TECHNICAL REPORT

ESPON

Towards an ESPON thesaurus? Some preliminary considerations for the thematic structuring of the ESPON database

MAIN RESULTS

- Construction of corporate thesaurus for structuring knowledge and facilitate information retrieval needs to comply with international guidelines
- Literature review shows a wide range of examples developed by international agencies and organisations on structured and controlled vocabulary
- Qualitative and quantitative text analysis improves the ability in defining standards and therefore to further advance on the harmonisation and coherence of concepts developed by ESPON
- For this purpose, it is important to determine links between different terms, namely by word co-occurrence analysis
- Lastly, we argue that such exercises are important to support the thematic structuring of the ESPON database by linking terms with data for information retrieval by potential users

ESPON 2013 DATABASE

MARCH 2011

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

19 pages

LIST OF AUTHORS

Nuno Madeira, University of Luxembourg
Geoffrey Caruso, University of Luxembourg

Contact

E-mail: nuno.madeira@uni.lu; Tel. +352 46 66 44 9691
E-mail: geoffrey.caruso@uni.lu; Tel. +352 46 66 44 6625

TABLE OF CONTENTS

Introduction

The thematic structuring of a geographical database aims at synthesising information through typologies and indicators that can address European political priorities and support applied research. The anonymous part of the application form specifies the main goals to be achieved by this research project. Within this context, the thematic dimension of data plays a crucial role, particularly due to the necessity of connecting geographical information to political objectives. Our goal is therefore to explore ways of progress on this issue and further advance on the structuring of the database into different themes and sub-themes that may support ESPON needs and its scientific objectives.

Over the last few years a great number of statistical agencies and international organisations have developed corporate glossary databases to integrate data and metadata in order to reinforce the communication and dissemination of standard concepts and benefit from the potentialities of terminology management tools for information retrieval by the end-users.

The seminal book *"Organizing Knowledge: An Introduction to Managing Access to Information"* says, in its most recent edition, that *"…organization of knowledge is the other face of information retrieval. The better organized that knowledge is, the easier it is to retrieve specific items of knowledge…"* (Rowley & Hartley, 2008: 4). The point of departure is therefore the relationship between data, information and knowledge. More concretely, how these concepts apply together for the formulation of a common set of terminology that supports the management of controlled vocabulary for storage and information retrieval?

According to these authors, depending on the environment and purpose of our needs different tools for organizing knowledge and information retrieval may apply. These tools have characteristics in common that improve the ability in defining standards and further advance in the harmonisation and coherence of concepts. One of the most interesting examples developed over the last decades is the thesaurus and its potentialities for alphabetical display of words and expressions that are linked by standardised semantic relationships. In the following chapter of this report some of these features are briefly described.

1. Main features for thesaurus construction

From the moment information and data started to be easily accessible online many international agencies and organisations demonstrated their interest in having a corporate thesaurus to structure knowledge facilitate information retrieval. In order to ensure a minimum structure several international guidelines have been designed such as those formulated by the ISO. From a strictly information retrieval sense, a thesaurus is defined in the current international standard, ISO 11179, as the vocabulary of a controlled indexing language, formally organised so that relationships between concepts are made explicit. Its aim is to match the vocabulary used by the indexer with the language of the searcher, and thus improve the retrieval of relevant information.

The following section introduces some basic concepts about thesaurus. According to Noguera-Iso et al (2005) a thesaurus is an organised collection of terms enriched with relations that are linked to one another by cross-referencing to the database. A term is a word (or expression) that represents a conceptual category and for that reason a thesaurus can be seen as a structured and controlled vocabulary of words either represented in monolingual or multilingual format.

The adoption of such semantic tools has the benefit of ensuring that a subject will be described using the same preferred term each time it is indexed and this will make it easier to find all information about a specific topic during the search process. This means that using a thesaurus improves search results due to the inclusion of information about the relationships of words or expressions that represent standardised relationship indicators. The thesaurus specifies which word or expression will be used as the preferred

(or authorised) term for a particular concept and then connects the non-preferred terms or synonyms to it. The purpose is to provide a wide range of paths to reach those terms. Ultimately, it works as a classification tool to ease the communication process for indexers and information seekers that need to speak a common language.

Against this background, it is appropriate to take advantage from certain initiatives that stimulate the implementation of international standards and guidelines for wide dissemination of data and metadata exchange. Both INSPIRE and SDMX initiatives have the benefit of improving harmonisation and coherence of data. Since the ESPON DB 2013 project combines two types of data and metadata: (1) territorial statistical metadata (i.e. NUTS, WUTS, LAU) and (2) spatial metadata (i.e. environmental data), it is crucial to integrate both categories.

Due to the particularities of the ESPON programme access to data and metadata will be made available only in English language. With this regard, it should be mention that the management of monolingual thesauri is less complex but at the same time reduces the effectiveness for non-speaking English users of using ESPON results and evidence.

The purpose of building semantic tools (i.e. thesauri) is to maintain not only relationships among surface terms but also be able to deal with ambiguities. To this end, the thematic structuring of the database must be based on ESPON themes, sub-themes and keywords. The purpose is two-fold: in one hand, it would structure the database into different themes that address ESPON lexical concepts developed over the previous programme and, on the other hand, it would allow a more comprehensive use of the database by potential end-users (i.e. researchers, academics, practitioners, policy-makers).

However, some difficulties emerge within this discourse due to the complexity of political priorities. At this stage it is not obvious how the thematic structuring of the ESPON database should be organised. For this same reason it is important to have a minimum knowledge of what type of thematic fields should be added to an hypothetic ESPON thesaurus when storing data (and metadata) provided by the other applied research projects. With this regard, text and lexical analysis software applications show clear advantages and the ESTI framework, as described in the *"ESPON Handbook for Data Collection, Harmonisation and Quality Control"*, can provide reasonable solutions for the automatic exploration of most of the data under investigation.

This brief overview is also important to understand that the structure of a thesaurus is generally defined a priori and terms are organised by cross-references that have generally three different types: (1) equivalence relationship; (2) associative relationship; and (3) hierarchical relationship. The following section provides a general description of these types of standardised relations (Nogueras et al, 2005; Severino, 2007).

The Equivalence Relationship

This relationship concerns relations among terms that are considered equivalent inside the thesaurus. In other words, it covers synonyms. When two or more terms express the same concept, one of these is selected as preferred term. The equivalence relationship is expressed by the following: U or USE which leads from a non-preferred term to synonyms or quasi-synonyms which act as preferred terms, and UF or USED FOR the reciprocal, which records entry terms leading to the synonyms (e.g. spatial planning is a synonym for spatial development).

The Hierarchical Relationship

The hierarchical relationship is the most important feature that distinguishes a thesaurus from an unstructured list of terms, such as glossary. It is based on position ranking, from a superior to a subordinate position. Terms that are superior in a category are at a super-ordinate level; terms that fall under or below a category are at a subordinate level. The broader term (BT) represents a class or a whole, while the narrower term (NT) refers

to components or part of the broader concept (e.g. Mediterranean countries is the broader term and Malta the narrower term).

In this sense, the hierarchical relationships links terms to other terms expressing more general and more specific concepts, hierarchically related terms are grouped under general subdivisions, which in turn are grouped into areas ok knowledge. Normally, hierarchical relationships are indicated by the prefixes BT (broader term), i.e. label for the super-ordinate descriptor, between a specific descriptor and a more generic descriptor; and NT (narrower term), i.e. label for the subordinate descriptor, between a more generic descriptor and a more specific descriptor.

The Associative Relationship

The associate relationship is used when two terms overlap in meaning. This relationship may be symmetrical or asymmetrical. In both cases, it indicates that a term has similarities with other concepts. However, some features must be distinguished.

A related term relationship alerts users to the fact that other information of interest may be classified under a different, but related, set of terms. Related terms relationships are only established between preferred terms. This type of relationship is symmetrical and therefore the relationship between terms is non-hierarchical (e.g. GDP is related to economic development and economic development is related to GDP). It means that similar terms (or related terms) are conceptually associated, but not equivalent or hierarchical related. The associative relationship is normally expressed by the abbreviation RT (related term). To what regards asymmetrical terms, the relationship is similar but not reciprocal. In other words, there is no related-term reference in the opposite direction (e.g. ageing population is related with demographic trends, but there searching for demographic trends is unlikely to be interested in ageing population).

2. Thesauri developed by international organisations

The following section describes some of the examples of online thesauri developed by international organisations, namely ILO (International Labour Organisation) and UNESCO (United Nations Educational, Scientific and Cultural Organisation), EnVoc (United Nations Environment Programme), GEMET (General Multilingual Environmental Thesaurus of the European Environmental Agency), Theseus (Statistical Office of the European Communities), and OECD Macrothesaurus (Organisation for Economic Co-operation and Development).

International Labour Organisation

The ILO thesaurus may be browsed in two ways: alphabetically and hierarchically. It derives from a list of more than 4000 terms in the field of economic and social development (e.g. labour and employment policy, vocational training, working conditions) used to index and retrieve information. Potential new terms are monitored to see how their use develops in the literature and discussions are held with indexers and subject specialists to decide on their inclusion in the ILO thesaurus. Every term is presented in English, French, and Spanish. Many are followed by definitions or explanatory notes. Besides, ILO Thesaurus is available electronically and it can be easily downloaded in the three above mentioned languages.

Figure 1. Print screen of the ILO Thesaurus online interface.
Source: www.ilo.org/thesaurus

United Nations Educational, Scientific and Cultural Organisation

The UNESCO thesaurus is a controlled vocabulary developed by the United Nations Educational, Scientific and Cultural Organisation which includes subject terms for several areas of knowledge (education, culture, social and human sciences), names of countries and groupings of countries (political, economic, geographic, ethnic and religious, and linguistic groupings). The terms are linked together by the three types of relationships described before, i.e. hierarchical, associative and equivalence relationships.

The UNESCO Thesaurus also includes scope notes which explain the meaning and application of terms. The same happens to French and Spanish equivalents of English

preferred terms. A more complete description of the structure of this thesaurus can be obtained in the report *"UNESCO Thesaurus: A Structures List of Descriptors for Indexing and Retrieving Literature in the Fields of Education, Science, Social and Human Science, Culture, Communication and Information"*. A general description of the structure for each thesaurus by themes, sub-themes and related terms is illustrated as an annex to this report and addresses a particular focus on themes potentially related to the ESPON database.

United Nations Environment Programme

EnVoc is a multilingual thesaurus developed by UNEP (United Nations Environment Programme) with a controlled and structured vocabulary for the use in indexing, storing and retrieval of environmental information. It contains categorized and alphabetical lists of subjects and keywords in context list and is available in the six official United Nations languages (i.e. Arabic, Chinese, English, French, Russian, and Spanish). Additionally a number of other governments have undertaken the translation into their national languages which testifies the importance given to environment terminology and the value of the thesaurus as a reference tool (UNEP, 1997).

This initiative has its origin on the Infoterra thesaurus of environmental terms also developed by UNEP and constitutes revised version that reflects emerging environmental concerns and new technologies, especially in the field of environmental information. The thesaurus has evolved from an unstructured list of keywords with very limited application to a structured environmental thesaurus with a broad application base comprising several different types of expertise (e.g. librarians, database developers, GIS specialists).

As a subject, environment is multidisciplinary by nature. Its study, therefore, covers a wide range of linkages to disciplines and consequently it would be very difficult, though not impossible, to conduct a thesaurus containing every environmental term. The terms listed in EnVoc are drawn from a relatively high level in the environmental terminology hierarchy. This strategy eliminates unnecessary duplication with more specialised thesauri and the end product is closely integrated, compact and more practical to use.

European Environmental Agency

The processes of exchanging environmental information have significantly increased around the world during the last years and that reality motivated the development of a standard multilingual vocabulary capable of promoting and facilitating information exchange among countries on key environmental issues. The project on global environmental thesaurus was designed by a joint cooperation of different institutions with profound experience in this field.

The example provided by EEA (European Environmental Agency) is very comprehensive because it created important synergies among different organisations that had previously worked on the definition of common terminologies.

Indeed, shortly after setting up EEA it was launched an initiative called GEMET (General Environmental Multilingual Thesaurus). The idea was to merge in a unique thesaurus the best terminologies of six multilingual thesauri, including EnVoc, in order to propose a reference thesaurus to different organisations and, at the same time, provide an agreed common language basis for the exchange of environmental information. The initiative assumed great relevance in satisfying three basic functions, namely: define a system of controlled term, build a multilingual dictionary, and develop a glossary. In the late 1990s the final version of GEMET was published by EEA. The current version is available in 27 languages, and contains over 6,000 descriptors and access is open to all users in several formats. It can be browsed and searched on-line, accessed trough Web services or downloaded as HTML or SKOS files (cf. www.eionet.europa.eu/gemet).

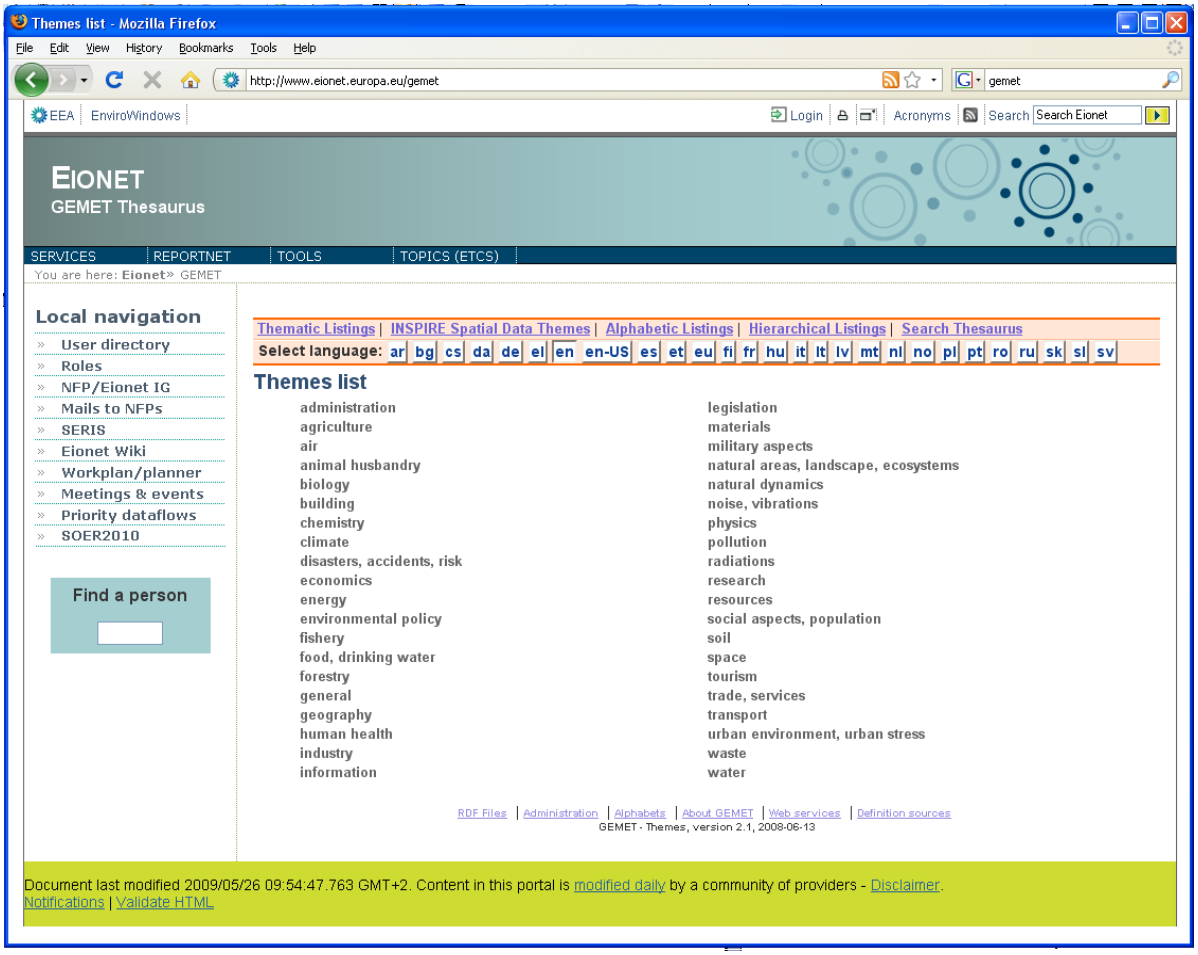

Figure 2. Print screen of the GEMET Thesaurus online interface.
Source: www.eionet.europa.eu/gemet

Organisation for Economic and Co-operation Development

This thesaurus has not been entirely developed by OECD. It represents a joint cooperation initiative in association with United Nations and the International Organisation of Labour to create some unity and consistency on economic and social development terms that are used in the international context. To this end, the constitution of the OECD Macrothesaurus used information from several sources and compiled the terms into a single scheme that would be useful to all constituents. Table 1 in Annex gives an overview of the structured word tree by themes and sub-themes.

Food and Agricultural Organisation of the United Nations

The AGROCOV Thesaurus (cf. www.fao.org/agrovoc) was developed by FAO (Food and Agricultural Organisation of the United Nations) and the Commission of the European Communities, in the early 1980s. The initiative was designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains (e.g. environment). Due to the potentialities for indexing and retrieving data at world-wide level it is available in the five official languages at FAO (i.e. English, French, Spanish, Chinese and Arabic).

The AGROVOC Thesaurus is currently structured as a traditional thesaurus and stored in a relational database. It consists of words and expressions (terms), organized in semantic relationships and used to identify or search resources. The main role is to standardize the indexing process in order to make searching simpler and more efficient, and to provide the user with the most relevant resources.

As a multilingual, structured and controlled vocabulary, AGROVOC is made up of terms, which consist of one or more words representing always one and the same concept. For each term, a word block is displayed, showing the semantic relationship to other words or expressions. Scope notes are also used to clarify the meaning and the context of terms (words and expressions).

The rationale behind the process of structuring this thesaurus is similar to the examples described in the previous sections of this technical report, particularly to what regards the semantic relationships between concepts. These relationships (i.e. equivalence, hierarchical, associative) provide the scope and structure for the thesaurus. However, such initiatives can greatly beneficiate from the add-value of using extended sets of relationships that go beyond the standardised relationships to perform more granular and more consistent indexing, and to enable more effective searching and browsing for users.

Ultimately, it is worth mentioning that AGROVOC is available free of charge and users are encouraged to propose new terms for inclusion in the database. Figure 3 illustrates a print screen of the online interface.

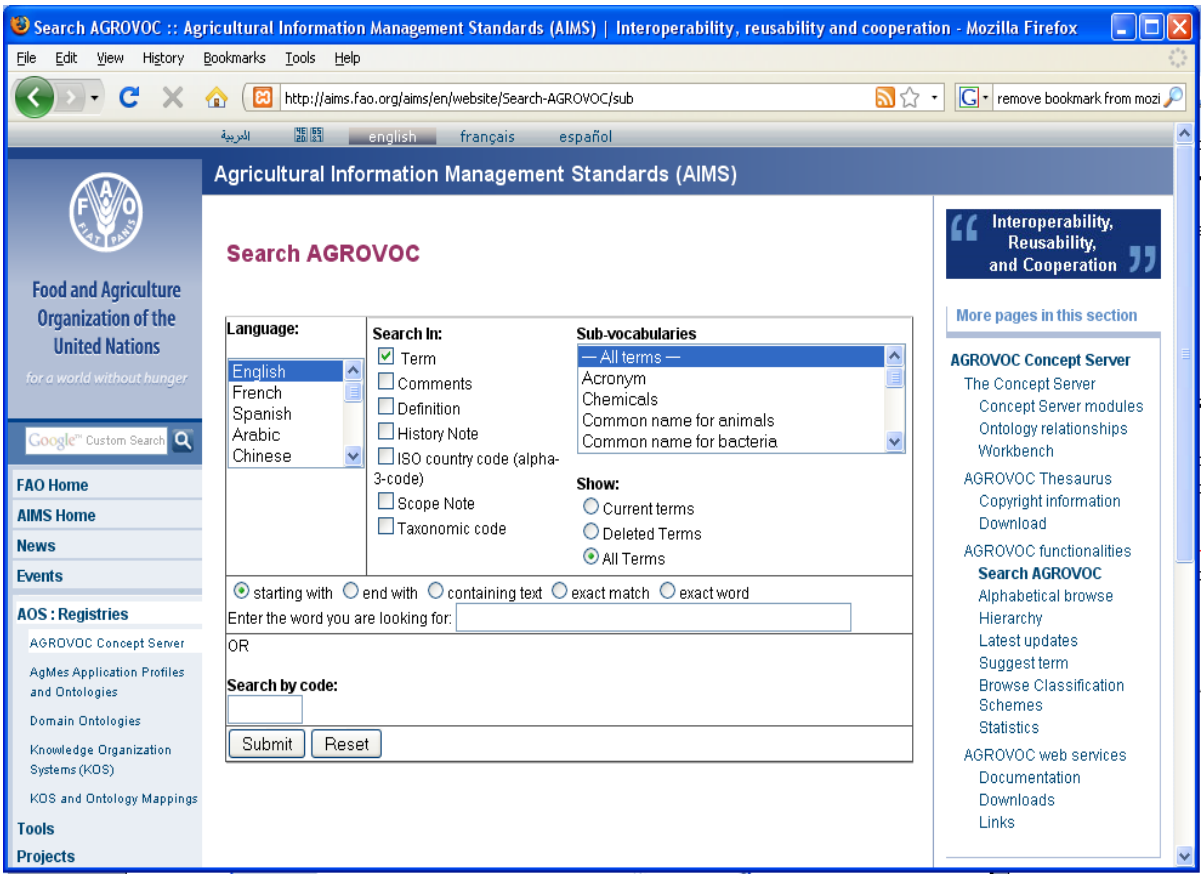

Figure 3. Print screen of the AGROVOC Thesaurus online interface.
Source: aims.fao.org/aims

Statistical Office of the European Communities

Eurostat, like many other international organisations, also developed a thesaurus-based system for consultation, indexation and retrieval of statistical data in order to help end-users to find information in its major reference database, *NewCronos*[1]. However, shortly after the beginning of this initiative the project was abandoned due to a negative priority given by Eurostat (De Norre; Groenez & Pellegrino, 2004). As an indicative reference, Figure 1 shows the linkages defined by Eurostat for statistical information systems.

[1] *NewCronos* is a unique database of macroeconomic and social statistics at EU level that includes data from candidate countries, EFTA countries, and major economic partners of the European Union (e.g. Japan, United States).

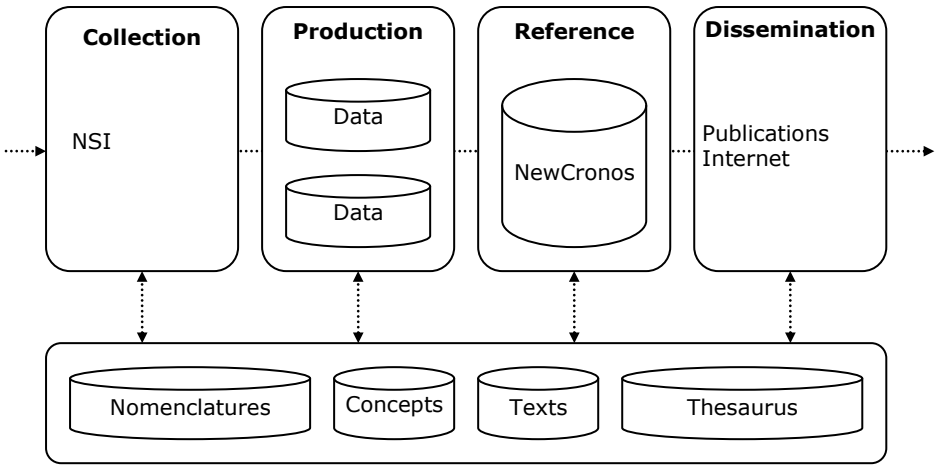

Figure 1. Eurostat's architecture for statistical information systems.

The initiative dates back the beginning of this decade and the logic behind the process has been described by De Norre & Groenez (2002) as a structured list of expressions used for indexing the content of data (i.e. statistical tables) and therefore ease the searching process by potential end-users. *Theseus*, as it was called, has been initially designed to be a multilingual thesaurus that would link major metadata of *NewCronos* to the contents of the thesaurus.

The agreement to further advance *Theseus* involved the development of an information management system and a web-interface for consultation. The content also required expertise from a group of specialists having the appropriate linguistic and content knowledge to regularly follow the evolution of the *NewCronos* database. In addition, it required expertise in adopting a more natural lexical instead of a rather technical language. For this reason, *Theseus* intended to suppress the ambiguities of natural lexical, in particular synonymy and homonymy.

The grounds for adopting a thesaurus were similar to the ones that lead other organisations to demonstrate the same approach, i.e. improving the information retrieval process by offering a multilingual tool to support potential users in searching data and information. To this end, *Theseus* structured its vocabulary in standardised semantic relationships and domains of knowledge. Therefore, the technical aspects of *Theseus* have been structured, at the first level, in semantic classes that would assemble all terms that have a specific meaning for a given sub-topic.

However, the purpose of this initiative goes far beyond the features offered by other online thesauri previously described in this report. In fact, this initiative also involved the capacity of linking metadata developed to define and describe statistical data and their structure to reference database *NewCronos*. This process was the basis for implementing the search facilities and three different approaches were considered in the indexation procedure.

The automatic indexation created too many problems and was abandoned without leaving any visible results. According to De Norre & Groenez (2002), automatic indexation was not compatible with high quality indexation and often required manual validation. The outputs obtained by this process were quite arguable.

The second approach followed a two-step procedure. Its rationale was analysed and developed by using a semi-automatic indexation that, firstly, compared metadata texts with *Theseus* keywords and, secondly, made use of information science expertise to determine appropriate links between metadata and keywords. The difficulties encountered during the test phase revealed that this option was not the best possibility. In general, the indexation process was too imprecise and did not take into account implicit concepts.

Finally, the third approach proved to be much more efficient. The manual indexation procedure provided some useful information on the quality of the metadata mostly due to the introduction of dictionaries defining both the *NewCronos* structures (i.e. themes, domains, collections, groups, subjects) and tables (i.e. titles).

Despite the efforts in finding a proper indexation procedure for the Eurostat thesaurus, the truth is that other major problems needed a solute on. One of those main issues raised during the indexation procedure was how to keep the index coherent in respect with the daily evolution of *NewCronos* and *Theseus*. In other words, it was necessary to find a proper logging and reporting procedure of the changes to support the work of information science specialists. Consequently, depending of the type of changes specialised staff could opt on the best suitable indexation procedure (i.e. manual, semi-automatic) to make the necessary updates which, in this case, were applied in a much more limited scope.

Ideally, *Theseus* proposed different search facilities based on the indexation. Some of the features included: search in via the index; selection of keywords or synonyms via thematic list of semantic categories, or permuted index (i.e. tracking the word in keyword expressions); combination of keywords (i.e. Boolean equations); and feedback to refine the search by offering context information derived from the thesaurus (i.e. semantic categories, hierarchical relations and scope notes).

The purpose of building a thesaurus seemed to be unquestionable due to the given importance attributed by Eurostat to its high-quality statistical information services, particularly when it comes to provide statistical information and promote harmonisation of statistical methods across Europe. Those ambitions anticipated as well some further technical developments regarding the integration with other tools developed by Eurostat, such as RAMON (Metadata Server), or CODED (Concepts and Definitions Database).

3. Towards an harmonised vocabulary at the EU level

The examples described previously demonstrate an increasing interest by international organisations in creating thesaurus for structuring their knowledge and support information retrieval.

The Eurovoc thesaurus is a multilingual and thematic thesaurus focusing in all fields which are of importance for the activities of the European institutions. It has been developed as a result of cooperation between the European Parliament, the European Commission and Publications Office with support from DG Information Society and Media. The thematic scope of Eurovoc is defined by 21 fields, divided into 127 microthesauri, and the last version contains 6645 descriptors, 6669 hierarchical and 3636 associative relationships.

Some fields are more developed than others due to the interests expressed by the European institutions. However, this situation does not mean that the thematic structuring has been organised using a logic method. On the contrary, the grouping of descriptors into fields is to a certain extent arbitrary. This means that certain descriptors can relate to two or more subject fields, but in order to make the thesaurus easier to manage and to limit its size it is generally accepted that limits have to be established according an hierarchy. Descriptors which could fit into two or more subject fields are assigned only to the field in which seems to be most natural for users.

In terms of languages and standards adopted by the Eurovoc thesaurus, the method is rather similar with other online thesauri developed by international organisations. It is published in all official languages of the European Community and complies with standards defined by the International Standards Organization, namely ISO 2788 for monolingual thesauri and ISO 5964 for multilingual thesauri. The structure of thesauri has been defined with regard to semantic relationships. Within the framework of the Eurovoc thesaurus, four different relationships may apply, namely: (1) micro-thesaurus, (3) equivalence, (4) hierarchical, (5) associative relationship and, finally, (5) scope note.

All descriptors are accompanied by a reference to a microthesaurus in order to show, under the abbreviation of MT, which microthesaurus or microthesauri they belong. For instance, within the field *04 Politics* it has been included several microthesauri, such as *0406 Political framework* or *0411 Political party* (cf. Table 2 in Annex).

The equivalence relationship between descriptors and non-descriptors is shown by the abbreviations: UF (used for), between the descriptor and the non-descriptor(s) it represents; USE between a non-descriptor and the descriptor which takes its place. The equivalence relationship covers relationships of several types: (a) genuine synonymity, or identical meanings; (b) near-synonymity, or similar meanings; (c) antonymy, or opposite meanings; (d) inclusion, when a descriptor embraces one or more specific concepts which are given the status of non-descriptors; because are not often used.

The hierarchical relationships between descriptors are shown by the abbreviations:

BT (broader term) between a specific descriptor and a more generic descriptor, together with a number showing the number of hierarchical steps between the specific descriptor and each broader term. Occasionally, certain descriptors in fields 72 Geography and 76 International Organizations are regarded as multi-thematic due to the presence of several broader terms at the next higher level, and descriptors with no broader terms are top terms. NT (narrower term) between a generic descriptor and a more specific descriptor, together with a number showing the number of hierarchical steps between the generic term and each narrower term.

The associative relationship between descriptors is shown by the abbreviation RT (related term) between two associated descriptors. The most significant features of the associative relationship are connected with its symmetrical approach, its incompatibility with the hierarchical relationship. Indeed, the descriptors under the same top term cannot be linked by an associative relationship, i.e. if two descriptor are linked by a hierarchical relationship there cannot be an associative relationship between them and inversely. Moreover, this semantic relationship can be interpreted in various forms (e.g. cause and effect, sequence in time or space, among others).

Ultimately, some descriptors are accompanied by scope notes, introduced by the abbreviation SN, and containing either a definition, if this clarifies the meaning of the descriptor or guidance on how to use the descriptor when indexing and formulating queries.

The Luxembourg-based Office for Official Publications of the European Communities (OPOCE) acts, on behalf of the European Commission, on management issues related to the Eurovoc thesaurus, especially regarding the necessary adaptations to the latest developments in which the Community institutions are active.

This special feature of Eurovoc is aimed at meeting the needs expressed by its users while keeping in mind the objective of covering in homogeneous manner, the fields of activity of the European Community. To this end, Eurovoc is subject to maintenance actions based on both internal and external needs and all users are invited to contribute to the development of Eurovoc by suggesting the addition, deletion or modification of descriptors, non-descriptors and semantic relationships.

Finally, it is worth mentioning that OPOCE plays a central role in the dissemination policy of the whole European Commission by proving tools, such as the web-based EU-Bookshop application, for searching and retrieving publications. The Eurovoc Steering Committee has officially adopted the version Eurovoc 4.3 on 28 November 2008 and shortly after that has been made available online (cf. http://europa.eu/eurovoc).

4. Text analysis applications for thematic structuring

In order to ensure a comprehensive thematic structuring of the ESPON database and therefore identify, retrieve and navigate statistical and geographical data it is essential to integrate spatial concepts developed by the previous ESPON programme. Within this

context, MAXQDA is a powerful tool for qualitative and quantitative data analysis. The simplest function of this application determines an index of word occurrences in a set of texts for statistical analysis.

The potentialities of this software have been initially tested on the ESPON report *"Europe in the world"*. This publication constitutes a synthesis of the Project 3.4.1 commissioned by the ESPON 2006 programme and it has approximately 30,000 words. However, for this exercise some words have been excluded due to lack of importance, namely defined and undefined articles, propositions, conjunctions, and numerals. The results achieved demonstrate an interesting occurrence of words that might be important to consider when defining themes, sub-themes and keywords for an ESPON thesaurus intended to store data developed by applied research projects. As an illustration, the following figure shows a random selection of some of the most common words found at least ten times in the above mentioned report.

Figure 2. Most commonly used words in the 'Europe in the World' report.
Note: This illustration has been created by using wordle, a web-based application that generates "word clouds" in order to give greater prominence to words that appear more frequently in the source text.

In a certain way these words represent the ESPON lexical concepts introduced by the applied research projects of the previous ESPON programme and consequently the role played by these words in meaning-construction of semantic units might have been extremely important in influencing the terminology adopted by other European institutions. Against this background, similar exercises of text analysis in documents and reports produced by other European institutions (e.g. EEA, EUROSTAT, OECD, European Commission, European Parliament, and Committee of the Regions) could further support the definition of a thematic structure for the database. A very important benefit of such exercises relates to the ability of defining standards that can lead to improved harmonisation and coherence of spatial concepts and, in doing so, preserve the overall meaning of terms for information retrieval by end-users.

A second approach to enrich our analysis and support the definition of an hypothetic ESPON thesaurus would be measuring the significance of certain terms (or expressions) to appear associated with other terms (or expressions). That is to say, determine co-occurrence of words. This approach could be performed by collecting a great number of reports from distinct sources and determine the relationships among words. The most common distribution of words could be used to accurately assess the significance and therefore contribute to the definition of possible relationships of words.

5. Final considerations

This technical report explores the structure of online thesauri developed by international organisations. These tools are aimed at easing access to knowledge for information

retrieval. To this end, the examples provided illustrate different approaches that vary according to the needs. However, the methods used for thesaurus construction are generally the same and, to a certain extent, represent joint cooperation initiatives among international organisations that cover similar fields or categories. The purpose of such initiatives is to create some consistency and unity. For instance, the constitution of the OECD Macrothesaurus corresponds to a joint initiative among United Nations, the International Labour Organisation, and the Organisation for Economic Co-operation and Development. In the same way, the development of GEMET is a result of the cooperation between the United Nations Environment Programme and the European Environmental Agency.

The purpose of having a specific thesaurus for the ESPON database would certainly contribute to generate consistency and enhance searches more effectively by end-users. In this sense, quantitative text analysis tools have the ability to support the definition of appropriate vocabulary for indexing terms. It has been argued that some of the lexical developed by the previous ESPON programme influenced the terminology adopted by other EU institutions and therefore similar exercises could further support the constitution of an ESPON thesaurus. Against this background, the ability to define standards and improve harmonisation and coherence of spatial concepts would preserve the overall meaning of words and expressions for information retrieval by potential end-users.

Next steps towards the thematic structuring of the ESPON database will involve qualitative and quantitative text analysis on various scientific reports (e.g. ESPON, INTERREG, INTERACT, DGs, CoR, EESC, European Parliament), determine links between different terms, namely by word co-occurrence analysis and types of relationships and, ultimately, link those terms with data for information retrieval (indicators, typologies) in order to structure the ESPON database.

References

De Norre, B. & D. Groenez (2002). *A Multilingual Thesaurus for Accessing Eurostat's Reference Databases*. Joint UNECE/EUROSTAT Work Session on Statistical Metadata. Luxembourg, 6-8 March 2002.

De Norre, B.; Groenez, D. & M. Pellegrino (2004). *Integrating Statistical Terminology Tools within Eurostat's Dissemination Policy*. Joint UNECE/EUROSTAT/OECD Work Session on Statistical Metadata. Geneva, 9-11 February 2004.

Diaz Munoz, P. (2008). The role of Statistical Data and Metadata Exchange in global statistical infrastructure. *Statistical Journal of the IAOS*, 25(2008): 47-54.

Felluga, B.; Plini, P.; Cunningham, G. & S. Lucke (2000). *The Global Environmental Thesaurus Project*. Global Conference on Access to Environmental Information. Dublin, 11-15 September 2000.

Nogueras, J. ; Zarazaga, F. & P. Muro (2005). *Geographic Information Metadata for Spatial Data Infrastructures. Resources, Interoperability and Information Retrieval*. Springer: Heidelberg.

OECD (2007). *Data and Metadata Reporting and Presentation Handbook*. Organisation for Economic Co-operation and Development: Paris.

Roe, S. & A. Thomas (2004). *The Thesaurus. Review, Renaissance, and Revision*. Routledge: London.

Rowley, J. & R. Hartley (2008). *Organizing Knowledge: An Introduction to Managing Access to Information*. Ashgate: London.

Severino, F. (2007). The term development in the thesauri of international organisations. *The European Journal of Development Research*, 19(2). 327-351.

UNEP (1997). *EnVoc Multilingual Thesaurus of Environmental Terms*. United Nations Environment Programme: Nairobi.

Williamson, N. & C. Beghtol (2004). *Knowledge Organization and Classification in International Information Retrieval*. Routledge: London.

Annexes

ILO Thesaurus	UNESCO Thesaurus
01 International relations 02 Social policy, social protection and social security 03 Economic development 04 Law, human rights, government and politics 05 Social sciences, culture, humanities and art 06 Education and training 07 Rural development, agriculture, forestry 08 Economic activities 09 Trade 10 Transport NT1 Transport NT1 Goods Transport NT1 transport Infrastructure NT1 Vehicles NT1 Air, sea, inland waterway, road and railway NT2 Mode of transportation NT2 Air transport NT2 Sea transport NT3 Port NT3 Waterway NT4 inland waterway NT 5 *canal* BT *Waterway* RT *Inland water transport* NT2 Inland water transport NT2 Railway transport NT2 Road transport NT1 Loading and packing NT1 International, inland and urban transport 11 Financing 12 Management 13 Labour and employment 14 Population, race relations and migration 15 Health and safety 16 Environmental sciences NT1 Ecology NT1 Natural resources NT2 Natural resources NT2 Resources development BT Natural resources NT1 Environment 17 Earth sciences 18 Research and science 19 Library and information science	Countries and Country Groupings Africa NT1 Africa NT2 Central Africa NT2 East Africa NT2 North Africa NT3 Algeria BT *Arab states* BT *French speaking Africa* BT *French speaking countries* BT *Islamic countries* BT *Maghreb* BT *Mediterranean countries* BT *North Africa* NT3 Egypt BT *Arab states* BT *French Speaking countries* BT *Mediterranean countries* BT *North Africa* NT3 Libya NT3 Morocco NT3 Sahara NT3 Tunisia NT2 Southern Africa NT2 West Africa NT1 Americas and the Caribbean NT1 Asia and the Pacific NT1 Economic Groupings NT1 Ethnic and Religious Groupings NT1 Europe NT1 Geographic Groupings NT1 Linguistic Groupings NT1 Political Groupings Culture Education Information and Communication Politics, Law and Economics Science Social and Human Sciences Urban Planning

Table 1. Brief comparison between ILO and UNESCO thesauri.
Source:

Note: Both examples illustrate a detailed overview of selected categories and its semantic relationships.

1. INTERNATIONAL COOPERATION; INTERNATIONAL RELATIONS
   1. International Cooperation
   2. International Relations
   3. International Organizations
   4. Countries and Regions
2. ECONOMIC POLICY; SOCIAL POLICY; PLANNING
   1. Economic Policy; Planning
   2. Social Policy
   3. Social Problems
   4. Social Services
3. ECONOMIC CONDITIONS; ECONOMIC RESEARCH; ECONOMIC SYSTEMS
   1. Economic research; Economics
   2. Economic Conditions
   3. Economic Systems
4. INSTITUTIONAL FRAMEWORK
   1. Law; Legislation
   2. Human Rights
   3. Politics
5. CULTURE; SOCIETY
   1. Social Sciences
   2. Culture
   3. Society
   4. Ethics; Religion
   5. Art
   6. Languages
   7. Communication
6. EDUCATION; TRAINING
   1. Educational Sciences
   2. Educational Development; Educational Policy
   3. EDUCATIONAL SYSTEMS
   4. EDUCATIONAL INSTITUTIONS
   5. CURRICULUM; TEACHING; LEARNING
   6. STUDENTS; TEACHING PERSONNEL
7. AGRICULTURE
   1. AGRICULTURAL ECONOMICS
   2. LAND ECONOMICS
   3. AGRICULTURAL ENTERPRISES
   4. AGRICULTURAL EQUIPMENT
   5. AGRICULTURAL PRODUCTION
   6. AGRICULTURAL RESEARCH
   7. PLANT PRODUCTION
   8. FORESTS
   9. ANIMAL PRODUCTION
   10. FISHERY
8. INDUSTRY
   1. INDUSTRIAL ECONOMICS
   2. INDUSTRIAL ENTERPRISES
   3. INDUSTRIAL ENGINEERING; INDUSTRIAL EQUIPMENT
   4. INDUSTRIAL PRODUCTION; INDUSTRIAL PRODUCTS
   5. INDUSTRIAL RESEARCH
   6. FOOD INDUSTRY
   7. WOODWORKING INDUSTRY; PULP AND PAPER INDUSTRY
   8. TEXTILE INDUSTRY; LEATHER INDUSTRY
   9. RUBBER INDUSTRY
   10. CONSTRUCTION INDUSTRY; CERAMICS INDUSTRY; GLASS INDUSTRY
   11. ENERGY
   12. CHEMICAL INDUSTRY
   13. MINING
   14. METALWORKING INDUSTRY
   15. ELECTRONICS; ELECTRICAL EQUIPMENT
   16. COMMUNICATION INDUSTRY
9. TRADE
   1. DEMAND; MARKET; CONSUMPTION
   2. PRICES
   3. MARKETING
   4. DOMESTIC TRADE
   5. INTERNATIONAL TRADE
10. TRANSPORT
   1. TRANSPORT ECONOMICS
   2. GOODS; PASSENGERS
   3. TRANSPORT INFRASTRUCTURE
   4. MEANS OF TRANSPORT
   5. MODES OF TRANSPORT
   6. LOADING; PACKAGING
   7. INTERNATIONAL TRANSPORT; URBAN TRANSPORT
   8. TRAFFIC
   9. FREIGHT
11. PUBLIC FINANCE; BANKING; INTERNATIONAL MONETARY RELATIONS
   1. PUBLIC FINANCE; TAXATION
   2. CURRENCIES; FINANCING
   3. INTERNATIONAL MONETARY SYSTEM
12. MANAGEMENT; PRODUCTIVITY
   1. ENTERPRISES
   2. ECONOMIC CONCENTRATION
   3. ENTREPRENEURS
   4. MANAGEMENT
   5. EQUIPMENT
   6. TECHNOLOGY
   7. PRODUCTION; PRODUCTIVITY
   8. PRODUCTS; PRODUCT DEVELOPMENT
   9. COST ACCOUNTING; PROFIT
13. LABOUR
   1. HUMAN RESOURCES
   2. EMPLOYMENT SERVICES; OCCUPATIONAL QUALIFICATIONS; PERSONNEL MANAGEMENT
   3. OCCUPATIONAL SAFETY
   4. DISMISSAL; LABOUR MOBILITY
   5. LABOUR RELATIONS
   6. WAGES; WAGE INCENTIVES
   7. LEISURE
   8. OCCUPATIONS
14. DEMOGRAPHY; POPULATION
   1. POPULATION DYNAMICS
   2. AGE GROUPS
   3. ETHNIC GROUPS
   4. HABITAT; RURAL AREAS; URBAN AREAS
   5. FERTILITY; FAMILY PLANNING
   6. MORTALITY
   7. MIGRATION
15. BIOLOGY; FOOD; HEALTH
   1. BIOLOGY; PARASITOLOGY; BIOCHEMISTRY
   2. ANATOMY; GENETICS; PHYSIOLOGY
   3. FOOD; NUTRITION
   4. MEDICINE; DISEASES
   5. PHARMACOLOGY; TOXICOLOGY
16. ENVIRONMENT; NATURAL RESOURCES
   1. ECOLOGY
   2. NATURAL RESOURCES
   3. DISASTERS; POLLUTION
   4. POLLUTION CONTROL; ENVIRONMENTAL ENGINEERING
   5. RESOURCES CONSERVATION
17. EARTH SCIENCES; SPACE SCIENCES
   1. METEOROLOGY
   2. CLIMATE
   3. GEOGRAPHY
   4. GEOPHYSICS; GEOLOGY; SOIL SCIENCES
   5. HYDROLOGY; WATER
   6. OCEANOGRAPHY
   7. SPACE SCIENCES
18. SCIENCE; RESEARCH; METHODOLOGY
   1. RESEARCH; SCIENCE
   2. ORGANIZATION OF RESEARCH
   3. RESEARCH METHODS; THEORY
   4. DATA COLLECTING
   5. EXPERIMENTS
   6. MEASUREMENT
   7. MAPPING
   8. MATHEMATICS; STATISTICAL ANALYSIS
   9. COMPARISON; EVALUATION
   10. FORECASTS; TIME FACTOR
19. INFORMATION; DOCUMENTATION
   1. INFORMATION
   2. DOCUMENTS
   3. TERMINOLOGY
   4. CONFERENCES

Table 2. List of themes and sub-themes of the OECD Macrothesaurus.

04 POLITICS
- 0406 political framework
- 0411 political party
- 0416 electoral procedure and voting
- 0421 parliament
- 0426 parliamentary proceedings
- 0431 politics and public safety
- 0436 executive power and public service

08 INTERNATIONAL RELATIONS
- 0806 international affairs
- 0811 cooperation policy
- 0816 international balance
- 0821 defence

10 EUROPEAN COMMUNITIES
- 1006 Community institutions and European civil service
- 1011 European Union law
- 1016 European construction
- 1021 Community finance

12 LAW
- 1206 sources and branches of the law
- 1211 civil law
- 1216 criminal law
- 1221 justice
- 1226 organisation of the legal system
- 1231 international law
- 1236 rights and freedoms

16 ECONOMICS
- 1606 economic policy
- 1611 economic growth
- 1616 regions and regional policy
- 1621 economic structure
- 1626 national accounts
- 1631 economic analysis

20 TRADE
- 2006 trade policy
- 2011 tariff policy
- 2016 trade
- 2021 international trade
- 2026 consumption
- 2031 marketing
- 2036 distributive trades

24 FINANCE
- 2406 monetary relations
- 2411 monetary economics
- 2416 financial institutions and credit
- 2421 free movement of capital
- 2426 financing and investment
- 2431 insurance
- 2436 public finance and budget policy
- 2441 budget
- 2446 taxation
- 2451 prices

28 SOCIAL QUESTIONS
- 2806 family
- 2811 migration
- 2816 demography and population
- 2821 social framework
- 2826 social affairs
- 2831 culture and religion
- 2836 social protection
- 2841 health
- 2846 construction and town planning

32 EDUCATION AND COMMUNICATIONS
- 3206 education
- 3211 teaching
- 3216 organisation of teaching
- 3221 documentation
- 3226 communications
- 3231 information and information processing
- 3236 information technology and data processing

36 SCIENCE
- 3606 natural and applied sciences
- 3611 humanities

40 BUSINESS AND COMPETITION
- 4006 business organisation
- 4011 business classification
- 4016 legal form of organisations
- 4021 management
- 4026 accounting
- 4031 competition

44 EMPLOYMENT AND WORKING CONDITIONS
- 4406 employment
- 4411 labour market
- 4416 organisation of work and working conditions
- 4421 personnel management and staff remuneration
- 4426 labour law and labour relations

48 TRANSPORT
- 4806 transport policy
- 4811 organisation of transport
- 4816 land transport
- 4821 maritime and inland waterway transport
- 4826 air and space transport

52 ENVIRONMENT
- 5206 environmental policy
- 5211 natural environment
- 5216 deterioration of the environment

56 AGRICULTURE, FORESTRY AND FISHERIES
- 5606 agricultural policy
- 5611 agricultural structures and production
- 5616 farming systems
- 5621 cultivation of agricultural land
- 5626 means of agricultural production
- 5631 agricultural activity
- 5636 forestry
- 5641 fisheries

60 AGRI-FOODSTUFFS
- 6006 plant product
- 6011 animal product
- 6016 processed agricultural produce
- 6021 beverages and sugar
- 6026 foodstuff
- 6031 agri-foodstuffs
- 6036 food technology

64 PRODUCTION, TECHNOLOGY AND RESEARCH
- 6406 production
- 6411 technology and technical regulations
- 6416 research and intellectual property

66 ENERGY
- 6606 energy policy
- 6611 coal and mining industries
- 6616 oil industry
- 6621 electrical and nuclear industries
- 6626 soft energy

68 INDUSTRY
- 6806 industrial structures and policy
- 6811 chemistry
- 6816 iron, steel and other metal industries
- 6821 mechanical engineering
- 6826 electronics and electrical engineering
- 6831 building and public works
- 6836 wood industry
- 6841 leather and textile industries
- 6846 miscellaneous industries

72 GEOGRAPHY
- 7206 Europe
- 7211 regions of EU Member States
- 7216 America
- 7221 Africa
- 7226 Asia and Oceania
- 7231 economic geography
- 7236 political geography
- 7241 overseas countries and territories

76 INTERNATIONAL ORGANISATIONS
- 7606 United Nations
- 7611 European organisations
- 7616 extra-European organisations
- 7621 world organisations
- 7626 non-governmental organisations

Table 3. List of themes and sub-themes of the Eurovoc Thesaurus.
Source: http://europa.eu/eurovoc.