

A Software Tool for Thesauri Management, Browsing and Supporting Advanced Searches

J. Nogueras-Iso, J.A. Bañares, J. Lacasta, J. Zarazaga-Soria

Computer Science and Systems Engineering Department
University of Zaragoza

{jnog | banares | jlacasta | javy}@unizar.es

ABSTRACT

Thesauri have been widely used to improve the precision and recall of information retrieval in digital libraries for decades. They provide a uniform and consistent vocabulary for indexing metadata ("description of the data holdings") and for supplying users with a suitable vocabulary for the retrieval. Therefore, a thesaurus management tool becomes a vital component in the development of any kind of digital library and this also extensible in the context of a Spatial Data Infrastructure (SDI), a digital library specialised in geographic information resources. This work presents a thesauri management tool with two main objectives: a basic management and browsing of thesauri; and providing some mechanisms to enhance cross-discipline interoperability between different thesauri.

INTRODUCTION

According to ISO (1986), a thesaurus is a set of terms that describe the vocabulary of a controlled indexing language, formally organized so that the a priori relationships between concepts (for example synonymous terms, broader terms, narrower terms and related terms) are made explicit. Thesauri have been widely used to improve the precision and recall of information retrieval in digital libraries for decades. They provide a uniform and consistent vocabulary for indexing metadata ("description of the data holdings") and for supplying users with a suitable vocabulary for the retrieval. Moreover, they can be used to expand users queries by automatically adding new terms to the query; or they can improve metadata interoperability for accessing information within networked knowledge organisation systems.

Therefore, a thesaurus management tool becomes a vital component in the development of any kind of digital library. This is also extensible in the context of a Spatial Data Infrastructure (SDI). The main objective of an SDI, apart from involving many other issues, is to provide the discovery, evaluation and access to spatial data for a community of users. And hence,

an SDI can be considered as digital library specialised in geographic information resources.

This work presents a tool which has been developed at the University of Zaragoza with two main objectives: a basic management and browsing of thesauri; and providing some mechanisms to enhance cross-discipline interoperability between different thesauri. Following section sketches the overall architecture of this tool. Then, section three presents the basic capabilities of the tool, basically the management, and browsing of thesauri according to the ISO norms for monolingual and multilingual thesauri (ISO (1986) and ISO (1985) respectively). Section four explains the enhanced capabilities of this thesaurus tool. These enhanced capabilities are mainly oriented to the semantic disambiguation of thesauri against an upper-level ontology as *WordNet* (Miller 1990). Around this semantic desambiguation, the tool provides facilities to expand automatically thesaurus terms with new terms from other thesauri having an equivalent meaning. Finally, this work ends with a conclusions and future lines section.

ARCHITECTURE

The tool has been developed in Java and it is deployed with two levels of operation: a simplified version which stores the thesaurus structure (only *BT*, *NT* relationships) on an Access 2000 database; and a complete version with full functionality that stores thesauri in an Oracle 9i database. The application allows to access to different databases via JDBC. The complete version takes advantage of the Oracle Intermedia Text package (CTX_THES) capabilities. This package implements the ISO-2788 norm for monolingual thesauri and also provides language translation relationships.

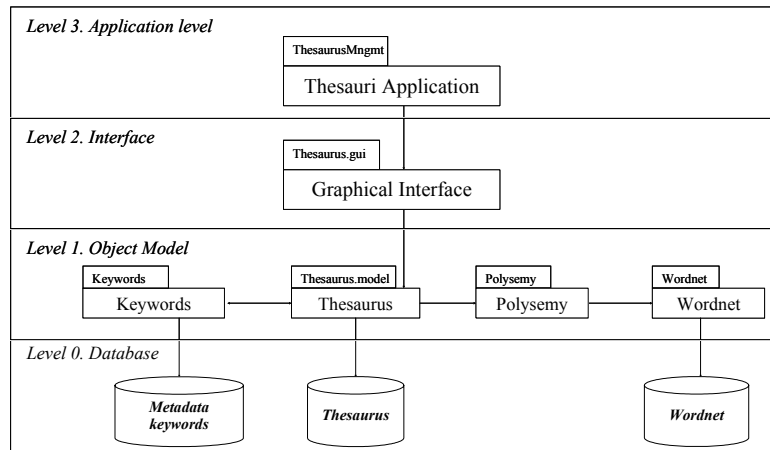


Fig. 1: Layered architecture of Thesaurus Tool

Fig. 1 shows the architecture of the application. As it can be observed, a four-layer architecture has been used to separate gradually the final user application from the lowest level, the database layer. The database layer contains *WordNet* lexical knowledge base, the thesauri stored in Oracle 9i, and a repository of metadata records (whose keywords may be expanded, see section ,Enhanced Capabilities'). The next level provides the object model that represents the information entities stored at the database level and the methods to access them. And in addition to this, this level also includes the classes for the extraction of polysemy, the semantic disambiguation of thesauri and the expansion of keywords due to the low-level services provided. The second level contains generic graphic interface components for the visualization of thesauri. And finally, the third level is the application level that provides the integrated tool for thesauri management, browsing, semantic disambiguation and keyword expansion functionality.

BASIC CAPABILITIES

This tool facilitates the edition of thesauri according to the ISO norms for monolingual and multilingual thesauri (ISO (1986) and ISO (1985) respectively). That is to say, it supports the definition of concepts and relationships between concepts including synonym terms (*SYN, USE*), broader terms (*BT*), narrower terms (*NT*), related terms (*RT*), preferred terms (*PT*)

and scope notes (*SN*) and language translations (*TR*). The visualization and browsing of thesaurus terms is possible with several graphical interface presentations. At present an alphabetical and a hierarchical presentation of thesauri is available. And to facilitate the discovery and visualization of terms, it is also possible to perform queries of "like" style queries. Another remarkable feature of the tool is its multilingual access support. Apart from the typical internationalisation of the application, the thesauri are browsed (if exists translation) according to the default system language or the language selected by the user.

Fig. 2 provides an overview of the graphical user interface of the application. There, it is shown how a user has displayed the UNESCO thesaurus (UNESCO 2003) and the GEMET thesaurus (GEMET 1999), selecting different presentations (alphabetical presentation for UNESCO and hierarchical for GEMET) and different languages (English for UNESCO and German for GEMET). Besides, the term *BIOCHEMISTRY* (or *BIOCHEMIE* in German) has been browsed in both thesauri, showing a different hierarchical path of terms in every case.

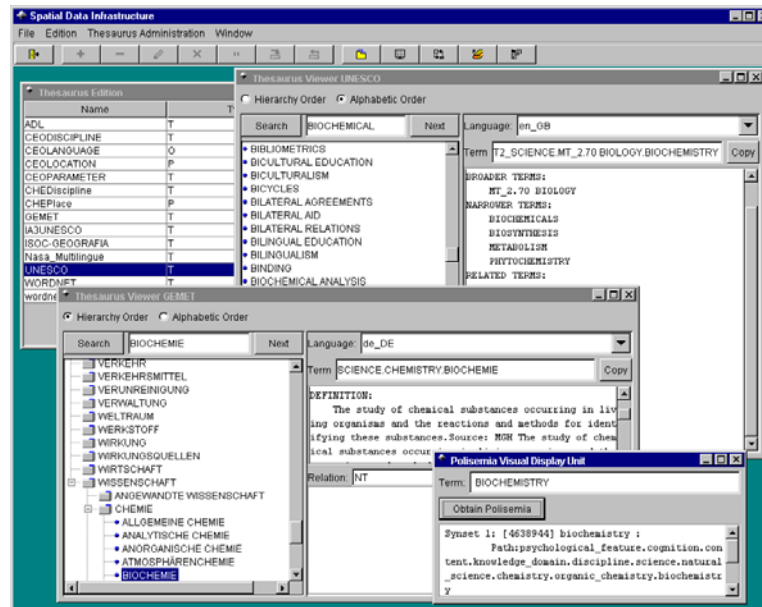


Fig. 2: Overview of the graphical user interface

As concerns import and export capabilities, this tool facilitates the exchange of thesauri by means of text files. These files may use two formats to encode the thesaurus phrases that belong to each thesaurus branch, where a branch represents the hierarchical tree (*BT/NT* relationships) whose root is a term with no broader terms in the thesaurus. With first format (see Fig. 5), each line contains the succession of narrower terms (separated by dots) from the branch root to the new term. Additionally, other relationships (*SYN*, *TR*, ...) are on the right side of the new term with a “+<relationship>+<value>”. Second format is similar to the previous one with the difference the branch terms are hierarchically numbered. The main limitation of these formats is that they are not standardized. Future improvements of this capability should be directed to the exchange of thesauri in interoperable formats like *RDFS/XML*.

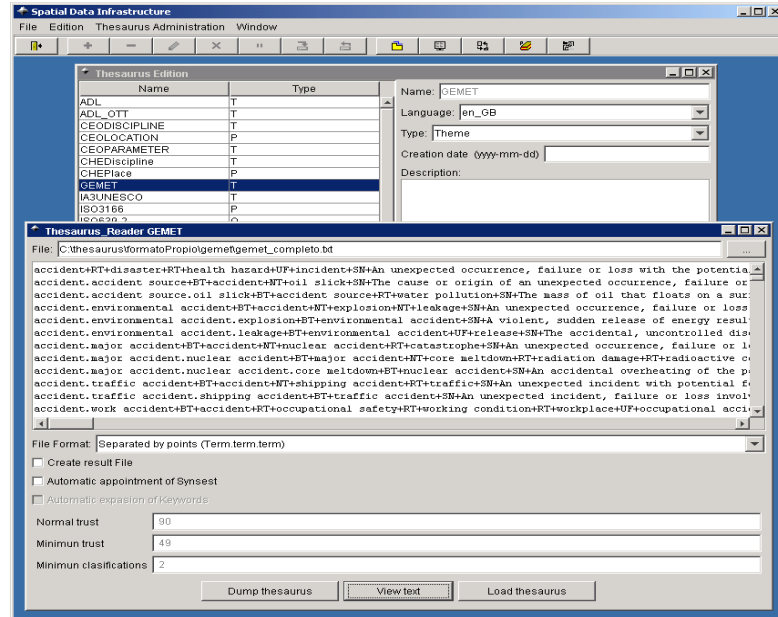


Fig. 3: Import/Export of thesauri

Finally, another feature of the basic operation of this tool is the incorporation of a users control system. Thanks to this, it is possible to deploy different versions of this application that prevent users without the required permission from updating the edition of thesauri and other restricted operations. For instance, we have included this tool as an additional component

to browse thesauri for a metadata edition tool (see Zarazaga, Lacasta et al. 2003) and there the user only has privileges to visualize the thesaurus terms, not to modify the content of the thesaurus.

ENHANCED CAPABILITIES

As it has been mentioned in the introduction, the enhanced capabilities of this tool are based on the semantic disambiguation of thesauri against the concepts of an upper level ontology (representation of concepts and their relationships in a particular application domain). Following subsections detail these enhanced capabilities.

Browsing of WordNet ontology

First of all, this tool allows the visualization of *WordNet* ontology, as if it were another thesaurus created by the tool (see Fig. 4). *WordNet* can be considered as an ontology of general level, which is structured in a hierarchy of *synsets*, where *synsets* are defined as set of synonyms representing a particular concept.

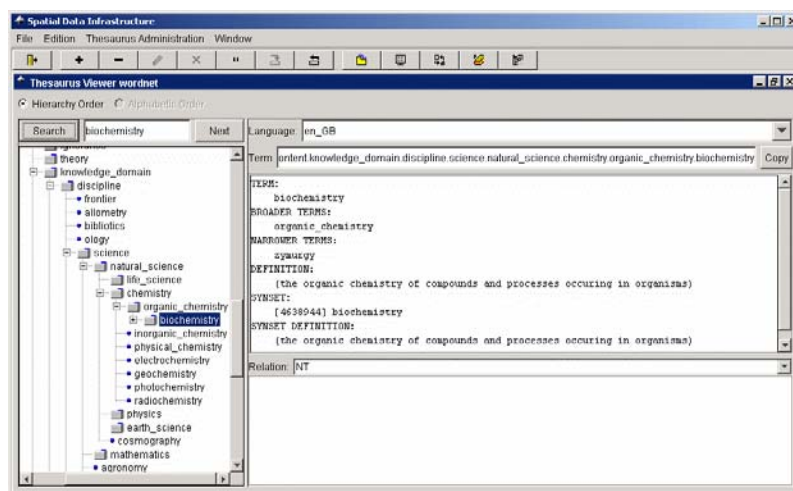


Fig. 4: Visualization of *WordNet* ontology

This functionality is provided by the Java *wordnet* package depicted in the architecture. This package facilitates the access to the libraries able to browse lexical database (see <http://www.cogsci.princeton.edu/~wn/> to

download *WordNet*). As the software of these libraries is implemented in C language and our application has been developed in Java, we had to implement the crosswalk that access to *WordNet* native libraries via JNI (Java Native Interface) and returns the information of *synsets* in the same way as the information related with thesauri created by the tool.

Search of polysemic senses in Wordnet

Given that this tool provides access to *WordNet*, it also facilitates the possibility to find the senses of a term (single word or set of related words) in *WordNet*. This functionality is provided at low level by the package *Polisemy*, which was presented in the architecture. Given a term, this package looks up it in the *WordNet* database and extracts all the possible *synsets*. In case a term is a compound term (more than one word) and is not directly included in *WordNet*, the *Polisemy* component would extract all *synsets* corresponding to each word in the compound term. Furthermore, this component uses morphological techniques to reduce the number of not-found words and to search the senses of adjectives which are associated with a noun. For instance, given the adjective *administrative*, the component will look the *synsets* associated with *administration*.

Fig. 5 displays the window that facilitates the extraction of *WordNet synsets* given a term or phrase. In this case, Fig. 5 shows the polisemic senses of term *administration*.

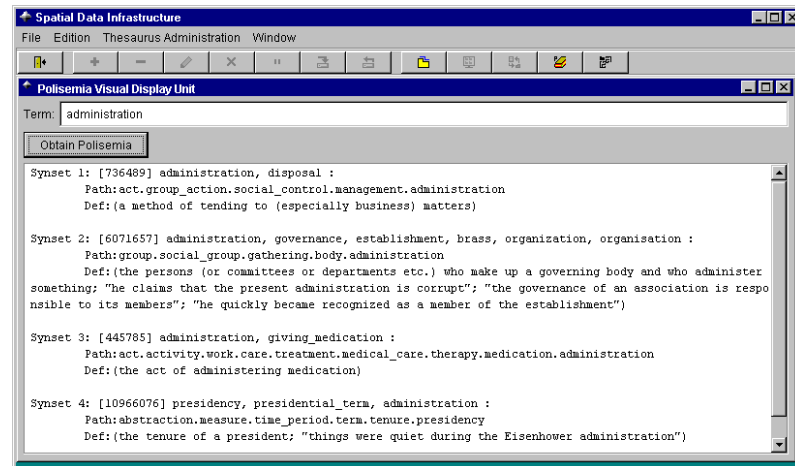


Fig. 5: Extraction of polisemic senses of term *administration*

Semantic disambiguation of thesauri

Perhaps, the most outstanding capability of this tool is the semantic disambiguation of thesauri in relation to *WordNet* concepts (or *synsets*). The tool applies an unsupervised disambiguation method that takes advantage of the thesaurus hierarchical structure (broader and narrower terms), which is used as the word context for a voting algorithm to find the closest sense. That is to say, given a thesaurus term with several *synsets* (or senses), we must determine what is the closest sense in relation with the rest of terms that conform a thesaurus branch (a branch represents the hierarchical tree of *BT/NT* terms whose root is a term with no broader terms). For that purpose, the method extracts the *synsets* of every term in the branch so as to confront each other and compute their semantic distance. For each ambiguous thesaurus term, the method will choose the *synset* having the highest degree of similarity with respect to the *synsets* of other terms in the branch. More details about the semantic disambiguation method can be found in (Mata, Bañares et al. 2002).

```

biosphere
It is not found in WordNet 1.6.
***
biosphere.anatomy
13.457159 alternative names for the body of a human being;
"Leonardo studied the human body"; "he has a strong physique"; "the
spirit is willing but the flesh is weak"
3.428149 the branch of morphology that deals with the structure of
animals
0.96582395 a detailed analysis; "he studied the anatomy of crimes"
Wrong assignment.
***
biosphere.anatomy.cell biology
13.984746 the basic structural and functional unit of all
organisms; cells may exist as independent units of life as in
monads or may form colonies or tissues as in higher plants and
animals
8.219941 a device that delivers an electric current as the result
of a chemical reaction
8.117833 small room in which a monk or nun lives
8.117833 a room where a prisoner is kept
7.248421 any small compartment; "the cells of a honeycomb"
1.5215139 a small unit serving as the nucleus of a larger political
movement
***
8.70637 all the plant and animal life of a particular region
4.149211 characteristic life processes and phenomena of living
organisms: "the biology of viruses"
3.0541077 the science that studies living organisms
***

```

Fig. 6: Tracing the disambiguation process. Cursive text has been manually included⁴.

The disambiguation method can be applied whenever a new thesaurus is imported. The *WordNet synsets* that are finally associated with the thesaurus terms are stored as a *TR* relationship (using *SYNSET* as language). This

⁴ Biosphere is not a noun of the WordNet 1.6, however it is included in the WordNet 1.7.

TR relationship is used to indicate the translation of a term and in this case *SYNSET* language it is interpreted as the disambiguation language. Fig. 6 shows a piece of the debug file which traces the disambiguation process whenever it is applied. Such file stores the possible *synsets* associated with each thesaurus term and the weight which was assigned by the disambiguation method. Finally, once the disambiguation method is applied, the user will be able to visualize the *synset* that was finally assigned (see Fig. 7) to each term in the same way as other relationships. In fact, we could update manually the *synset* assigned to a thesaurus term.

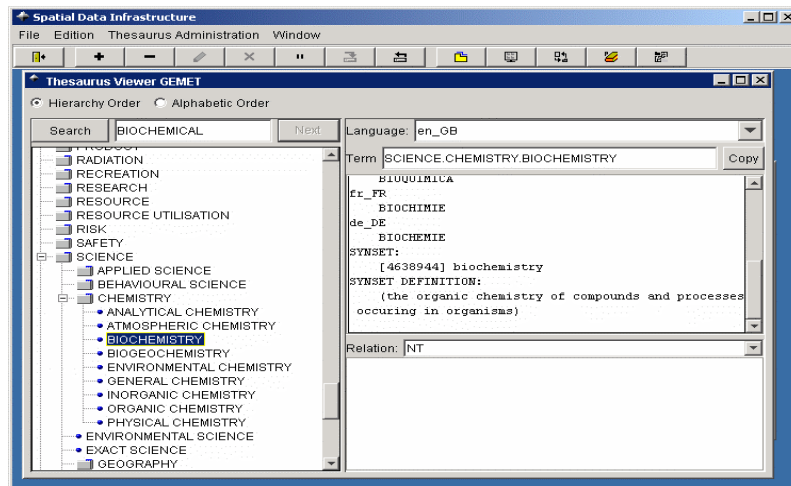


Fig. 7: Disambiguated senses are also shown by the thesaurus viewer

Keywords expansion

The final capability of this tool is the automatic expansion of a set of keywords. Given a set of keywords belonging to an initial set of thesauri, this tool will obtain a set of related terms belonging to a new thesaurus. This functionality is based on the semantic disambiguation of thesauri and is oriented to enhance classifications and description of resources. For instance, digital libraries can make profit of this method to expand the keywords stored in their metadata records and improve the performance of discovery services. On the other hand, although metadata creators usually select terms from well recognised thesauri, very different thesauri may have been used if metadata records describe resources from separate application domains.

Therefore, this method facilitates an homogenous classification of all meta-data records. The expansion will generate terms in other thesauri, which share a similar sense with the ones selected by metadata creator. This way, a user query containing a term in a specific domain, do not need to be enhanced with synonyms or semantic equivalent terms in other domains since, reducing as well the response time of discovery services. Additionally, this automatic expansion method could also be applied to the user query.

The method to expand the keyword section is based on a basic routine which estimates the probability to expand an original keyword section with *a new term* belonging to *a new different thesaurus*, not used in the original keyword section yet. This basic routine is composed of two main steps. First step is the collection of all the synsets corresponding to the terms, which were selected by metadata creator. As a result of this first step, we obtain *an initial collection of synsets*. Secondly, a comparison between the synsets of the *new term* and the *initial collection of synsets* is performed. This comparison consists of the computation of a reliability percentage for the *new term*, which is calculated as the number of synset coincidences divided by the number of synsets of the *new term* and multiplied by 99. The reason to use a final factor of 99 and not 100 is to obtain a maximum reliability percentage of 99 for automatically expanded terms, reserving uniquely a 100-reliability percentage for the terms which were originally selected by metadata creators. If this reliability percentage is greater than a threshold reliability percentage, which was defined previously by the user who performed the expansion, this *new term* is added in the keyword section.

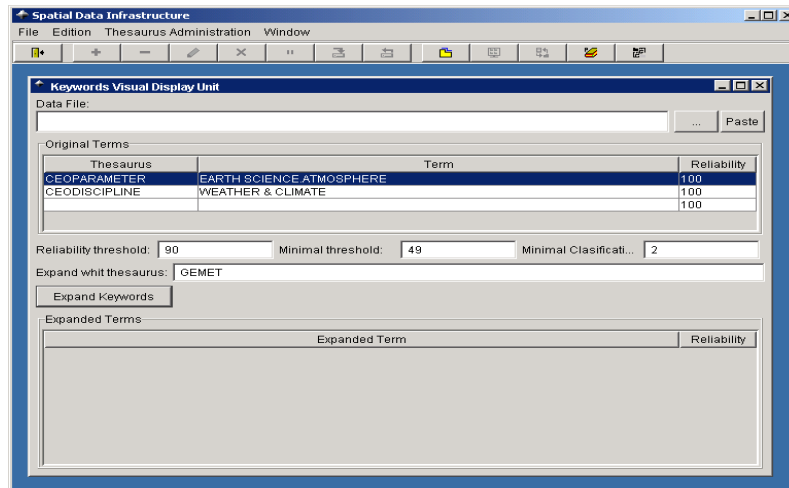


Fig. 8: Keywords expansion window

Fig. 8 displays the window that facilitates the expansion of keywords. As an example of this capability, the expansion of terms appearing in Tab. 1 (also introduced in Fig. 8) will be shown. These terms could correspond to the manual classification of a resource, which is included within the keywords section of a metadata record. The terms were selected from CEODISCIPLINE (a controlled list of 30 terms proposed to identify disciplines) and CEOPARAMETER (a controlled list of 1037 terms proposed to identify the types of features contained in a geospatial data resource) thesauri that are defined in (CEO 1999).

Tab. 1: Manually introduced classifications

Thesaurus	Original term	Reliability
CEOPARAMETER	earth science→ atmosphere	100
CEODISCIPLINE	weather & climate	100

Tab. 2 shows the results of the expansion method for the input terms is in Tab. 1, all of them having a reliability value over 49. Summing up, 11 new terms were found, which belonged to three better structured thesauri: GEMET (the General General European Multilingual Thesaurus (GEMET 1999), the Alexandria Digital Library Feature Type Thesaurus (ADL 1999, referenced as ADL-FTT in Tab. 2), and a list thematic keywords proposed by the "NASA Master Directory" project (NASA 1996).

Tab. 2: Terms automatically expanded

Thesaurus	Expanded term	Reliability
GEMET	atmosphere	99
	climate	
	climate→ weather	
	climate→weather→ weather condition	
	climatic issue	
	climatic issue→ weather	
	climatic issue.weather→ weather condition	
ADL-FTT	regions→ climatic regions	49.5
NASA	atmospheric science	49.5
	atmospheric science→ atmospheric pressure	
	atmospheric science→ atmospheric temperature	

CONCLUSIONS AND FUTURE LINES

This paper has presented a flexible tool to manage thesauri and providing enhanced functionality for improvement of classifications. In fact, this tool has been incorporated as an additional component of a metadata edition tool (Zarazaga-Soria, Lacasta et al. 2003) that enables metadata creators to select the appropriate term for the distinct metadata fields.

Future lines of this work will be directed to the creation of a Thesaurus Web Service providing some of the functionality offered by this tool. This Web service will reuse the software components presented in this paper in order to offer on-line thesaurus browsing, *WordNet* polysemy extraction or keywords expansion.

As a last remark, it should be mentioned that we are also using the semantic disambiguation of thesauri to test different information retrieval strategies for geographic data catalogs. These geographic catalogs manage metadata records compliant with some metadata scheme that includes a keywords section or subject metadata element. Besides, it is assumed that the terms used to fill the content of this keyword section are selected from disambiguated thesauri. Therefore, it is possible to index metadata records according to a unified system, which is precisely the synsets associated with the disambiguated thesaurus terms. And finally, thanks to this homogenous indexing of metadata records, classical information retrieval algorithms can be applied to compute the similarity between the metadata records and the concepts expressed by the user queries.

ACKNOWLEDGEMENTS

The basic technology of this work has been partially supported by the Spanish Ministry of Science and Technology through the project TIC2000-1568-C03-01 from the National Plan for Scientific Research, Development and Technology Innovation, and by the project P089/2001 from the Aragón Government. Javier Lacasta's work has been partially supported by a grant from the Aragón Government and the European Social Fund.

REFERENCES

- ADL (2003): *Homepage of the Alexandria Digital Library Project*. <http://www.alexandria.ucsb.edu> (last accessed, May 2003).
- CEO (1999): *Recommendations on Metadata. Describing the data, services and information you have available! (version 2.0)*. A User Guide provided by the Center for Earth Observation Programme (CEO programme) of the European Commission. February 1999.
- GEMET (1999): *General European Multilingual Thesaurus. Version 2.0*. ETC/CDS 1999. <http://www.mu.niedersachsen.de/cds/> (last accessed, May 2003).
- ISO (1985): *ISO-5964. Documentation: Guidelines for the establishment and development of multilingual thesauri*.
- ISO (1986): *ISO-2788. Documentation: Guidelines for the establishment and development of monolingual thesauri*.
- Mata, E., Bañares, J.A., Gutierrez, J., Muro-Medrano, P.R., Rubio, J. (2002): *Semantic Disambiguation of Thesaurus as a Mechanism to Facilitate Multilingual and Thematic Interoperability of Geographical Information Catalogues*. Proc. of the 5th AGILE Conference: 61-66. Palma de Mallorca (Spain), April 2002.
- Miller, G.A. (1990): *WordNet: An on-line lexical database*. International Journal of Lexicography, 3(4) (Special Issue).
- NASA (1996): *Draft geospatial thematic keywords from the NASA Master Directory in short and long format for CSDGM of FGDC*. Version available from <http://www.fgdc.gov/clearinghouse/reference/refmat.html> (last accessed, May 2003).
- UNESCO (2003): *UNESCO Thesaurus*. United Nations Educational, Scientific and Cultural Organization. Available from <http://www.ulcc.ac.uk/unesco/> (last accessed, May 2003).

Zarazaga-Soria, F.J., Lacasta, J., Nogueras-Iso, J., Torres, M.P., Muro-Medrano, P.R. (2003): *A Java Tool for Creating ISO/FGDC Geographic Metadata*. Proceedings of GI-days 2003. Münster (Germany), 26-27 June 2003.