Building a domain ontology from glossaries: a general methodology

Loris Bozzato, Mauro Ferrari, and Alberto Trombetta

Dipartimento di Informatica e Comunicazione
Università degli Studi dell'Insubria
Via Mazzini 5, 21100, Varese, Italy

Abstract. We propose a general methodology to build up a domain ontology from one or more domain glossaries. The particular feature of this methodology is in the parallel construction of a domain ontology and a complete domain terminology. In this paper we fully describe the methodology phases and we apply them to a real-world example from the medical domain.

1 Introduction

Nowadays, ontologies have become a relevant representation formalism and many application domains are considering their adoption. This attention claims for methods for reusing domain knowledge resources in the development of domain ontologies. Accordingly, in this paper we discuss a general methodology to create a domain ontology from one or more domain glossaries. Before explaining the actual steps of the methodology, let us introduce some of the ideas that led to its composition. Basically, the methodology is to be manually carried out by the ontology developers with the help from domain experts. Some steps of the methodology cannot be easily automatized because an explicit *interpretation* of the semantics of terms is required as, e.g., in the initial phase of clustering. In analogy to other proposals for the translation of thesauri [9, 20], the translation is filtered by the interpretation of the domain experts. Hence, the proposed methodology provides a way for defining a domain ontology from a domain glossary and not a simple translation. A particular feature of our methodology is that, while building the domain ontology, one also reconstructs a *complete terminology* for the domain of interest. However, these two representations are distinct: the terminology should not necessarily be complete with respect to entities external to the domain (or to entities that are not essential or specific for the examined domain), while the ontology should even represent concepts without a specific name in the terminology and be complete with respect to the represented domain. Another feature of our methodology is that it allows one to build up a *complete* domain ontology (i.e., without undefined objects) starting from a possibly incomplete glossary: this is obtained by a sort of *saturation* on the glossary. We restrict our attention to input glossaries with unstructured terms, as opposed to what is assumed, for example, in thesauri [9, 16, 20]. We define a

glossary as a list of (possibly lexicographically ordered) *lemmas*; each lemma is composed by a *term* and a *textual definition* that provides the meaning of the lemma; *references* to others lemmas can appear in the definition. Our definition of glossary is coherent with the one of [1, 18] and the definition of lexicon in [8].

2 Example: an ontology for orthodontic terminology

In order to illustrate our methodology, we will employ a real-world motivating example, taken from an ongoing collaborative effort between our Department and the Department of Clinical and Biological Sciences. In fact, this project also constituted the motivation for the definition of our methodology. The objective of the project is in the definition and representation of a standard terminology for *orthodontics*: the motivation for this research can be found in the complexity of the orthodontic technical terminology and in the lack of a common reference standard for such terminology.

The complex relations among orthodontic terms suggested a structured representation as an *ontology*: this representation allows to formally define semantic relationships among domain concepts and explore implicit connections by automated inference. As for the representation formalism, we have employed *OWL (Web Ontology Language)* [21]. Its choice has been dictated by the well-founded semantics of the language and its wide acceptance as a de-facto standard.

The basis for the development of the domain terminology has been identified in the *AAO Orthodontic Glossary* [15], a simple yet quite comprehensive glossary developed by the *American Association of Orthodontists*[1]. The particular form of the initial data led to the present proposal for a general ontology development methodology from one or more glossaries.

In the next section, we will formulate a running example based on terms extracted from the original glossary that covers the first steps of development of the actual project ontology.

3 Methodology specification

In this section we present the phases of our methodology using the following schema. For each *phase* we describe the required activities and the *input* and *output* documents, possibly enriched with an example of their structure. In some cases, we also provide some possible extensions or guidelines for the presented step. In the following examples, starting from a minimal (and incomplete) fragment of the initial glossary, we will concentrate on the development of the cluster for `AnatomicPart` and we show how to combine this with the cluster for `AbnormalCondition` in the conceptual schema.

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

Phase 1 – Clustering

Group *terms* of the original glossary in *clusters*. To define clusters, follow the *is-a* direction from terms to clusters: that is, every term should be seen as instance of a "type" defined by a cluster. Clusters do not have to be disjoint: however, terms must be classified in their most specific clusters. Some cluster inclusions can be suggested in this phase.

Input: *original glossary* (one or more).

Example:

Gingiva The tissue that surrounds the teeth, consisting of a fibrous tissue that is continuous with the periodontal ligament and mucosal covering.
Ligament, periodontal See periodontal ligament.

Output: *table of clusters*. A grouping by cluster of the terms from the original glossary.

Example:

Cluster	Anatomic Part	
Description	Element that is part of the human anatomy.	
Elements	*Term*	*Source*
	Gingiva	O
	Ligament, periodontal	O
	Periodontal ligament	O

An alternative way to define clusters of terms can be represented by a specialization of a general superstructure: the structure could be either a *foundational ontology* (as, for example, the well-known *DOLCE* ontology [11]) or a general high level ontology that is specific for the domain of interest.

As for the structure of the documents, the table of clusters should contain a textual definition for each cluster, in order to clarify the intended meaning of the grouping. It is also desirable to maintain a column of information about the source glossary of each term. When suggesting inclusions between clusters, this should be specified in the table of clusters by indicating the superclusters of each cluster.

Phase 2 – Saturation

Find every term in the *definitions* that does not appear as a lemma in the glossary. Add the new terms to the glossary, providing them with a description: the description can also be a reference (e.g., *see...*) to present terms. Classify the new terms in the previously found clusters.

Input: original glossary, table of clusters.

Output: updated table of clusters, updated glossary.

Example:

Cluster	Anatomic Part	
Description	Element that is part of the human anatomy.	
Elements	*Term*	*Source*
	Gingiva	O
	Ligament, periodontal	O
	Periodontal ligament	O
	Tissue	A
	Tooth	A

In the output document for this phase it is useful to keep track of the added terms: in the given example, in the *Source* column we specify with "A" that a term has been *added*, and with "O" if it belongs to the *original* terms. During this phase, it is possible to enrich the original glossary with other terms not contained in the original glossary.

Phase 3 – Relationship identification

Find *relationships* between terms, extracted from term definitions. Proceed a cluster at a time, searching for the distinctive relations of each cluster. Classify the relations under their respective cluster. An early informal description of the meaning and features (e.g., intended range and domain) of each relation should also be given.

Input: glossary, table of clusters.

Output: *table of relations*. Analogously to the table of clusters, it contains a grouping of the identified relations for each cluster and a description of their meaning.

Example:

Anatomic Part			
Relation	*Range*	*Domain*	*Description*
hasContinuity	Anatomic Part	Anatomic Part	Continuity between two anatomic parts.
isSupportOf	Anatomic Part	Anatomic Part	Relates a part with its supporting parts.
isSurroundingOf	Anatomic Part	Anatomic Part	Relates a part with its surrounding parts.
isPartOf	Anatomic Part	Anatomic Part	Relates a part to the parts it is direct component of.

Phase 4 – Disambiguation

Group terms with their synonyms (e.g., *see...*). Determine, among the synonyms, the *preferred term* to represent each *concept*.

Input: table of clusters.
Output: *table of concepts*. Refinement of the table of clusters in which terms are grouped to their synonyms.

Example:

Anatomic Part	
Concept	*Synonym*
Gingiva Periodontal ligament Tissue Tooth	Ligament, periodontal

Phase 5 – Class grouping

Find concepts that represent *subclasses* of the cluster they are in. Group concepts belonging to the new found classes. In general, try to create a partition of the cluster, possibly adding new classes at the same level of the found subclasses.

Input: table of concepts.
Output: table of concepts, updated with class groupings.

Example:

Anatomic Part		
Class	*Concept*	*Synonym*
Tissue	Gingiva Periodontal ligament	Ligament, periodontal
(Generic)	Tooth	

Phase 6 – Conceptual modelling

Using a graphical modelling formalism (such as E-R diagrams or UML models), proceed through the following steps:

1. *Skeleton schema*: group clusters in *areas*. Find abstract relationships between them, by generalization of the previously found relations.
2. *Refinement*: separately develop every area, starting from the subclasses found in the previous phase. Define inclusions between classes and relations and refine features and restrictions of the relations.
3. *Integration*: build the final model by combining area schemas and refining relationships among areas.

Input: table of concepts, table of relations.
Output: *ontology conceptual model*. A graphical representation of the conceptual structure of the domain ontology.

Example:
A skeleton schema for the clusters `AnatomicPart` and `AbnormalCondition` is shown in Figure 1. In Figure 2 we show the refinement for the class `AnatomicPart` and for its relation `hasContinuity`. By joining this with the refinement for `AbnormalCondition` (not shown), we obtain the final integration schema in Figure 3.

We do not put a constraint on the graphical language used to develop this phase, as long as it is enough usable and expressive to develop the ontology conceptual model: in our examples we have adopted an UML metamodel inspired to the one of [2].

Phase 7 – Schema representation

Represent class and property hierarchies in the final representation language (OWL, description logics).

Input: ontology conceptual model.
Output: *ontology schema*. Final representation of the class and property taxonomies.

Phase 8 – Ontology representation

Represent class instances as individuals; encode textual definitions for each concept by means of property values and restrictions.

Input: ontology conceptual model, ontology schema, glossary.
Output: *domain ontology*. Final representation of the domain ontology, complete with information about classes, relations and individuals.

The previous phase should be formalized to explain how to encode the textual definitions in terms of relations and logical connectives. This formalization clearly depends on the expressivity of the final representation language and its constructors.

Phase 9 – Annotation

Annotate each class and individual with information derived from glossary: the most relevant information is preferred term, synonyms, textual definition.

Input: domain ontology, table of concepts, glossary.
Output: *annotated domain ontology*.

Example:
An example of a domain ontology enriched with an external annotation is given in Figure 4.

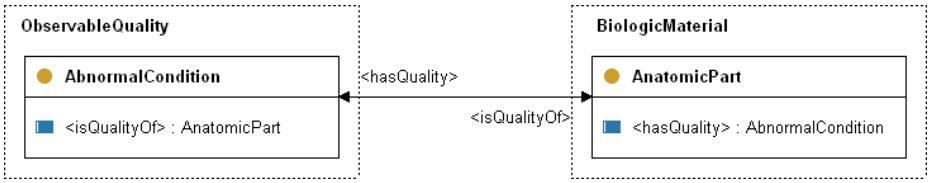

Fig. 1. Skeleton schema

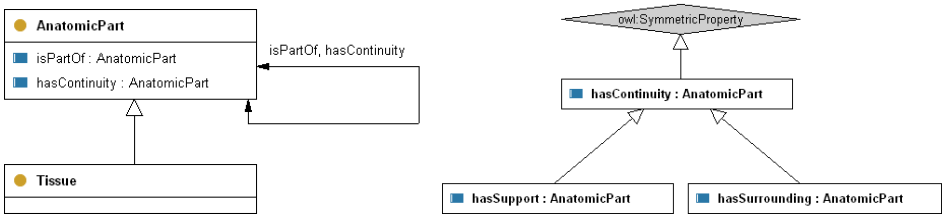

Fig. 2. Refinement for `AnatomicPart` and relation `hasContinuity`

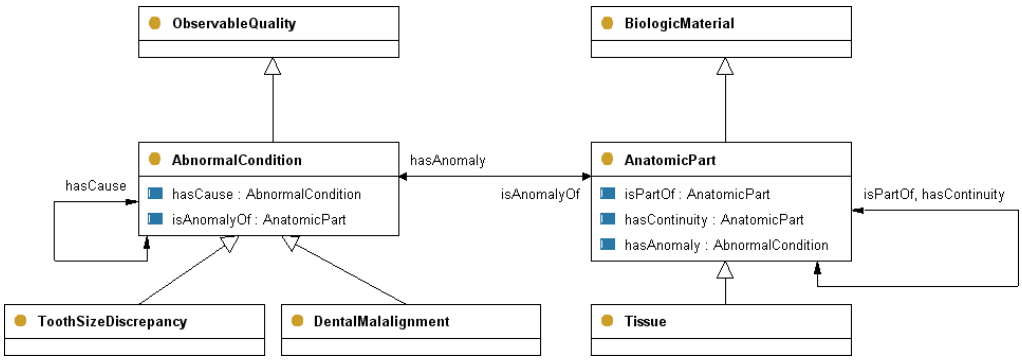

Fig. 3. Integration schema

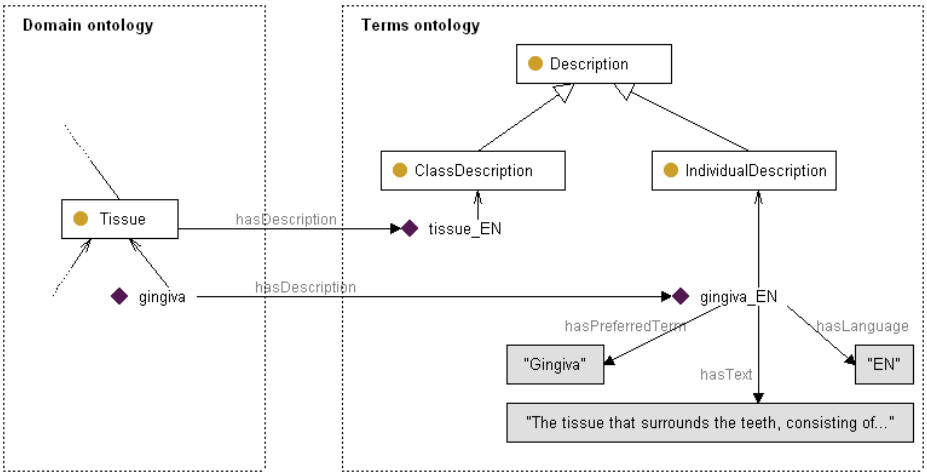

Fig. 4. Example of a domain ontology external annotation

With the last example we have shown a possible way to annotate the final domain ontology: we propose to include the annotations in a separate ontology, in which the *description* of each concept appear as an individual. For every description, the information about the described concept (such as its preferred term and synonyms) is linked as datatype properties. In this way, it is easy to uniformly associate different descriptions for each concept (for example, distinct by language or target users), regardless of the representation of the concept as a class, individual or property.

4 Related works

As ontologies have been applied in the modeling of domains of ever increasing complexity, a principled approach to their design is necessary. As such, the field of *ontology engineering* has become a rich and articulated research area (see, for example, the effort in the development of a general methodology for ontology networks in the EU funded project *NeOn*[2]). We give a brief outline of the field in this section: for more complete and detailed surveys, see [6, 13, 18].

Most of the proposals of general methodologies for ontology development are inspired by a top-down approach: well-known examples are the *DOLCE* methodology [11] and *Ontology Development 101* [12]. A slightly different approach is taken by *Methontology* [7], which differentiates the activities revolving around ontology development not only on a technical view, but also in the managing and supporting activities. Such attention on collateral activities is shared for example with *OTKM* [19], which develops the idea of *ontology requirements specification*

[2] `http://www.neon-project.org/`

document (ORSD). Recent methodology proposals, like *DILIGENT* [14], also introduce methods for ontology development in a collaborative setting.

Such general approaches are very interesting in that they show the benefits of a structured and domain independent approach when dealing with ontology development. However, it is widely accepted that taking into account relevant domain features from the beginning of the ontology development process yields more effective results: as such, more specialized, domain dependent methodologies have been proposed. Moreover, as emphasized in [18], the *reuse* of existing domain knowledge in ontology development allows to reduce the effort for the representation of the domain and permits to build ontologies based on already consensuated and well accepted knowledge sources. In fact, also from what has been noted in Section 1, our work can be located in the ontology engineering subfield dealing with *non ontological resource reuse and reengineering*, in particular in the definition of methodologies for reengineering a given set of (more or less structured) vocabularies or thesauri. Notable works in this area are [9, 16, 20], while in the related field of lexicon to ontology translation we cite [3–5, 10]. In a wider scope, by the definitions given in [18], this work can be seen as a proposal for a non ontological resource *reengineering pattern* for glossaries: this view also suggest that our methodology can be "plugged-in" in a more general framework for ontology development and lifecycle.

5 Conclusion and future works

In this paper we have discussed a methodology for the construction of a domain ontology on the basis of one or more (possibly incomplete) glossaries: we have also discussed how this methodology allows to build a complete terminology for the domain of interest. As a motivating example, we have applied it to a medical domain case. In the following, we summarize some of the directions for future work, both for the original project and for the methodology refinement. As for the orthodontic terminology development project, our next step is to complete the ontology with terms not yet considered, also by solving some current representation issues. Before proceeding to the publication and diffusion of the final ontology, the quality of the ontology has to be checked by domain experts. Also to help in this publication and valuation activity, we are currently developing a web application for the consultation of the ontology: the application will allow users to explore the ontology structure starting from terms in the final glossary. Our future work in the methodology refinement should include a formal definition for transformations from textual descriptions of concepts to their logic representation. Similarly, we need also to define formal definitions for methods for extracting relations from textual descriptions of concepts. Other interesting directions of work concern the analysis on the potential automatization of some phases and, on the ontology evaluation side, the study of the formal properties of the resulting ontologies. The first can involve the use of NLP techniques for clustering or saturation phases. Both automatization and evaluation can benefit from the application of our methodology in different domains.

References

1. ISO 1087-1:2000. Terminology. Vocabulary. Part 1: Theory and application.
2. S. Brockmans and P. Haase. A Metamodel and UML Profile for Networked Ontologies - A Complete Reference. Technical report, Universität Karlsruhe (TH), Apr. 2006.
3. P. Buitelaar, M. Sintek, and M. Kiesel. A Multilingual/Multimedia Lexicon Model for Ontologies. In *ESWC 2006*, pages 502–513, 2006.
4. F. Busa, N. Calzolari, A. Lenci, and J. Pustejovsky. Building a Semantic Lexicon: Structuring and Generating Concepts. In *Computing Meaning*, pages 29–51. Kluwer Academic Publishers, 2001.
5. P. Cimiano, P. Haase, M. Herold, M. Mantel, and P. Buitelaar. LexOnto: A Model for Ontology Lexicons for Ontology-based NLP. In *Proceedings of the OntoLex07 Workshop, held in conjunction with ISWC'07*, 2007.
6. M. Cristiani and R. Cuel. Methodologies for the Semantic Web: state-of-art of ontology methodology. *SIGSEMIS Bulletin*, 1(2):103–112, 2004.
7. M. Fernández-López, A. Gómez-Pérez, and N. Juristo. METHONTOLOGY: From Ontological Art Towards Ontological Engineering. In *Spring Symposium on Ontological Engineering of AAAI*, pages 33–40, 1997.
8. G. Hirst. Ontology and the Lexicon. In Staab and Studer [17], pages 209–230.
9. E. Hyvönen. Miksi asiasanastot eivät riitä vaan tarvitaan ontologioita? (why thesauri are not enough but ontologies are needed?). *Tietolinja*, (2), 2005. In Finnish.
10. A. Lenci. Building an Ontology for the Lexicon: Semantic Types and Word Meaning. In *Ontology-Based Interpretation of Noun Phrases*, pages 103–120, 2001.
11. C. Masolo, S. Borgo, A. Gangemi, N. Guarino, and A. Oltramari. Ontologies library (final). WonderWeb Deliverable D18, ISTC-CNR, Padova, Italy, Dec. 2003.
12. N. F. Noy and D. McGuinness. Ontology development 101: A guide to creating your first ontology. Technical Report KSL-01-05, Stanford KSL, 2001.
13. H. S. Pinto and J. P. Martins. Ontologies: How can They be Built? *Knowl. Inf. Syst.*, 6(4):441–464, 2004.
14. H. S. Pinto, S. Staab, and C. Tempich. DILIGENT: Towards a fine-grained methodology for Distributed, Loosely-controlled and evolving Engineering of oNTologies. In *ECAI 2004*, pages 393–397, 2004.
15. D. R. Poulton, C. J. Burstone, T. M. Graber, L. E. Keso, and D. L. Turpin. AAO Orthodontic Glossary. On-line version: `http://www.braces.org/healthcareprofessionals/dentists/glossary/`.
16. D. Soergel, B. Lauser, A. Liang, F. Fisseha, J. Keizer, and S. Katz. Reengineering Thesauri for New Applications: The AGROVOC Example. *Journal of Digital Information*, 4(4), 2004.
17. S. Staab and R. Studer, editors. *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, 2004.
18. M. C. Suárez-Figueroa. D5.4.1. NeOn Methodology for Building Contextualized Ontology Networks. NeOn Project Deliverable D5.4.1/v1.0, NeOn, Feb. 2008.
19. Y. Sure, S. Staab, and R. Studer. On-To-Knowledge Methodology (OTKM). In Staab and Studer [17], pages 117–132.
20. M. van Assem, M. R. Menken, G. Schreiber, J. Wielemaker, and B. J. Wielinga. A Method for Converting Thesauri to RDF/OWL. In *ISWC 2004*, volume 3298 of *LNCS*, pages 17–31. Springer, 2004.
21. F. van Harmelen and D. L. McGuinness. OWL Web Ontology Language Overview. W3C Recommendation, W3C, Feb. 2004. `http://www.w3.org/TR/2004/REC-owl-features-20040210/`.