(Draft) Proceedings of the 10th EC-GI&GIS Workshop: ESDI: State of the Art. 2004, p. 125-125.

IMPROVING MULTILINGUAL CATALOG SEARCH SERVICES BY MEANS OF MULTILINGUAL THESAURUS DISAMBIGUATION

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

ABSTRACT

Multilinguality is an important aspect for the creation of public services in countries like Spain, with four official languages (Spanish, Catalonian, Basque and Galician), and overall, if these services are aimed for a European audience with a big number of official languages. Thus, an initiative for creating a catalog service at the Spanish or at the European level must take into account the necessity of supporting metadata written in a variety of languages. When a user submits a query request to a catalog, (s)he should obtain the data resources that verify the restriction specified by the user and with independence of the language used in the metadata records describing these resources. Our R&D group has developed a strategy for managing these multilingual metadata aspects that face up the problem from three points of view: the use of multilingual controlled lists, the use of multilingual and interrelated thesauri, and the word sense disambiguation of large text descriptive elements. These three substrategies, integrated within a catalog server, are being currently tested in several spatial data infrastructure projects.

KEYWORDS: Cross-language Information Retrieval, Geographic Metadata, Catalogs

INTRODUCTION

Geographic information is vital for decision-making and resource management in diverse areas (natural resources, facilities, cadastres, economy...), and at different levels (local, regional, national or even global). The fact is that thousands of organizations and agencies (all levels of government, the private and non-profit sectors, and academia) throughout the world spend billions of euros each year producing and using geographic data (Somers, 1997; Groot and McLaughlin, 2000). And in parallel to this increasing production of geographic information, it has been encouraged the creation of networked solutions to facilitate the discovery, evaluation and access of geographic data. Spatial Data Infrastructures (SDI) provide this framework for the optimization of the creation, maintenance and distribution of geographic information at different organization levels (e.g., regional, national, or global level) and involving both public and private institutions (Nebert, 2001).

One of the main components of an SDI is a geographic catalog that enables users, or application software, to find the information that already exists within a distributed computing environment. According to (Kottman, 1999), geographic catalogs are a solution to publish descriptions of geospatial data and enable searches across multiple servers. These descriptions of geospatial data are called metadata ("data about data") and their content structure is established by recognized organizations (FGDC, 1998; ISO, 2003). The use of indexed and searchable metadata provides a selected and

disciplined vocabulary against which intelligent geospatial queries can be performed, thus enabling the understanding among users from the same or different geographic information communities.

One feature to take into account regarding catalog services development is that it does not seem reasonable to think about the development of different catalogs that work like standalone nodes and are accessed only by client applications developed by the same company or with the same technology. On the contrary, in order to promote the sharing of geographic information throughout the maximum number of users, it is necessary to create distributed networks of catalogs that use a standardized mechanism for catalog querying, thus enabling enterprise and technological independence. A successful example of such a network is the National Geospatial Data Clearinghouse project, which was developed by the Federal Geographic Data Committee (FGDC) as a key component of the U.S. National Spatial Data Infrastructure. The nodes of this network conform to the ANSI/NISO Z39.50 information and retrieval protocol, which has been widely used since the beginning of the 1990s for the construction of OPACs (Online Public Access Catalogs). And although the Clearinghouse project had originally a national character, many servers from other countries (e.g. Canada, Australia, South Africa or Uruguay) have adhered to the initiative. A more recent initiative for the standardization of catalog services is the one proposed by the OpenGIS Consortium (OGC). Integrated by more than 250 companies, government agencies, and universities, OGC's mission is to promote the development and use of advanced open systems standards and techniques in the area of geo-processing and related information technologies delivering spatial interface specifications that are openly available for global use. And one of these specifications is the OGC Catalog Interface Implementation Specification (Nebert, 2002). This specification does not provide enhanced capabilities, in comparison with the Clearinghouse that uses the Z39.50 protocol (one of OGC profiles is even compatible with Z39.50), but from a broader perspective, this specification can be more useful because it has been conceived as a part of an integrated and interoperable architecture of geographic information services.

However, one aspect that has not been enough exploited within the development of catalog services is their multilingual capabilities. Multilinguality is an important aspect for the creation of public services in countries like Spain, with four official languages (Spanish, Catalonian, Basque and Galician), and overall, if these services are aimed for a European audience with a big number of official languages. Thus, an initiative for creating a geographic data catalog service at the Spanish or at the European level must take into account the necessity of supporting metadata written in a variety of languages. When a user submits a query request to a catalog, (s)he should obtain the data resources that verify the restriction specified by the user and with independence of the language used in the metadata records describing these resources. There are a lot of geographic information resources that are catalogued using only one language. But users that make their queries in one language may be interested in existent resources that have been described in another language. The user is more interested in the resource (map, image or multimedia resource in general) rather than in the metadata describing the resource. Thus, catalogs must provide users with the mechanisms facilitating the multilingual search without forcing cataloguing organizations to describe their resources in all the possible languages. The process of metadata creation in several languages is complex and expensive: specialized stuff is required with abilities not only in geographic information but also in language translation. Furthermore, any metadata update must be also performed in all languages.

Therefore, the objective of this paper will be to present the way to extend the basic catalog services with multilingual information retrieval capabilities. In addition to the word indexing of metadata records, we aim at providing an indexing of metadata records and user queries in a language

neutral way. Our multilingual information retrieval approach will focus on three types of metadata elements: elements whose domain value is restricted to a list of values; elements containing keywords and categories that metadata standard guidelines recommend to select from controlled vocabularies such as thesaurus or taxonomies; and large text descriptive elements. And for each type of metadata element, we have proposed an indexing system independent of the language used for writing the metadata content. In the case of elements restricted to a list values, this indexing will be the numerical code of these values. And in the last two cases, the indexing will be based on the conceptual indexing of element values with respect to the concepts of an upper-level lexical ontology like WordNet. Thanks to this language independent indexing of metadata records, user queries can be specified in any language and the discovery services will be able to find resources catalogued in other languages.

The rest of the paper is structured as follows. The following section revises the state of the art in multilingual information retrieval. Then the third section describes our approach to multilingual information retrieval. And the last section is devoted to some conclusions and comments about additional issues that must be considered to provide multilingual search services.

STATE OF THE ART IN CROSS-LANGUAGE INFORMATION RETRIEVAL

Traditionally, the Information Retrieval problem has been understood as the process that given a query (expressing the information needs of a user) and a collection of documents returns an ordered list of documents, which are supposed to be relevant with respect to the query. And as a special case of information retrieval, cross-language information retrieval (also known as multilingual information retrieval) deals with the problem of finding documents that are written in other languages than the one used in the queries. During the last decade, multilingual information retrieval has acquired great importance and nowadays it is considered as a multidisciplinary research area that combines several aspects from Natural Language Processing, Information Retrieval and Digital Libraries Research.

The origins of cross-language information retrieval as a separate discipline can be found in 1996 with the organization of the first specific workshop for the systematic comparison of cross-language information retrieval systems within the Special Interest Group on Information Retrieval (SIGIR) of the ACM. And since then, numerous international activities have been organized periodically: the Text Retrieval Conference (TREC) created in 1997 a special cross-language information retrieval track; the NII-NACSIS Text Collection for IR workshop (NTCIR), created in 1998, includes a track for the comparison of multilingual systems working with English and Asian languages; and the Cross-Language Evaluation Forum (CLEF) was created in 2000 for the evaluation of multilingual information retrieval systems working with European languages.

(Oard, 1998) revises different approaches that are used for the cross-language information retrieval problem. And in general, the cross-language information retrieval strategies can be categorized in three main blocks: the translation of queries; the translation of documents; and the conceptual indexing of documents and queries in a language independent manner. Let us see some details about these categories.

The first block of strategies is focused on the translation of user queries to the different languages used in the collection of documents managed by the information retrieval system (e.g., digital library, catalog, …). This is perhaps the most popular approach because it demands much less effort than translating or processing the documents in any way. These strategies depend, overall, on the types of

resources used for the translation: bilingual dictionaries, automatic translation programs, thesauri, or analysis of corpora (cross-related collections of documents in different languages).

On the contrary, the second block of approaches deals with the translation of documents. The advantages of this second block of approaches are the following: translations are more precise because they count on a wider context to determine the sense of each word; and that the translation errors affect in a lower degree to the retrieval effectiveness (an error in the translation of a document term is less problematic than an error in the translation of a query term). However, translation of documents requires a high computational cost and great amounts of storage space. Therefore, most of these strategies do not obtain readable translations; they are just focused on obtaining translated terms that are later used to apply information retrieval strategies.

Finally, the third block of strategies aims at indexing document and queries in some common representation. This common representation is independent of the language used in queries or documents. For instance, (Gilarranz et al., 1997) proposes the use of EuroWordNet as the indexing system for multilingual environments. WordNet (Miller, 1990) is an English lexical database whose basic object is a set of strict synonyms called synset, which represents one underlying lexicalized concept. And apart from these basic objects, WordNet also provides semantic relations (synonymy, hyponymy, meronymy, etc.) among these synsets. Evolving from WordNet, EuroWordNet (Vossen, 1998) was developed as its multilingual version consisting of cross-related WordNets in several languages (French, German, Spanish, Dutch, Italian, Czech, Estonian and English). It includes the semantic relations among words in different European languages and the multilingual relations among the concepts in different languages. The main advantages of using the EuroWordNet Interlingual Index are the following: it is more scalable than the translation approaches when the number of languages increases; and it avoids traditional problems such as the identification of synonym terms or word senses (these problems are intrinsic to the method). However, the main drawback is that automatic disambiguation techniques are still not mature enough to assure the required accuracy.

DESCRIPTION OF THE CROSS-LANGUAGE INFORMATION RETRIEVAL APPROACH

As mentioned in the introduction, apart from the basic monolingual discovery services, we have extended these services with additional multilingual capabilities. From the cross-language information retrieval strategies described in previous section, we could say that we have opted for the last category of strategies, i.e. those strategies using a language independent representation of queries and documents. Figure 1 shows our proposal for the retrieval model adding the multilingual capabilities.

The first process applied in the retrieval model is the analysis and indexing of metadata records and user queries. Both the elements of a metadata records and the restriction over those elements in the user queries must be appropriately analyzed and indexed. As it can be observed in figure 1, apart from the classical monolingual indexing of records and queries, we have also added the indexing of these records in a language independent representation. On one hand, the classical indexing consists of the word indexing of text elements, the spatial indexing for those elements containing geographic locations expressed as geometries and the indexing of date type elements. And on the other hand, the extended indexing is particularized for three types of elements: elements whose domain value is restricted to a list of values; elements containing keywords and categories that metadata standard guidelines recommend to select from controlled vocabularies such as thesaurus or taxonomies; and large text descriptive elements. Hence, for each type of metadata element we have proposed an

indexing system independent of the language used for writing the metadata content. Additionally, it must be remarked that, except for the spatial and the date indexes, this indexing process also computes the frequency of each index term and its inverse frequency that will be later used to obtain the relevance of each record with respect to the query.

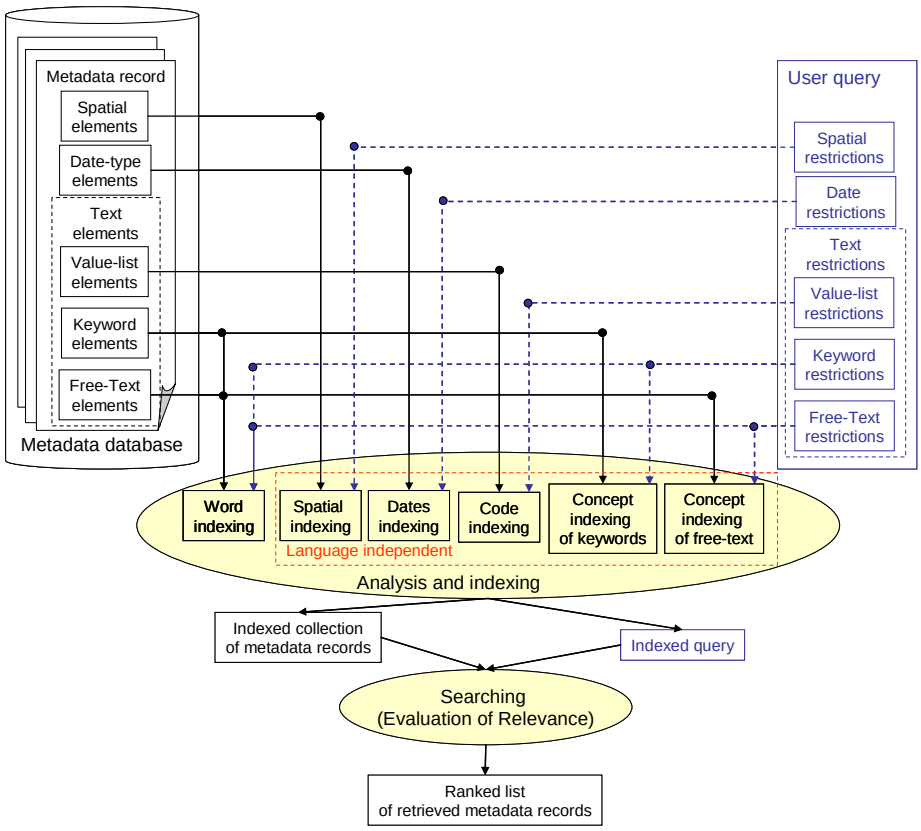


Figure 1: The process of retrieving information

The second step of the process is the searching process, in other words, the process in charge of evaluating the relevance of each metadata record and the user query. The inputs for this process are the indexed collection of records, which is performed off-line, and the indexed query which is obtained on-line. For the computation of this relevance value, this process applies an adapted version of the cosine formula of the vector-space retrieval model (Salton, 1971). The vector-space retrieval model proposes a framework in which partial matching is possible and it is characterized by the use of a weight vector representing the importance of each index term with regard to a metadata record (document). The angle formed between the query vector and the metadata record vector determines the proximity between the user information need and the metadata record that is evaluated. Thus, the computation of the cosine of this angle gives an idea of the degree of similarity.

Next subsections will detail the special indexing of text metadata elements that facilitate the cross-language information retrieval.

Indexing of value-list elements

CSDGM (FGDC, 1998) and ISO19115 (ISO, 2003) geographic metadata standards define numerous metadata elements whose domain is restricted to finite list of values, which has been translated to several languages by the national organizations for standardization that have adopted these standards. For instance, Table 1 shows the *MD_GeometricObjectTypeCode* list of values used in ISO19115 to specify the name of point or vector objects used to locate zero, one, two, or three dimensional spatial locations in the described datasets. Our proposed information retrieval model will index this type of metadata elements and the user restrictions on these elements by means of their numerical codes. This way, both records and queries will use the same numerical code.

Code	English value	Spanish value
001	complex	Complejo
002	composite	Compuesto
003	curve	Curva
004	point	Punto
005	solid	Sólido
006	surface	Superficie

Table 1: Example of list of values to fill the *geometricObjectType* metadata element

Indexing of keyword elements

In the case of elements containing the topics, subject and keywords of the resource, we propose the use multilingual and interrelated thesauri. Thesauri provide a specialized vocabulary for the homogeneous classification of resources and for supplying users with a suitable vocabulary for the retrieval. However, if a catalog aims at providing access to the general public (not only constrained to the community of experts that documented the resources in the catalog), it is not reasonable to assume that casual users will use the same query terms as the keywords used in metadata records and in the same language. Furthermore, if the catalog contains descriptions of resources from different application domains, metadata creators have probably used different thesauri (increasing the heterogeneity of keywords).

In order to fill the semantic gap between user queries, metadata records and language heterogeneities, we propose the interrelation of thesauri by means of their disambiguation with respect to multilingual lexical ontologies. Some of the thesaurus terms are polysemic, i.e. they have several senses. The problem of thesaurus disambiguation will consist in determining which one of the senses of an ambiguous term must be used in a particular context defined by the related terms (broader and narrower terms) of this ambiguous term. In particular, this work proposes the disambiguation of thesaurus terms with respect to WordNet (Miller, 1990), a large-scale lexical database developed from a global point of view that can provide a good kernel to unify, at least, the broader concepts included in distinct thesauri. Although WordNet is an English lexical database, it serves for our purposes because we have initially started with the disambiguation of multilingual thesauri such as the GEneral Multilingual Environmental Thesaurus (GEMET) or the UNESCO thesaurus. GEMET

(http://www.mu.niedersachsen.de/cds/) is a thesaurus for the classification of environmental resources that has been developed by the European Environment Agency and the European Topic Centre on Catalogue of Data Sources together with international experts. It contains a core terminology of 5,400 environmental terms (with their definitions) and the terms have been translated into 19 languages. On the other hand, the UNESCO thesaurus (http://www.ulcc.ac.uk/unesco/) includes a structured list of general descriptors in English, French and Spanish for indexing and retrieving literature in the fields of education, science, social and human science, culture, communication and information.

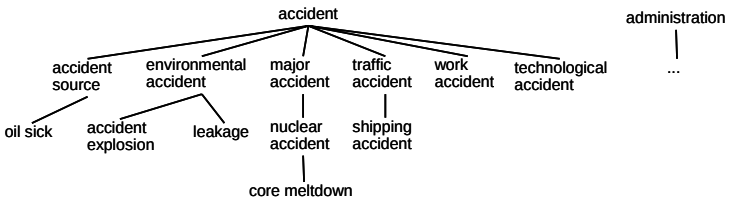


Figure 2: Example of thesaurus branches

Anyway, our final intention is to move towards the use of EuroWordNet (Vossen, 1998), which, as mentioned in the second section, was developed as the multilingual version of WordNet. Using this multilingual lexical database, we could also disambiguate the terms belonging to monolingual thesauri. The main problem of this multilingual lexical database is that this resource is the result of a European project already finished in 1999. Despite the potential possibilities of this resource for multilingual information retrieval and natural language processing, there is not a coordinated strategy for the maintenance of this resource and it depends greatly in the individual efforts of the members (universities and research laboratories) of the consortium that participated in the project.

Our thesauri disambiguation method can be classified as an unsupervised disambiguation method and applies a heuristic voting algorithm that makes profit of the hierarchical structure of both WordNet and the thesauri. Whereas thesaurus hierarchical structure provides the disambiguation context for terms, the hierarchical structure of WordNet enables the comparison of senses from two related thesaurus terms. WordNet is structured in a hierarchy of synsets which represent a set of synonyms or equivalent terms. The initial step of the disambiguation process is to divide the thesaurus into branches (a branch corresponds to a tree whose root is a term with no broader terms and that is constituted by all the descendants of this term in the "broader term/narrower term" hierarchy). The branch provides the disambiguation context for each term in the branch. Secondly, the disambiguation method finds all the possible synsets that may be associated with the terms in a thesaurus branch. And finally, a voting algorithm is applied where each synset related to a thesaurus term votes for the synsets related to the rest of terms in the branch. The main factor of this score is the number of subsumers in synset paths (the synset and its ancestors in WordNet). The synset with the highest score for each term is elected as the disambiguated synset. Table 2 shows the final score of synsets for the branch accident in figure 2. For the sake of clarity, some terms and their corresponding synsets have not been shown. A more detailed explanation of the algorithm to obtain the score can be found in (Mata et al., 2001). And (Nogueras-Iso et al., 2004) details an information retrieval system that makes profit of this thesauri disambiguation for the indexing of keywords sections of metadata and adapts the vector-space retrieval model to obtain the relevance of each metadata record with respect to the user queries.

Term	Subterm	Synset path	Score
accident			
		event->happening->trouble->misfortune->mishap->accident	3,143
		event->happening->accident	2,560
accident->accident source			
		event->happening->trouble->misfortune->mishap->accident	2,304
	accident	*event->happening->accident*	1,873
		>reference	0,713
		entity->object->location->point->beginning	0,705
		entity->object->artifact->facility->source	0,685
		entity->life_form->person->communicator->informant	0,397
		entity->life_form->person->creator->maker->generator	0,397
		psychological_feature->cognition->content->idea->inspiration->source	0,186
	source	*abstraction->relation->social_relation->communication->written_communication->writing->document->source*	0,009
accident->accident source->oil slick			
		entity->object->film->oil_slick	0,214
. . .			

Table 2: Disambiguation of a thesaurus branch

A special remark concerning the indexing of user restrictions on elements containing keywords and specified in other languages than English is that they will be done by means of the use of multilingual thesauri like GEMET and UNESCO. That is to say, until the definitive incorporation of multilingual lexical databases such as EuroWordNet, values specified in user restrictions will be looked up in the terms of GEMET or UNESCO and their disambiguated synset will be used for the indexing of user queries.

Indexing of free text elements

Finally, we have also proposed the conceptual indexing of metadata elements that contain large text descriptions such as the typical *abstract*, *purpose* or *supplemental information* elements. That is to say, similar to the disambiguation of thesauri, we perform the word sense disambiguation of free-text metadata elements with respect to the concepts of a lexical database.

The disambiguation method applied is equivalent to the one used in previous subsection but this time using the surrounding words in the metadata element as the disambiguation context. As lexical database it has been also used WordNet as an initial step. The final intention again is to move towards the use of EuroWordNet. But at present, the retrieval system enables at least the conceptual indexing of English descriptive elements.

In order to illustrate this free text disambiguation, let us consider the following text for the disambiguation of the polysemic word “palm”:

> “Cordyline terminalis Tricolor, a cabbage **palm**, has lance-shaped leaves impressively streaked with creamy white, pink and red.”

Similar to the disambiguation of thesauri, the disambiguation method finds all the possible synsets that may be associated with the surrounding words of “palm” in the text. And then, a voting algorithm is applied where each synset related to a surrounding word votes for the synsets related to “palm”. The main factor again of this score is the number of subsumers in synset paths (the synset and its ancestors in WordNet). And in addition, other factors such as the distance of the surrounding words with respect

to the ambiguous word are also taken into account. Table 3 shows the final score of the possible senses (synsets) for the ambiguous word "palm". As it can be observed, the correct sense ("any plant of the family "Palmae") was successfully selected with the highest score.

Synset nr.	Synset path	Definition	Score
8882567	entity>life_form>plant>vascular_plant>woody_plant>tree>palm	any plant of the family Palmae	0,4816
4312128	entity>part>body_part>area>palm	the inner surface of the hand from the wrist to the base of the fingers	0,2984
5029668	abstraction>relation>social_relation>communication>message>approval>award>decoration	an award for winning a championship	0,1001
9819435	abstraction>measure>linear_measure>linear_unit>palm	a linear unit based on the length or width of the human hand	0,0647

Table 3: Disambiguation of the polysemic word "palm"

CONCLUSIONS

This paper has presented an extension of geographic catalog search services to address cross-language information retrieval issues. As international markets and trans-national information networks interact, the access to information written (or described) in many languages is becoming increasingly relevant. And this is particularly relevant in the context of geographic information and the development of public services at the European level: users are interested in finding geographic information resources independently of the language that has been used at every national organization to describe this resource. Spatial queries restricting the spatial extent of a geographic information resource or restrictions about dates are inherently language independent. However, for other metadata elements containing text values, some kind of pre-processing and translation techniques must be applied to provide cross-language retrieval. Our multilingual approach has been focused on achieving a language independent representation for three types of these text metadata elements: the numerical indexing of elements restricted to a set of values; and the conceptual indexing of elements containing keywords and large text descriptions. The conceptual indexing has been mainly based on the use of multilingual thesauri and their disambiguation against the concepts of the WordNet upper-level lexical database.

Currently, this strategy is being tested in different SDI projects developed at the University of Zaragoza. In particular, we are involved in the development of multilingual portals for the Environment Department of the Galicia region in Spain, and the Spanish National Spatial Data infrastructure. Additionally it must be remarked that, apart from the information retrieval model, other aspects must be considered to fulfil a complete cross-language solution. On one hand, according to the language specified by the user, the GUI components (labels, buttons, value lists, …) must be displayed in the appropriate language. And on the other hand, although not providing the translation of metadata records, the HTML reports showing complete metadata contents should at least display the labels in the appropriate language. For these requirements, Java internationalization techniques and XML technologies (including XSLT) have been used to dynamically internationalize the software components, load Web pages contents stored as XML documents, and apply the appropriate stylesheets to display the required portal style and with the appropriate language for text labels.

As future lines of the work presented in this paper, the cross-language information retrieval approach should be fully integrated with a multilingual lexical database. Some initial steps have been

already taken with the use of EuroWordNet, the multilingual version of WordNet. However, this resource has not been updated to the last versions of WordNet and alternative approaches should be studied. In addition, it should be interesting the connection of discovery services to gazetteers providing toponyms with different translations.

ACKNOWLEDGEMENTS

The basic technology of this work has been partially supported by the Spanish Ministry of Science and Technology through the project TIC2003-09365-C02-01 from the National Plan for Scientific Research, Development and Technology Innovation. The work of J. Lacasta (ref. B139/2003) has been partially supported by a grant from the Aragón Government and the European Social Fund.

BIBLIOGRAPHY

FGDC (1998). Content Standard for Digital Geospatial Metadata, version 2.0. Document FGDC-STD-001-1998, Federal Geographic Data Committee (FGDC), Metadata Ad Hoc Working Group.

Gilarranz, J. , Gonzalo, J. and Verdejo, M.F. (1997). An approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database. *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*.

Groot, R. and McLaughlin, J. (2000). *Geospatial Data Infrastructure: concepts, cases and good practice*. Oxford University Press, New York, USA.

ISO (2003). Geographic information - Metadata. ISO 19115:2003, International Organization for Standardization (ISO).

Kottman, C. (1999). The OpenGIS Abstract Specification. Topic13: Catalog Services (version 4). OpenGIS Project Document 99-113, OpenGIS Consortium Inc.

Mata, E.J., Ansó, J., Bañares, J.A., Muro-Medrano, P.R. and Rubio, J. (2001). Enriquecimiento de tesauros con wordnet: una aproximación heurística. *Proc. of IX CAEPIA, Gijón*, pp. 593-602

Miller, G.A. (1990). Wordnet: An on-line lexical database. *Int. J. Lexicography 3*.

Nebert, D. (2001). Developing Spatial Data Infrastructures: The SDI Cookbook v.1.1. Global Spatial Data Infrastructure, available at http://www.gsdi.org.

Nebert, D. (2002). OpenGIS Catalog Services Specification, Version 1.1.1. OpenGIS project document 02-087r3, Open GIS Consortium Inc.

Nogueras-Iso, J., Lacasta, J., Bañares, J. A., Muro-Medrano, P. R. and Zarazaga-Soria, F. J. (2004). Exploiting disambiguated thesauri for information retrieval in metadata catalogs. *Lecture Notes on Artificial Intelligence (LNAI) 3040*. In Press.

Oard, D.W. (1998). A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. *Proceedings of the Third Conference of the Association for machine Translation in the Americas*, pp. 472-483.

Salton, G. (ed.) (1971). *The SMART retrieval system - Experiments in Automatic Document Processing*. Prentice Hall, Inc., Englewood Cliffs, NJ

Somers, R. (ed.) (1997). Framework Introduction and Guide. Federal Geographic Data Committee (FGDC), available at http://www.fgdc.gov/framework/frameworkintroguide.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities*, vol. 32, no. 2-3 (Special Issue on EuroWordNet), pp. 73-89.

Improving Multilingual Catalog Search Services by means of Multilingual Thesaurus Disambiguation

J. Nogueras-Iso, F.J. Zarazaga-Soria, J. Lacasta, R. Tolosana, P.R. Muro-Medrano
jnog, javy, jlacasta, rtolosana, prmuro@unizar.es
Computer Science and Systems Engineering Department,
University of Zaragoza, Zaragoza, Spain

Multiliguality is an important aspect for the creation of public services in countries like Spain, with four official languages (Spanish, Catalonian, Basque and Galician languages), and overall, if these services are aimed for a European audience with a big number of official languages. Thus, an initiative for creating a catalog service at the Spanish or at the European level must take into account the necessity of supporting metadata written in a variety of languages. When a user submits a query request to a catalog, (s)he should obtain the data resources that verify the restriction specified by the user and with independence of the language used in the metadata records describing these resources.

Our R&D group has developed an strategy for managing these multilingual metadata aspects that face up the problem from three points of view: the use of multilingual controlled lists, the use of multilingual and interrelated thesauri, and the word sense disambiguation of large text descriptive elements.
Firstly, for those metadata elements whose content is restricted to an enumeration of values, it is proposed the use of multilingual controlled lists, that will enable the expansion of user queries on those elements.
Secondly, we propose the use multilingual and interrelated thesauri for those elements containing the topics, subject and keywords of the resource. Thesauri provide a specialized vocabulary for the homogeneous classification of resources and for supplying users with a suitable vocabulary for the retrieval. However, if a catalog aims at providing access to the general public (not only constrained to the community of experts that documented the resources in the catalog), it is not reasonable to assume that casual users will use the same query terms as the keywords used in metadata records and in the same language. Furthermore, if the catalog contains descriptions of resources from different application domains, metadata creators have probably used different thesauri (increasing the heterogeneity of keywords). In order to fill the semantic gap between user queries, metadata records and language heterogeneities, we propose the interrelation of thesauri by means of their disambiguation with respect to multilingual lexical ontologies. In particular, this work proposes the disambiguation of thesaurus terms with respect to WordNet (and its multilingual version EuroWordNet), a large-scale lexical database developed from a global point of view that can provide a good kernel to unify, at least, the broader concepts included in distinct thesauri.
And thirdly, we also propose the word sense disambiguation of those metadata elements that contain large text descriptions such as the typical *abstract*, *purpose* of *suplemental information* elements. The word sense disambiguation would also make profit of multilingual lexical ontologies (e.g., EuroWordnet), enabling the indexing of metadata records by concepts that are independent of the language used.

These three substrategies have been integrated within a metadata edition tool and a catalog sever, which have been built at the University of Zaragoza and are currently used in several spatial data infrastructure projects. The final version of this paper will present the development of this strategy and its inclusion in the software components.

(Draft) Proceedings of the 10th EC-GI&GIS Workshop: ESDI: State of the Art. 2004. p. 125-125

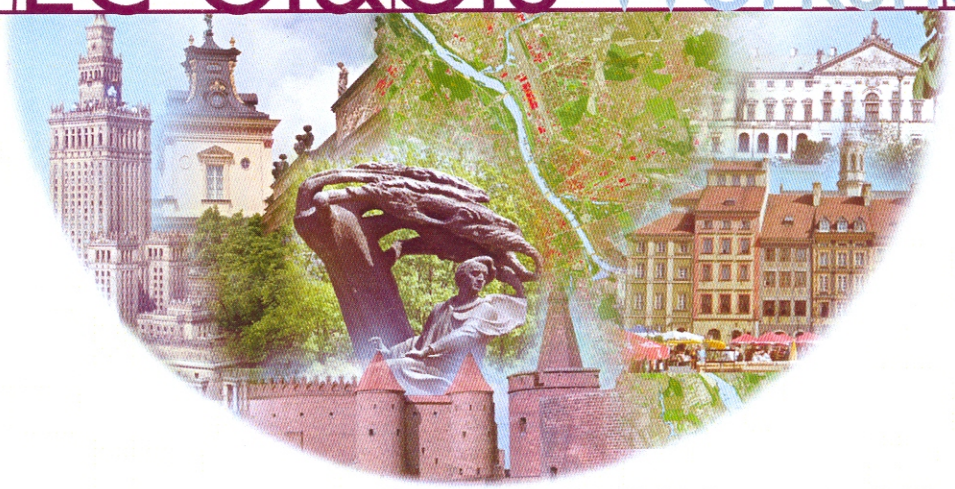
Institute for Environment and Sustainability

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Abstracts

ESDI: State of the Art

10th EC-GI&GIS Workshop



Warsaw, Poland
23-25 June 2004

INSTYTUT GOSPODARKI PRZESTRZENNEJ I MIESZKALNICTWA
INSTITUTE OF SPATIAL MANAGEMENT AND HOUSING

IGiK
Institute of Geodesy and Cartography

City of Warsaw

ASSOCIATION OF POLISH CITIES

CENTRUM GOSPODARKI PRZESTRZENNEJ

ESRI

hp invent

INTERGRAPH

(Draft) Proceedings of the 10th EC-GI&GIS Workshop: ESDI: State of the Art. 2004, p. 125-125.