A Model and Method to Terminologize Existing Domain Ontologies

Dagmar Gromann

A Model and Method to Terminologize Existing Domain Ontologies

Dagmar Gromann

Vienna University of Economics and Business
Welthandelsplatz 1, 1020 Vienna, Austria
dgromann@wu.ac.at

Abstract. In today's shrinking world, the need to automatically process multilingual knowledge becomes increasingly pressing, particularly in specialized communication. Domain ontologies enable automated computing of structured knowledge, but feature little and mostly English natural language content. Terminological resources, on the other hand, provide rich multilingual data, but differ in their distribution format, data semantics, representation language, and approach to terminology science. This diversity makes it difficult to interchange their data and link them to ontologies. To address this issue, we propose a terminology interchange model that supports sharing terminologies and linking them to ontologies. The proposed model is a formalized ontology on terminology that represents the main elements of terminology science derived from ISO TC 37 standards and best practices. A methodology for applying the model to domain ontologies and merging the resulting terminological resource with available multilingual terminological data is proposed.

Keywords: Terminology, Domain Ontologies, Standards, Metamodeling, Data Interchange and Integration

1 Introduction

Ontologies are crucial to writing consistent and formalized definitions within a specialized domain, relying on formal semantics [7, 13]. Human users, natural language processing applications, semantic indexing, and information retrieval based on ontologies additionally require natural language content in the ontology [13, 5]. However, few resources provide rich multilingual natural language information and formal semantics. If available at all, natural language content is usually restricted to annotation properties, such as `rdfs:label`, and to the English language [5]. To address this problem, the proposed modular terminology interchange model (T-Mint) represents the main elements of concept-oriented terminologies as a formal ontology, a so-called 'terminology of terminology' [4]. It is designed to facilitate the data integration and interchange of terminological data and link them to domain ontologies.

In line with ISO TC 37 standards, the proposed terminology interchange model consists of a conceptual, language, and term level. To those three main

levels, administrative and descriptive information is added by re-using data categories from ISOcat[1]. To make the proposed model itself highly re-usable, it is represented as a modular ontology. Its three main modules are the core-structure module, representing the three main levels, a data category selection module, and a sub-term module. Thereby, connecting the core structure with any customized data category subset is facilitated. Instances of the proposed model provide dereferenceable terminological data - such as terms, definitions, contexts, etc. - that can be related, queried, and further annotated.

Established solutions to terminological modeling in ontologies frequently restrict natural language data to annotation properties, whereby they cannot be related, annotated, or used for reasoning. Section 2 specifies current practices and similar models at the terminology-ontology interface. To be truly useful, any such terminological metamodel needs to be based on requirements of the community. This is why Section 3 briefly summarizes the most important requirements taken from standards, best practices, and current modeling practices. The proposed model is detailed in Section 4. A methodology for applying it to a domain ontology and interchanging its data with other terminologies is suggested in Section 5 prior to some concluding remarks.

2 Background

RDFS's label properties link synonymous strings to an ontological entity. Because such annotation properties cannot be annotated or related, no information other than XML language tags can be added. The Simple Knowledge Organization System (SKOS) model [8] allows for a differentiation between 'preferred', 'alternative', and 'hidden' labels. With its SKOS-XL extension, terms are instances of OWL Full class and can thus be related to each other. However, modern terminological resources require a higher complexity of terminological data and modeling decisions than provided by these vocabularies. Furthermore, all references to 'terminology' were removed in the current version of the W3C SKOS recommendation for a reason [15]. In fact, SKOS is a representation format for controlled vocabularies targeted towards human users [8] and not machine-readability [14] or representing terms in use.

Descriptive and administrative properties are frequently added as metadata. Metadata vocabularies tend to define all properties in their vocabulary, even if this leads to duplication. For instance, depending on the vocabulary notes can be represented as `rdfs:comments`, `skos:note` [8], `omv:description` [10] to name a few. Without explicitly mapping one to the other their content is not detected as identical by automated programs. Tao et al. [14] tackle this problem by proposing terminology guidelines on vocabularies and not just their metadata. ISO TC 37 handled this problem by introducing data categories for metadata and data, which have been collected in the ISOcat repository.

The objective of the proposed terminological interchange model is to represent a large variety of terminological data in general and for an arbitrary ontol-

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

ogy. To meet this goal, the proposed model is based on standardization efforts within the terminology community, in particular the Terminological Markup Framework (TMF) (ISO 16642:2003) and TermBase Exchange Format (TBX) (ISO 30042:2008). Together TMF and TBX define the basis for a family of terminological markup languages, but have no mechanism for relating terminological data to ontologies.

TERMINAE [2] presents the most comprehensive and well-established approach for modeling terminological information in ontologies, but focuses mainly on learning informal ontologies from text and represent terms as annotation properties. Thus, a separate manipulation of terms and concepts in TERMINAE is not granted. Other terminological models for ontologies reduce natural language content to natural language definitions [12, 11].

3 Terminology Metamodel Requirements

A metamodel on the elements of terminology science needs to address the requirements of its intended users and community. Requirements to the terminology interchange model derive from a systematic analysis of the state of the art in terminological modeling. First, standardization efforts of the terminology community were analyzed, in particular TMF, TBX, and ISO 704:2009 on terminology work. Second, an extensive literature review focused on theoretical methods (e.g. [1, 4, 7, 15]), guidelines (e.g. [14, 9]), and existing analysis of terminological resources (e.g. [3]). Lessons learned and limitations of existing models (e.g. [11]) and available resources (e.g. SNOMED CT) helped gather principles and requirements for the proposed model. The major requirements identified from this systematic analysis are:

1. **Multi-Purpose and Multi-Domain Applicability:** While the representation format itself needs to be independent of any specific purpose, the model it represents should be applicable to various purposes, such as indexing, natural language processing, capturing findings, etc., and domains, such as finance, biomedicine, law, etc. This entails that the model needs to be highly re-usable for various settings.
2. **Concept Orientation:** The meaning of a concept is unique in its concept system. Terms related to the concept have at least one and not more than one meaning. Although subordinate concepts inherit characteristics from superordinate ones, differentia between parent and child need to exist.
   (a) *Concept Permanence:* Concepts and terms can be manipulated separately so that the concept system is maintained even if terms evolve.
   (b) *Unique Non-Semantic Identifier:* Reclassification of concepts or polyhierarchies make hierarchical numbering (e.g. 1010 is subclass of 10) difficult. Thus, the unique identifier needs to be void of any (implicit) meaning.
   (c) *Hierarchical Ordering:* Each concept has at least one parent. Top-concepts are children of a unique empty top concept that facilitates extending the resource. Polyhierarchies may only be added when qualifying the subsumption relation (IS A) with a subdivision criterion, e.g. IS_Aagent.

3. **Term Autonomy:** All terms (abbreviations, symbols, variants, etc.) can be documented with all, i.e., unlimited number of, necessary term-related details and data, including term use and context. This means that the representation format needs to provide for extensions at any time.
4. **Accessibility:** Terminological data need to be accessible to humans and machines. This means that definitions and descriptions of meanings need to be formalized as well as represented in natural language.
5. **Interoperability:** Any terminological metamodel needs to be available in a format that facilitates data interchange across applications and resources.

4 Terminology Interchange Model

The proposed terminology interchange model (T-Mint) illustrated in Fig. 1 represents a metamodel for terminology. One terminological data collection aggregates several terminological entries and can reference one domain ontology. The proposed model is represented as a modular ontology, each module being a logically consistent separate ontology with alignments across the modules. The proposed model consists of three major modules: the core-structure module (CSM), the data category selection module (DCSM), and the sub-term module (STM). Components of the core-structure module correspond to the Terminological Markup Framework (TMF) metamodel and are further described by data categories, which are derived from ISOcat and represented in the data category selection module. The data category selection module currently represents data categories as specified in the default subset of TBX with some adaptations based on technological change and best practices. Terms that consist of several words, so-called multi-component terms, can be separated into term components and sub-terms. Elements in in Fig. 1 marked as `<<auxiliary>>` are abstract and only included for demonstrative purposes, but have no corresponding class or relation in the actual model implementation.

Terminological Data Collection The terminological data collection is realized as a resource, i.e., an ontology file containing terminological data. Terminological entries in the collection are subclasses of a unique empty top concept `owl:thing`. Administrative details are added to the file directly by means of re-using pre-existing meta-data from repositories, such as the Dublin Core[2] elements `dc:creator` or `dc:title`.

Terminological Entry It represents the conceptual level of a terminology and is equivalent to `Meaning` of the semiotics.owl[3]. Each entity of the input ontology might be referenced by one terminological entry and represents the meaning of this entry. Thus, the object property `reference` - re-used from the ontolex lexicon model for ontologies[4] with a relation of the same name - is modeled as functional with the class `TerminologicalEntry` in its domain.

[2] `http://dublincore.org/`
[3] `http://www.ontologydesignpatterns.org/cp/owl/semiotics.owl`
[4] `http://www.w3.org/community/ontolex/wiki/Main_Page`

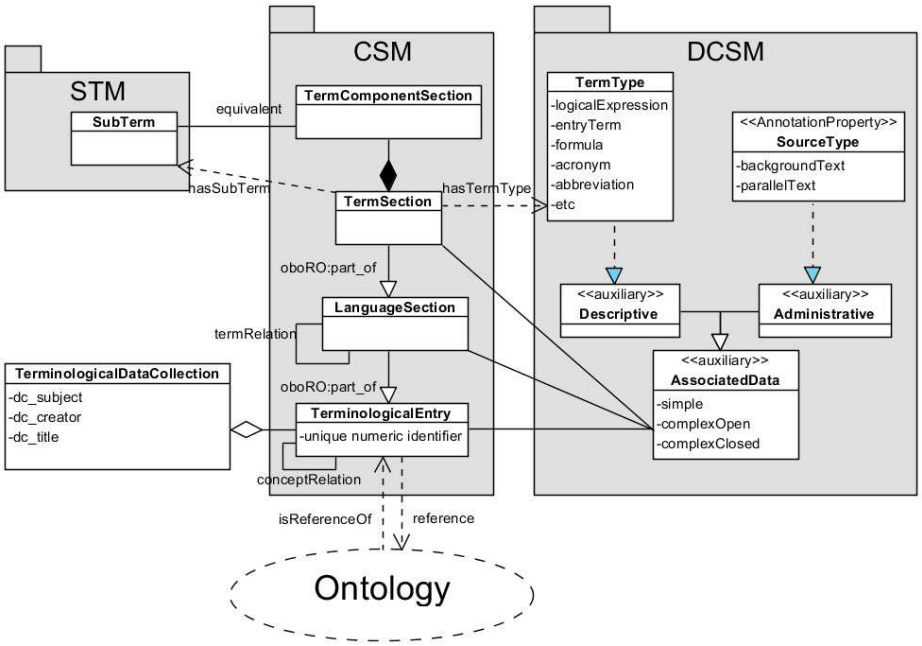

Fig. 1. The Terminology Interchange Model (T-MINT)

Each terminological entry can be related to other terminological entries by means of concept relations, such as `antonymConceptOf`.

Language Section Synonymous terms in one language are all grouped into one language section, which is a part of a terminological entry. The relation `oboRO:part_of` is re-used from the OBO Relation Ontology[5], which is transitive, reflexive, and anti-symmetric. Terminological entries can be monolingual with one or multilingual with multiple language sections.

Term Section Each term section consists of one term in one language and is a `oboRO:part_of` a language section. The class term section is instantiated by an actual term, e.g. 'securities' is the label of an `owl:Individual` of the class term section. Term sections can be related to other term sections by means of term relations, some of which are predefined. For instance, if two terms are spelling variants of each other, they are either related by `variant` or its subproperty `spellingVariant`.

Term Component Section This section represents individual components of multi-component terms. Thereby, descriptive and administrative details can also be added to a component of a term. Each component has to be linked to at least one term.

Descriptive Data Categories Simple data categories (DCs) are atomic and can neither be assigned a value nor contain other DCs, so they are modeled

[5] `http://obofoundry.org/ro/`

as individuals. Complex closed DCs provide picklists of predefined values in the form of simple DCs, such as `contextType` with a value 'associated context'. They are modeled as classes with an enumeration of their individuals by means of `ObjectOneOf`. Complex open DCs have a predefined datatype but open data value, so they are modeled as data type properties specifying the data type, which mostly are plain literals or integers. For instance, `geographicalUsage` is a complex open DC and has the datatype restriction plain literals.

Administrative Data Categories Administrative data are represented as annotation properties of the core-structure components the data category belongs to. For instance, `sourceType` can be added to the conceptual, language, and term level of the core-structure module. For this purpose, the proposed ontology re-uses existing annotation properties of various repositories, such as the Dublin Core[6] or the Ontology Metadata Vocabulary (OMV)[7]. Some administrative data can alternatively be modeled as relation, for instance, `conceptOrigin` can relate the concept in the terminological data collection to the concept in the original resource if the latter has a dereferenceable URI.

Sub-Terms Sub-terms are components of multi-component term which themselves are terms in the same domain as the term they are part of. For example, 'liquidity risk symptom' in the financial domain might be dissected to 'liquidity', 'risk', 'symptom' and combinations thereof. While 'liquidity' could still pertain to the financial domain, 'symptom' on its own is less likely to be included in a terminology on finance. So the former might be considered a sub-term because it is domain-specific. Sub-terms can be identical to term components, but term components do not need to be domain-specific. This idea of creating a repository of sub-terms has the objective of providing a full account of terms within a domain ontology.

The proposed terminology interchange model relates terminological data to domain ontologies and allows for a description of their semantics in more detail. Furthermore, it constitutes a terminological data collection which can be re-used for other purposes, such as natural language processing. Terminologies using T-Mint allow for terms and definitions to be addressed and annotated directly, related to each other, and be manipulated separate from terminological entries. It has to be noted here that T-Mint is not merely a syntactic conversion of the TMF meta-model and TBX-Default. Its representation in the Web Ontology Language (OWL) required the addition of formal semantics and several modeling decisions, which were based on best practices, existing models, and existing resources.

An increasing tendency to model terminologies and linguistic models based on description logic could be observed [3]. One reason for this might be that high-quality available DL reasoners can be used for automated consistency checking, e.g. to avoid terminological cycles. Furthermore, OWL makes the open world

[6] `http://dublincore.org/`
[7] `http://omv2.sourceforge.net/`

assumption, which means that knowledge not represented or inferred is simply unknown and not wrong or false. Thereby, OWL facilitates a modification or extension of existing resources. This is why the proposed model is represented in the DL-based EL subprofile of OWL 2, which has been chosen for its scalability, ample tool support, and successful record with biomedical terminological ontologies. It has to be noted that OWL 2 EL is specifically orientated towards resources with high classification needs, which is the case for terminological resources with very large subsumption hierarchies.

The proposed model itself is domain-independent and adequate for various purposes and domains. Each of its entries is identified by a unique numeric identifier to ensure concept permanence. Concepts and terms are related but modeled separately so that evolving terms can be changed without having to alter the concept. Its modular structure allows for the creation of customizable subsets of data categories as well as unlimited addition of new categories to ensure term autonomy and multi-purpose adequacy. While the human-driven aspects are ensured by the usage of natural language names, the machine-readability requirement is met by the chosen representation language. While OWL is fit to facilitate syntactic interoperability, semantic interoperability relies on formal and explicit representations of meaning provided in T-Mint and by reference to a formal ontology. Moreover, the strong emphasis on re-using existing vocabularies contributes towards the model's interoperability with existing models and resources, e.g. resources using Dublin Core, the OBO Relation Ontology, GOLD, or semiotics.owl.

5 Methodology

This section presents methodologies for two possible use cases of the proposed model, namely generating a terminological resource and merging it with other, informal terminologies. Both cases serve the purpose of showing potential application scenarios of T-Mint, but still have to be validated and evaluated in a separate paper.

5.1 Terminologizing a Domain Ontology

One of the main goals of T-Mint is to be used to generate terminological resources to describe domain ontologies. It represents the meta-structure of a terminological resource and needs to be instantiated for the specific domain. A top-down methodology for creating an instance of the T-Mint model is described below.

Methodology This methodology presupposes a formal domain ontology represented in OWL as input. The methodology requires an ontology editing method, such as the OWL API[8], for manipulating the input ontology and the created T-Mint instance, and one off-the-shelf NLP tool for POS tagging and tokenization.

[8] `http://owlapi.sourceforge.net`

- Random generation of integer as building block for unique non-semantic URI to create terminological entry and use the `reference` relation to establish connection to a top hierarchy class of the input ontology
- Use language tag from RDF or SKOS label to create a language section - the URI for the language section uses the number created above and adds the `xml:lang` tag to it, e.g. 1423en for English
- Extraction of label from RDF, SKOS, or URI fragment if no label is available and creation of a term section identified by a number added to the language section identifier, e.g. 1423en1
- Tokenization produces subcomponents of the label, which are linked to the term section
- Part of speech tags are represented as complex closed data categories, extracted from the created tags, and related to the term components
- Identified noun phrases are suggested as sub-terms, which have to be evaluated manually for their reference to the domain of the input ontology
- Definitions are, if available, instantiated and related to the terminological entry

The described methodology produces a collection of terminological entries - of which one entry is exemplified below - that depend on the input ontology for formal semantics and concept relation. Frequently, ontologies do not feature any natural language definitions. In such cases, a generation of the natural language definition based on the formal definition is one alternative. One way to achieve this goal is a combined method of ontology verbalization and ontology design patterns as we describe elsewhere [6].

```
finance:1423 tmintCore:reference ontology:LiquidityRiskSymptom ;
                                 rdf:type tmintCore:TerminologicalEntry ;
finance:1423en rdf:type tmintCore:LanguageSection ;
                                 oboRO:part_of  finance:1423 ;
finance:1423en1 rdf:type tmintCore:TermSection ;
                                 tmintDCSM:literalForm "Liquidity risk symptom"@en ;
                                 oboRO:part_of  finance:1423en1 ;
                                 tmintDCMS:hasPartOfSpeech gold:NounPhrase ;
                                 tmintDCMS:hasTermType tmintDCMS:entryTerm ;
finance: liquidity  rdf:type tmintCore:TermComponent ;
                                 tmintCore:isComponentOf finance:1423en1 ;
                                 tmintDCMS:hasPartOfSpeech gold:Noun ;
```

5.2 Merging existing resources with T-Mint instances

Existing multilingual and concept-oriented terminological resources, such as EuroTermBank[9], are frequently available as alphabetical listings in formats restricted to human readability. Nevertheless, if they contain proper terminological definitions, the superordinate concept of a terminological entry can be extracted from the natural language definition. The superordinate concept of the terminological entry in the T-Mint instance is obtained by way of the subsumption hierarchy of the input ontology. Merging the entry of the existing resource with the entry of the T-Mint instance creates a machine-readable multilingual resource.

[9] `http://www.eurotermbank.com`

Methodology The following components can be used to evaluate whether two concepts can be considered equivalent:

- All terms - entry as well as semantically related terms
- Term components
- POS tags assigned to terms and term components
- Nouns, adjectives, and verbs extracted from natural language definitions
- Superordinate concepts extracted from definition and subsumption hierarchy

Instead of string-matching each of these components from an entry of the T-Mint instance to the potential matching entry in another resources separately, we suggest representing them as vectors. Vector Space Models are used to measure the similarity of keywords in documents in information retrieval. As the merging process can be compared to a keyword search, the Vector Space Model is used as a means of similarity measure of both entry vectors. Each dimension of the vector represents one term or expression from the list above. Its value depends on the frequency of occurrence of the term. The dot product of both vectors is divided by the product of their norms. The threshold for merging entries should be lower for exact matches of entry terms than for partial matches. Merging means adding new term sections based on the extracted content of the resource grouped by language to the terminological entry of the T-Mint instance. Thereby, more languages are added to the ontology by reference.

6 Conclusion

In this paper we propose a terminological metamodel build on standards and best practices from the terminology community. Established vocabularies frequently provide terminological data as informal RDF resources and/or annotation properties (e.g. SKOS). Both cases do not support the relation of terminological data, their annotation, and machine-readability. This is why, the proposed model is a formalized ontology that allows modeling highly complex terminological data in relation to a formalized domain ontology. We provide a methodology for terminologizing domain ontologies as well as for merging available multilingual terminologies building on the proposed terminology interchange model. The major motivation for terminologizing ontologies is the ability to relate, annotate, and directly address natural language elements while maintaining the reference to formal semantics. First experiments still need to be extended to provide a full evaluation and validation of the proposed method and model.

References

1. Antia, B.E., Budin, G., Picht, H., Rogers, M., Schmitz, K.D., Wright, S.E.: Shaping Translation: A View from Terminology Research. Meta: Translators' Journal 50 (2005)

2. Aussenac-Gilles, N., Szulman, S., Depres, S.: The Terminae Method and Platform for Ontology Engineering from Texts. In: Buitelaar, P., Cimiano, P. (eds.) Ontology Learning and Population: Bridging the Gap between Text and Knowledge, vol. 167, pp. 199–223. IOS Press, Amsterdam, The Netherlands (2008)
3. Bodenreider, O., Smith, B., Kumar, A., Burgun, A.: Investigating Subsumption in SNOMED CT: An Exploration into Large Description logic-based Biomedical Terminologies. Artif. Intell. Med. 39(3), 183–195 (2007)
4. Cimino, J.J.: Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. Meth. Inf. Med. 37(4-5), 394–403 (1998)
5. Garcia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the Multilingual Web of Data. Web Semant.: Sci., Serv. and Agents on the World Wide Web 11(0) (2012)
6. Gromann, D.: Terminology-Based Patterns for Natural Language Definitions in Ontologies. In: Presutti, V., Gruninger, M. (eds.) In Proceedings of the Workshop on Ontology and Semantic Web Patterns (4th Edition) - WOP2013. CEUR Workshop Proceedings, Sydney (2013)
7. Madsen, B.N., Thomsen, H.E.: Terminological Concept Modelling and Conceptual Data Modelling. Int. J. Metadata Semant. Ontol. 4(4), 239–249 (2009)
8. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009. (2009)
9. Montiel-Ponsoda, E., Vila Suero, D., Villazón-Terrazas, B., Dunsire, G., Escolano Rodríguez, E., Gómez-Pérez, A.: Style Guidelines for Naming and Labeling Ontologies in the Multilingual Web. In: Baker, T., Hillmann, D.I., Isaac, A. (eds.) Proceedings of the International Conferece on Dublin Core and Metadata Applications. pp. 105–115. Dublin Core Metadata Initiative (2011)
10. Palma, R., Hartmann, J., Haase, P.: OMV - Ontology Metadata Vocabulary for the Semantic Web Version 2.4.1. Technical Report (2009)
11. Reymonet, A., Thomas, J., Aussenac-Gilles, N., IRIT-Melodi, U.: An Ontological and Terminological Meta-model for Semantic Information Retrieval. In: WS 2 Workshop Extended Abstract, 9th International Conference on Terminology and Artificial Intelligence. pp. 28–29 (2011)
12. Roche, C.: Ontoterminology: How to Unify Terminology and Ontology into a Single Paradigm. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). ELRA, Istanbul, Turkey (2012)
13. Seppälä, S., Ruttenberg, A.: Survey on Defining Practices in Ontologies: Report Summary. In: Seppälä, S., Ruttenberg, A. (eds.) Proceedings of the International Workshop on Definitions in Ontologies (DO 2013). vol. 1061. CEUR Workshop Proceedings (2013)
14. Tao, C., Pathak, J., Solbrig, H.R., Wei, W.Q., Chute, C.G.: Terminology Representation Guidelines for Biomedical Ontologies in the Semantic Web Notations. J. Biomed. Inform. 46(1), 128–138 (2013)
15. Wright, S.E., Summers, D.: Crosswalking from Terminology to Terminology: Leveraging Semantic Information across Communities of Practice. In: Calzolari, N., Sasaki, F., Teich, E., Witt, A., Wittenburg, P. (eds.) LREC 2008 Workshop: Uses and usage of language resource-related standards. pp. 21–29. ELRA/ELDA, Marrakech, Marocco (2008)