Enriching thesauri with ontological information: Eurovoc thesaurus and DALOS domain ontology of consumer law

Maria Angela Biasiotti[1] and Meritxell Fernández-Barrera[2]

[1] CNR-ITTIG,Via de Barucci, 20,
50127 Florence (Italy)
biasiotti@ittig.cnr.it,
[2] European University Institute, Via dei Roccettini, 9,
I-50014 San Domenico di Fiesole (FI) Italy
Meritxell.Fernandez@EUI.eu

Abstract. This paper analyses the semantic shortcomings of thesauri in comparison with ontologies in the framework of the trend to building KOS that enable IR by concept search instead of textual search. A particular case study in the domain of the consumer law is presented, in which the differences in terms of semantic depth between the Eurovoc thesaurus and the DALOS ontology are analysed. Moreover the paper analyses the existing technical solutions for semantically enriching thesauri, and explores which would be the possibilities in the case of the Eurovoc thesaurus taking into account that a great number of documents have already been indexed with its descriptors.

Key words:
Thesauri, Ontologies, Legal semantics, Conceptual Information Retrieval

1 Framework

The purpose of the Semantic Web approach is to make web content machine-processable in order to develop new functionalities beyond the mere display of data, such as enabling a better access to relevant information contained in web documents. One of the main problems of the WWW is information overload [8] and the limited software performance in extracting useful information for users. It is thus desirable to develop improved techniques for accessing quickly not only relevant documents, but the bits of information embedded in them that specifically match the user queries.

The paper presents an analysis of the different semantic depth in which a specific legal domain (the consumer law domain) is represented by different KOS: the Eurovoc thesaurus and the DALOS domain ontology of consumer law. Given the shortcomings of Eurovoc and after having reviewed the state of the art in semantic enrichment of thesauri, a transition procedure to support the shift from a traditional KOS, like Eurovoc, towards a full-fledged and semantically rich KOS is suggested.

Before analysing the two different representation models adopted by Eurovoc and by Dalos in the Consumer Protection field we will briefly outline which are the main characters and features of these considered models.

2 Case study. Representation of the consumer law domain

2.1 Knowledge Organization Systems

As to the different metadata structures that can channel the addition of semantics to legal information, we can mention several ways of attaching meaning to the information contained in the web, like catalogs, thesauri and frames.

More specifically and in brief ontologies are controlled vocabularies expressed in an ontology representation language (OWL), whereas taxonomies or semantic nets are a collection of controlled vocabulary terms organized into a hierarchical structure, and thesauri are networked collections of controlled vocabulary terms. Each term in a taxonomy is in one or more parent-child relationships to other terms in the taxonomy, while in a thesaurus associative relationships are used in addition to parent-child relationships. In more detail the latter can be defined as a classification tool to assist libraries, archives or other centres of documentation to manage their records and other information.

This functionality is achieved by establishing paths between terms. The establishment and development of a thesaurus is generally arranged in accordance with the standards of ISO (International Standards Organization), which are especially recognised at international level.

In this context, the main difference between thesauri and ontologies is actually the degree of semantic precision with which they describe contents. On the one hand, a thesaurus embodies a terminological representation of a domain (a particular lexicalisation of a conceptualisation) which is not as semantically complete as the formal conceptual representation provided by an ontology, and its limited structure makes it therefore unsuitable for advanced semantic applications [12]. In particular, the relationships linking the terms (the controlled vocabulary to represent concepts) in a thesaurus (BT, NT, RT) are usually not enough for a deep analysis of the semantics of the indexed documentation. On the other hand, ontologies provide a deeper conceptual representation of the domain (with a richer set of relationships between concepts like *part of*, *instance of*, *role*, among others, depending on the semantic domain) and can therefore enhance better access to the content of specialised documents.

In this framework, the purpose of this paper will be to explore the suitability of different KOS for representing the semantics of a specific legal domain and to analyse how a semantically simpler KOS (a thesaurus) can be enriched. We will show the need to move towards this trend by analysing two different knowledge representation models provided for the consumer protection domain.

2.2 Representation of Consumer Law Domain

The legal domain chosen for our case study is the consumer law domain. It is chosen just as a preliminary study to analyse the semantic depth of Eurovoc as regards legal issues. However, in further work it would be desirable to extend the analysis to other legal domains.

From a textual analysis of the the Consumer law discipline, namely *Consumer Protection*, it arises that relevant concepts to be considered among others are: *Advertising, misleading advertising, commercial communication, Market surveillance, inspection, disclosure, Commercial activity, Contract, selling price, unit price, consumer goods, product, raw material, products sold in bulk, agricultural products, finished product, services, all inclusive services, information society services, financial services, producer, buyer, consumer, trader, contract, unfair-terms, credit-agreement* and so on. In particular we will analyse the representation of this particular domain provided by two KOS: (i) the Eurovoc thesaurus and (ii) the Dalos ontology of consumer law.

Case study set-up In this section the specific representation of the consumer protection domain provided by the two considered models will be offered, outlining shortcomings and needs.

The first KOS considered, Eurovoc, has not a proper framework for the consumer protection law within sector Law. It sets it within sector 20 devoted to Trade, in the Micro Thesaurus Consumption. Therefore the consumer protection is just the NT of the top term consumer and has some RT relationship with other relevant concepts such as *advertising, producer's liability, publishing of prices* and so on.

The *consumer* descriptor has then hierarchical relationships with other relevant terms such as *consumer information, European consumer information agency, consumer movement, product quality, product designation, product life, product safety, defective product.*

From the EUROVOC scenario it emerges that the description provided by Eurovoc of *Consumer* has few constraints with respect to the Consumer law protection. In fact, Eurovoc pertains to the category of traditional thesauri, structured on hierarchical and synonymy relations, with a rigid structure and inter-lingual relations but a poor semantics. Its scope is broad (European policy issues), and the components devoted to the normative domain are very weak in precision and granularity. As the focus is on socio-economic issues, depth in law is quite low and the structure is not appropriate to EU law. One of its main limitations is therefore that it is not suitable for the indexation and the search of specialised documents.

It is indeed quite flat and unconsistent with respect to the domain: the consumer is a top term and has only some hierarchical relationship with other descriptors such as *consumer information, product safety*, and a few others.

Generally speaking the Eurovoc model presents:

- *lack of expressivity and granularity*: the concepts considered are very few with respect to those emerging form the legal sources (for instance, a non-descriptor is foreseen for *consumers rights*, but no descriptor exists to refer to a specific instance of this concept, like *withdrawal right*);

- *lack of semantics*: the thesaurus relationships are semantically overloaded, in the sense that the same relationship is used to express different semantic links where three different possible ontological relations have been identified for the relation NT, and two different relations for RT.
- *lack of legal orientation*: relationships are fixed and inexpressive of the domain relationships and meaning. They are limited to BT, NT, RT, UF and are therefore not specific to the domain. For instance, the link between *Consumer and consumers right* is firstly, not direct, since *consumerss right* is a non-descriptor for *consumer protection*, which is a NT of Consumer. Secondly, the relation is not specific to the domain, it is just a generic hierarchical relationship, whereas in a deeper semantic model the representation could be that *Consumer has-right-towards* someone else.

As Eurovoc was drafted exclusively for manual indexing and retrieval purposes, the lack of semantic precision generates frequent inconsistencies among several hierarchical and synonymy relations so that it is mainly suitable for retrieving related terms. In the same way as all existing thesauri it is focused on documentation and lacks sufficient granularity for semantic access to EU law.

The ontology of consumer law developed in the frame of the DALOS project[3] provides a formal representation of the classes of the world entities involved in the domain of consumer law (*agents, actions, legal roles, legal effects*).

Relevant concepts are all present as captured directly from the legal sources by the NLP techniques. They are well identified into five classes Agent, Quality, Region, Event and Object and linked to each other by some significant relationships. The contextualization provided by the Dalos domain ontology is consistent with respect to the domain since:

- there is a higher conceptual expressivity and granularity, as proven by the big number of domain specific concepts: for instance *withdrawal-right; damage-compensation; consumer-complaint.*
- relationships between concepts have a legal orientation (are specific to the domain).
- and relationships are not semantically overloaded, that is, a semantic relation is not used with more than one meaning, unlike RT and NT in Eurovoc.

Comparing the two different approaches in representing the other relevant concepts of the Consumer protection law, such as advertising, we obtain the following scenario:

The two models considered produce two different representations as to granularity, expressiveness, constraint and consistency.

Descriptors in Eurovoc are not consistent with respect to the specific domain as they are identified independently for the concepts embedded in the domain itself. They are identified according to a top-down approach by an expert who

[3] `http://www.dalosproject.eu/`

considers those terms relevant for describing the domain without any reference to the legal texts. Whereas the Dalos ontology has been built according to a bottom-up approach which enables experts to take into due consideration the richness, the concepts and relations arising directly from the text, that are the sources of law ruling on the Consumer protection field.

Indeed Eurovoc pretends to describe the consumer protection domain with around 23 terms, whereas the DALOS ontology aims at doing the same with over 100 concepts.

Eurovoc shortcomings as regards the representation of the consumer law domain could thus be tackled by enriching the simplified conceptualisation represented by its terminology with deeper conceptual relations specific to the domain. In the following section we will briefly review some technical solutions for designing a more complete semantic model building on a pre-existing semantic resource.

3 Techniques for enriching semantically lexical resources

From an analysis of the literature dealing with the enrichment of metadata resources with a deeper semantics it is possible to identify different techniques. On the one hand, those that rely on the combination of pre-existing resources and on the other hand those that simply restructure a pre-existing resource.

The techniques belonging to the first trend (highlighted in [9]) maintain the autonomy of both resources and therefore avoid an actual *merging*. They are the following:

- *restructuring* a computational lexicon following ontological principles, focusing thus on the lexical resource as the final output and using the ontology merely as a guiding tool. The resulting lexical resource respects certain ontological restrictions but does not become a full-fledged ontology.
  Some of the suggestions for a better semantics of lexical resources include: making a proper use of the is-a link so that it expresses not only lexical relations but ontological ones (-this would amount to using the is-a relation only to link entities that share similar identity criteria-); avoiding the confusion between concepts and instances, avoiding the subsumption of types by roles as derived from the ONTOCLEAN methodology- and not mixing different levels of generality, among others [5].
- *populating* an ontology with lexical information, which amounts to mapping lexical units to ontological entries;
- *aligning* an ontology with a lexical resource, that is, combining the restructuration of the computational lexicon according to ontological-driven principles and its mapping with the ontological resource.

With regard to the second trend, we can mention those techniques that simply draw on a single pre-existing resource, namely a thesaurus and exploit the possibility of transforming it into a more complex knowledge representation structure, that is, into an ontology. The main difference from the previous techniques is that in this case the input consists just of one initial semantic resource (the thesaurus) and that the nature of the resulting resource changes: the thesaurus not only gains in semantic precision and structure, but it becomes a full-fledged ontology, that is, a rigid semantic representation of the domain with a hierarchy of classes and corresponding slots or properties that increases in constraint density[4] (a higher set of relationships linking its terms). Several works have shown concerns on the connection between thesauri and ontologies ([1]; [6]) and discussed their different semantic scope. A general perspective on the possibility of converting thesauri into ontologies is given by [13] and specific examples of thesaurus reengineering are [10].

4 Applicability of existing techniques to eurovoc: approaches for improvement

As to the techniques that could be applied to achieve our goals, several possibilities arise.

4.1 Mapping of EUROVOC to pre-existing legal ontologies

The first one, corresponds to the mapping of lexical entries and ontological classes (populating the classes of the ontology with the thesaurus terms). This may in some cases be feasible, where in Eurovoc there exists a term *advertising* and in Dalos ontology a concept with the same name. In this case an equivalence relationship could be added to the Dalos ontology to link the Dalos *advertising* concept and corresponding term in Eurovoc, creating a direct link between the two resources. However, in some other cases populating the ontology with Eurovoc classes might be more difficult, where Eurovoc has no corresponding term for the Dalos concept *product*.

Indeed, since Eurovoc is a general thesaurus it therefore does not match the semantic requirements of a specialised domain such as the consumer law, and there will be cases in which it will not be possible to find an Eurovoc term corresponding to a class of the DALOS ontology. This is why it will be necessary, in the process of mapping the ontology to the thesaurus, to enrich the semantics of Eurovoc adding more terms. This would amount to making some kind of restructuring of the thesaurus, but in this respect, we have to take into account an important restriction, namely, that Eurovoc is currently used to index documents and in order not to loose this indexation it is necessary to maintain current Eurovoc terms. This imposes limitations on a possible restructuration of the thesaurus, for it will not be possible to delete any descriptors even if semantic consistency might require to do so in some cases.

A further difficulty of following the approach of using the thesaurus to populate a pre-existing ontology is that it would be necessary to identify ontologies

[4] The notion of constraint density is introduced by [9] as the *density of the "network of constraints" that holds between the concepts.*

for several specialised domains of the law, if the whole structure of Eurovoc referring to the legal field wants to be expanded semantically. Dalos would be the option for the consumer protection domain but it might turn up more difficult to find ontologies of other specific legal areas, such as international law or environmental law.

4.2 Reengineering Eurovoc into a formal ontology

A different option would be to use merely the Eurovoc thesaurus and transform it into a full-fledged ontology. This would require providing the thesaurus with a well structured semantics and to represent it in a highly expressive language like OWL. In order to do that, it would be necessary to:

- increase specificity and granularity by adding more classes corresponding to the legal domain;
- refining the relationships existing between the different terms of the thesaurus (BT, NT, RT, UF, USE) according to the semantics of the domain (adding therefore other types of relations);
- checking ontological consistency (ontological constraints).

5 Results

The proposed approach tries to meet the new trend arising from the Semantic Web towards the development of legal KOS able to allow the user to search by concept instead of searching by words. This implies the use of tools able to expand the query from a semantic point of view, meaning that the concept identified is surrounded by other concepts semantically linked to it. In this framework, the paper has analysed the semantic scope of a traditional KOS (Eurovoc) used for the indexation and retrieval of legal information in the EU institutions and national parliaments.

The paper has provided some concrete evidence on the semantic shortcomings of the Eurovoc thesaurus for the representation of the legal domain by analysing how it represents the particular domain of the consumer protection law in comparison to DALOS domain ontology. The main findings are, firstly, that in Eurovoc there is a lack of semantic granularity, since many relevant concepts of the domain are not represented by its descriptors; relations linking terms are semantically overloaded and therefore only shallowly expressive; and relations are not specific of the legal domain, but generic (RT, BT, NT, UF). Taking into account that currently one of the main functionalities of the Eurovoc thesaurus is to be used as an indexing and searching tool of legal documentation (it is for instance used by Eur-lex, the gateway for accessing European Union law), it arises the need of equipping the thesaurus with more powerful conceptual structures specific to the legal domain in order to improve search and legal information retrieval.

Secondly, the paper assesses which are the technical possibilities, according to the state of the art, for enriching semantically the Eurovoc thesaurus. Two methods are highlighted as feasible solutions: on the one hand, using Eurovoc to populate pre-existing ontologies on specific legal domains; on the other hand, transforming Eurovoc into a full-fledged ontology by enriching its conceptual structure and expressing it using formal representation languages.

6 Further work

The paper presents some preliminary conclusions as to possible directions to solve the problem of the semantic limitations of a current indexing and retrieval tool used at EU and national level for legal documentation. However, further research is foreseen in order to implement the project:

- A whole assessment of the semantic representation of the legal domain by Eurovoc: building on the case study presented in this paper that analyses the representation of the domain of the consumer law, further analysis of Eurovoc semantic representation of other legal domains is required.
- Analysis of the benefits that the proposed approach would bring about: run experiments to measure the degree of improvement of information retrieval tasks by the use of an ontological structure instead of Eurovoc structure and analyse the benefits for the various users of Eurovoc (EU institutions, national parliaments, private users with licence).
- Analysis of the costs of implementing the approach: in terms firstly, of the KOS reengineering costs: and secondly, the adaptation of current information systems to the new KOS.

References

1. Arano, S.: Thesauruses and ontologies [on line]. Hipertext.net, 3, (2005) <http://www.hipertext.net> [Consulted: 07/01/09]. ISSN 1695-5498
2. Benjamins, Casanovas, P., Gangemi, A., Selic, B. (Eds.): Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications. Springer Verlag. Berlin Heidelberg 5, (2005)
3. Coulthard, Malcolm and Johnson, Alison: An Introduction to Forensic Linguistics. Language in Evidence. London and New York Routledge, (2007)
4. Cross, P., Brickley, D., Koch, T.: Conceptual relationships for encoding thesauri, classification systems and organised metadata collections and a proposal for encoding a core set of thesaurus relationships using an RDF Schema. Available in: http://www.desire.org/results/discovery/rdfthesschema.html, (2000)
5. Gangemi, A., Guarino N., Oltramari, A.: Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top Level In Formal Ontology in Information Systems. Proceedings of FOIS2001, eds. C. Welty and S. Barry, 285-296. New York: Association of Computing Machinery, (2001)
6. García Jiménez, A.: Instrumentos de representacin del conocimiento: tesauros versus ontologías. Anales de Documentación, 7, 79–95, (2004)
7. Hirst, D.: Ontology and the Lexicon Handbook on Ontologies. Information Systems. Springer, (2003)

8. Lazonder et al.: Differences between Novice and Experienced Users in Searching Information on the World Wide Web. Journal of the American Society for Information Science, 51(6), 576–581, (2000)
9. Oltramari, Prévot, Borgo: Theoretical and practical aspects of interfacing ontologies and lexical resources. Proc. of the 2nd Italian Semantic Web workshop SWAP 2005 (Semantic Web Applications and Perspectives), Trento, (2005)
10. Qin, J., Paling, S.: Converting a controlled vocabulary into an ontology: the case of GEM. Information Research, 6, 2. Available in: `http://informationr.net/ir/6-2/paper94.html`, (2000-01)
11. Sartor, G.: Legislative information and the web. Biasiotti et al.: Legal Information Management of Legislative Documents, (2008)
12. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J, Katz, S.: Reengineering thesauri for new applications: the AGROVOC example. Journal of Digital Information, 4(4), (2004)
13. Wilson, M.: Migrating from Thesauri to Ontologies. Available in: `http://www.w3c.rl.ac.uk/ukofficepasttalksindex.html`, (2002)