontology concepts. Connecting these resources by means of formal languages, such as the Resource Description Framework (RDF)[5] and the Simple Knowledge Organization System (SKOS)[6], enables the preservation of the original classification ID for all terms and their variants, as well as the concepts they are associated with.

At the end of the day, nothing can be said against still opting for a new, centralized and unique terminology in case the linking mechanisms reveal consistent overall similarities and/or suggest the possibility of an integrative re-organization of the various knowledge sources.

2 Research Background

Term banks initially portrayed terms as lexical units [8], overloading the term with different meanings. Gradually, a concept-oriented approach developed, emphasizing the relationship of one concept per term entry [3]. Recent developments view terminological resources as expert systems, focusing on a knowledge-oriented approach [8]. For instance, César et al. [12] harmonize a wide variety of standards regarding the improvement of software processes with a focus on terminology. Ontologies are applied to the task of eliminating inconsistencies on a semantic and conceptual level, implicitly harmonizing the terminology [12].

The TermSciences initiative [17] establishes semantic relations among medical terminologies, by means of TMF-compliant metadata. Ontologies or high-level terminologies serve the unification process of different resources. Nevertheless, the project centers around merging, grouping, restructuring resources, converting term-centered representations to concept-oriented ones. Our proposal focuses on the benefit of different conceptualizations, i.e. ontological, terminological, lexical, to the process of harmonization with a very clear emphasis on terminology rather than controlled vocabularies and a preservation of its integrity and origination.

Several models exist to account for the terminological dimension of ontologies such as ontoterminology [16], termontography [14], or the Terminae method [15]. Whereas the latter two focus on the establishment of one terminology for or in combination with an ontology, the former emphasizes the differences. Roche et al. [16] highlight the importance of separating the linguistic and the conceptual dimension of terminology and ontology, as terms cannot simply be reduced to the textual content of `rdfs:label` or `rdfs:comment` annotation properties without any linguistic layer.

The model for the integration of conceptual, terminological and linguistic objects in ontologies (CTL) [1] uses the Terminae method [15] and the *LexInfo* metamodel [4] to obtain a modular and multi-layered linguistic annotation of

[5] `http://www.w3.org/RDF/`

[6] `http://www.w3.org/2004/02/skos/`

ontology labels, further detailed in [2]. Expanding on the CTL model [1] and formalizing the approach, we focus on separating the lexical, syntactic, terminological and (domain) semantic levels into adequate resources, linking them with RDF and SKOS. Lexical and syntactic descriptions will be provided using *lemon*, a Lexicon Model for Ontologies [11]. The *lemon* model offers a formal representation of linguistic information to be associated with the word forms contained in the `rdfs:label` annotation property of ontology classes, and with a clear referential mechanisms to ontology classes, thus defining the semantic of such linguistic expressions by their references to concepts.

3 Industry Classification Systems

Industry classification systems aim at providing a comparison of companies across nations. Due to numerous and often competing classification systems, the resulting overlapping and inconsistent terminologies require harmonization on a conceptual and term level, including the harmonization of the linguistic properties of the tokens building the term. In the following, we suggest a linking approach for harmonizing two major industry classification systems.

The Global Industry Classification Standard (GICS) represents a taxonomy of industry sectors developed by MSCI and Standard & Poor's[7]. The GICS structure consists of 10 sectors, 24 industry groups, 68 industries and 154 subindustries into which all major companies have been categorized. The ten main industries are: Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Healthcare, Financials, Information Technology, Telecommunication Services, Utilities.

Similar to the GICS, the Industry Classification Benchmark (ICB) developed by Dow Jones and FTSE[8] consists of four major levels. The system is organized into 10 industries, 20 supersectors, 41 sectors and 114 subsectors. The ten main industries are: Oil & Gas, Basic Materials, Industrials, Consumer Goods, Healthcare, Consumer Services, Telecommunications, Utilities, Financials and Technology.

In comparison, both systems classify a company according to its principal business, apply four major levels to their structure and have a comparable number of subcategories. In both cases the categories are organised in a hierarchical tree. Intermediate nodes are labelled with short natural language strings and the leaf nodes are equipped with (partly lengthy) definitions. Both systems are delivered in several languages

One major difference is to be found in the consumers section. GICS differentiates between staples and discretionary containing both goods and services, whereas ICB distinguishes consumer goods from consumer services. As this regards the top-level classification, it is an important aspect to be considered in

[7] See respectively `http://www.msci.com/products/indices/sector/gics/` and `http://www.standardandpoors.com/indices/gics/en/us`

[8] See `http://www.ftse.com/Indices/Industry\_Classification\_Benchmark/index.jsp`

the harmonization strategy. Naturally, the terms used to designate equivalent categories differ substantially.

4 Three-layered Model for Harmonization

Conceptual structures in an ontology differ from those in terminologies. The ontology links on the basis of domain knowledge, whereas the terminology links on a linguistic and language-related background. The combination of both types of information seems to be beneficial to the process of harmonization. Our model illustrated illustrated in Fig. 1 utilizes terminologies – complying with the Terminological Markup Framework (TMF) [7] – in combination with ontologies to create a net of labels interlinked with SKOS and RDF(S). In order to clearly distinguish between terminological and morpho-syntactic information, we additionally include a lexicon level to be represented using *lemon*.

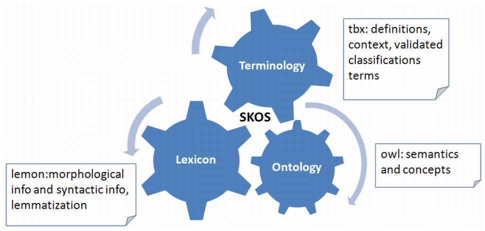
Terminology
tbx: definitions, context, validated classifications terms
SKOS
Lexicon
Ontology
owl: semantics and concepts
lemon:morphological info and syntactic info, lemmatization

Fig. 1. Three-layered model for harmonization

Each component of the system represents different aspects of the net of labels. Firstly, the lexicon mainly provides information on basic lexical, morphological and syntactic information. Secondly, the terminology (in TMF) as such represents the validated terms and "soft" variants [9] such as synonyms, acronyms and orthographic variants. Finally, ontologies provide the (domain) semantic layer. The suggested layered model allows thus to state whether a term varies morphologically or semantically. The resulting net of labels contains the original classification ID of each term, whether it is a preferred term or normalized form, etc., rich linguistic information and a thorough conceptual basis provided by the ontology.

In detail, within the process of creating the terminologies we apply general principles such as concept orientation [3], term consistency, etc. to validate the classifications' terminology. The harmonization strategy is a two-fold contrastive approach considering the conceptual level of terminology and its designations. Term harmonization either refers to the designation of one concept by terms or the establishment of equivalences across languages or term variations in one language [5].

5 Harmonizing Industry Classification Systems

Subsequent to obtaining the multilingual taxonomies from the respective web presences of the industry classifications, we utilized the source data to create terminologies and ontologies, lexicalizing the latter. This entire process abides to the current ISO standards for terminology (ISO1087, ISO704) and harmonization [5], proposing an extension of the latter.

5.1 From Source Data to Terminology

Based on the resources provided by ICB and GICS we created one TermBase eXchange (TBX) [6] format term base for each classification, which allows for a semi-formal representation of the multilingual terminology and for a validation of the classifications' terminology. The initial analysis of the input data necessitated the harmonization of terms on several levels. At times designations provided pleonastic information as illustrated in the following example:

```
<termEntry id="ICB1779">
  <descrip type="subjectField">mining</descrip>
  <descrip type="definition">ICB sector</descrip>
  <langSet xml:lang="en">
       <descrip type="definition">Companies producing and exploring platinum,
         silver and other precious metals not defined elsewhere.</descrip>
       <tig>
              <term>Precious Metals</term>
              <termNote type="partOfSpeech">noun</termNote>
       </tig>
  </langSet>
</termEntry>
```

[Simplified TermBase eXchange (TBX) example of the ICB terminology.]

As the definition clearly classifies platinum as precious metal, it represents a case of pleonasm. Thus, the entry was adapted to ”Precious Metals” in the term base. Similarly, the use of homonymous designations for different categories on the same hierarchical level has to be avoided in the terminologies, such as the ICB classification containing two sibling sectors both defining mining.

Concept orientation refers to the fact that each term entry contains the full terminological data for the respective concept [3]. GICS designates a mining category ”Steel,” but the definition clearly states that it classifies ”Producers of iron and steel and related products” - only referring to steel could infringe the integrity of this terminological entry. Additionally, term consistency is often an issue in combination with concept orientation. In contrast to its sibling, the ICB subsector ”Exploration & Production” does not refer to its supersector *Oil & Gas Producers* in its designation. On the basis of the definition provided it can be adapted to ”Oil & Gas Exploration & Production” in order to improve both consistent terminology and concept orientation.

The presented methodology clearly employs a bottom-up approach, analyzing the leaf nodes first. This initial analysis represents a prerequisite step for the actual harmonization on a terminological and conceptual level.

5.2 Harmonization Steps

The process of concept harmonization usually precedes the process of term harmonization [5]. In case the concepts are equivalent, a correspondence between them can be established. For instance, the definition "Residential Retail Estate Investment Trusts (REITs)" can be aligned directly in both classifications by `skos:exactMatch` as they are orthographically and semantically identical. However, most cases are more complicated.

Lexical information are represented by the *lemon* model. Although the representation of term variation is not the primary objective of the lexicon-ontology model, it is generally possible [13]. *lemon* creates sense objects that refer to one ontology concept (semantic by reference). The whole *lemon* entry is used to refer to a concept, not the canonical or the alternative form of the term. But one would like to be able to state that a term used in a category of a classification system is an alternate form of a term that is used in a category of another classification, while the two categories can be related by an equivalence relation. In *lemon* two lexical entries have to be created for this purpose.

TMF neither provides a solution to this problem of including two terms in one term entry while preserving the original source by means of the reference ID of both terms as they are used in their respective classification system. TBX allows for the inclusion of synonyms in an entry and also variants, but each entry has one ID. As with lemon, two entries are needed to establish the equivalence or relation between the terms by means of a cross-reference.

The objective is to obtain two equivalent and equal terms referring to their original ID and to establish the harmonization by means of relations. Thus, it is up to the user to decide to which term entry the information extracted by the ontology-based system should be mapped. The harmonization is accomplished by means of relations utilizing SKOS and RDF(S), as illustrated below `mfo` meaning "Multilingual Financial Ontology".

```
tbx:ICB rdf:type skos:ConceptScheme.
mfo:ICB rdf:type skos:ConceptScheme.

lemon:full_line_insurance rdf:type skos:Concept;
   lemon:canonicalForm [lemon:writtenRep "Full line insurance"@en  ] ;
   lemon:reference <http://icb.org/ICB8532>  ;
   skos:inScheme mfo:ICB ;
   skos:inScheme tbx:ICB.

tbx:GICS rdf:type skos:ConceptScheme.
mfo:GICS rdf:type skos:ConceptScheme.

lemon:multi_line_insurance rdf:type skos:Concept;
   lemon:canonicalForm [lemon:writtenRep "Multi-line insurance"@en  ] ;
   lemon:reference <http://gics.org/GICS40301030> ;
   skos:inScheme mfo:GICS ;
   skos:inScheme tbx:GICS.

<http://icb.org/ICB8532> skos:closeMatch <http://gics.org/GICS40301030>.
```

[Linking the labels of a GICS and an ICB concept, by means of SKOS.]

The example shows how the ontology concept points to the terminology, which in turn is linked with the lexicon. The `closeMatch` indicates that the two concepts are sufficiently aligned to be used interchangeably. And so the associated labels (lemon entries that refer to the concepts) can be interlinked. We can not apply the skos matching mechanisms directly to the lemon entries, since we want to establish a semantic interoperability, and not a string-based one. The aspect of multilingualism represents an additional challenge, as terms in different languages might not be truly harmonized within one entry, even if it is less an issue with such a standardized representation of terms.

Finally, we created frequency lists for each classification and found that several phrases or words are only mentioned in the definition, but not in the designations of the classifications. Whereas the ICB definitions contain the term "company" 62 times, it is not to be found once in the designations of the classification. Similar statistics apply to manufacturer, producer, distributor to name but a few. Due to the predominance of company, we decided to add the term to the ontology and apply it to labels where no other business activity is predominant. In case of several types of business activity, consistency calls for the use of company again. However, the major basis for this decision is provided by the definition. One example of GICS is the subsector "Aluminum," which as such can clearly not be identified as ontologically valid or conceptually sound, as it does not provide any information on company. Thus, we decided to introduce a superordinate node for the concept *company.*

6 Conclusion

Confronted with a variety of competing schemes in the field of industry classification, we investigated the possibility to harmonize their respective terminology, also for the benefit of a multilingual information extraction task, which has to map textual data in the financial domain to concepts described in such classification systems. We opted for an approach that proposes a three-fold model, clearly separating lexical, (morpho-)syntactic, terminological, and (domain) semantic levels. Using SKOS and RDF(S), we designed intra-model relations by interlinking the lexicon entries, the terms, and concepts in and betweeen each resource. These links preserve the original source information and thus document the role of terminology within the process of harmonization. As an addtional result we see the emergence of a net of conceptual labels that can be organized independently from the ontological sources in which they were introduced.

Acknowledgements. The DFKI part of this work has been supported by the Monnet project (Multilingual ONtologies for NETworked knowledge), co-funded by the European Commission with Grant No. 248458.

References

1. Declerck, T., Lendvai P.: Towards a standardized linguistic annotation of the textual content of labels in knowledge representation systems. In: The seventh international conference on Language Resources and Evaluation. LREC-10, Malta (2010)
2. Declerck, T., Lendvai, P., Wunner, T.: Linguistic and Semantic Features of Textual Labels in Knowledge Representation Systems. In: Harry Bunt (ed.): Proccedings of the Sixth Joint ISO - ACL/SIGSEM Workshop on Interoperable Semantic Annotation, Oxford, United Kingdom, ACL-SIGSEM (2011)
3. Bassey, A., Budin, G., Picht, H. Rogers, M., Schmitz, K.D., Wright, S.E.: Shaping Translation: A View from Terminology Research. Translators' Journal 50:4 (2005)
4. Buitelaar, P., Cimiano, P. Haase, P., Sintek, M.: Towards linguistically grounded ontologies. In: Proceedings of the 6th European Semantic Web Conference, pp. 111-125, Springer Berlin/Heidelberg (2009)
5. ISO 860: Terminology work - Harmonization of concepts and designations (2005)
6. ISO 30042: Systems to manage terminology, knowledge, and content - TermBase eXchange (TBX) (2008)
7. ISO 16642: Computer applications in terminology - Terminological markup framework (2003)
8. Vasiljevs, A., Gornostay, T., Skadina, I.: From Terminology Database to Platform for Terminology Service. In: Proceedings of the CHAT 2011, Vol. 12, pp. 16-21, NEALT Proceedings Series (2011)
9. Delpech, E., Daille, B.: Dealing with lexicon acquired from comparable corpora: validation and exchange. In: Proceedings of the TKE 2010, pp. 211223, Dublin, Ireland (2010)
10. Bizer, C., Heath, T., Berners-Lee T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS). 5:3, 1-22 (2009)
11. McCrae, J., Spohr, D., Cimiano, P.: Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. The Semantic Web: Research and Applications. Volume 6643 of LNCS, pp. 245-259. Springer, Berlin (2011)
12. César, P., Pino, F.J., García, F., Piattini, M., Baldassarre, T.: An ontology for the harmonization of multiple standards and models. Computer Standards & Interfaces, 34: 1, pp. 48-59 (2012)
13. Montiel-Ponsoda, E., Aguado-de-Cea, G., McCrae, J.: Representing term variation in *lemon*. WS 2 Workshop Extended Abstracts, TIA 2011, pp. 47-50, Paris (2011)
14. Temmerman, R., Kerremans, K.: Termontography: Ontology Building and the Sociocognitive Approach to Terminology Description. CIL17, Prague (2003)
15. Aussenac-Gilles, N., Szulman, S., Despres, S.: The Terminae Method and Platform for Ontology Engineering from Texts. In: Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. IOS Press, pp. 199–223, IOS, Amsterdam (2008)
16. Roche, C., Calberg-Challot, M., Damas, L., Rouard, P.: Ontoterminology: A new paradigm for terminology. In: International Conference on Knowledge Engineering and Ontology Development, pp. 321-326, Portugal (2009)
17. Khayari M., Schneider, S., Kramer, I., Romary, L.: Unification of Multi-Lingual Scientific Terminological Resources Using the ISO 16642 Standard, The TermSciences Initiative. In: Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine (LREC 2006), Genoa (2006)