Design and evaluation of a semantic enrichment process for bibliographic databases

Javier Lacasta[a,*], Javier Nogueras-Iso[a], Gilles Falquet[b], Jacques Teller[c], F. Javier Zarazaga-Soria[a]

[a] *Computer Science and Systems Engineering Dept., Universidad de Zaragoza, Spain*
[b] *Centre universitaire d'informatique, Université de Genève, Switzerland*
[c] *LEMA-University of Liege, Belgium*

Abstract

The limited semantics of thesauri and similar knowledge models hinder the searching and browsing possibilities of the bibliographic databases classified with this type of resources. This work proposes an automatic process to convert a knowledge model into a domain ontology through the alignment with DOLCE, an upper level ontology. This process is facilitated by an intermediary alignment with Wordnet, a lexical model. The process has been tested with the thesauri and bibliographic databases of Urbamet and the European Urban Knowledge Network. The Urbamet model has been used to create an atlas of urban related resources with advanced search capabilities.

Keywords: Ontologies, Digital libraries, Semantic Web, Data and knowledge visualization

1. Introduction

In the information retrieval context (IR), the resources of a collection are frequently classified and searched using concepts from thesauri and other simple knowledge models. However, since they reflect the vision of those who created and maintain them, they are not homogeneous and may contain

*Corresponding author

Email addresses: jlacasta@unizar.es (Javier Lacasta), jnog@unizar.es (Javier Nogueras-Iso), Gilles.Falquet@unige.ch (Gilles Falquet), Jacques.Teller@ulg.ac.be (Jacques Teller), javy@unizar.es (F. Javier Zarazaga-Soria)

heterogeneous concepts and relations [1]. For example, in the Urbamet thesaurus[1] *car* is not a sub-concept of *vehicle*. This lack of semantics limits their usability for IR. Thesauri may be used to expand queries by including the *narrower* concepts of query terms, but are disqualified for logical inference [2, 1]. Their relations are too generic, cannot be easily interpreted, and their unclear hierarchies hinder the browsing through the concepts.

Lauser [3] highlights the advantages of using ontologies with respect to thesauri. To facilitate the generation of new ontologies, Lacasta et al. [4] describe a process to transform the thesaurus used to index a collection into an ontology. They propose the use of a manual alignment between the concepts of a thesaurus and the categories in the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [5] to precise the meaning of the thesaurus concepts and *broader/narrower* relations. The main issue with this approach is the need of a manual alignment. This step reduces the applicability of the process to big models because the manual work required is not affordable.

This paper proposes an automatic alignment process which can be used to replace the manual approach described in Lacasta et al. [4]. It reduces the required human intervention to the revision of the generated ontology, increasing in this way the size of the models that can be transformed.

Performing this alignment automatically is not trivial because of the terminological gap between any thematic thesaurus and DOLCE. They do not share concepts, so traditional alignment techniques based on the identification of exact or close equivalences are not applicable [6]. The proposed process fills this gap using the WordNet lexical database [7] as an intermediate structure that allows the connection of the specific concepts in a thesaurus with the abstract classes in DOLCE. WordNet is ideal for this task as it provides conceptual-semantic and lexical relations between general and thematic concepts from the natural language.

We have tested the process with the thesauri used to classify the European Urban Knowledge Network[2] (EUKN), and the Urbamet[3] bibliographical databases. Additionally, to demonstrate the improved search possibilities that the generated models provide, we describe a thematic atlas system that

[1] `http://www.urbamet.com/thesaurus/thesaurusurbamet.htm`

[2] `http://www.eukn.org/E\_library`

[3] `http://www.urbamet.com/`

uses thesauri enriched with DOLCE relations to access the Urbamet collection.

The rest of the paper is structured as follows. Section 2 introduces the proposed formalization method. Then, section 3 analyzes the quality of the models generated when applying the process to EUKN and Urbamet thesauri. Section 4 shows the applicability of these models for improving search systems. Section 5 reviews other formalization works in the literature and compares them with the proposed approach. Finally, this paper ends with a discussion about the process, some conclusions, and an outlook on future work.

2. Alignment-based method for the formalization of thesauri

The alignment of a thematic thesaurus and DOLCE requires to deal with the abstraction gap between them. DOLCE is a formal ontology focused on describing data types and general relations independent of the context [8]. It provides three main abstract categories: *Perdurants*, which comprise events, processes, phenomena, activities and states; *Endurants*, focused on entities that maintain their identity along the time (e.g., physical objects, social objects); and *Qualities*, understood as entities that can be perceived or measured (e.g., color, shape). On the contrary, thematic thesauri contain very specific terminology. Because of this, it is unlikely that exact (or partial) equivalences can be found between them. The proposed alignment process searches for specialization relations that classify the thesaurus concepts as subtypes of DOLCE classes.

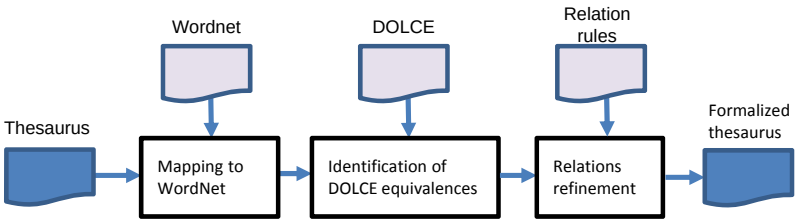

Figure 1: Thesaurus formalization process

WordNet facilitates the identification of these specialization relations. It is a lexical database of English that groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical

relations providing a *hypernym/hyponym* hierarchy of semantically related concepts. This hierarchy can be used as connector between the abstract concepts of DOLCE and the specific terminology in a thematic thesaurus.

The process is divided in three subtasks (see Figure 1): the mapping between the thematic thesaurus and WordNet; the use of the WordNet *hypernym/hyponym* relations to deduce the alignment between the source thesaurus and DOLCE; and the use of these alignments to refine the thesaurus *broader/narrower* relations. The following subsections describe each subtask in detail.

2.1. Mapping to WordNet

WordNet concepts are much more specific than those in DOLCE. This simplifies the alignment with a thematic thesaurus. However, they are still too generic to make the alignment a simple lexical matching process. A more elaborate process that analyzes the meaning of the involved concepts is required.

The matching process shown in Figure 2 solves this issue by analyzing the lexical terms in the concept labels and their definitions to identify subsumption relations. The process consists of three consecutive steps: *lexical mapping with WordNet*, *thesaurus context based disambiguation* and *resource collection based disambiguation*. Each step is only executed if the previous step does not identify a single alignment.

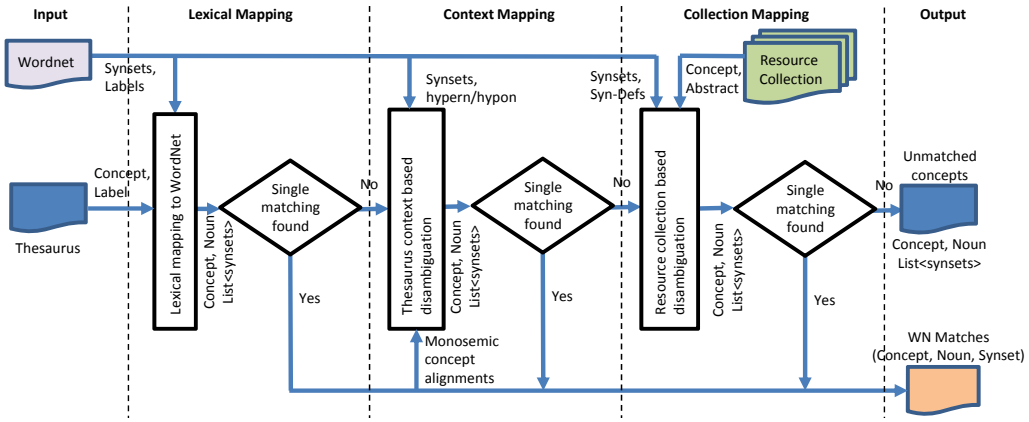

Figure 2: Thesaurus-WordNet mapping process

2.1.1. Lexical mapping to WordNet

This step identifies exact correspondences with WordNet and subsumption relations in concepts with no exact correspondence. Algorithm 1 shows how this matching is performed. It obtains the labels of the concept to match (preferred and alternatives), transforms them into singular form using a simplified version of Porter stemming algorithm [9] to deal with number issues (invocation to *getSingularForm* function in Alg. 1), and searches for exact concordance of the processed labels in WordNet (*getSynset* function in Alg. 1). The identification of concordances can be performed using JWNL, a library that allows the access to WordNet[4].

If no exact equivalence is found, an alternative approach is used. Thesaurus construction methodologies promote the definition of terms as nouns restricted by one or several adjectives or prepositional phases [10]. These nouns are more general and by construction it can be established a subsumption relation between them and the original labels. For example, the concepts "public baths" and "equipment for senior citizens" can be considered as subtypes of "bath" and "equipment". The extraction of the noun in each concept label is performed using a lexical tagger [11] (*getNoun* function in Alg. 1) which uses a set of rules (*nounIdRules* in Alg. 1) based on the thesaurus term definition guidelines. These nouns are matched with WordNet in the same way as the complete concept labels (*getSynsets* function).

```
input  : concept //Concept to match
output : senses //Set of WordNet senses matched
senses ← ∅;
for label ∈ labels(concept) do
    singLabel ← getSingularForm(label);
    labelSenses ← getSynsets(singLabel);
    if labelSenses = ∅ then
        noun ← getNoun(singLabel,nounIdRules);
        labelSenses ← getSynsets(noun);
    end
    senses ← senses ∪ labelSenses;
end
return senses;
```

Algorithm 1: Lexical mapping to WordNet

If a concept/noun is polysemic, it will correspond to several senses (synsets) in WordNet (the different meanings a term has in natural language). This is

[4] `http://sourceforge.net/projects/jwordnet/`

a problem because only one of them is correct. The other senses may have different *hypernym* relations in WordNet and lead to an incorrect alignment with DOLCE. Figure 3 depicts an example of this sense alignment problem. "High school" concept is mapped to the WordNet "school" concept, which has 7 different senses (3 are shown in the figure). Depending on the selected sense, "high school" can be classified as a "process" or as a "group" instead of as an "institution".

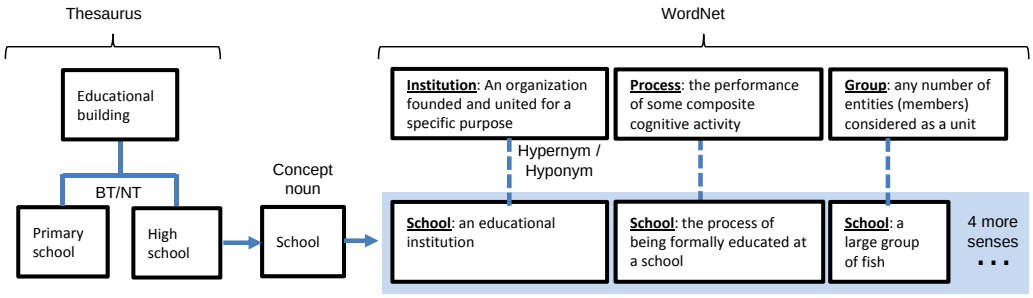

Figure 3: Example of sense disambiguation problem

The following two steps apply heuristics that use additional knowledge to disambiguate the correct sense.

2.1.2. Thesaurus context based disambiguation

Thesauri are homogeneous descriptions of the knowledge in a certain area of interest. Since all the contained concepts have related meanings, one may reasonably suppose that all the concepts sharing a noun use it with the same sense. This circumstance allows us to assume that already defined monosemic matches can be used to identify the sense of all other thesaurus concepts sharing its noun. Algorithm 2 describes this process. First, it obtains all the labels in the thesaurus concepts containing the noun to disambiguate (invocation to *getLabelsUsingNoun* function in Alg. 2) that have been matched with a single WordNet sense in the previous step. Then, it matches the WordNet hypernyms of these labels (*getHypernyms* function in Alg. 2) with the set of possible senses of the noun. The senses with more matches (they are hypernyms of more disambiguated concepts) are selected as new set of possible synsets (*sensesWithMoreOccurrences* function in Alg. 2), but only if the result is a single sense the label is considered disambiguated. If there is no concordance, the original set of synsets is not modified.

input : noun *//Noun to be disambiguated*
possibleSenses *//Set of possible senses associated with the noun*
thesaurusLabels *//Labels in the thesaurus*
disambiguatedLabels *//Labels monosemicaly matched with WordNet in the previous step*
output : reducedSenses *//Reduced set of senses assigned to the noun*

labels ← getLabelsUsingNoun(noun,thesaurusLabels) ∩ disambiguatedLabels;
reducedSenses ← getHypernyms(labels) ∈ possibleSenses;
if *reducedSenses* ≠ ∅ **then**
  reducedSenses ← sensesWithMoreOccurrences(reducedSenses);
else
  reducedSenses ← possibleSenses;
end
return reducedSenses;

Algorithm 2: Thesaurus context based disambiguation

Figure 4 shows an example of how this process is applied. It depicts some Urbamet concepts that include "sport" as noun ("sports", "water sport", "winter sport" and "equestrian sport"). From the 7 different WordNet senses of "sport", the one that is a *hypernym* of "water sport" (it has a previous monosemic matching) is selected. As a consequence, the other concepts in the set ("sports", "winter sport", "equestrian sport") are identified as "activities" and not as "living things".

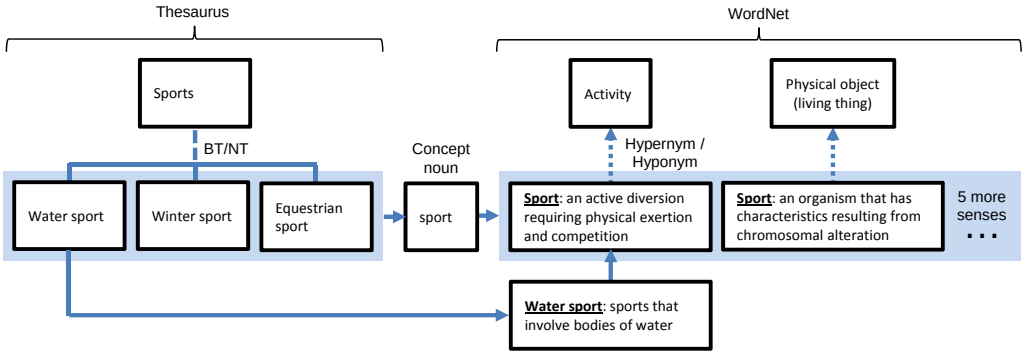

Figure 4: Example of thesaurus context based disambiguation

2.1.3. Resource collection based disambiguation

Those concepts that are not matched with a single WordNet synset are submitted to an alternative disambiguation process. It uses a bibliographical database classified with the thesaurus as the disambiguation context. The

process identifies the resources classified with a concept containing the noun to disambiguate and extracts the nouns contained in their description (e.g., abstract section). The nouns extraction can be performed using GATE, a software library that facilitates natural language processing tasks[5]. Then, it compares these names with those in the WordNet definitions of the possible synsets.

The nouns in the abstract set are matched with the nouns in each candidate WordNet sense definition using the direct cosine similarity measure described in equation 1. It calculates the similarity degree between a WordNet synset s and a thesaurus concept c as a value between 0 (disjoint sets) and 1 (equivalent sets). The sense with the highest similarity degree with respect to the concept is the selected one. In the formula, $SN(s)$ stands for the set of nouns in the definition of synset s; $AN(c)$ is the set of nouns in the collection of abstracts classified with the concept c; and $occur(n, X)$ describes the number of occurrences of the noun n in the set X.

$$Sim(s,c) = \frac{\sum_{n_i \in SN(s) \cap AN(c)} (occur(n_i, SN(s)) * occur(n_i, AN(c)))}{\sqrt{\sum_{n_i \in SN(s)} (occur(n_i, SN(s))^2)} * \sqrt{\sum_{n_i \in AN(c)} (occur(n_i, AN(c))^2)}}. \quad (1)$$

Figure 5 shows the application of this matching process to the Urbamet "Industrial landscape" concept. The search for the noun "landscape" in WordNet yields 4 senses. Then, the resources in the collection classified with a "landscape" concept ("Industrial landscape", "Rural landscape", "Vegetal landscape"...) are located (140 articles). Next, the resource abstracts and the WordNet sense definitions are processed to extract their common nouns (e.g., the "expanse" term appears 7442 times). Finally, the similarity formula (equation 1) is applied and the sense more similar to the abstracts is identified (an expanse of scenery...).

2.2. Identification of DOLCE equivalences

Once the thesaurus concepts are linked to a WordNet sense, the WordNet *hypernym/hyponym* hierarchy is used to identify specialization relations with DOLCE.

This process requires an alignment between the abstract WordNet concepts (in the firsts levels of the *hypernym/hyponym* hierarchy) and DOLCE.

[5] `http://gate.ac.uk/`

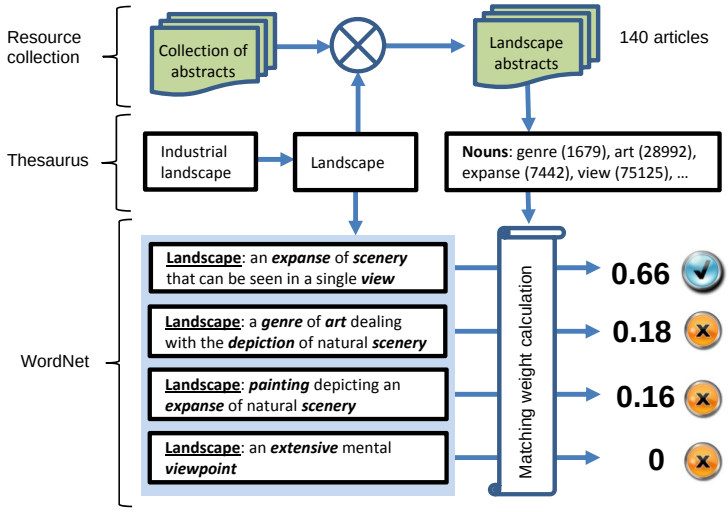

Figure 5: Example of resource collection based disambiguation

In the literature, we have only identified the alignment proposed by Gangemi et al. [12], and it was discarded because it is not broadly available and it is not compatible with the last versions of WordNet. Therefore, we have generated our own WordNet-DOLCE alignment. It takes the collection of WordNet senses and searches for exact equivalences of their labels in DOLCE. To improve the number of matches identified, gender/number issues are corrected using Porter stemming algorithm.

As a result, 75 out of the 208 DOLCE classes have been matched, including among them the desired main DOLCE categories (activity, agent, process ...). The non-aligned elements are very specific, and it is difficult that any thesaurus concept can be aligned to them (e.g., socially constructed person, production workflow execution). Moreover, many of the unmatched concepts are specializations of other matched ones. Thus, even if the correct alignment cannot be found, a more general one can be provided.

```
input  : sense //WordNet sense aligned with a thesaurus concept
         dolceMatch //WordNet-DOLCE alignment
output : dolceID //Identifier of the DOLCE class aligned
senses ← getHypernyms(sense);
for sns ∈ senses do
    dolceID ← dolceMatch.getValue(sns);
    if dolceID ≠ ∅ then
        return dolceID;
    end
end
```

Algorithm 3: Identification of DOLCE equivalences

Algorithm 3 shows how the senses matched with the thesaurus concepts are aligned with DOLCE. First, the *getHypernyms* function obtains all the hypernyms of a sense ordered by distance to the sense (this function uses the JWNL API). Then, it reviews the synsets in the list searching for a match with DOLCE (the map dolceMatch contains the previously defined WordNet-DOLCE alignment). The first match identified is the one used as alignment between the thesaurus concept associated to the WordNet sense and DOLCE. This approach guarantees that the most specific alignment available is used. If the *hypernym/hyponym* branch has no sense aligned with DOLCE, the thesaurus concept is left unmapped. Figure 6 shows the formalization result of the "winter sport" concept. This concept has been previously aligned with the "Active diversion..." WordNet sense (see Figure 4). When going up in their *hypernym/hyponym* hierarchy, first it is found "recreation", which has no DOLCE mapping, and then "activity", which corresponds to the DOLCE "activity" concept. Therefore, the "winter sport" thesaurus concept is classified as a kind of DOLCE "activity".

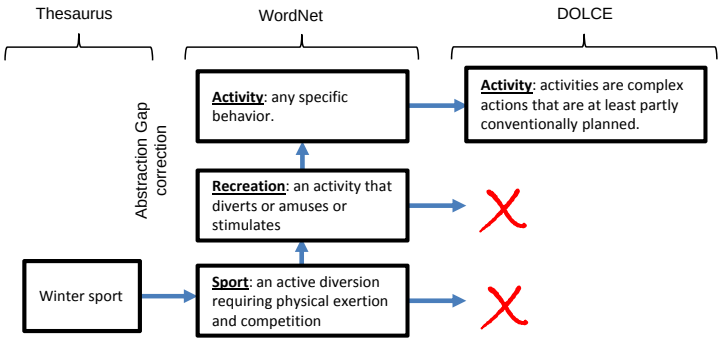

Figure 6: Example of mapping between a thesaurus concept and DOLCE through WordNet

Table 1: Rules to infer relations in urban models

Pairs of DOLCE classes identified as superclasses of two thesaurus concepts holding a BT/NT relation	**Inferred relation**
(activity → physical/abstract-quality) (geographical/physical/information-object → abstract-quality) (rational-agent → abstract-quality) (regulation → abstract-quality) (plan → abstract-quality)(physical-quality → abstract-quality) (physical-quality → physical-quality)	has-quality
(activity → rational-agent) (activity → information/physical-object) (activity → regulation) (activity → principle) (phenomenon → geographic-object)	participant
(abstract-quality → abstract-quality) (activity → plan) (phenomenon → activity) (geographic-object → geographic-object) (regulation → plan)	part
(plan → activity) (rational-agent → information-object) (rational-agent → physical-object) (rational-agent → plan) (norm → system-design)	generic-dependent
(physical-object → physical-object) (rational-agent → rational-agent) (regulation → regulation) (information-object → information-object)	subclass-of
(physical-object → activity) (physical-object → plan)	instrument-of
(activity → activity)	result-of

2.3. Relations refinement

The last part of the formalization step consists in refining the original concept relations (hierarchical and associative) in terms of the DOLCE relations. We use subclass inference rules to determine these relations. For example, since two DOLCE *physical-objects* hold a *part-of* relation, the *broader/narrower* relation between two thesaurus concepts classified as *physical-objects* is redefined as a *part-of* relation.

DOLCE provides several possible relations between two classes. For example, the relation between a *geographical-object* and a *physical-object* can be of type *part* or of type *subclass*. To automate the selection of the correct one, a set of predefined rules that determine the relation to use for each pair of DOLCE classes in a given context is used. Table 1 shows a summary of the rules defined in the context of urbanism for the models described in the experiment section. For example, the selected relation between a *physical-object* and an *abstract quality* is *has-quality*.

As an illustrative example of this relation refinement, Figure 7 shows how these rules are applied to the 10 narrower relations of the *Environmental sustainability* concept in EUKN thesaurus. In the alignment process, the *Environmental sustainability* concept and three of its children have been tagged as a DOLCE *activity*; the other 7 have been classified as DOLCE *physical-quality*. Following the rules, two *activities* hold a *result-of* relation between them. In the case of an *activity* and a *physical-quality*, the relation

is *has-quality*. This improved model provides a richer knowledge structure with more advanced search capabilities. For example, it allows inferring that *Environmental sustainability* is the *result-of* *Waste management and recycling*, *Environmental education*, and *Green public procurement*, and that it can be measured (*has-quality*) through the *Water quality*, the *Air quality* and so on.

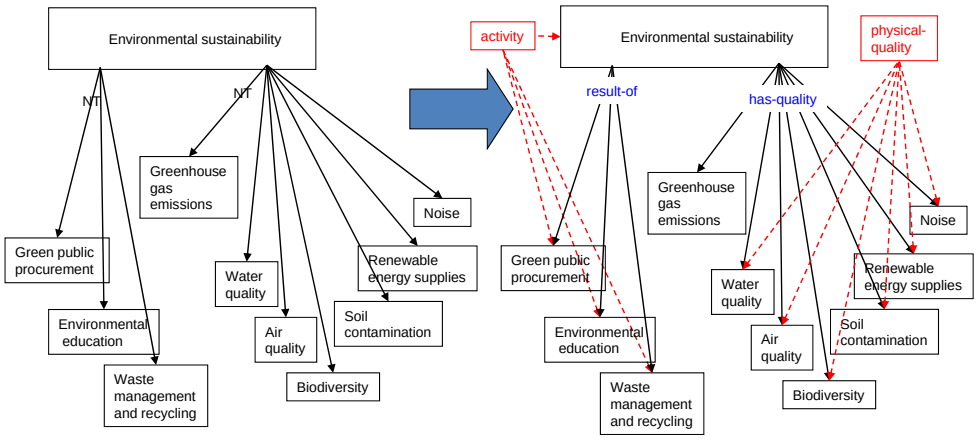

Figure 7: Transformation of the *Environmental sustainability* concept and its narrower concepts

3. Evaluation of the process results

The formalization process has been applied to EUKN and Urbamet thesauri. The use of EUKN allows comparing the quality of the result with respect to the manual alignment proposed in Lacasta et al. [4]. The use of Urbamet is oriented towards the validation of the process with bigger models.

This section starts describing the features of Urbamet and EUKN collections and their thesauri. Then, it analyzes the quality of the alignment obtained between each thesaurus and WordNet. Finally, it describes the quality of the alignment between the two thesauri and DOLCE.

3.1. The Urbamet and EUKN collections and their thesauri

The CDU (French Center of Urban Documentation) and EUKN are two organizations focused on enhancing the exchange of urban knowledge and expertise between scientists, practioners and decision-makers. Designed and maintained by the CDU, the Urbamet database was initially developed in

1969 to facilitate the sharing of knowledge between professionals in urban planning, housing and transport in France. It gathers bibliographic notes issued internally in the Center as well as those proposed by the 97 French departmental administrations in charge of equipment (roads, technical services and planning issues) and major urban planning agencies in France.

EUKN started in 2004 as a pilot project of different European states. Nowadays, it is an intergovernmental knowledge network that acts as hub for existing networks of urban practitioners, researchers and policy-makers at all governmental levels. It collects information provided by national "knowledge nodes", which identify and select relevant experiences, publications or networks to be publicized at a European level via EUKN.

The structure of the collections and the thesauri used to classify them is shown in Table 2. EUKN and Urbamet thesauri (263 and 3,844 concepts respectively) are basic knowledge organization models that do not contain definitions, scope notes or related terms. The EUKN collection consists of 3,253 articles (*#Articles* column in Table 2), which contain on average a single reference to its thesaurus (*#Concepts/Article* column) and each concept is cited by 8 resources on average (*#Articles/Concept* column). Urbamet is a bigger database with almost 250,000 bibliographic notes and an estimated growth of 8,000 new notes each year. From these 250,000 notes, the 9,684 notes published during the period 2005-2006 have been used in our experiments. They are better classified than EUKN resources: each article contains more than 8 references to concepts, and each concept has around 4 citations. However, they are not completely well constructed and integrated. On the one hand, the thesauri include replicated concepts and they have an unclear hierarchical structure. On the other hand, the collections only use a subset of the thesauri for classification: 59% in the case of EUKN, and 73% in the case of Urbamet (*%Thes Used* column in Table 2). Additionally, the collections include other external terms to describe the resources (132 in EUKN, 312 in Urbamet).

3.2. Mapping of the Urbamet and EUKN thesauri with DOLCE

The application of the formalization process to EUKN and Urbamet thesaurus has required their transformation into a common format: SKOS [13] for thesauri and RDF/Dublin core [14] for the bibliographic record describing each article. Dublin Core is a standard vocabulary for resource description. Their elements are generic and they are used for describing a wide range of resources. SKOS provides a standard way to represent knowledge organization

Table 2: Comparison of Urbamet and EUKN thesaurus and collections

Thesaurus	Concepts	PrefLab(en)	AltLab(en)	BT/NT	RT	Defs
Eukn	263	263	0	262	0	0
Urbamet	3844	3844	504	3821	0	0

Collection	#Articles	% Thes Used	#Concepts/Article	#Articles/Concept
Eukn	3253	59.31%	1.10	7.95
Urbamet	9684	73.57%	8.74	4.30

systems such as thesauri using the Resource Description Framework (RDF). This facilitates its use in distributed, decentralized metadata applications.

Additionally, since the Urbamet collection abstracts are in French, we had to translate them into English. This has been automatically done using the Microsoft Bing Translator Web Service[6]. With respect to the analysis of the results, we have manually measured the quality of the thesaurus-WordNet alignment and the WordNet-DOLCE alignment. However, while the EUKN model has been completely reviewed, the size of Urbamet has obliged us to select a representative branch of the whole thesaurus: the 208 concepts of the "urban planning development" branch.

3.2.1. Thesaurus-WordNet mapping results

Table 3 shows the polysemy degree of the EUKN and Urbamet concepts and their extracted nouns (for those without exact non-ambiguous correspondence in WordNet). The table contains the number of concepts in each thesaurus that have 0, 1 or N senses in WordNet (*# concepts* column) and the corresponding percentage (*% concepts* column). Since the a-priori probability of getting the correct sense is the inverse of the number of senses of a concept, we found that 43.50% of EUKN concepts and 30.28% of Urbamet would be correctly matched.

Having into account that an alignment is only considered correct when a single (and correct) sense is returned, it can be observed that the alignment coverage increases in EUKN from a 20.91% to a 64.25% and in Urbamet from a 9.61% to a 88.94% (see Table 4 shown latter).

Table 4 shows the improvement obtained when the improved mapping

[6] `http://www.microsofttranslator.com/`

Table 3: Senses in WordNet of EUKN and Urbamet concepts

Senses	EUKN		Urbamet	
	# concepts	% concepts	# concepts	% concepts
0	13	4,94	13	6,25
1	55	20,91	20	9,61
2	54	20,53	19	9,13
3	46	17,49	38	18,26
4	25	9,50	39	18,75
5	15	5,70	10	4,80
6	30	11,4	25	12,01
7	4	1,52	13	6,25
8	5	1,90	1	0,48
9	10	3,80	13	6,25
10	0	0	5	2,40
11	5	1,90	5	2,40
12	1	0,38	4	1,92
>=13	0	0	3	1,44

Probability of selecting the correct sense:
EUKN: 43.50% - Urbamet: 30.28%

process is applied (concept is considered aligned when a single sense is obtained). It shows the number of concepts processed (*Conc* column), the number of those aligned (*Conc Aligned* column), the percentage of concepts aligned (*% Thes Align* column), the number of concepts with a correct alignment (*Conc Corr Align* column), the percentage of success (*% Corr Align* column) and the percentage of the thesaurus/branch correctly aligned (*% Thes CAlign* column). It can be observed that the quality of the alignments is high (83% and 87%) but the coverage could be improved (especially in EUKN). Even though, the final percentage of concepts correctly aligned is increased up to 53% and 77% of the concepts (vs. the initial 30.28% and 43.50%).

Table 4: Thesaurus-WordNet alignment results

	Conc	Conc Align	% Thes Align	Conc Corr Align	% Corr Align	% Thes CAlign
EUKN	263	169	64.25%	141	83.43%	53.61%
Urbamet	208	185	88.94%	161	87.02%	77.40%

3.2.2. Thesaurus-WordNet-DOLCE mapping results

Using the thesaurus-WordNet alignments as an intermediate step, the final thesaurus-DOLCE alignment results are shown in Table 5. The table shows how many of the correct Thesaurus-WordNet mappings of each approach (*WN Align* column) end in a correct DOLCE alignment. It contains the total of correct alignments (*DC Align* column) and the corresponding percentage (*% Align* column). It can be observed that around 59% of EUKN concepts and 75% of URBAMET concepts aligned with WordNet are correctly aligned with DOLCE. The others are wrongly aligned or unaligned. The right side of the Table 5 shows the final results of the complete thesaurus-DOLCE alignment process. It shows the percentage of concepts correctly aligned (*% T Corr* column), incorrectly aligned (*T Incorr* column), and not aligned (*% T not* column).

Table 5: WordNet-DOLCE and final alignment results

	WN Align	DC Align	% Align	-	% T Corr	% T Incorr	% T not
EUKN	141	83	58.86%	-	31.55%	24.71%	43.72%
Urbamet	161	120	74.53%	-	57.69%	22.21%	20.19%

The process works significantly better for Urbamet (58%) than for EUKN (32%) because of the differences in thesaurus-WordNet alignment coverage and because EUKN concepts are assigned to WordNet areas with worse DOLCE alignment. The lack of mappings between EUKN and WordNet can be explained analyzing the structure of the thesaurus concept terms and its collection. On the one hand, the EUKN thesaurus is more heterogeneous than the Urbamet model. The Urbamet thesaurus has been created by documentalists, and it has been improved and refined since its creation in 1969. On the contrary, EUKN thesaurus is more recent (started in 2004) and frequently includes multiple concept terms that are difficult to align (e.g., Production & manufacture, Universities & spin-offs). On the other hand, the number of EUKN resources and their abstract length is smaller with respect to Urbamet (the context is not so rich). Additionally, about 40% of EUKN thesaurus concepts have been never used for classifications (no context can be derived from abstracts if these concepts are polysemic). With respect to the WordNet-DOLCE alignment difference, the problem is the lack for correct WordNet-DOLCE mappings for certain WordNet areas (more used in EUKN). For example, in EUKN, 78.30% of the "activities" and 72.90%

of the "physical-objects" are correctly identified, but this is reduced to a 36.30% of the regulations and to a 4.70% of the rational-agents. It is caused by the way the WordNet-DOLCE alignment is performed. WordNet polysemy makes that 289 WordNet senses share labels with 75 classes in DOLCE and not all these senses are correct for alignment (they do not have a compatible meaning). The use of one incorrect WordNet-DOLCE matching in the process leads to an erroneous Thesaurus-DOLCE alignment. A better WordNet-DOLCE mapping would greatly improve the alignment quality of this step, leaving the thesaurus-WordNet coverage problem as the main issue to solve.

The results obtained in the replacement of BT/NT relations by DOLCE relations are shown in Table 6. It has not been possible to apply the automatic process to all the relations (*#BT/NT* column) because the process requires that the two involved concepts are aligned with DOLCE. The distribution of the unmapped concepts along the thesauri structure hinders the conversions capabilities reducing the number of processable relations to a 14% and 34% respectively (see *#RToForm* and *%RToForm* columns in Table 6). However, the quality of those relations that it has been possible to formalize is quite good. Columns *#Corr*, *%Corr*, *%Corr* and *%Not* inform about the number of correctly derived relations and corresponding percentages with respect to the number of relations refined. In EUKN all the transformed relations are correct. In the case of Urbamet, the results seem to be more discrete (65%). However, the problem is not the errors (4.2%), but the lack of a suitable relation in DOLCE (30.8%). This is caused by the spatial thematic of the selected Urbamet branch: DOLCE does not provide relations between a spatial region and the rest of the concepts in the ontology.

Table 6: Relations refinement

	#BT/NT	#RToForm	%RToForm	#Corr	%Corr	%Incorr	%Not
EUKN	262	37	14.1%	37	100%	0%	0%
Urbamet	207	71	34.3%	46	65%	4.2%	30.8%

4. Applicability of the formalized models to provide a different view of bibliographic databases

To test the applicability of the formalized model in an information retrieval context we have designed a facet based search component that pro-

vides access to the Urbamet collection as a thematic atlas. It combines the spatial references provided by the bibliographic records of Urbamet with the thematic view provided by the DOLCE categories in the formalized thesaurus. The objective is to facilitate the selection of a spatial area of interest and show the categories of documents (DOLCE classes) referencing to the selected area (see Figure 8, step 1). As a result, the user obtains the different topics of the selected category (thesaurus concepts, step 2), the documents associated to these topics (step 3) and other topics related to each selected one (step 4).

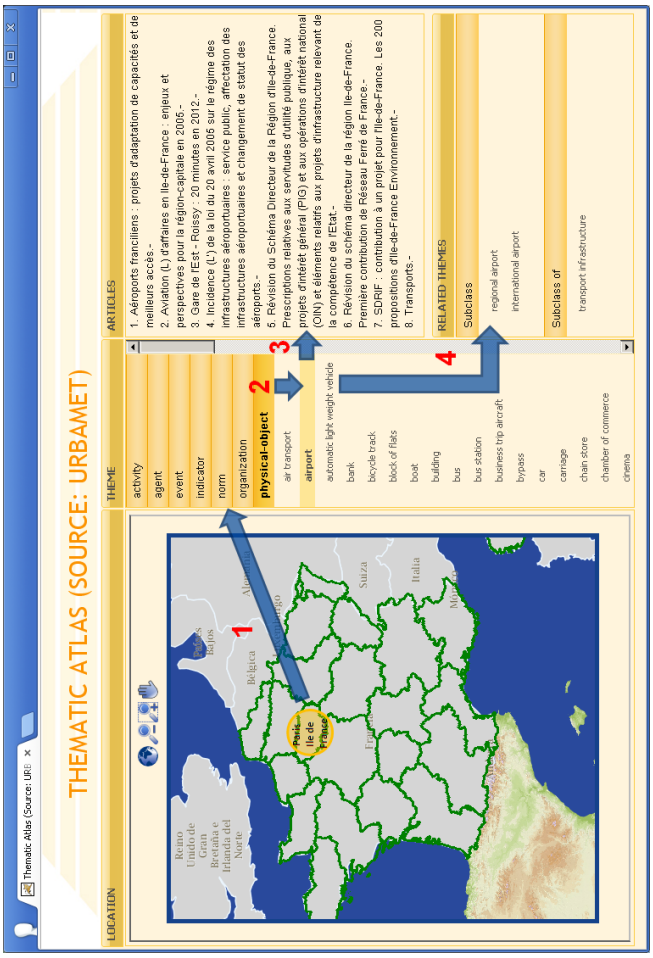

Figure 8: Prototype of the thematic atlas derived from the Urbamet database

The construction of the prototype has required the transformation of the

Urbamet collection descriptions into a semantic repository that integrates the formalized Urbamet thesaurus (for thematic access) and a jurisdictional ontology (for spatial access). The original descriptions (with title, abstract, location...) have been transformed into a Dublin Core RDF model. With respect to the Urbamet thesaurus, the original XML files have been transformed to SKOS format using the conversion methodology described in Lacasta et al. [15]. Then, our proposed formalization process has been used to generate an enriched OWL file. The spatial model extends the jurisdictional domain ontology described in Lopez-Pellicer et al. [16]. This ontology provides an OWL model aligned with DOLCE that describes the different jurisdictional divisions of a country. It has been completed with the French jurisdictional instances provided by the Second Administrative Level Boundaries data set project [17] and the municipalities provided by the Institut Geographique National[7] (Urbamet collection is mainly focused on France). Additionally, since the collection contains some resources making references to countries outside France, the list of countries provided by ISO-3166 standard [18] has been added to the model.

The integration of the collection with the thematic and spatial ontologies is done by replacing the original textual references contained in the bibliographic records describing the resources with identifiers (URIs) referencing to the ontologies. The result is stored in a Jena[8] semantic repository and accessed through a SPARQL end point provided by the Fuseki[9] library.

The resulting system facilitates the browsing through the thesaurus in a more precise manner. For example, it indicates that the concept *international airport* is a subclass of *airport* (see Figure 8). This is a more specific definition than the general *broader* relation in the original Urbamet thesaurus. Another advantage is that it opens the possibility of executing queries that would be difficult to perform using a classic relational database. Figure 9 shows the SPARQL queries required to perform the navigation steps indicated in Figure 8. It can be observed how they allow the relation of elements not directly connected in the stored model to provide a navigation that integrates the DOLCE classes as categories of the thesaurus concepts. It also provides relations between the concepts that go beyond the original *broader/narrower*

[7] `http://www.ign.fr/`

[8] `http://jena.apache.org/`

[9] `https://jena.apache.org/documentation/serving\_data/`

relationships (they are replaced with the different types of relations provided by DOLCE).

```
(1) Select distinct ?dolceClass where {
        ?resUri dc:coverage <http://jdo/france/ile-de-france>.
        ?resUri dc:subject ?urbTheme.
        ?urbTheme rdfs:subClassOf ?dolceClass}

(2) Select distinct ?urbTheme where {
        ?resUri dc:coverage <http://jdo/france/ile-de-france>.
        ?resUri dc:subject ?urbTheme.
        ?urbTheme rdfs:subClassOf
            <http://www.loa-cnr.it/ontologies/DOLCE-Lite.owl#physical-object>}

(3) Select distinct ?resUri where {
        ?resUri dc:coverage <http://jdo/france/ile-de-france>.
        ?resUri dc:subject <http://www.urbamet.com/thesaurus/airport>}

(4) Select distinct ?relation ?related where {
        <http://www.urbamet.com/thesaurus/airport> ?relation ?related.
        {?relation rdfs:subPropertyOf
            <http://www.loa-cnr.it/ontologies/DOLCE-Lite.owl#immediate-relation>} union
        {?relation rdfs:subPropertyOf
            <http://www.loa-cnr.it/ontologies/DOLCE-Lite.owl#immediate-relation-i>} union
        {?relation rdfs:subPropertyOf
            <http://www.loa-cnr.it/ontologies/DOLCE-Lite.owl#mediated-relation>} union
        {?relation rdfs:subPropertyOf
            <http://www.loa-cnr.it/ontologies/DOLCE-Lite.owl#mediated-relation-i>}}
```

Figure 9: Queries required to perform the browsing steps displayed in Figure 8

5. Related work

The process described in this paper aligns automatically a thesaurus with DOLCE ontology. The objective is to provide a better classification of the thesaurus concepts and to refine its relations. The better identification of these elements is required in different contexts related to information retrieval. Concepts and relations in a thesaurus may not be properly defined. As a consequence, their ambiguity may increase the difficulty to find the information classified according to these concepts. A better identification of these elements helps to provide better search systems for already available collections of resources.

This section reviews the works focused on adding semantics to thesauri and other knowledge models and compares them with the approach proposed in this paper.

The most simple transformation approaches are manual. They provide high quality alignments but at expenses of a high transformation cost. Tudhope et al. [19] describe how to specialize the associative relations of the Art and Architecture Thesaurus (AAT) into richer subtypes through the analysis of sample extracts of AAT Editorial Related Term Sheets and the AAT editorial rules. Wielinga et al. [20] focus on tagging AAT hierarchy with unique identifiers and slots corresponding with their main terms and synonyms to generate RDF concepts. Golbeck et al. [21] describe the transformation process of the National Cancer Institute (NCI) thesaurus into OWL format. They describe what happens when converting concepts and defining the original roles of the concepts as OWL restrictions in properties. In the same line, Chun and Wenlin [22] describe the conversion process of the Chinese Agricultural Thesaurus from a relational database into RDF format, but it does not add any additional semantics.

Semiautomatic processes simplify the concept conversions and reduce their cost. This is the case of Soergel et al. [23] and Kawtrakul et al. [24], who identify patterns between concept categories to establish transformation rules. These rules can be automatically applied to *broader/narrower* and *use/use-for* relations to generate more appropriate ones. In a similar way, Khosravi and Vazifedoost [25] propose a re-engineering process based on rules that allow the transformation of the ASFA Persian thesaurus relations. Finally, Hepp and de Bruijn [26] describe an algorithm that derives OWL classes from thesaurus concepts and their *broader/narrower* relations. It creates two ontology classes per 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. Then, it inserts *subClassOf* relations between the classes in the original hierarchy context.

Focusing on knowledge models different from thesauri, Aleksovski et al. [27] propose a method to match lists of terms using different disambiguation and heuristic techniques and pre-existent upper or domain formal ontologies. van Damme et al. [28] also show how folksonomies and other unstructured vocabularies can be used to construct ontologies. They describe an approach for deriving ontologies from folksonomies based on the statistical analysis of the folksonomies, the use of online lexical and semantic web resources, the application of ontology matching (and mapping) approaches, and the computer assisted revision of the results. Vatant [29] proposes the use of OWL/RDF to define constraints on topics, associations, roles and other knowledge objects manipulated by a Topic Map to be able to validate if a topic map commits to

an ontology. Finally, Sridharan et al. [30] propose a topic map approach that incorporates semantic annotations to construct a multi-level ontology-driven topic map that facilitates an effective visualization, classification and global authoring of e-learning resources.

From a general perspective, our proposed process is similar to those previously described as they also present disambiguation procedures to detect the equivalences on the basis of names, definitions, and relations of the concepts in the models to align. However, there are some relevant differences. On the one hand, it focuses on formalizing thesauri, a specific kind of knowledge model that has a standardized structure of properties and relations. This reduces the expected kind of relations and allows the use of more specific techniques to refine the meaning of those relations. On the other hand, the automatic matching process described in this paper does not focus on aligning models with similar terminology, but it is centered on finding relations between a thesaurus and an ontology with a different level of abstraction (DOLCE). The proposed process uses the subsumption relation of an intermediate ontology (WordNet) to avoid the semantic gap and connect the concepts between the source thesaurus and the ontology.

We use WordNet as an intermediate model because it contains a hierarchy of concepts with a level of abstraction taht allows connecting general DOLCE classes with specific thematic thesaurus concepts. An alternative to WordNet as intermediate model could be the SUMO ontology [31]. It is mapped to WordNet and has a similar level of abstraction in the concepts contained. Therefore, it may work in a similar way as WordNet. Another alternative is YAGO [32], a knowledge base which integrates concepts extracted from Wikipedia, WordNet, and GeoNames. This knowledge base maintains WordNet hierarchical relations, but it includes much more concepts. This may be an advantage because it probably increases the matching coverage. On the other hand, as it is automatically generated, it contains errors, which can reduce the quality of the generated match. Other general ontologies such as DBpedia [33] are not appropriate because they lack the kind of hierarchical structure required to perform the alignment process described in this paper.

Finally, the automatic nature of our proposed process presents some limitations. It mainly depends on the correct identification of equivalences between the thesaurus concepts and WordNet. An error in this identification will generate an incorrect DOLCE class and the inferred relations will not be the appropriate. As it can be observed in the results discussed in section 3, the quality greatly depends on how these mappings have been established.

Currently, the percentage of correctly matched concepts does not allow a complete direct use of the formalized thesaurus in a working environment. However, it is a starting point that can be refined into a more complete ontology, instead of having to generate it from scratch.

6. Conclusions and outlook on future work

This paper has described a formalization approach for thesauri that reduces the heterogeneity of their concepts and relations. The process enriches the thesaurus concepts and relations through an alignment with DOLCE. To fill the semantic gap between the models to align, the WordNet structure is used. The process starts with a thesaurus-WordNet alignment step, which uses the concept labels, their structure and a collection associated to the thesaurus to establish the alignment. Then, it follows a WordNet-DOLCE alignment step, which uses the WordNet *hypernyms/hyponyms* hierarchy to generalize the thesaurus concepts into the DOLCE level. It finishes with a relations improvement step, which uses the created alignment to replace the thesaurus *broader/narrower* relations with other more specific relations from DOLCE.

The quality of the generated models has been evaluated with two collections of urban resources: EUKN and Urbamet. The results have shown the possibilities of the process but also the areas of future improvement. On the one hand, the thesaurus-WordNet alignment process needs to be improved with additional alignment heuristics to increase the coverage in collections with a small volume of articles, or very short abstracts. Additionally, it must be taken into account that WordNet is only available in English. This is a problem if we need to formalize a thesaurus not available in English, or to use a bibliographic database in other languages as the context for disambiguation. Yago, SUMO or alternative knowledge models could be considered instead of WordNet. On the other hand, a better WordNet-DOLCE alignment is required. The current approach only provides direct lexical equivalences and it does not work for some kinds of concepts since some semantic gaps between WordNet and DOLCE still need to be filled. There are concepts with the same labels that have a different meaning (and should not be related) and concepts with different labels where a specialization relation could be defined.

In spite of the problems previously described, the generated model is an improvement in terms of cost with respect to performing the alignment

manually. The cost of reviewing the generated results is more reduced than the creation of a new alignment from scratch. On the one hand, the revision of each match only requires a comparison of the meaning of the two involved concepts. To reduce the time required, it can be facilitated with an adequate tool that shows the available information about the two concepts in their respective models. On the other hand, if a match is found incorrect (or there is no match), the required work is not increased with respect to the manual approach: review the DOLCE ontology in search of the suitable alignment.

An alternative to improve the quality of the model would be to perform the manual revision after each step of the process. This would eliminate the accumulation of errors that hinder the final results. However, it would require to review the thesaurus-Wordnet alignment, and the WordNet-DOLCE alignment. Therefore, it is not clear that this approach has any advantage in cost with respect to the manual approach. Moreover, our efforts are not oriented in this direction. Our final objective is to continue improving the different steps in the proposed process in order to generate a better model that requires less revision effort.

As future work, we will consider the possibility of using the different components of the process to construct a semiautomatic suggestion system that integrates the formalization and revision processes. Instead of filtering the concepts with multiple WordNet senses, the objective would be to calculate all the WordNet-DOLCE alignments and leave to a human being the final selection of the most suitable one (or the insertion of another one). This intermediate approach is expected to reduce the intellectual workload of establishing manual mappings (a reduced list is provided), while maintaining the quality of a supervised process.

To test the applicability of the generated models, a facet based search component that combines thematic and spatial features as an atlas has been proposed for Urbamet collection. Whereas the Urbamet thesaurus has been used as thematic model, a preexistent jurisdictional ontology has been used for spatial features. This application has shown how formalized models can facilitate searches in a collection based on their metadata descriptions and the advantages that the application of inference produces (e.g., in transitive "is-a" or "part-of" relations). Similarly, it has shown the browsing benefits from a more abstract access focused on DOLCE categories (activities, events, rational-agents ...) with extended relations between the resources. Additionally, in the future, these approaches could be extended with the use of non-transitive relations. For example, the user could as the system

for anything "performed by" an "agent" to obtain the resources classified as "activities" or "processes" (between others).

Another area of work would be the integration of other thesauri and knowledge models used in a processed collection. For example, in the case of temporal information, the ontology proposed by Gutierrez et al. [34] and Hurtado and Vaisman [35] could be used after it is aligned with DOLCE. Similarly, authority information (authors, organizations, departments...) could be integrated. In this case, repositories such as the International Virtual Authority File[10] could be used, but it is necessary to formalize them and include richer relations (e.g., "part-of", "works-in" or "collaborates-with") to manage their heterogeneity (e.g., European Union = U.E. = E.U.). Finally, it could be also interesting to determine if the process can be used to refine more heterogeneous hierarchical knowledge models such as taxonomies or even concept networks such as topic maps.

7. Acknowledgements

This work has been partially supported by the Spanish Government through the project TIN2012-37826-C02-01 0. The collaboration between European partners has been supported by the COST Action C21 Towntology.

References

[1] D. H. Fischer, From thesauri towards ontologies?, in: Structures and relations in knowledge organization - 5th International ISKO Conference, Lille (France), 18–30, 1998.

[2] D. Soergel, Indexing Languages and Thesauri: Construction and Maintenance, Melville Pub. Company, 1974.

[3] B. Lauser, From thesauri to Ontologies. A short case study in the food safety area in how ontologies are more powerful than thesauri, Agricultural Information and Knowledge Management Paper, Food and Agriculture Organisation of the United Nations, Rome (Italy), 2004.

[4] J. Lacasta, J. Nogueras-Iso, J. Teller, G. Falquet, Transformation of a keyword indexed collection into a semantic repository: applicability to

[10] `http://viaf.org/`

the urban domain, in: International Conference on Theory and Practice of Digital Libraries, vol. 6966 of *LNCS*, Berlin (Germany), 372–383, 2011.

[5] P. Mika, D. Oberle, M. Sabou, A. Gangemi, Foundations for Service Ontologies: Aligning OWL-S to DOLCE Alignment to Foundational Ontologies, in: Proceedings of the Thirteenth International World Wide Web Conference, New York (USA), 563–572, 2004.

[6] M. Ehrig, Ontology Aligment: Bridging the Semantic Gap, Semantic Web and Beyond: Computing for Human Experience, Springer, 2007.

[7] M. A. Hearst, WordNet: An Electronic Lexical Database, chap. Automated Discovery of WordNet Relations, MIT Press, 131–152, 1998.

[8] N. Guarino, Formal Ontologies and Information Systems, in: Proceedings of FOIS'98, Trento (Italy), 3–15, 1998.

[9] C. van Rijsbergen, S. Robertson, M. Porter, New models in probabilistic information retrieval, British Library Research and Development Report 5587, British Library, London (UK), 1980.

[10] International Organization for Standardization, Thesauri and Interoperability with other Vocabularies, ISO 25694, International Organization for Standardization (ISO), 2010.

[11] H. Cunningham, D. Maynard, K. Bontcheva, Text Processing with GATE, University of Sheffield Department of Computer Science, 2011.

[12] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, Sweetening WORDNET with DOLCE, AI Magazine 24 (3) (2003) 13–24.

[13] A. Miles, S. Bechhofer (Eds.), SKOS Simple Knowledge Organization System Reference, W3C Candidate Recommendation, W3C, 2009.

[14] International Organization for Standardization, Information and documentation - The Dublin Core metadata element set, ISO 15836, International Organization for Standardization (ISO), 2003.

[15] J. Lacasta, J. Nogueras-Iso, F. J. Zarazaga-Soria, Terminological ontologies. Design, Management and Practical Applications, chap. A representation framework for terminological ontologies, Springer, 25–53, 2010.

[16] F. J. Lopez-Pellicer, J. Lacasta, A. Florczyk, J. Nogueras-Iso, F. J. Zarazaga-Soria, An Ontology for the representation of Spatio-Temporal Jurisdictional Domains in Information Retrieval Systems, International Journal of Geographical Information Science 26 (3) (2011) 579–597.

[17] S. Ebener, Z. E. Morjani, Y. Guigoz, The Second Administrative Level Boundaries data set project (SALB). A working platform for improving data sharing, in: 4th EnviroInfo Conference, Geneva (Switzerland), Poster, 2004.

[18] International Organization for Standardization, Country Codes, ISO 5964, International Organization for Standardization (ISO), 2010.

[19] D. Tudhope, H. Alani, C. Jones, Augmenting Thesaurus Relationships: Possibilities for Retrieval, Journal of Digital Information 1.

[20] B. J. Wielinga, A. T. Schreiber, J. Wielemaker, J. A. C. Sandberg, From Thesaurus to Ontology, in: Proceedings of the 1st international conference on Knowledge capture, Victoria (Canada), 194–201, 2001.

[21] J. Golbeck, G. Fragoso, F. Hartel, J. Hendler, B. Parsia, J. Oberthaler, The national cancer institute's thesaurus and ontology, Journal of Web Semantics 1 (1) (2003) 1–5.

[22] C. Chun, L. Wenlin, From agricultural thesaurus to ontology, in: 5th Agricultural Ontology Service (AOS) Workshop, Beijing (China), 27–29, 2004.

[23] D. Soergel, B. Lauser, A. Liang, F. Fisseha, J. Keizer, S. Katz, Reengineering Thesauri for New Applications: the AGROVOC Example, Journal of Digital Information 4 (4).

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

[25] F. Khosravi, A. Vazifedoost, Creating a Persian ontology through thesaurus reengineering for organizing the Digital Library of the National Library of Iran, in: International Conference on Libraries, Information and Society, ICoLIS 2007, Petaling Jaya (Malaysia), 19–36, 2007.

[26] M. Hepp, J. de Bruijn, GenTax: A Generic Methodology for Deriving OWL and RDF-S Ontologies from Hierarchical Classifications, Thesauri, and Inconsistent Taxonomies, in: Proceedings of the 4th European Semantic Web Conference (ESWC 2007), vol. 4519 of *LNCS*, Springer, Innsbruck (Austria), 129–144, 2007.

[27] Z. Aleksovski, M. Klein, W. ten Kate, F. van Harmelen, Matching Unstructured Vocabularies using a Background Ontology, Lecture Notes in Computer Science 4248 (2006) 182–197.

[28] C. van Damme, M. Hepp, K. Siorpaes, FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies, in: Bridging the Gap between Semantic Web and Web 2.0 Workshop. ESWC 2007, Innsbruck (Austria), 57–70, 2007.

[29] B. Vatant, Ontology Driven Topic Maps, in: XML Europe 2004, Amsterdam (Netherlands), 2004.

[30] B. Sridharan, H. Deng, B. Corbitt, An ontology-driven topic mapping approach to multi-level management of e-learning resources, in: 17th European Conference on Information Systems, Verona (Italy), 1187–1198, 2009.

[31] I. Niles, A. Pease, Towards a standard upper ontology, in: 2nd International Conference on Formal Ontology in Information Systems, Ogunquit (USA), 2–9, 2001.

[32] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, in: 16th international conference on World Wide Web, Banff (Canada), 697–706, 2007.

[33] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, DBpedia - A crystallization point for the Web of Data, Web Semantics: Science, Services and Agents on the World Wide Web 7 (3) (2009) 154–165.

[34] C. Gutierrez, C. Hurtado, A. Vaisman, Temporal RDF, in: Proceedings of the European Conference on the Semantic Web (ESWC '05), Heraklion (Greece), 93–107, 2005.

[35] C. Hurtado, A. Vaisman, Reasoning with Temporal Constraints in RDF, Lecture Notes in Computer Science 4187 (2006) 164–178.