A Methodological Framework for Ontology and Multilingual Termontological Database Co-evolution*

Christophe Debruyne[1], Cristian Vasquez[1],
Koen Kerremans[2], and Andrés Domínguez Burgos[2]

[1] Semantics Technology and Applications Research Lab (STARLab),
Vrije Universiteit Brussel
{chrdebru,cvasquez}@vub.ac.be
[2] Centre for Special Language Studies and Communication, Erasmushogeschool Brussel
{Koen.Kerremans,Andres.Dominquez.Burgos}@ehb.be

Abstract. Ontologies and Multilingual Termontology Bases (MTB) are two knowledge artifacts with different characteristics and different purposes. Ontologies are used to formally capture a shared view of the world to solve particular interoperability and reasoning tasks. MTBs are general, contain fewer types of relations and their purposes are to relate several term labels within and across different languages to categories. For regions in which the multilingual aspect is vital, not only does one need an ontology for interoperability, the concepts in that ontology need to be comprehensible for everyone whose native tongue is one of the principal languages of that region. Multilinguality provides also a powerful mechanism to perform ontology mapping, content annotation, multilingual querying, etc. We intend to meet these challenges by linking both methods for constructing ontologies and MTBs, creating a virtuous cycle. In this paper, we present our method and tool for ontology and MTB co-evolution.

Keywords: Ontology Engineering, Multilingual Termontology Bases, Ontology Evolution.

1 Introduction

A computer-based, shared, agreed formal conceptualization is known as an ontology. Ontologies constitute the key resources for realizing a Semantic Web. The problem is not so much what ontologies are, but how they come to be. Methods are needed to support communities in reaching the meaning agreements necessary for semantic interoperability between two or more autonomously developed information systems for a particular goal. In previous work [5], we introduced a formalism for hybrid ontology engineering. In hybrid ontologies, concepts are both described in terms of natural language and formal descriptions. To this end, the ontologies are complemented with a glossary, containing the natural language descriptions. Just like the ontology, this glossary is in function of the community's semantic interoperability requirement. The natural language descriptions are used to drive the formal descriptions of these concepts.

* This work was partially funded by the Brussels Institute for Research and Innovation through the Open Semantic Cloud for Brussels Project.

P. Herrero et al. (Eds.): OTM 2012 Workshops, LNCS 7567, pp. 220–230, 2012.

In general, ontologies are used to capture a shared view of the world in a formalism to solve particular interoperability and reasoning tasks. Multilingual termontology bases (MTBs) are terminology bases in which ontological information is made explicit. They contain fewer types of relations and their purposes are to relate several term labels within and across different languages to categories. MTBs are used in a way that surpasses their 'traditional' role as terminological dictionaries for human users (e.g. by translators who often need to verify the right verb or adjective to combine with a given noun a set of candidate words with similar meanings). Given one or several predefined functions, MTBs need to represent (in natural language) those items of knowledge that are considered relevant for supporting specific tasks (e.g. domain modeling), applications (e.g. information extraction tools) or users (e.g. domain experts). These new needs have defined new methods in terminology analysis, new types of information to be included in MTBs as well as new ways of visualizing or representing this information. See for instance: [17,1,3,2,4].

The special linguistic resource in hybrid ontologies was not meant for capturing appropriate linguistic (syntactic, morphological, semantic and pragmatic) information of the natural language descriptions as well as to cope with multiple languages. In hybrid ontologies, we assume one community to agree on one common language. The linguistic resource is rather meant for facilitating meaning agreements on the formal description of concepts. MTBs do take into account this information. In line with the mission statement of the Ontology-Lexica Community Group[1], we bring hybrid ontologies and MTBs together in this paper. This enables – amongst others – capturing how elements in the (hybrid) ontology are realized in multiple languages, enable multilingual querying of the annotated datasets and provide additional documentation and information while consulting the (hybrid) ontology. We furthermore show how the methods for developing hybrid ontologies and MTBs are furthermore driven by each other. In the architecture that we will present, both methods are connected by means of SPARQL services.

2 Related Work

In ontology engineering, it is important that the community members first agree on the meaning of the concepts to be represented before formally describing them. Such agreements among community members can be reached on the basis of discussions in natural language, which will eventually lead to the creation of natural language definitions of concepts to which all members have contributed and agree. Quite a few surveys on the state of the art in ontology engineering methods exist [8,15,16]. However, we noted that for ontology development (for the Semantic Web), relatively few methods take into account a special linguistic resource for natural language definitions: DOGMA [10] and GOSPL [5,6], HCOME [12] and UPON [7]. Other methods do mention the idea of drawing inspiration from existing linguistic resources, but do not treat these resources as an integral part of the method. In the case of HCOME, it should be noted that users were seemingly not able to update this resource.

[1] http://www.w3.org/community/ontolex/

3 Method

In this section, we explain the ontology engineering and multilingual termontological database methods adopted in this paper and present how both methods can co-evolve. We will explain how these two representations with distinct degrees of precision (the first serving a specific purpose and thus containing only the relations needed for this purpose, the latter a general description with general relations). The two are thus complementary and - as we will explain - can benefit from each other's construction process. As we will adopt a method for hybrid ontology engineering, we will use the words hybrid ontology and ontology interchangeably.

3.1 Ontology Engineering Method

In hybrid ontologies, communities are promoted to first-class citizens, part and parcel of the formalism, such that the interactions within the evolving community leads to a series of change-operations applied to the ontology. The evolution of the interactions thus has a direct impact on the evolution of the ontology. The natural language aspect is vital, as the closer the link between human communication and the resulting system and/or business communication, the more likely such systems will work as intended by their various stakeholders. Concepts are described both informally by means of natural language descriptions stored in a *glossary* and formally by means of binary fact-types coming from a community and grounded in natural language called *lexons* [13], e.g., <*Cultural Domain Community, Event, starting on, end of, Date*>. A fact type is the collection of objects linked by the predicate. A fact is an element of the population of the fact type, e.g. "Event A" starting on "2012-05-22". A series of social processes have been defined which allows the ontology to evolve with the community and the community's agreements. Table 1 contains some of the social processes defined in GOSPL. A lexon is more "understandable" by people without training than, for instance, ontology languages such as OWL. The goal of the DOGMA framework is not to invent another ontology language, but rather to present a formalism for developing ontologies from which "implementations" in other formalisms can be distilled.

Table 1. Social processes in GOSPL

Social process: Request to …		
Remove gloss from lexon or term	Add lexon	Add constraint
Change gloss of lexon or term	Remove lexon	Change role hierarchy
Add gloss to lexon or term	Remove constraint	Remove synonym
Change supertype of term	Add synonym	

Fig. 1 summarizes the different processes in GOSPL. Starting from co-evolving communities and requirements, the informal descriptions of key terms have to be gathered before formally describing those concepts. Constraints and application commitments to these formal descriptions can be expressed in commitment languages, such as Ω-RIDL [18]. During the processes from creating the glossary to committing to the ontology, the communities can make agreements on gloss-equivalences (an agreement that

two descriptions refer to the same concept) and synonyms (an agreement that two terms refer to the same concept). The ontology, and the data described with those commitments can then be re-internalized by the community for another iteration. This allows communities to gradually build up the domain that needs to be captured by the ontology. Knowledge proposed by the *goal-driven* community typically comes from their existing autonomously developed and maintained information systems. In that sense, (hybrid) ontologies are rarely built from scratch. By capturing the social processes leading to ontology evolution, we do not only store the changes in the ontology, but also register the thought processes of communities that have lead to those changes.

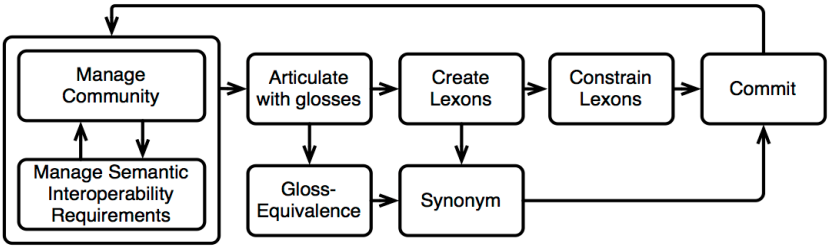

Fig. 1. The GOSPL method

3.2 Termontology Mining Method

In [11] a method called termontography was described for developing multilingual termontological databases, i.e. terminological resources which make explicit both terminological/linguistic and ontological/conceptual information. Termontological databases are in this respect similar to terminological knowledge bases, as defined by [14]. Unlike TKBs, termontological databases following the termontographic approach rely on a predefined categorization structure, which is used to identify and classify terminology in different natural languages. The categorization structure corresponds to an ontological structure listing the important categories and their relationships that are derived from a user requirements analysis. These relationships are not strictly confined to the typical hierarchical relationships (i.e. partitive and generic relationships) found in thesauri or taxonomies but can also encompass generic relationships expressing like causality, location and function. In the termontographic approach, the categorization level interconnects with a linguistic level, providing a wide range of information about the use of any term in natural language to express the meaning to which it is associated at the category level. The category level can be used by terminographers to classify multilingual terminology. It supports the processes of terminology analysis and management. The linguistic level provides additional information for analysis and production of human language.

This detailed information has always been relevant for human purposes. Translators, for instance, often need to verify the right verb or adjective to combine with a given noun within a set of candidate words with similar meaning. The framework can also be useful within the realm of natural language processing to overcome the gap between natural language and ontologies. Take for instance the term “opera”.

Even within one short text, opera can refer to either the building or the music genre. When we read that an opera is performed, we know it is about a piece of that music. When we read an opera is "re-furbished", we know the word denotes something different as compared to when we say that it is "being visited". Very often the correct meaning can be derived from looking at the correlation between the ambiguous term and some context words. For instance 'opera' and 'visit' vs. 'opera' and 'perform'. Still, in other cases, determining the real meaning of opera could above all depend on the syntactic paradigm used, as in "he was at the opera" (we cannot say the verb "be" and its possible meaning is determining our judgment here).

Fig. 2 shows how the TermontoPlatform supports the extraction and modeling of an MTB according to termintological principles. In this process, a terminologist and a human domain expert first define the knowledge domain they want to tackle. They then proceed to select document collections with textual information relevant for that domain. Using automatic means -term extraction- and human knowledge they proceed to compile a list of seed terms that stand for the basic conceptual items in that domain. They agree on a taxonomic model where these terms stand for the categories. Using expert knowledge and yet again the modules for term extraction, they identify the relationships they want to identify in the texts. They verify whether the extraction modules have the right rules to identify the kind of patterns - often but not exclusively verbal phrases - expressing those relationships. To do that they examine primarily documents where the seed terms are particularly relevant. They can modify and add new linguistic patterns to detect these relationships. Once this is done, they run the mining modules on the domain documents. The extracted terms and possible relationships are then verified, modified if needed, and exported. New data can be imported from other databases by adapting a filter module that converts other sources into data types of the MTB.

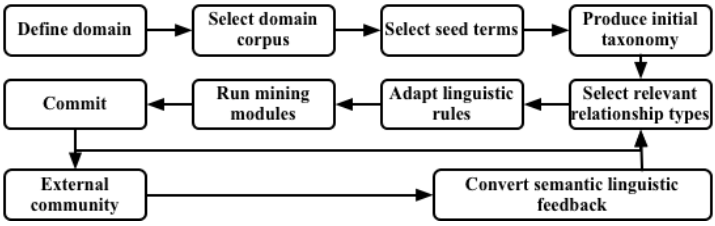

Fig. 2. A general view of the MTB process: ontological and other updates from the external community are taken over to perform a new mining and modeling cycle. The Termonto Platform supports humans through different steps. The mining modules in the middle of the process use the incoming rules, predefined seed terms and statistic criteria to extract possible terminological and ontological-relevant information that terminologists validate for internal and external consumption.

The MTB distinguishes a semantic level from a linguistic level. Labels attached to categories are not per se "terms" in the linguistic way: they are only tags that help identify categories more easily as the unique identifiers at semantic levels. A category label in the MLB model can exist for one or more languages. The same label can be visualized in different languages but only once for a given category in one specific language. The categories themselves can be connected to one or more terms within a context.

The term "opera" within a specific context has a meaning and is attached to the category with the English label "opera as piece". In other contexts that term is attached to two other different categories. In Fig. 3 you can see how two terms, opera and opera house, are attached to the category opera_house. These are synonyms. The stars represent preferred synonym. At the linguistic level, each term within a context can be attached to one or more lexical relations, which can be seen as links to other co-occurring terms. These other terms can help in the analysis of texts to detect which context and hence which concept is meant.

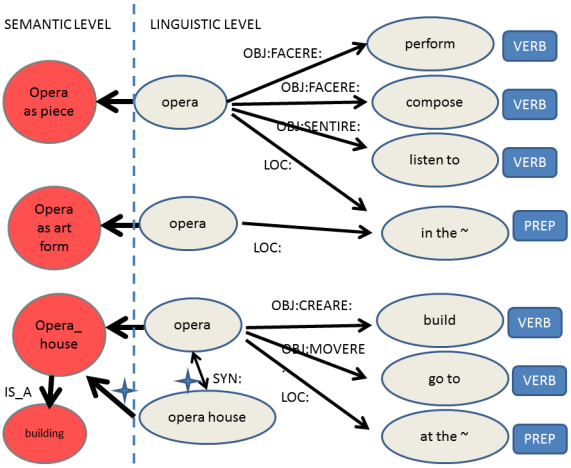

Fig. 3. A fragment of the linguistic data of a MTB: here, terms associated with categories such as opera as art piece, as art form or opera house. Special lexical relations linking to collocate words to these terms can help in disambiguation processes. They can also be used to identify possible raw data for identifying semantic relations.

3.3 Ontology and MTB Co-evolution

Fig. 4 depicts the co-evolution of the ontology and the MTB. Communities can use the MTB as a starting point to build up or adapt the ontology to–their needs by re-internalizing the ontology and start interacting. The social processes (for reaching agreements) result in several ontology evolution operators during externalization. Externalization here means the process of a community formally "writing" down their thoughts in artifacts that describe the universe of discourse. The social processes and changes in the ontology can be queries to refine or steer the crawling processes of the MTB. The interactions on terms are available through an API, table or as RDF. The crawler can be steered by analyzing the number of activity of terms in the ontology within a time window. The link between the ontology and the MTB is captured in externalization and is used to provide links from the implemented ontology to the MTB.

In the previous Section, we described some of the data types present in the MTB. The structure between the Categories and Terms allow us to query the MTB based on the term labels in the ontology's facts. This can be used to retrieve glosses from the category definitions and information about the fact's term labels from the MTB term

entries. Category relations that connect two categories in the MTB even serve as inspiration to add facts to the ontology; i.e. a linguistic knowledge "graph" on the side. The next section will go into more detail on how this is precisely implemented and how the MTB is queried.

As can be seen from Fig. 2 and Fig. 4, the TermontoPlatform can also import data from other knowledge systems. As Fig. 2 shows, a module is used to determine how the import process takes place. That means the module needs to know what data types are relevant and to what data types or rules they can be mapped. "Terms" as used in GOSPL are not necessarily reliable to determine linguistic terms. They need to be verified by a terminologist before they can overwrite or serve as a synonym. Still, they can be used to improve the term detection in the text miner modules. The lexons can be mapped to correct and expand the semantic level of the MTB. New lexons that represent a semantic relationship between two objects or instances can be used in the module for semantic relation detection (Fig. 2) in order to identify new linguistic patterns in the domain documents that can be selected to add yet new rules for further automatic detection of semantic relations.

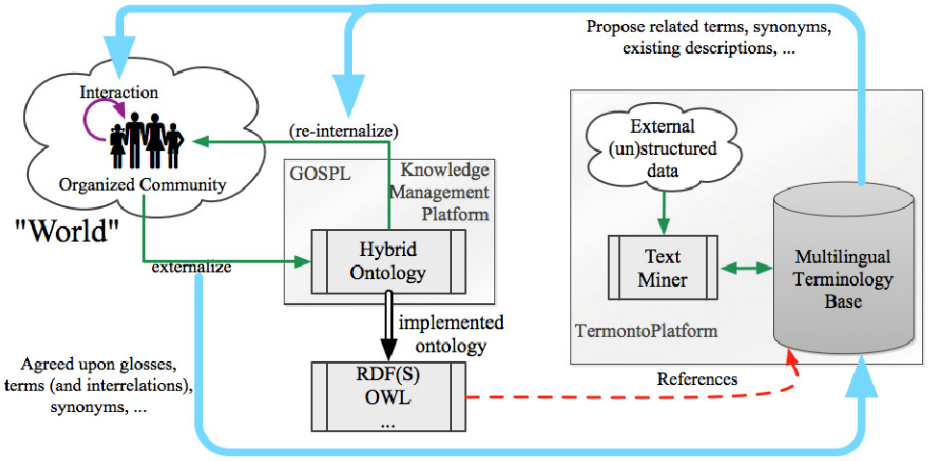

Fig. 4. Co-evolution of the ontology and the MTB. The MTB is used as a source of inspiration in the community interactions that will result in ontology evolution operators. The hybrid ontology can - at any time - be transformed into other formalisms referring to the MTB for documentation and additional context.

4 Tool

The GOSPL [6] prototype is the tool we have adopted for hybrid ontology engineering. The core of GOSPL is a series of services for hybrid ontology engineering to which multiple (types of) clients can connect. Services include: community management, ontology retrieval, starting and resolving discussions involving requests, etc. Discussions involve around requests to evolve the ontology. The interface depicted in Fig. 5 is a screenshot of one such client.

The MTB's schema was lifted to an ontology using the GOSPL platform. This ontology was then used to publish the MTB as RDF triples on the Web. There are currently 20430 triples available. The SPARQL endpoint is then used to link the knowledge management platform with the MTB. Fig. 6 contains an RDF description of the English term opera.

The community working on the ontology using GOSPL with an integrated multilingual termontology base can - at each time - ask for an implementation of the hybrid ontology in OWL. The translation of a fact-oriented formalism into another formalism has already been presented, for instance in [9]. The generation of the OWL implementation of the ontology now incorporates links with the MTB as well as additional annotation for documenting purposes next to the labels and glosses of the hybrid ontology.

The benefits of this approach on the information service level are twofold. First, it gives pointers to the MTB for multilingual understanding of some of the facts and concepts in the ontology. Secondly, it enables querying the information not only via the ontology, but also via the concepts and terms of the MTB, allowing for querying using synonyms (within one language and across languages). The community has been externalizing their perception of reality through modeling processes driven by glosses, of which a subset is provided by the MTB. Thanks to this link between the

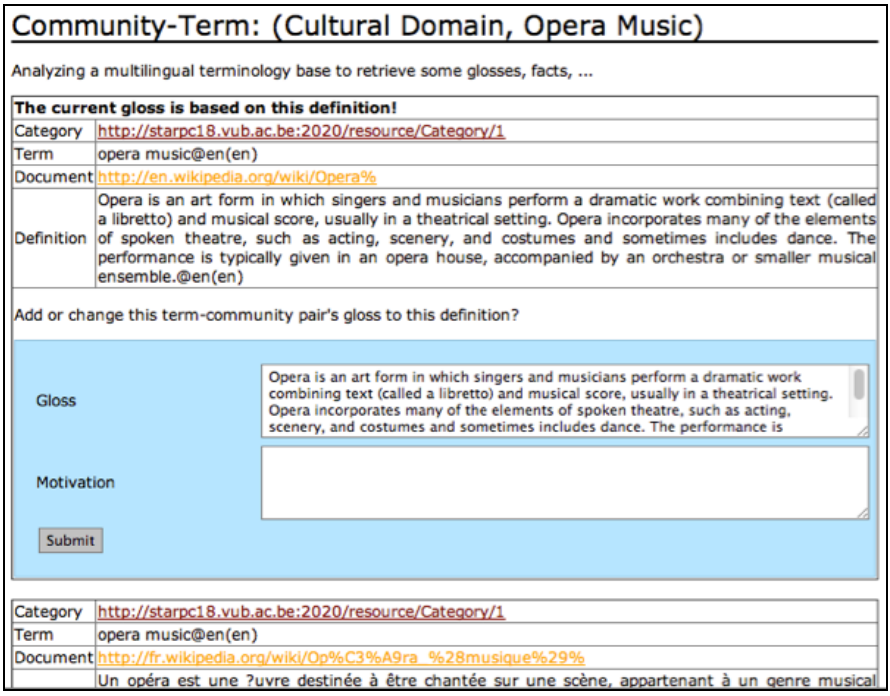

Fig. 5. The GOSPL [6] prototype linked with the MTB. Users can ask for a list of definitions that they can adopt as a gloss. Note that the user is shown on which definition the current gloss is based on (if the gloss came from a definition from the MTB). Also note that a user can alter the definition, but a link with the original definition is kept.

Property	Value
is ont:Category_with_Term of	<http://localhost:2020/resource/Category/169>
is ont:Collocation_of_Term of	<http://localhost:2020/resource/Collocation/15>
is ont:Collocation_of_Term of	<http://localhost:2020/resource/Collocation/16>
is ont:Collocation_of_Term of	<http://localhost:2020/resource/Collocation/17>
is ont:Collocation_of_Term of	<http://localhost:2020/resource/Collocation/18>
is ont:Linguistic_Context_of_Term of	<http://localhost:2020/resource/LinguisticContext/51>
ont:Term_with_Gender	none
ont:Term_with_Gramatical_Number	singular
ont:Term_with_Syntactical_Pattern	noun
ont:Term_with_Term_Label	opera (en)
rdfs:label	opera (en)
rdf:type	ont:Term

Fig. 6. Triples of a term in the MTB. Namespaces have been omitted

purpose driven ontology and more general MTB, the community will potentially understand this query. The benefits of this approach on the methodological level are threefold. First, it helps users discover new facts and textual description from the MTB, facilitating ontology development. Different communities and their corresponding ontologies evolve towards each other by agreeing on the relations (e.g., equality) of their concepts. Secondly, the discussions provide points of interest for the MTB development by - for instance - looking at terms in the ontology that are the source of an active discussion but for which no gloss has been provided. And thirdly, the glosses not originating from the MTB are used as a source to feed the MTB. The MTB is a powerful tool for automating parts of ontology mapping, as it takes also into account language variation, i.e. different ways in which users express concepts in different languages.

5 Case

The work we presented here is the result of one of the cases in the Open Semantic Cloud for Brussels project in the cultural domain. The goal of this case is to annotate a relevant datasets on events (concerts, theater, etc.) in various venues (e.g., opera houses) in Brussels. The various examples in this paper originate from this use case. Ultimately, the goal would be that a user is able to take a picture of a venue and present the user with, for instance, the current shows taking place. The heterogeneous data sources from the different autonomously developed information systems already motivated the use semantic technology and ontologies.

Next to annotating the heterogeneous data sources, a problem is the multilingual nature of Brussels. Even though the ontology development can be done with a community of stakeholders having different languages, the use of that data needs to be accessible to most users, not only to build services on top of that data but also to understand the concept and facts in the ontology. By linking ontologies with MTBs, queries expressed in different languages about the same concepts can be formulated. Currently, the MTB is being developed while the community in the cultural domain is working on the ontology. One of the results is facilitating the mapping of concepts when multiple ontologies have linkages to the same MTB.

6 Conclusions

In this paper, we presented a methodological framework for ontology and multilingual termontological database co-evolution. Ontologies stem from the community's need to exchange information for a particular purpose and are the result of a series of meaning agreements. In hybrid ontology engineering, those agreements are driven by the natural language descriptions formulated by the community. Multilingual termontological databases are general-purpose multilingual thesauri with general semantic relations (e.g., subsumption, part-whole, etc.), storing variances in meaning across languages. The hybrid ontology engineering processes can use the MTB as a useful source for natural language descriptions and general relations between concepts, and the MTB construction processes can benefit from the hybrid ontology engineering activities to pinpoint the communities topics of interest, e.g., to steer the mining process or adapt the seed terms.

To this end, we have published the MTB as RDF on the Web to facilitate the integration with the GOSPL knowledge management platform. Information on the interactions between the communities is provided via an API or as RDF. Those interactions are annotated by means of SIOC. These descriptions are then used to mine the points of interest of the community by for instance pointing to the terms and lexons that more strongly engaged people in interacting with one another than other terms or lexons.

References

1. Aussenac-Gilles, N., Condamines, A., Szulman, S.: Prise en compte de l'application dans la constitution de produits terminologiques. Actes des 2e Assisses Nationales du GDR I3 (2002)
2. Cabré, M.T.: El principio de poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en Terminología (I). IBÉRICA 16, 9–36 (2008)
3. Collet, T.: What's a term? An attempt to define the term within the theoretical framework of text linguistics. Linguistica Antverpiensia 3, 99–111 (2004)
4. Faber, P. (ed.): A Cognitive Linguistics View of Terminology and Specialized Language. Mouton De Gruyter (2012)
5. Debruyne, C., Meersman, R.: Semantic Interoperation of Information Systems by Evolving Ontologies through Formalized Social Processes. In: Eder, J., Bielikova, M., Tjoa, A.M. (eds.) ADBIS 2011. LNCS, vol. 6909, pp. 444–459. Springer, Heidelberg (2011)
6. Debruyne, C., Reul, Q., Meersman, R.: GOSPL: Grounding ontologies with social processes and natural language. In: Latifi, S. (ed.) ITNG, pp. 1255–1256. IEEE Computer Society (2010)
7. De Nicola, A., Missikoff, M., Navigli, R.: A software engineering approach to ontology building. Information Systems 34, 258–275 (2009)
8. Gomez-Perez, A., Fernandez-Lopez, M., Corcho, O.: Ontological Engineering with examples from the areas of Knowledge Management. In: e-Commerce and the Semantic Web. Springer-Verlag New York, Inc., Secaucus (2003) ISBN: 1852335513
9. Hodrob, R., Jarrar, M.: Mapping ORM into OWL 2. In: Alnsour, A., Aljawarneh, S. (eds.) ISWSA, p. 9. ACM (2010)

10. Jarrar, M., Meersman, R.: Ontology Engineering - the DOGMA Approach. In: Dillon, T.S., Chang, E., Meersman, R., Sycara, K. (eds.) Advances in Web Semantics I. LNCS, vol. 4891, pp. 7–34. Springer, Heidelberg (2008)
11. Kerremans, K., Desmeytere, I., Temmerman, R., Wille, P.: Application-oriented terminography in financial forensics. Terminology 11(1), 83–106 (2005)
12. Kotis, K., Vouros, A.: Human-centered ontology engineering: The hcome methodology. Knowledge Information Systems 10(1), 109–131 (2006)
13. Meersman, R.: The use of lexicons and other computer-linguistic tools in semantics, design and cooperation of database systems. In: CODAS, pp. 1–14 (1999)
14. Meyer, I., Skuce, D., Bowker, L., Eck, K.: Towards a new generation of terminological resources: an experiment in building a terminological knowledge base. Presented at the, Nantes, France (1992)
15. Simperl, E.P.B., Tempich, C.: Ontology Engineering: A Reality Check. In: Meersman, R., Tari, Z. (eds.) OTM 2006. LNCS, vol. 4275, pp. 836–854. Springer, Heidelberg (2006)
16. Siorpaes, K., Simperl, E.: Human intelligence in the process of semantic content creation. World Wide Web 13(1-2), 33–59 (2010)
17. Temmerman, R.: Towards New Ways of Terminology Description: The Sociocognitive-Approach. John Benjamins Publishing Company, Amsterdam (2000)
18. Verheyden, P., De Bo, J., Meersman, R.: Semantically Unlocking Database Content Through Ontology-Based Mediation. In: Bussler, C.J., Tannen, V., Fundulaki, I. (eds.) SWDB 2004. LNCS, vol. 3372, pp. 109–126. Springer, Heidelberg (2005)