

Generating an urban domain ontology through the merging of cross-domain lexical ontologies

J. Lacasta, J. Noguera-Iso, F.J. Zarazaga-Soria, P.R. Muro-Medrano
Computer Science and Systems Engineering Department, University of Zaragoza,
María de Luna 1, E-50018 Zaragoza, Spain
{jlacasta, jnog, javy, prmuro}@unizar.es

Abstract: In order to classify resources, digital libraries have traditionally used different types of lexical ontologies, which describe the terminology used in an area of knowledge. This paper analyzes how lexical ontologies covering different areas of knowledge can be merged to generate an enriched urban terminology. This work proposes a method to combine these different perspectives into a single network of urban related concepts. The objective of this network is to facilitate a draft for a more formal (non lexical) urban domain ontology.

Key words: Urbanism, Lexical Ontologies, Ontology Mapping

1. Introduction

Urbanism is usually defined as the study of cities including their economic, political, social and cultural environment. As it can be observed from this definition, this discipline could be considered as an intersection of different domain areas such as economics, politics culture or civil engineering. One way to represent the knowledge behind urbanism is by means of the use of ontologies. The term ontology is used in information systems and in knowledge representation systems to denote a knowledge model, which represents a particular domain of interest. According to (Gruber, 1993) an ontology is “an explicit formal specification of a shared conceptualization”. Therefore, given the multidisciplinary character of urbanism, the development of an urban domain ontology requires a revision of all the aforementioned cross-domain areas, capturing the concepts directly involved with the built environment of urbanism.

The purpose of this paper is to reproduce this exercise of revising and merging the knowledge from different domains in order to obtain a better definition of the urban domain. In particular, this work proposes a method for the definition of an urban domain ontology through the merging of thesauri representing the knowledge behind different domains. A thesaurus is a lexical ontology that 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 (ISO, 1986). The applicability of thesauri for search and retrieval in digital libraries has promoted the creation and diffusion of

These different types of analysis are based in only two different procedures: the analysis of the linguistic similarity between labels and the analysis of the relation between classes. The differences between the available mapping procedures are the techniques used to identify the similarities and the types of analysis considered.

Lexical ontologies have some particularities in their structure with respect to other types of ontologies. They consist of a set of lexical concepts that share a reduced set of property types and relation types. For example, the thesauri structure described in ISO-2788 (ISO, 1986) and ISO-5964 (ISO, 1985) standards is reduced to alternative labels for a lexical term in different languages and a reduced set of possible relation types (*narrower-broader* and a general *related* type that provides little semantic).

As commented in (W3C, 2005), representing each concept of a lexical ontology as a different class produce several problems for its use in resource classification. Therefore, the most usual way to model lexical ontologies is represent each concept as an instance of a general “Concept” class, which define the available types of properties and relations. In this model, instances of a class can be directly used as values of properties in the description of a resource, and therefore, it has been used to create many different lexical ontology formats (Miles et al., 2005; Lauser et al., 2006; Miller, 1990). Its main drawback is that the provided ontological structure is very poor (only one class). The generalized use of this model to represent thesauri makes unnecessary the use of mappings techniques to relate their structure (they are equivalent). Therefore, all the mapping work has been focused in the analysis of similarity between instances.

An additional problem to compare lexical ontologies is the format used to represent them. For decades, the evolution of digital libraries has encouraged the use of lexical ontologies describing the terminology of an area of knowledge in the form of taxonomies, classification schemes or thesauri, promoting in that way the creation and diffusion of well-established collection in different domains. However, the lack of standardization has produced a huge variety of incompatible formats that increase the complexity of the comparison process.

3. Method for the generation of an urban domain ontology

Urbanism can be considered as an intersection of different domain areas such as economics, politics culture or civil engineering. In this context, the process to develop an urban domain ontology, providing explicit and formal specification of the knowledge behind the urbanism discipline, makes necessary to revise all these cross-domain areas and capture all the relevant concepts.

This section describes the process to capture the structure of relations between urban concepts through the analysis and comparison of cross-domain lexical ontologies with a thesaurus structure. The result obtained is a network of related urban concepts that shows the relevance of concept relations. Figure 1 remarks the different steps of the process, showing the inputs and the produced results. Four different tasks can be highlighted and are described in detail in the following subsections: the harmonization of the interchange format used for thesauri, the

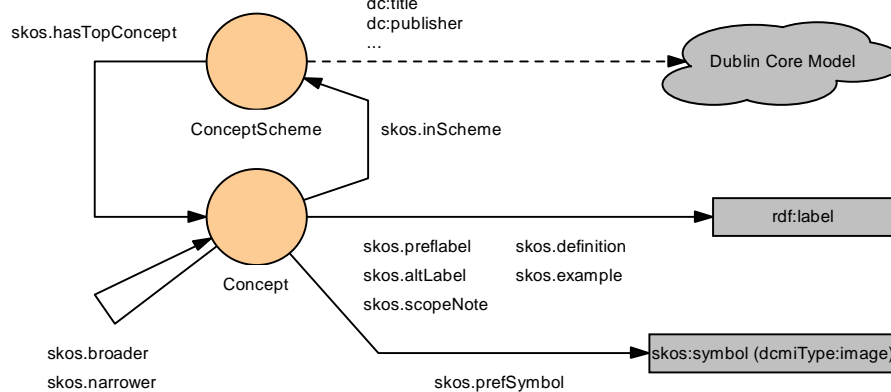


Figure 2: SKOS-Core Model

In order to facilitate the transformation of the different analyzed thesauri into SKOS, we have developed a customizable tool (developed in Java) that converts several file formats, according to a traditional thesaurus model, into the SKOS representation format. Figure 3 shows the mapping established by the tool between a classical thesaurus model and the SKOS representation model. The Unified Modelling Language (UML) notation is used for the representation of the two models. As it can be observed from the figure, the following transformations are applied:

- A concept scheme is created to represent the source thesaurus.
- Each thesaurus preferred term generates a new concept in the SKOS representation (except if it is not used for classification).
- Each translation derives a new preferred label in the language of the label.
- Each term related by a UF/USE relation (synonymy relation) is converted into an alternative label of the related concept.
- The RT relations between terms are converted to *skos:related* relations between the corresponding concepts. The same happens with the BT and NT relations that are converted to *skos:broader* and *skos:narrower*.
- The description of a term is converted into the definition of its associated concept.
- The concepts whose associated term is marked as TT are included in the concept scheme as top terms.
- Another important item is the URI that must be created for each SKOS concept. It has to be generated by using the information provided by the source format. In the example, the term value used as the preferred label of the concept can be converted into an URI by adding an *http://* prefix.

It is worth noting that this tool can be customized to different source formats (text files, relational databases). It provides a common infrastructure for the parsing of source formats and the writing of SKOS output files (in RDF format). Therefore, the support of a new source format is reduced to the development of a new plug-in

set of terms of a thesaurus specialized in urbanism is used as seed for this search. In addition, the relations between the concepts present in the urban thesaurus are used in the next step as a base for the construction of urban domain ontology.

As commented previously, from the available mapping techniques, only the analysis of the values of the properties of the instances is useful. Here, the linguistic similarity between the preferred and alternative labels has been considered for the mapping. The analysis of other properties as definition, scope notes and relations is left for future work.

In the mapping process, every concept of every thesaurus (including the urban one) is compared with every concept of the other thesauri to find equivalences. Two concepts are considered equivalent when at least one of the labels of a concept (preferred and alternatives) is equal to a label in the other concept. Here, the use of multilingual thesauri has the advantage of having labels in different languages to compare (the labels used to describe two concepts may differ in a language but be equivalents in other one). In order to improve the results, plurals, accents and capital letters have been removed. This approach could be enriched with misspellings detection, stemming and word order analysis among others, but given the strict rules used to define the labels used in a thesaurus no much improvement would be expected.

Equation 1 measures the relevance of the mapping obtained between two concepts. The higher the number of labels two concepts share from the total they have, the higher is their equivalence. This equation can be applied to each obtained mapping, but it is used here to analyze the quality of the thesaurus mapping with respect to the urban one showing the relevance that each different knowledge area gives to urbanism (see the experiments in section 4).

$$probabilityOfEquivalence = \frac{2 * numberOfMatchedLabels}{totalNumberOfLabelsInTheTwoConcepts} \quad (1)$$

Each set of mapped concepts is grouped into a cluster (group of equivalent concepts), which is identified with one of the URIs of the original concepts. Figure 4 shows a simplified example of the cluster generated for the “Zonas Urbanas” concept (“Urban areas”) mapping different thesauri and considering only the Spanish labels. In the example, it can be seen that the “Area Urbana” concept of GEMET is included in the cluster thanks to the presence of this label in the concepts of EUROVOC and AGROVOC. In addition, the relevance of the mappings is included to show that some of them are stronger than others.

Not all the clusters obtained in the mapping process are useful; many contain concepts not related to urban terminology. Therefore, only the clusters that contain a concept from the urban thesaurus and those with at least a concept directly related (*broader*, *narrower* and *related* relations) to another one in a cluster of the first case are stored. The rest are considered as not relevant to urbanism and they are pruned from the system. To maintain the consistency, the relations of the remaining concepts with the deleted ones are also eliminated.

the source thesauri. See figure 5 in the experiment section as an example of the obtained network.

In many situations, it is not interesting to have a network with all the existent relations but only the most important ones. 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.

3.4. Serialization and visualization of the urban domain ontology

For the serialization of the generated structure, we have proposed the use of the Web Ontology Language (OWL) (Bechhofer et al., 2004) and XTM format (Pepper and Moore, 2001). On the one hand, OWL is a widely accepted language for the definition of formal ontologies based on RDF. On the other hand, XTM is a format for the exchange of topic maps with an emphasis on the find-ability of information. We have selected XTM because of its advantages for the visualization and navigation through the generated network of concepts. It can be easily visualized by a wide range of tools compliant with this format. For instance, we have selected the TMNAV tool created in the TM4J project (TM4J, 2001) but other tools could also have been used.

4. Testing the method in the urban domain

The process described previously has been used to generate a network of urban concepts using GEMET, AGROVOC, EUROVOC and UNESCO as thematic thesauri. These thesauri provide a shared conceptualization in the areas of economics, politics, culture and environment: EUROVOC is a multilingual thesaurus covering the fields in which the European Communities are active (it provides a means of indexing the documents in the documentation systems of the European institutions and of their users); GEMET is a thesaurus for the classification of environmental resources developed by the European Environment Agency and the European Topic Centre on Catalogue of Data Sources; AGROVOC is a specialized thesaurus for the classification of geographic information resources (with special focus on agriculture resources), which has been created by the Food and Agriculture Organization of the United Nations; and UNESCO is a general purpose thesaurus for use in the indexing and retrieval of information in the UNESCO Integrated Documentation Network. The different origins and objectives of these thesauri provide different views of the urban terminology they contain.

From the available thesauri about urbanism, URBISOC (Alvaro-Bermejo, 1988) was selected as a basis for the filtering of urban terminology. URBISOC has been developed by the Spanish National Research Council to facilitate classification at bibliographic databases specialized in scientific and technical journals on Geography, Town Planning, Urbanism and Architecture. This thesaurus contains around 3,600 different concepts labelled in Spanish.

These five thesauri have been published in completely different representation formats. UNESCO and AGROVOC are stored in a database format, but each one with

commented in the description of the process and shown later, the less relevant ones could have been deleted to obtain a smaller structure.

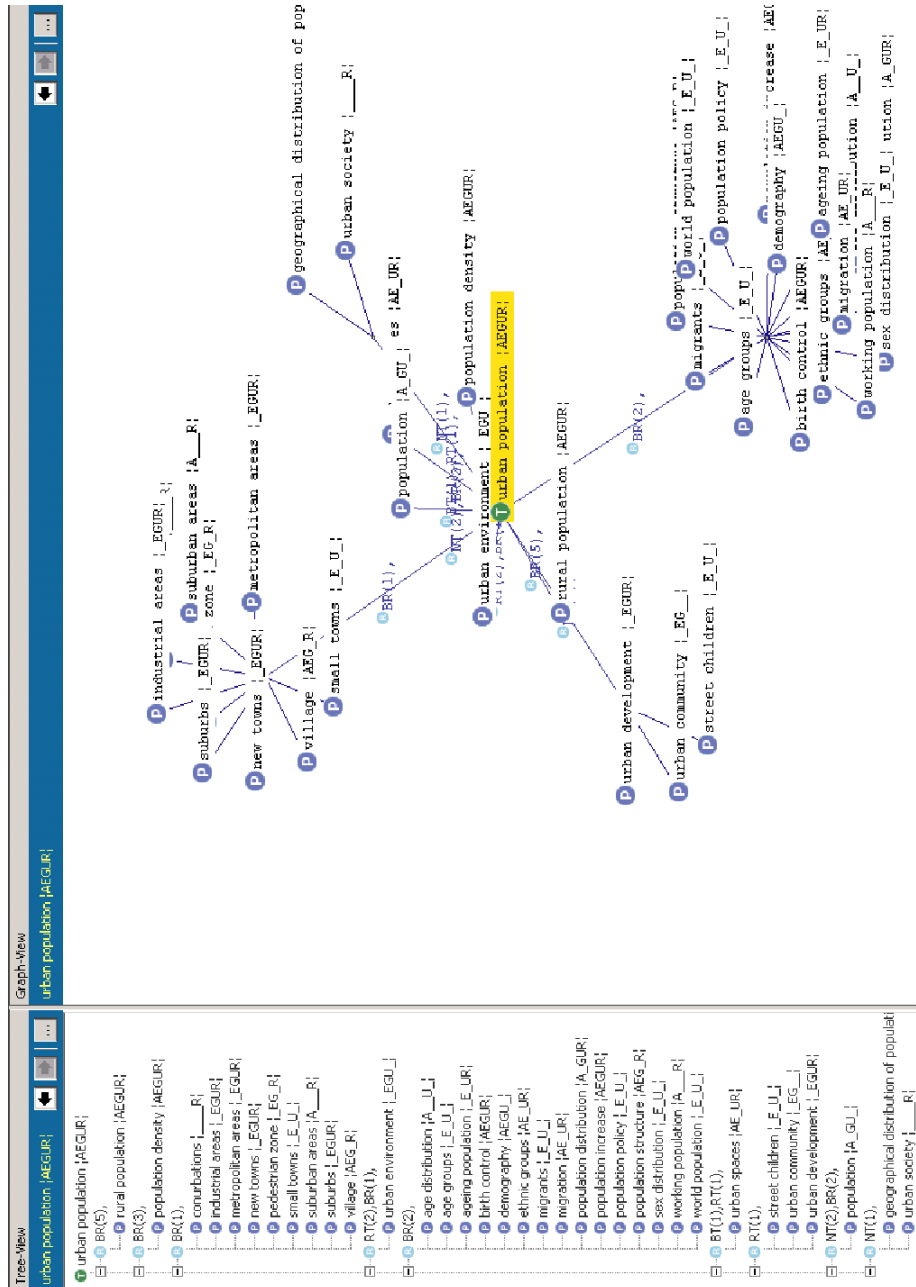


Figure 5: Visualization of a part of the generated urban domain ontology

In the experiment, we found that URBISOC, the specific thesaurus for the filtering of concepts related to urbanism, contains very generic concepts in the top part of the broader-narrower hierarchy. Therefore, these general terms had to be removed before using it as a filtering mechanism.

As regards the mappings established between the source thesauri to obtain the clusters of concepts, we could observe that urban terminology is a relevant part of the analyzed thesauri (up to 18% in UNESCO). Future work will improve the used mapping techniques by having into account the structure provided by the relations between the concepts.

In addition, we have shown, through the experiment, how to reduce the size of the generated network of concepts by pruning the less relevant relations. This pruning is able to reduce the size of the network from 6,200 concepts to only the 1,353 more related. A future improvement as concerns the relations between clusters is to take into consideration the *grandparent* and *grandchildren* relations between thesaurus concepts. The objective is increasing the relevance values of some of the existent relations. For example, two concepts in a thesaurus can be directly related through a narrower relation, however, in other one they may be related through an intermediate concept.

The urban domain ontology obtained as a result of the method proposed has several advantages in comparison with the thesauri used as source and the thesaurus used for filtering in the following areas:

Consensus and focus: The concepts of the resulting network have been selected by consensus thanks to the mappings among the different sources, removing those concepts that are neither common nor focused on urbanism.

Relations: With respect to the relation structure, the total number of available relations is bigger than the existent ones in each of the original sources. Besides each relation has a weight that indicates its relevance. As future work, the semantics of these relations should be enriched. The information provided by definitions, examples, and naming patterns in the properties of the original concepts should help to refine the current relations (e.g., broader relations could be refined as *part of*, *instance of* or *generalization* relations).

Multilingual support: Thanks to the combination of different sources of knowledge with multilingual support, the output network is enriched with alternative terminology in different languages.

Formalism: Since the output network has been generated using a formal language such as OWL, we have increased its usability, facilitating the work with reasoning engines.

Finally, it must be noted that, apart from serving as a first draft for an urban domain ontology, the generated network of urban concepts can be directly applied in information retrieval systems for resource classification, thematic indexing or query expansion.

Acknowledgements

This work has been partially supported by the University of Zaragoza through the project UZ 2006-TEC-04.

- Stumme, G., Maedche, A., September 2001. Ontology merging for federated ontologies on the semantic web. In: In Proceedings of the International Workshop for Foundations of Models for Information Integration (FMII-2001). Viterbo, Italy.
- TM4J, 2001. Homepage of the TM4J proyect.
<http://tm4j.org/>
- UNESCO, 1995. UNESCO Thesaurus: A Structured List of Descriptors for Indexing and Retrieving Literature in the Fields of Education, Science, Social and Human Science, Culture, Communication and Information. UNESCO Publishing, Paris.
<http://www.ulcc.ac.uk/unesco/>
- W3C, April 2005. Representing classes as property values on the semantic web.
<http://www.w3.org/TR/swbp-classes-as-values/>