(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

Ontology learning from thesauri: an experience in the urban domain

Javier Nogueras-Iso[1], Javier Lacasta[1], Jacques Teller[2], Gilles Falquet[3], Jacques Guyot[3]

[1]Computer Science and Systems Engineering Dept., University of Zaragoza, Spain
{jnog, jlacasta}@unizar.es

[2]LEMA, Université de Liège, Belgium
jacques.teller@ulg.ac.be

[3]Departement de systemes d'information,
Université de Genève, Switzerland
{Gilles.Falquet, guyot}@cui.unige.ch

Abstract. Ontology learning is the term used to encompass methods and techniques employed for the (semi-)automatic processing of knowledge resources that facilitate the acquisition of knowledge during ontology construction. This chapter focuses on ontology learning techniques using thesauri as input sources. Thesauri are one of the most promising sources for the creation of domain ontologies thanks to the richness of term definitions, the existence of a priori relationships between terms, and the consensus provided by their extensive use in the library context. Apart from reviewing the state of the art, this chapter shows how ontology learning techniques can be applied in the urban domain for the development of domain ontologies.

1. Introduction

The activity of knowledge acquisition constitutes one of the most important steps at the beginning of the ontology development process. This activity is essential in all the different methodologies for ontology design as a previous step to the conceptualization and formalization phases. As its name indicates, this activity is devoted to gather all available knowledge

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

resources describing the domain of the ontology and to identify the most important terms in the domain (Gandon, 2002).

To alleviate the work of knowledge acquisition there is an emerging interest in the study of methods and techniques for the (semi-)automatic processing of knowledge resources. The main aim of this automatic processing, known as ontology learning (Gómez-Pérez et al., 2003; Antoniou and van Harmelen, 2004), is to apply the most appropriate methods to transform unstructured (e.g., text corpora), semi-structured (e.g., folksonomies, HTML pages) and structured data sources (e.g., databases, thesauri) into conceptual structures (Gómez-Pérez and Manzano-Macho, 2003). The methods of ontology learning are usually connected with the activity of ontology population which also relies on (semi-)automatic methods to transform unstructured, semi-structured and structured data sources into instance data (i.e., instances of ontology concepts).

Among all the knowledge resources to be used as an input for ontology learning, thesauri, hierarchical classification standards and such taxonomies are likely the most promising sources for the creation of domain ontologies at reasonable costs (Hepp and de Bruijn, 2007). A thesaurus defines a set of terms describing the vocabulary of a controlled indexing language, formally organized so that the a priori relationships between concepts (e.g., synonymous terms, broader terms, or narrower terms) are made explicit. Additionally, the applicability of thesauri for search and retrieval in digital libraries has promoted the creation and diffusion of well-established thesauri in many different domains. Therefore, thesauri reflect some degree of community consensus and contain, readily available, a wealth of category definitions plus a hierarchy.

During the last years and even within the context of digital libraries and information retrieval, there is a general consensus about promoting the use of more elaborated ontologies. Ontologies with formal *is-a* hierarchies, frame definitions or even general logical constraints can improve the performance of retrieval systems. As (Fisher, 1998) remarks, the advantage of doing this transformation work between models is that combining formal ontologies with concept-oriented lexical databases can cover a spectrum of functionality which in principle includes all the traditional services of a classical thesaurus, and can offer more. (Soergel et al, 2004) remark that we need to change the use of thesauri into other more formal when at least one of the following requirements is needed:

- Improved user interaction with thesauri on both the conceptual and the term level for improved query formulation and subject browsing, and for more user learning about the domain.

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

- Intelligent behind-the-scenes support for query expansion, both concept expansion and synonym expansion, within one language and across languages.
- Intelligent support for human indexers and automated indexing/categorization systems.
- Support for artificial intelligence and semantic Web applications.

This chapter will be devoted to review the state of the art in ontology learning from thesauri, and to show how these techniques can be applied to practical examples in the urban domain. This domain is a quite interesting one since, although quite technical, it usually involves a number of different scientific disciplines, from architecture to law or transport engineering. Accordingly, it is not obvious to delineate a priori the set of concepts to be captured in an urban ontology which is well reflected by the intricacy and looseness of existing urban thesauri. Furthermore, urban systems have evolved quite rapidly over the last decades due to growing environmental concerns and metropolisation processes. Urban knowledge is hence a very "active" material, which constitutes a key challenge for existing thesauri and further justifies their transformation into formal ontologies.

This chapter will present two use cases showing the experience of transforming two urban thesauri employed in two closely-related bibliographic databases, called URBAMET and URBISOC. On the one hand, URBAMET is a bibliographic database developed by the French Centre for Urban Documentation for indexing bibliographic notes. The first version of the thesaurus for this bibliographic database was released in 1969 and it contained 2,300 terms. Nowadays, it contains around 4,200 terms and has been used for indexing some 230,000 technical documents related to urban development. On the other hand, URBISOC was developed by the Spanish National Research Council for the indexing of scientific and technical journals on Geography, Town Planning, Urbanism and Architecture. The thesaurus created for this database contains around 3,600 different concepts labelled in Spanish.

The rest of this chapter is organized as follows. Section 2 analyzes existent methods for ontology learning classified according to the type of source data. Then, section 3 describes experiences of transforming sources to ontologies in the urban domain. Finally, this chapter ends with some conclusions and ideas for future work.

2. State of the art in ontology learning from thesauri

Among the works related to the transformation of thesauri into ontologies, we must first cite a set of works that transform thesauri from its na-

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

tive format into Semantic Web languages such as RDF, OWL or SKOS (a W3C initiative for the representation of knowledge organization systems such as thesauri, classification schemes, subject heading lists, taxonomies, and other types of controlled vocabulary). The output of these methods can not be categorized as a formal ontology because the relations between concepts are still ambiguous, but at least it is a step forward. We move from the term-based approach recommended in ISO standards, in which terms are related directly to one another, into a concept-based approach. In the concept-based approach concepts are interrelated, while a term is only related to the concept for which it stands; i.e. a lexicalization of a concept.

Some examples of these works are (van Assem et al., 2004) and (van Assem et al., 2006) that describe the methods applied for transforming MeSH (Medical Subject Headings) and GTAA (Common Thesaurus for Audiovisual Archives) and IPSV (Integrated Public Sector Vocabulary) thesauri into RDF/OWL and SKOS. (Golbeck et al., 2003) describes the conversion of the NCI (National Cancer Institute) thesaurus into OWL format. Every thesaurus concept is translated into an OWL class. This thesaurus holds specific roles or relations (not usual BT/NT, USE/UF, RT) that are translated into specific RDF properties. (Wielinga et al., 2001) describes the transformation of the Art and Architecture Thesaurus (AAT) into an ontology expressed in RDFS. The full AAT hierarchy was converted into a hierarchy of concepts, where each concept has a unique identifier and slots corresponding with the main term and its synonyms.

A second set of works are more ambitious and try to transform the ambiguous BT/NT relations of thesauri into more formal relations such as *is-a* or *part-of* hierarchies. The ISO 2788 guidelines for monolingual thesauri contain a differentiation of the hierarchical relation into generic, partitive and instance relations. However, because the main purpose of thesauri was to facilitate document retrieval, the standards allow this differentiation to be neglected or blurred. But in contrast to thesauri, ontologies are designed for a wider scope of knowledge representation and need all these logical differentiations in relationships (Fisher, 1998). As stated in (van Assem et al., 2006) a major difference between thesauri and ontologies is that the latter feature logical *is-a* hierarchies, while in thesauri the hierarchical relation can represent anything from *is-a* to *part-of*. (Fisher, 1998) identifies several cases where this no differentiation of the BT/NT relation may be a source of fallacies or problems when transforming a thesaurus into an ontology. In particular this work focuses on the problems of identifying subsumption and instance relations behind the ambiguous BT/NT.

For instance, (Hepp and de Bruijn, 2007) describes an algorithm called GenTax to derive an RDF-S or OWL ontology from most hierarchical classifications available in the SKOS exchange format. This algorithm, implemented in the tool SKOS2Gentax[1], derives OWL classes from the instances of SKOS concepts and their broader and narrower relations. The algorithm assumes that SKOS concepts can be used in different contexts with varying semantics of the concepts and their relations. The algorithm has two main steps. Firstly, it creates two ontology classes per SKOS concept: one for the context of the original hierarchy, and a related second class (subclass of the first one) for the narrower meaning of the concept in a particular context. Secondly, GenTax inserts *subClassOf* relations between the classes in the original hierarchy context. However, since SKOS broader and narrower relations are translated by default to an is-a taxonomy, the output of the algorithm requires many corrections.

Other works use natural language processing to refine the hierarchical relation of thesauri. For example, (Clark et al., 2000) describes the experience of transforming a technical thesaurus (Boeing's technical thesaurus) into an initial ontology. In particular, this work introduces algorithms for enhancing the thesaurus connectivity by computing extra subsumption and association relations. An important characteristic of technical thesauri is that many concept names are compound (multi-word) terms. They implemented a graph enhancement algorithm for this task, which automatically inferred these missing links using word-spotting/natural language processing technology. Additionally, they also used natural language processing to refine the RT relation into finer semantic categories.

Another remarkable work with the aim of automating the refinement of relations is the one of (Soergel et al., 2004; Kawtrakul et al., 2005). It introduces a semi-automatic approach for detecting problematic relations, especially BT/NT and USE/UF relations, and suggesting more appropriate ones. Upon the experience obtained with the transformation of AGROVOC into an ontology, their approach is mainly based on the identification of patterns and the establishment of rules that can be automatically applied. The method is based on three main ideas. Firstly, they try to find expert-defined rules. Assuming that concepts are associated with categories (e.g., geographic term, taxonomic term for animals ...), experts may define rules that can be generally applied to transform BT/NT relations of concepts under the same category into *is-a* or *part-of* hierarchies. Secondly, they propose noun phrase analysis to detect *is-a* hierarchies. If two terms in a BT/NT relation share the same headword, this relation can be transformed into *is-a*. Alternatively, if two terms are in the same hierarchy of hypernyms in Wordnet, their relation is also transformed into *is-a*. Thirdly, in the case of RT relations, which usually are under-specified relations, refinement rules, acquired from experts and machine learning, are

[1] http://www.heppnetz.de/projects/skos2gentax/

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

applied. If we identify a particular case of conversion of an RT relation between two terms, we may derive a general rule for the hypernyms of these two particular terms and apply it again to all their hyponyms related through RT.

3. Applicability of ontology learning in the urban domain

3.1. The URBAMET use case

This subsection presents a methodology for the analysis of an urban thesaurus that should lead to the incremental development of a shared urban ontology.

URBAMET (URBAMET, 2009) is a bibliographic database created and maintained by the French Centre for Urban Documentation. The corpus currently includes 280'000 documents and is fed with an additional 8'000 documents each year. Originally designed in 1969 with 2'300 terms, the URBAMET thesaurus currently includes 4'200 terms, which are used to index the document corpus. It is a hierarchy of terms with 24 main themes (top level categories). Figure 1 shows the main themes and an excerpt of the hierarchy of sub-domains in the field of transportation.

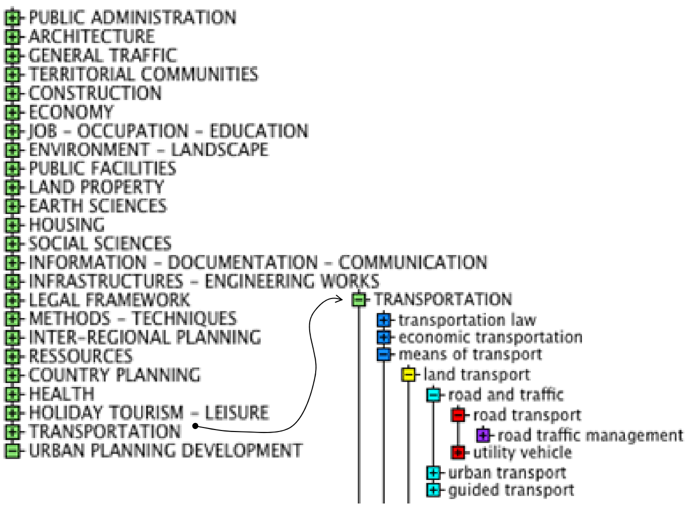

Fig. 1. Main themes (domains) of the URBAMET thesaurus

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

It can be observed on the figure that the terms in URBAMET denote either concepts or sub-domains. For instance, the term “utility vehicle” may denote a concept that has an intension (the properties of a utility vehicle) and an extension (the set of all utility vehicles). Conversely, the term “road and traffic” can hardly denote a concept: it is difficult to figure out what is an instance of “road and traffic”. Moreover “road and traffic” cannot be considered as a specialization of its parent term “land transport”. Hence, the URBAMET thesaurus, at least on the first levels, is mostly a hierarchy of sub-domains. As a consequence, it does not provide an immediate starting point or structural backbone for the construction of an urban ontology.

Since the thesaurus cannot be directly used to build an ontology, the proposed methodology relies on the existing thesaurus and the indexed document corpus. The document classification induced by the thesaurus is analyzed with an automated document classifier. This tool operates on document contents. Initially a training corpus is used to teach the classifier the class concepts. Then the tool can start classifying other documents. The analysis is performed in the following steps:

1. Extracting the corpus
2. Building up the training catalogue
3. Training and validating the classifier
4. Generating and analyzing the confusion matrix (list of mistakes made by the classifier)
5. Generating the Top-50 terms (list of the most classifying terms)

For the creation of the training corpus, around 10,000 abstracts, together with their manual assigned themes, have been extracted from URBAMET. This means about 70 indexed words per document and a final vocabulary of about 18,000 words (stems). Then, the classifier built a neural network by reading the training files and applying the Winnow learning technique. Figure 2 depicts an excerpt of the neural network classifier employed to analyze the correspondence between the set of terms in the abstracts of the URBAMET database and the main themes (domains) assigned to the documents. The neural network contains weighted arcs from a word or pair of words to a domain. The weight of term *i* for domain *j* represents how strongly *i* draws a document to *j*. The neural network was trained with 80% of the corpus, using the remaining 20% for testing purposes. As a result of the performance of the generated classifier, the classifier discovers the main domain of each tested document with probability: 59% for the first proposed domain; 16% for 2nd choice; 7 % for 3rd choice. Hence, the classifier has a probability of 82% to find the correct class in the first 3 proposals (a random choice would give a 23% probability).

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

In general, it can be stated that the classifier is effective: the URBAMET classification corresponds to the text contents. However, to detect possible problems and restructure the domains, the methodology proposes an analysis based on the creation of confusion matrices. The objective is to find domains which are poorly classified. Figure 3 shows 2 excerpts of the complete in-out 24x24 matrix. Each cell M_{ij} represents the percentage of document in domain i classified in j. Ideally M_{ii} should be 100%. On the one hand, this confusion matrix indicates not clearly separated domains. For instance, see the confusion between *Traffic* and *Transportation* in figure 3 (top part). Probably, it would be a good idea to merge the domains and create new subdomains. On the other hand, this matrix also shows orthogonal or transverse domains. For instance, *Legal framework* and *Methods* are orthogonal to the other domains (see bottom part of figure 3). Documents are rarely only about *Law* or *Methods*, they usually present legal aspects of *Urbanism*, *Transportation*, etc.

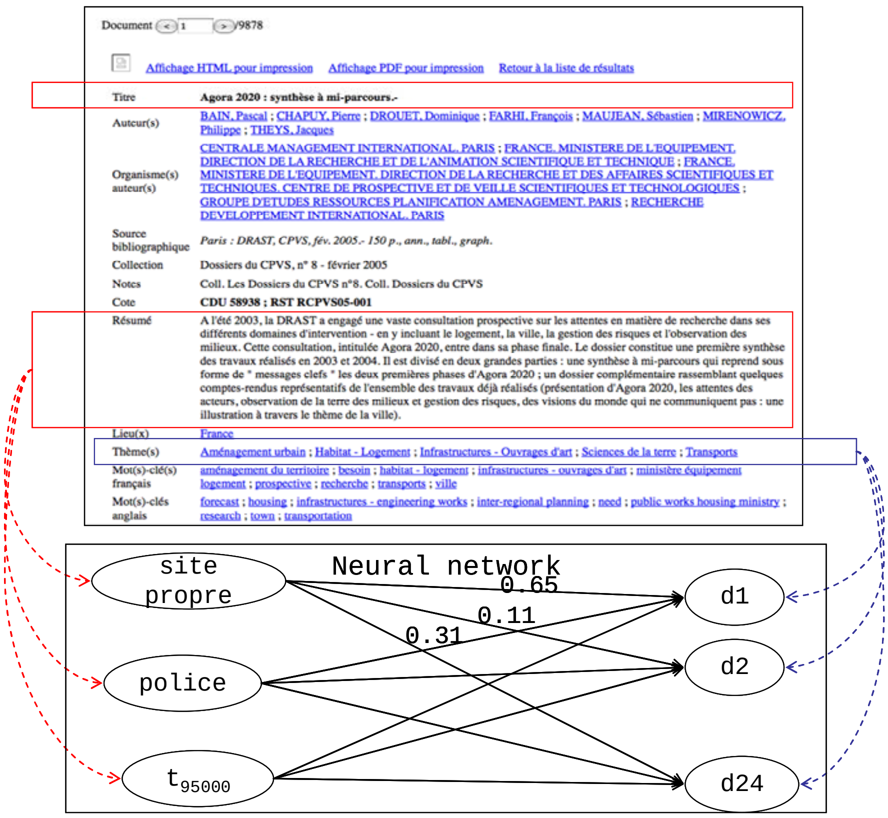

Fig. 2. Neural network classifier for the URBAMET thesaurus

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

Neural networks can be criticized because of the lack of explanation on the fact that the classifier chose a particular class (as compared to rule-based engines which can explain their reasoning). However, it is always interesting to analyze, for each class, the list of the most heavily weighted terms. This list is a selection of the "champion terms" of the domain. They are good candidate to designate concepts of the domain.

Comparing this list with the thesaurus terms provided interesting insights into each domain. For instance, among the 50 most classifying terms of the Environment domain, 34 were not found in the thesaurus. This probably indicates that this domain has evolved and that this evolution is not reflected in the current thesaurus.

In \ out	Transportation	Traffic	Tourism	…
Transport	45%	24%	3%	
Circulation	10%	40%	1%	
Tourism	1%	1%	49%	

In \ out	Legal	Methods	Urbanism	Infra…
Legal	8%	3%	5%	3%
Methods	2%	4%	4%	13%
Urbanism	17%	14%	24%	4%
Infrastructure	2%	11%	1%	22%

Fig. 3. Some results obtained from the confusion matrix

This study shows that even if a thesaurus is far from having a clear ontological structure, it can be exploited to create an ontology if it is considered together with the document corpus it indexes. In fact, such thesauri are the superposition of a domain/sub-domain hierarchy and some ontological elements. Some terms, like means of transport designate at the same time a domain (everything that is related to means of transport) and a concept (a system for transporting people or objects). The confusion between domains and concepts and between different URBAMET domains may be related to the incremental development of the thesaurus, which currently reflects diverging rationales and methodologies. A methodology for extracting an ontology from such a thesaurus must literally *extract* the ontological elements, as was done by selecting the most classifying terms of each domain.

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

3.2 The URBISOC use case

This subsection presents the work done to transform an urban thesaurus into a more formalized model. The urban thesaurus employed as use case has been the one developed by the Spanish National Research Council to facilitate the classification of a bibliographic database called URBISOC, which is specialized in scientific and technical journals on Geography, Town Planning, Urbanism and Architecture (Alvaro-Bermejo, 1988). This thesaurus, called URBISOC from now on, contains around 3,600 different concepts labelled in Spanish. It is very close to URBAMET in its scope and use though it has been developed in a much shorter period and its design has been trusted to a reduced number of domain experts in charge of safeguarding its consistency.

Apart from the formalization objective, there were two additional goals for the transformation of this thesaurus. On the one hand, we wanted to convert it into a multilingual resource. On the other hand, we wanted to enrich it with more concepts. It must be taken into account that urbanism can be considered as an intersection of different domain areas such as economics, politics culture or civil engineering.

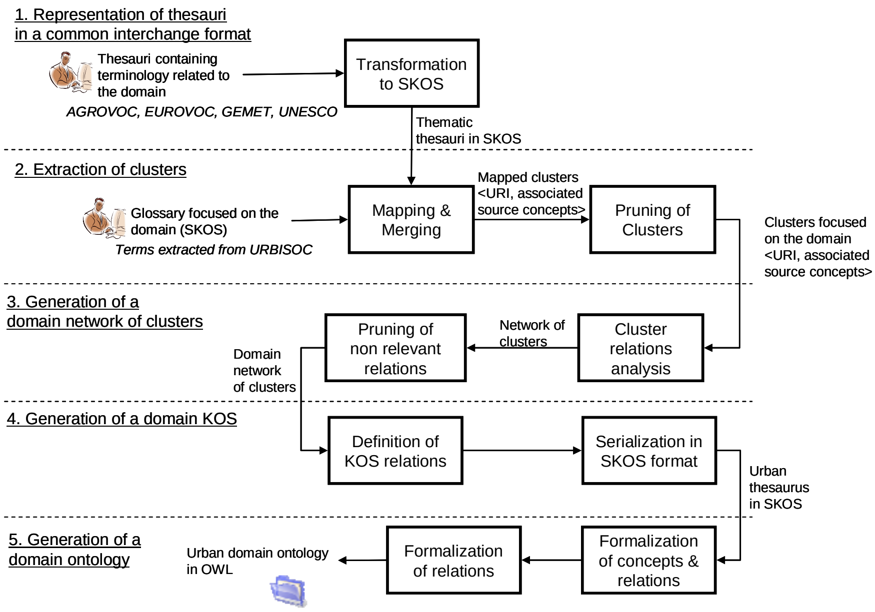

Fig. 4. Workflow for the generation of an urban domain ontology

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

The transformation methodology proposed is based on the merging of source thesauri containing concepts from cross-domain areas. The method takes as input a set of different thesauri and obtains as a result a more consistent and formalized ontology. Figure 4 remarks the 5 main steps involved in this process, showing the inputs and the produced results:

1. Representation of input thesauri in a common format. This task is devoted to the transformation of the input thesauri into SKOS (Nogueras et al, 2007), a W3C initiative for the representation of Knowledge Organization Systems. The thesauri used as input for the method are: GEMET (the GEneral Multilingual Environmental Thesaurus of the European Environment Agency)[2] , AGROVOC (the FAO Agricultural Vocabulary)[3] , EUROVOC (the European Vocabulary of the European Communities)[4] and the UNESCO thesaurus[5] . They provide a shared conceptualization in the areas of economics, politics, culture and environment.
2. Extraction of clusters. This is the main step and consists in the detection of intersections between concepts in the different input thesauri, through the analysis of their lexical similarities and making profit of their multilingual support. Each set of mapped concepts is grouped into a cluster, which is the name given to a concept in the output ontology. A cluster represents a group of equivalent concepts and is identified with one of the URIs of the original concepts. But previous to this and because top terms of input thesauri are usually very generic, we must identify core concepts specific to the knowledge area in the cross-domain thesauri. Thus, a reduced set of terms in the knowledge area, which is extracted from URBISOC (the *urban planning* term and the recursive chain of *related* and *narrower* terms), is added as another input in the merging process to focus on the domain.
   Additionally, not all the clusters obtained in the mapping process are useful; many clusters contain terms not related to the desired domain. Therefore, only the clusters that contain a concept from the selected list of terms and those with at least one concept directly related (through *broader*, *narrower* and *related* relations) to another one in a cluster of the first case are kept. The rest are considered as not relevant and they are pruned from the system.
3. Generation of a domain network of clusters. This step consists in connecting the clusters previously extracted. The relations between the con-

[2]http://www.eionet.europa.eu/gemet
[3]http://www.fao.org/aims/ag_intro.htm
[4]http://europa.eu/eurovoc/
[5]http://www.ulcc.ac.uk/unesco/

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

cepts assigned to the different clusters are converted into relations between the clusters that contain them. The relations between clusters are labelled with: the types of relations, which are derived from the original types of relations between concepts; and a weight that represents the number of occurrences for each original relation type between the concepts of the inter-related clusters.
   Besides, it must be noted that the output network may be still too complex and/or contain spurious clusters. Therefore, a process to prune the less relevant relations has been created. This process receives as input the complete network of concepts and a weight threshold to determine if a relation is maintained. All the relations with a weight below the threshold are pruned. After the pruning, all the clusters that do not have at least one relation with another one are also eliminated.
4. Generation of a new thematic thesaurus. The generation of the thesaurus consists in taking the clusters of the network and organizing them into a hierarchical model. The clusters are transformed into concepts of the new thesaurus; one of the labels of the original concepts within the cluster is selected as preferred label. With respect to the thesaurus structure, each relation is marked with the type that has more occurrences. Additionally, those concepts that do not have *broader* relations are marked as top terms. Finally, the generated structure is reviewed to verify that the BT/NT relations structure does not contains cycles. If any cycle is found, it is removed by replacing the BT/NT relation that generates the cycle by a *related* relation.
5. Formalization of the thematic thesaurus. In order to transform the obtained thesaurus into a formal model the following tasks have been performed: transformation of each thesaurus concept into a class, identification of relations with higher semantics (*is-a*), and serialization into OWL format. The transformation of the thesaurus concepts into OWL classes requires the transformation of their identifiers, and the registration of their preferred and alternative labels as *rdfs:label* properties. With respect to the relations, to determine which narrower relations can be transformed into *is-a* relations, the following heuristic has been used: "a narrower relation is transformed into an *is-a* relation if the related concepts contain the same headword (substantive) in at least one of their labels (preferred or alternatives) in any of the available languages". The relations that are not transformed are left as they were and must be manually converted.

Table 1 shows the results obtained from the formalization process. Each row represents a possible output ontology according to the weight threshold used for pruning non relevant relations in the third step of the process.

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

For each output ontology, table 1 informs about: the number of concepts, the number of *RT* relations, the number of original BT/NT relations obtained in step 4, and the percentage of *is-a* relations derived from these *BT/NT* relations in step 5.

Table 1. Features of the output ontologies

Weight threshold	Nr concepts	Nr *RT*	Nr *BT/NT*	Nr *is-a*	% *is-a*
1	4698	13992	4297	890	20 %
2	2568	4150	1480	698	47 %
3	1514	1266	857	455	53 %
4	1082	566	552	318	57 %
5	681	302	317	195	61 %

The output ontologies cannot be considered as a final work, but they are a helpful resource for ontologists and experts in the domain. On the one hand, the ontology with weight thresshold 1 can be used to explain the relation between *urban planning* (seed concept in URBISOC to focus on the domain, see step 2) and other concepts that, at first sight, might seem far related. For instance, Figure 5 shows three possible paths to explicate the connection between the *recycling* concept and the *urban planning* concept. On the other hand, as the thresshold increases, the derived ontologies help to discard spurious concepts, which are only considered in some of the original input thesauri. However, it must be taken into account that the increase of weight threshold implies as well a decrease of contributions from other domains. On the positive side, it can be observed that an increasing threshold leads to a higher ratio of *BT/NT* relations that can be clearly identified as *is-a* relations.

Path 1:	**Path 2:**	**Path 3:**
recycling	**recycling**	**recycling**
RT	*BT*	*RT*
returnable container	waste treatment	water reuse
BT	*BT*	*RT*
containers	environmental policy	water management
RT	*NT*	*BT*
transport	land use planning	environmental policy
RT	*NT*	*NT*
urban transport	**urban planning**	land use planning
RT		*NT*
urban planning		**urban planning**

e

Fig. 5. Three possible paths to understand the connection between *recycling* and *urban planning*

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

4. Conclusions

Although to build high-quality ontologies, some kind of manual processing is always required, there are ontology learning methods that can alleviate the task of ontology construction. This chapter has been devoted to present different ontology learning methods that make profit of existent thesauri for building ontologies. In general, we must say that there are not industrial applications for ontology construction. Quite the opposite, depending on the application domain and the availability of sources, ontologists must choose the best ontology learning method in each case.

Additionally, this chapter has shown two different use cases in the context of the urban domain. On the one hand, the URBAMET use case has demonstrated the use of automated classification, with a neural network, for: evaluating the quality of the thesaurus hierarchy (in terms of concept overlap and confusion), finding parts that must be re-structured; and identifying new emerging terms that correspond to new concepts already present in the documents but not yet introduced in the thesaurus. On the other hand, the URBISOC use case has presented a method that takes as input a set of different thesauri and obtains, as a result of a merging and pruning process, a more consistent and formalized ontology with multilingual support.

It has not yet been possible to apply the same methodology to both thesauri because required data was not available in the same way in the two use cases. However, applying a neural network classifier approach to both URBAMET and URBISOC document corpus would provide very interesting information about possible convergences and divergences between the content of these bibliographic databases and the ontology resulting from their analysis.

Comparing the concepts obtained in both use cases would further constitute a promising avenue for building multi-lingual ontologies in the urban domain. Multi-lingual issues are especially challenging in this domain. Urban conceptualisations are traditionally based on a mix between science and practice, which makes it hard to find perfect matches for high-level concepts like planning documents or types of interventions between different regions.

Finally ontologies resulting from automatic classification exercises may be submitted to a review amongst a sample of domain experts as well as by persons in charge of the indexation of new documents. This would certainly provide additional information about the relevance of designed methodologies for end-users as well as about possible divergences between domain experts and thesauri designers/managers.

(Draft) Ontology Theory, Management and Design: Advanced Tools and Models. 2010, p. 247-260.

Acknowledgements

This work has been partially supported by the COST UCE C21 Action (Urban Ontologies for an improved communication in Urban Civil Engineering projects) of the European Science Foundation.

References

Alvaro-Bermejo C (1988). Elaboración del Tesauro de Urbanismo URBISOC. Una Cooperación Multilateral. In *Encuentro Hispano-Luso de Información Científica y Técnica. II*, Salamanca.

Antoniou G, van Harmelen F (2004). *A Semantic Web Primer*, chapter Ontology engineering, pp 205–222. Massachusetts Institute of Technology.

Clark P, Thompson J, Holmback H, Duncan L (2000). Exploiting a thesaurus-based semantic net for knowledge-based search. In *Proc 12th Conf on Innovative Application of AI (AAAI/IAAI'00)*, pp 988–995.

Fisher DH (1998). From thesauri towards ontologies? In el Hadi WM, Maniez J, Pollitt SA, (eds), *Structures and relations in knowledge organization: proc. 5th Int. ISKO Conference*, pp 18–30, Lille, France.

Gandon F (2002). *Distributed Artificial Intelligence and Knowledge Management: ontologies and multi-agent systems for a corporate semantic web*. Scientific Philosopher Doctorate Thesis In Informatics, INRIA and University of Nice - Sophia Antipolis, Doctoral School of Sciences and Technologies of Information and Communication (S.T.I.C.)

Gómez-Pérez A, Fernández-López M, Corcho O (2003). *Ontological Engineering*, chapter Methodologies and Methods for Building Ontologies. Springer-Verlag, London (United Kingdom).

Gómez-Pérez A, Manzano-Macho D (2003). A survey of ontology learning methods and techniques. Deliberable 1.5, OntoWeb Consortium.

Golbeck J, Fragoso G, Hartel F, Hendler J, Parsia B, Oberthaler J (2003). The National Cancer Institute's thesaurus and ontology. *Journal of Web Semantics*, 1(1):1–5.

Hepp M, de Bruijn J (2007). Gentax: A generic methodology for deriving OWL and RDF-S ontologies from hierarchical classifications, thesauri, and inconsistent taxonomies. In *LNCS, Proceedings of the 4th European Semantic Web Conference (ESWC 2007)*, volume 4519, pp 129–144, Innsbruck, Austria. Springer.

Kawtrakul A, Imsombut A, Thunkijjanukit A, Soergel D, Liang A, Sini M, Johannsen G, Keizer J (2005). Automatic Term Relationship Cleaning and Refinement for AGROVOC. In *Workshop on The Sixth Agricultural Ontology Service*, Vila Real, Portugal.

Lacasta J, Nogueras-Iso J, López-Pellicer FJ, Muro-Medrano PR, Zarazaga-Soria FJ (2007). ThManager: An Open Source Tool for creating and visualizing SKOS. *Information Technology and Libraries (ITAL)*, 26(4):40–53.

Soergel D, Lauser B, Liang A, Fisseha F, Keizer J, Katz S (2004). Reengineering Thesauri for New Applications: the AGROVOC Example. *Journal of Digital Information*, 4(4):1–19.

Studer R, Benjamins VR, Fensel D (1998). Knowledge engineering: principles and methods. *Data & Knowledge Engineering*, 25(1-2):161 – 197.

URBAMET (2009). URBAMET home page. http://www.urbamet.com/ Retrieved on 10 Feb. 2009

van Assem M, Malaisé V, Miles A, Schreiber G (2006). A Method to Convert Thesauri to SKOS. In *Proceedings of the 3rd European Semantic Web Conference (ESWC-06)*, Budva, Montenegro.

van Assem M, Menken MR, Schreiber G, Wielemaker J, Wielinga B (2004). A method for converting thesauri to RDF/OWL. In McIlraith SA, Plexousakis D, van Harmelen F, (eds), *Proceedings of the Third International Semantic Web Conference (ISWC 2004)*, Hiroshima, Japan. Springer.

Wielinga BJ, Schreiber AT, Wielemaker J, Sandberg JAC (2001). From Thesaurus to Ontology. In *Proceedings of the 1st international conference on Knowledge capture*, pp 194 – 201, Victoria, British Columbia, Canada.