(Draft) Data & Knowledge Engineering. 2007, vol. 63, nº 3, p. 947-971. ISSN 0169-023X.

A Web Ontology Service to facilitate interoperability within a Spatial Data Infrastructure: applicability to discovery [1]

J. Lacasta, J. Nogueras-Iso, R. Béjar, P.R. Muro-Medrano, F.J. Zarazaga-Soria

Computer Science and Systems Engineering Department, University of Zaragoza, María de Luna 1, E-50018 Zaragoza, Spain

Abstract

Ontologies are used within the context of Spatial Data Infrastructures to denote a formally represented knowledge that is used to improve data sharing and information retrieval. Given the increasing relevance of semantic interoperability in this context, this work presents the specification and development of a Web Ontology Service (WOS), based on the OGC Web Service Architecture specification, whose purpose is to facilitate the management and use of lexical ontologies. Additionally, this work shows how to integrate this service with Spatial Data Infrastructure discovery components in order to obtain a better classification of resources and an improvement in information retrieval performance.

Key words: Spatial Data Infrastructures, Web Services, Ontologies, Information Retrieval, Metadata

1 Introduction

The term ontology is used in information systems and in knowledge representation systems to denote a knowledge model, which represents a particular

Email addresses: jlacasta@unizar.es (J. Lacasta), jnog@unizar.es (J. Nogueras-Iso), rbejar@unizar.es (R. Béjar), prmuro@unizar.es (P.R. Muro-Medrano), javy@unizar.es (F.J. Zarazaga-Soria).

[1] This work has been partially supported by the Spanish Ministry of Education and Science through the projects TIN2006-00779 and TIC2003-09365-C02-01 from the National Plan for Scientific Research, Development and Technology Innovation.

domain of interest. A body of formally represented knowledge is based on a conceptualization: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them. And an ontology provides "an explicit formal specification of a shared conceptualization" [1], i.e. it facilitates a formal notation interpretable by machines that enables a shared and common understanding of a domain.

Within the geospatial community the use of ontologies as knowledge representation mechanism is acquiring an increasing relevance for the development of Spatial Data Infrastructures (SDIs) [2,3,4]. According to the Global Spatial Data Infrastructure Association Cookbook [5], "the term Spatial Data Infrastructure (SDI) is often used to denote the relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data". These data, also known as geographic information (GI) or geospatial data, describe phenomena associated directly or indirectly with a location with respect to the Earth surface. Traditionally, these data were the core component of Geographic Information Systems (GIS), which is the term commonly used to refer to the software packages that allow to capture, store, check, integrate, manipulate, analyse and display them. However, the potential of spatial data as an instrument to facilitate decision-making and resource management in diverse areas (e.g., natural resources, facilities, cadaster or agriculture) of government or private sectors has led to the evolution of GIS into the broader concept of SDI. Governments start considering SDIs as basic infrastructures for the development of a country, becoming as relevant as the classical ones (e.g., electricity, water, gas, transport or telecommunication infrastructures) [6]. As mentioned in [5], "The SDI provides a basis for spatial data discovery, evaluation, and application for users and providers within all levels of government, the commercial sector, the non-profit sector, academia and by citizens in general". From the technical perspective, the widespread use of the SDI concept has meant an important revolution in the geographic information community, moving from monolithic and stand-alone applications towards a dynamic and cooperative environment of services and applications. The European Committee for Standardization (CEN) defines the SDI concept as a platform-neutral and implementation-neutral technological infrastructure for geospatial data and services, based upon non-proprietary standards and specifications [7].

One of the main aims in SDIs is to facilitate the so-called geospatial resource access paradigm in a dynamic and cooperative environment where interoperability plays a crucial role. As defined in the Global Spatial Data Infrastructure Cookbook [5], this paradigm represents an end-to-end communication between users and providers/brokers of geographic information where "successive iterations of resource discovery via metadata catalogs, followed by resource evaluation (such as Web Mapping Services), lead to data access either: direct as data sets, or indirect via data access services". However, one of the main bar-

riers for this ideal cooperation is the heterogeneity that must be faced when distributed systems cooperate to fulfil any of these steps (i.e., discovery, access or evaluation). In order to provide seamless interoperability in any of these scenarios, SDI-based initiatives must deal with the challenge of overcoming the syntactic and semantic heterogeneities that may arise in the systems (system services and data accessed through these services) participating in these distributed scenarios.

In such a situation, the use of standards and recommendations proposed by different standardization organizations (ISO-TC211 [2], CEN-TC287 [3]) and other community consortiums (Open Geospatial Consortium, W3C, ...) has supposed a very significant step for the foreseen interoperability, at least to solve the most basic problems of syntactic interoperability. However, as the implementation of standards and specifications is still open for the interpretation of developers, important semantic differences remain. Let us think only in the additional barrier of multilinguality derived from the establishment of a European SDI with an increasing number of official languages [8]. Moreover, the geospatial community, as other communities, expects to make profit of resources developed in other domains not necessarily using same specifications and standards. Thus, taking into account this extremely open environment, it can be understood that ontology-based solutions for interoperability should play an essential role in SDI technology. On the one hand, considering the resource discovery scenario, ontologies can be used for modelling metadata schema models and the controlled vocabularies that are used to fill the content of metadata records. On the other hand, ontologies can be used for a formal representation of the conceptual models of the data that are visualized, accessed, and processed along the evaluation and exploitation phases. The role of ontologies and the state of art for their applicability in SDIs are analyzed further in section 2.

In particular, this work focuses on the use of ontologies for discovery scenarios, i.e. classification of resources and information retrieval. In order to facilitate discovery, national and international organizations have defined standards [9,10,11,12] that establish the structure of descriptions (metadata) for geographic information, services or locations in a gazetteer. In this context, as shown in [13], selecting the appropriate vocabularies represents an important challenge in terms of interoperability. Therefore, terms of controlled vocabularies (controlled lists, taxonomies, thesauri...) are frequently recommended to harmonize data and metadata of an SDI and to improve quality of query results.

[2] International Organization for Standardization (ISO), technical committee for Geographic information/Geomatics

[3] European Committee for Standardization (CEN), technical committee for Geographic Information

However, despite the advantages derived from the use of a controlled vocabulary, certain problems of ambiguity inherent to language persist. This ambiguity is mainly caused by different semantic relations between the terms of a language such as polysemy, homonymy, meronymy, hypernymy or hyponymy. These semantic relations are especially problematic when SDI users try to search data from several sources (with different cataloguing criteria) and their queries do not contain the same terms as the metadata, or when queries are expressed in a language not used in the metadata. Lexical ontologies have proven to be useful to deal with these ambiguity problems providing structure and semantic to the controlled vocabulary and allowing to inter-relate them. In order to use them efficiently they have to be managed uniformly.

The objective of this paper is two fold. First, it proposes and describes the architecture of a new centralized ontology service, called Web Ontology Service (WOS), which enables uniform management of lexical ontologies (including discovery services) and gives ontology-based support to SDI components. One of the main features of this service is its full integration with the rest of components of a typical SDI, following and extending standard interfaces used in the geospatial community. It has been designed to be integrated within the OGC Web Service Architecture (WSA) [14], a standardized architecture for an SDI provided by the Open Geospatial Consortium (OGC) [4], a non-profit, international, voluntary consensus standards organization that is leading the development of standards for geospatial and location based services. The second objective of this paper is to explore the uses that this ontology service can provide to the different discovery components in an SDI, showing examples of service functionality improvements.

The rest of this paper is structured as follows. Section 2 describes the role of ontologies for geospatial resource access. Section 3 shows the related work in ontology management. Section 4 describes the architecture of the WOS service. Section 5 indicates the uses of WOS in SDI discovery. Finally, this work ends with a conclusions and future work section.

2 Analyzing the role of ontologies in the geospatial resource access paradigm

According to the language used to express the ontologies, it is usual to classify them into: lexical/terminological ontologies (glossaries, controlled vocabularies, taxonomies, thesauri), implementation-driven ontologies (conceptual schemas, knowledge bases) and formal ontologies (depending on language expressivity). All of these more or less formalized ontologies can facilitate

[4] http://www.opengeospatial.org/

the interoperability in the different scenarios involved in the resource access paradigm shown in figure 1 [5].

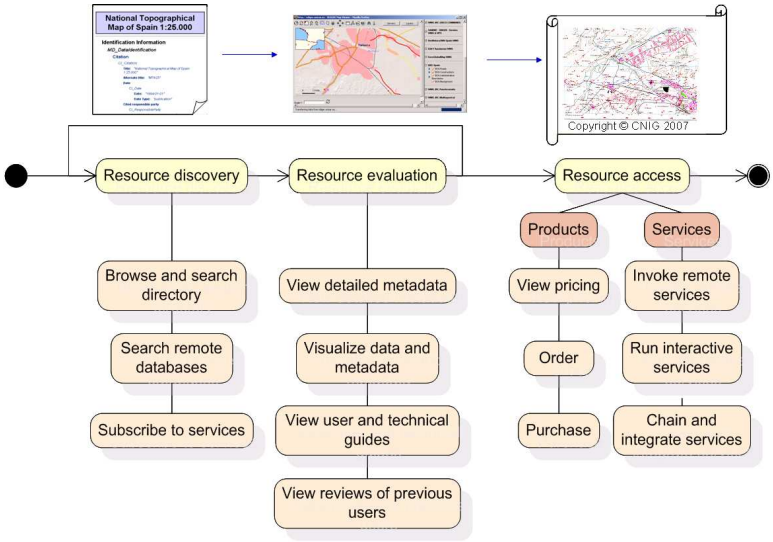

Figure 1. Geospatial Resource Access Paradigm, modified after [5].

As concerns **resource discovery**, some of the most remarkable problems that affect the interoperability and cooperation of discovery systems are metadata schema heterogeneity and content heterogeneity [15].

As regards the problem of metadata schema heterogeneity [16], given that a metadata schema is a model that contains a set of concepts with properties and relations to other concepts, their structure can be modelled as an ontology, where metadata records are instances of this ontology [17]. This kind of ontologies may be used to profile the metadata needs of a specific geospatial resource and its relationships with metadata of other related geospatial resources, or to provide interoperability across metadata schemas. Transformations of metadata between two different standards could be solved by systems that observe commonalities of two ontologies and automatically detect the metadata element mappings. An example of this kind of mappings can be seen in [18], where different metadata standards are used to describe geo-services. These metadata standards are modelled as ontologies using F-Logic and semantic technologies are used to match the ontologies.

For the problem of metadata heterogeneity [19] ontologies facilitate classification of resources and information retrieval. Metadata try to exactly describe information resources to enhance information retrieval, but this improvement depends greatly on the quality of metadata content. One way to enforce the

[5] To illustrate the Geospatial Resource Access paradigm, the figure shows the process initiated by a user (e.g., citizen, or local administration) to discover, evaluate and finally have access to a "National Topographic Map" (distributed by a National Mapping Agency).

quality is the use of selected terminology for some metadata fields in the form of lexical ontologies. These ontologies are used to describe contents but also allow computer systems to reason about them. This role of ontologies is even more significant in the case of developing a European SDI. In such an SDI, a strategy for cross-language information retrieval must be developed. Member states are not expected to provide translation for each metadata record they produce. Therefore, a European SDI catalog must tackle the problem of finding resources independently of the language used for metadata and data creation. Therefore, cross-language information retrieval strategies could consider either the automatic translation of queries to all possible languages, or the indexing document and queries in some common and language independent representation. In any of these cases, lexical ontology resources play a significant role for implementing these strategies [20,21].

Regarding **resource evaluation**, an SDI must facilitate the task of viewing detailed metadata, and must provide enough means to visualize the data appropriately. In this scenario, one could consider multilinguality and resolution level as main problems for system interoperability.

In the case of viewing metadata in a specific language required by the user, one may face the problem of having to translate it. Once again, metadata ontologies and lexical ontologies may facilitate the work in two important aspects. Firstly, a metadata ontology may provide the labels, in the appropriate language, for the elements of the metadata schema. Secondly, lexical ontologies may be used in the task of automatic translation of metadata to increase accuracy of translations.

Regarding the case of portrayal services for data visualization, one must face as well the problem of resolution level and "culture and linguistic adaptability". On the one hand, the resolution level affects portrayal of data because not all the features are meaningful at a particular zoom level. For instance, at a city scale level, it is worth visualizing the features of the urban transport network (streets, avenues, squares, ...). However, these urban network features are not meaningful for a road network at national level. On the other hand, culture and linguistic adaptability may influence the results offered by portrayal services. Although the visualization of data seems language independent, SDI developers must consider the internationalization of legends and the display of internationalized attribute information if necessary. For instance, the BALANCE project [22] uses external XML files to provide the translations of Web Mapping Services (WMS) capability documents, which are used by the client for the translation of WMS data layers names. Moreover, during the phase of resource evaluation, other multilingual and multinational issues must be taken into account, e.g. the selection of the correct Spatial Reference System, or the appropriate symbology according to cultural traditions of each country. Thus, one could seriously consider the creation of an ontology of features visualized

through portrayal services defining for each feature: the range of scales most appropriate for visualization, its textual label in every language, the most appropriate reference system for a geographic area, or the appropriate symbol (image) for rendering this feature on a map.

Finally, the **resource access** and further processing may benefit as well from the use of ontologies to facilitate data sharing and system development. Once again, ontologies help to define the meaning of features contained in geo-spatial data and they can provide a "common basis" for semantic mapping, e.g. to find similarity between two features that represent the same object but that have been defined using different languages. For instance, ISO/TC211 (technical committee for Geographic Information/Geomatics) has proposed several standardization items (19109 [23], 19110 [24], 19126 [25]) to create data dictionaries defining features and attributes that may be of interest to the wider international community. For example, [26] describes a system to interrelate features provided by different GI services to give a unified view to the final user, or [27], which provides communication between web services using an ontology based infrastructure. Other works like [3] even propose the creation of software components from diverse ontologies as a way to share knowledge and data. Furthermore, it is also usual in GI context to hear about extending the metaphor of Spatial Reference Systems (i.e., referencing things to some point on the ground) with the definition of Semantic Reference Systems [4]. The idea is that apart from spatial reference systems commonly used in maps and Geographic Information Systems (GIS), non-spatial components of geographic information should conform to some kind of semantic referencing.

As mentioned in the introduction, the focus of this work is on the use of lexical ontologies for discovery scenarios (i.e. classification of resources and information retrieval). Therefore, sections 3 and 4 will describe existing problems and proposals for a better management of lexical ontologies. Due to the multidisciplinary character of SDIs and its applicability to a wide range of application domains, there is a great variety of lexical ontologies with very different levels of specificity, language coverage (i.e., from monolingual to multilingual thesauri covering more than 20 languages), formalization (i.e., from simple glossaries to well-structured thesauri) or size (e.g., AGROVOC thesaurus [28] contains more than 16,000 concepts). Thus, an SDI discovery system must rely on an efficient and robust ontology management service to filter and select the most appropriate ontology for each specific context.

3 Related work in the management of lexical ontologies

Traditionally, the first approach in information community to manage lexical ontologies has been to create different ad-hoc web services that provide access

to a particular ontology. Some examples of this kind of service are the General Multilingual Environmental Thesaurus (GEMET) [29], the Agriculture vocabulary (AGROVOC) [28] of the Food and Agricultural Organization of the United Nations (FAO) or the Alexandria Digital Library Feature Type Thesaurus [30]. The Canadian Geospatial Data Infrastructure project [31] advanced in 1999 that an SDI would need a centralized ontology service with the objective of providing a mechanism to maintain lexical ontologies when the number to manage would increase. In 2004 they published a prototype of a web service, the Multilingual Geospatial Ontology (M3GO), with some limitation in the relations that it could manage and the ways to identify ontologies.

Another example in the modelling of ontologies and the specification of services is the Simple Knowledge Organization System (SKOS) project [32] that belongs to the World Wide Web Consortium (W3C) Semantic Web Activity [33]. This project has proposed a model to represent lexical ontologies using the Resource Description Framework (RDF) [34] syntax (see the SKOS-Core model in figure 2). This model has facilitated a de-facto standard for the exchange of concepts (*skos:Concept* resources), their properties (e.g., preferred labels or alternative labels with *skos:prefLabel* and *skos:altLabel* RDF properties) and the relations between concepts (e.g., *skos:broader*, *skos:narrower* or *skos:related* RDF properties). Additionally, the SKOS project has published a prototype of a web service to provide access to their ontologies, whose interface basically enables retrieval of terms and some types of relations among these terms. This prototype service could be also considered as a centralized service, but it does not exploit yet the use of ontology metadata descriptions proposed in the SKOS model.

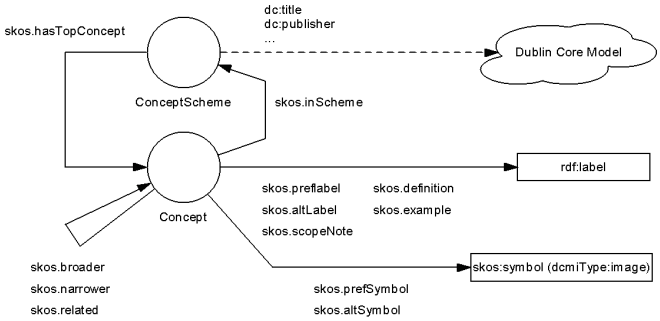

Figure 2. SKOS-Core model

There are also complex infrastructures such as KAON [35] or Ontolingua [36], which have been designed to share ontologies in general contexts and provide an API to access the ontologies stored in their repositories. The problem with these systems is again the lack of a search service to find the desired ontology. Users of the KAON tool must identify their desired ontology by means of a URI. And in the case of Ontolingua users must browse a plain list of ontologies (name plus a short description) until they find the ontology of their interest.

In summary, although it has been identified that the use of a centralized service is a step forward to facilitate access and management of ontologies in complex information infrastructures, the lack of standardization in access interfaces and exchange formats has limited its benefits. SKOS intends to unify the interchange format for lexical ontologies, but for ontology services there is still no consensus about their interface and functionality. Besides, one of the main drawbacks of current interfaces is that they do not offer proper discovery services for ontologies. Although it could be interesting to discover the more appropriate ontology for a specific geographic area or application domain, present services only facilitate access to an ontology by means of an agreed name. In addition, in the context of an SDI, it must be taken into account that the service must be integrated within a broader infrastructure of services (e.g. the OGC Web Service Architecture). Thus, some additional restrictions must be considered.

4 Architecture of the WOS service

The architecture of the WOS service consists of three layers as it is shown in figure 3. Firstly, the repository layer stores the ontologies (concepts and metadata describing the whole ontology) managed by the service and the concept core used for the interconnection of ontologies. Secondly, the application layer provides access to ontology concepts and their metadata. And thirdly, the service layer provides a web service wrapper to enable the access of web clients.

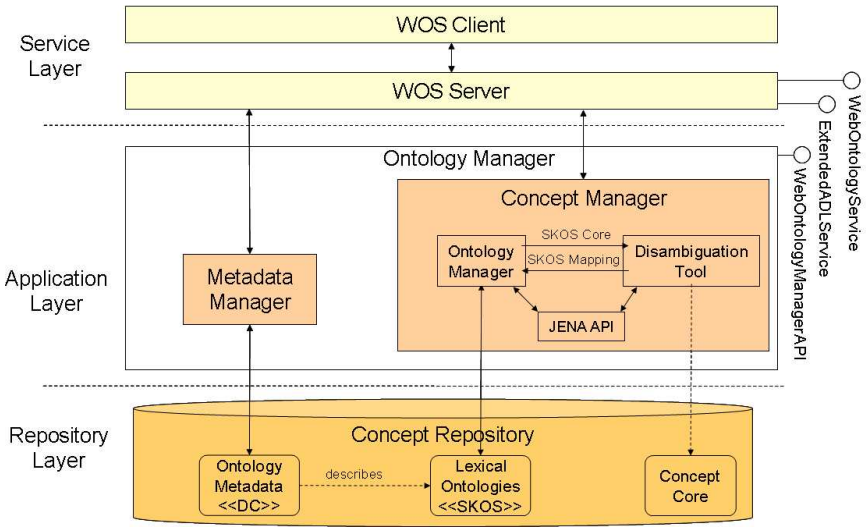

Figure 3. WOS Architecture

The following subsections describe the core components of the repository and application layers (section 4.1), and the external interface offered by the WOS

service (section 4.2).

4.1 Core components

In the repository layer, the SKOS model has been selected for the storage and exchange of ontologies. As stated in section 3, SKOS is a RDF based model that has been created specifically to manage lexical ontologies for the W3C Semantic Web project. Widely accepted within the digital library community, SKOS provides a very reach machine readable language for representing knowledge organization systems such as subject heading lists, taxonomies, classification schemes, thesauri, folksonomies, and other types of controlled vocabularies. In addition, if it were necessary, one could easily adapt SKOS vocabulary to fit more formal ontology languages such as OWL (Web Ontology Language) [37]. As both SKOS and OWL are based on RDF, it would be possible to define SKOS resources, properties and relations in terms of OWL constructs. By means of inheritance one could establish an almost 1:1 mapping between SKOS resources and OWL classes (*owl:Class*), between SKOS properties and OWL data type properties (*owl:DataTypeProperty*), and between SKOS relations and OWL object properties (*owl:ObjectProperty*).

The access to RDF SKOS documents storing ontologies is provided in the application layer through *Jena* [6] . *Jena* is a popular library that simplifies the manipulation of RDF documents, storing them in text files or in a relational database. One important advantage of using *Jena* is that it has an open model that can be extended with specialized modules to provide other ways of storage such as the *Jena-Sesame adapter* [7] , which provides access to *Sesame* [8] databases.

A fundamental aspect in the repository layer is the description of ontologies. Metadata for describing ontologies are considered as basic information to be facilitated to clients. These metadata, depicted in figure 3 as *Ontology Metadata*, are managed by the *Metadata Manager* component. The reason for this metadata-driven interface is that centralized ontology storage is not enough to manage them efficiently. Ontologies must be described and classified to facilitate the selection of the most adequate ontology for each situation. The lack of metadata describing them makes very difficult the identification of ontologies provided by other services, producing a low reuse of them in other contexts. Metadata are used in search processes to facilitate ontology retrieval, allowing users to search them not only by an agreed *name*, but also by the *application domain* or the associated *geographical area* among other descriptors.

[6] http://jena.sourceforge.net/
[7] http://sjadapter.sourceforge.net/
[8] http://www.openrdf.org/

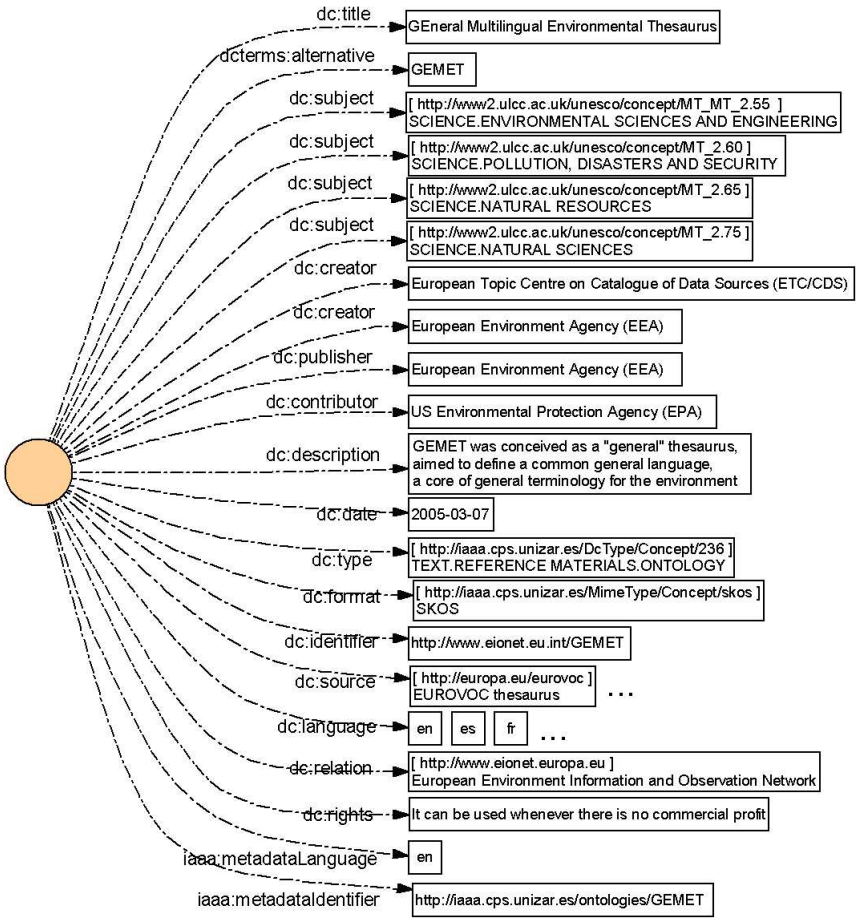

Figure 4. Metadata describing the GEMET thesaurus

For the purpose of describing ontologies in our service, a metadata profile based on Dublin Core [10] has been created [9]. Dublin Core has been used as basis of this profile because of its extensive use in the metadata community. It provides a simple way to describe a resource using very general metadata terms, which can be easily matched with complex domain-specific metadata standards. Additionally, Dublin Core can be also extended to define application profiles for specific types of resources. Following the metadata profile hierarchy described in [38], the application profile for the description of ontologies refines the definition and domains of Dublin Core elements, as well as it includes two new elements (metadata language and metadata identifier) to identify appropriately the metadata records describing ontologies. This profile has been defined using the IEMSR format [39]. IEMSR is an RDF based format created by the JISC IE Metadata Schema Registry project to define metadata application profiles. Figure 4 shows an example of ontology metadata for the description of the GEMET thesaurus. The RDF metadata is displayed as a hedgehog graph (reinterpretation of RDF triplets: resources,

[9] http://iaaa.cps.unizar.es/presentaciones/docontologydc_en.html

named properties and values). The purpose of these metadata is not only to simplify discovery, but also to identify which ontologies are useful for a specific task in a machine-to-machine communication (e.g., ontologies that cover a restricted geographical area or with a specific thematic).

In addition to the *Metadata Manager* and the *Jena API*, the application layer integrates a disambiguation mechanism (*Disambiguation Tool*) that enables the alignment of lexical ontologies with respect to a core upper-level ontology (the concept core displayed in figure 3). At present, WordNet [40] has been used as upper-level ontology. WordNet is structured in a hierarchy of synsets, defining a synset as a set of strict synonyms representing one underlying lexicalized concept. We have used the name "disambiguation" for this alignment method because the label of a concept in the ontology may be polysemic with respect to the possible synsets that may contain this label in Wordnet. Thus, the objective of this disambiguation tool consists in determining which one of the synsets of WordNet can be aligned to the real concept in the lexical ontology. A future step is to extend the tool for the disambiguation of ontologies in multiple languages, using a multilingual upper-level ontology (e.g., EuroWordnet [41]).

The disambiguation mechanism is based on an unsupervised technique applying a heuristic voting algorithm that makes profit of the hierarchical structure of both WordNet and the lexical ontology. Whereas the hierarchical structure (broader and narrower relations) of the lexical ontology provides the disambiguation context for concepts, the hierarchical structure of WordNet synsets (also organized in *is-a* relations) enables the analysis of meaning similarity between surrounding concepts. The initial step of the disambiguation process is to divide the lexical ontology into branches (a branch is a tree whose root is a top concept with no broader concepts and contains all the descendants of this concept in the "broader/narrower" hierarchy). The branch provides the disambiguation context for each concept in the branch. Secondly, the disambiguation method finds all the possible synsets that may be associated with the concepts in one branch. And finally, a voting algorithm is applied where each synset related to a concept votes for the synsets related to the rest of concepts 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 concept is elected as the most liable disambiguated synset (according to the scores, a liability probability is assigned to each possible synset). A full detailed description of the technique can be found in [15, Chap.4].

The disambiguation component, designed as an independent module, receives an input in SKOS format and returns the disambiguation with respect to the upper-level ontology using the SKOS-Mapping model [42]. This model is an RDF extension of SKOS that is used to describe exact, major and minor map-

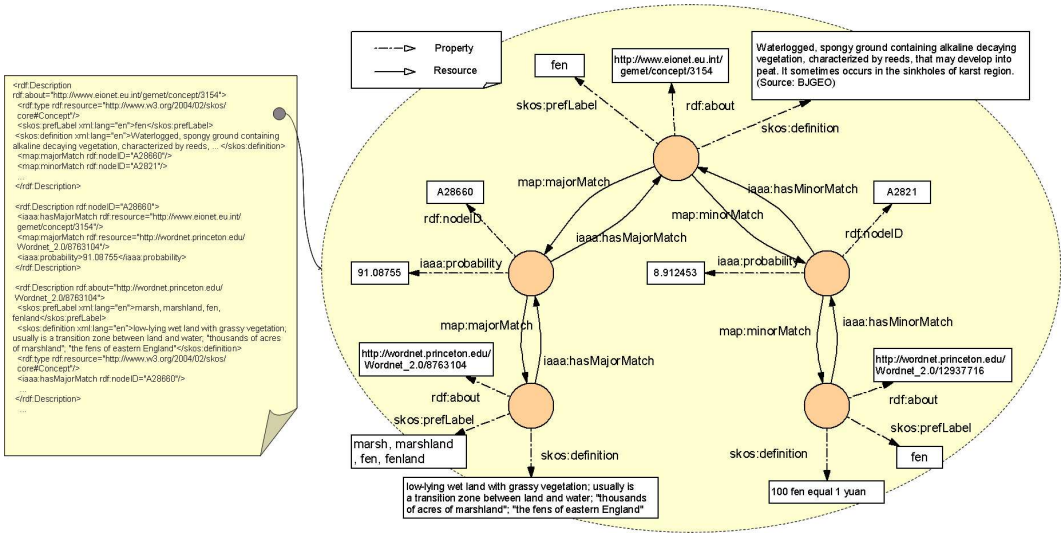

Figure 5. SKOS-Mapping extension

pings between two lexical ontologies (i.e. the ontology to disambiguate and the upper level ontology used as disambiguation base). Since the disambiguation algorithm can not assure a 100% exact mapping, only the major and minor mapping properties are used. The disambiguation algorithm returns, for each concept, a list of possible mappings with the upper level ontology. The one with the highest probability is assigned as the major mapping and the rest as minor mappings. SKOS-Mapping model has been extended by adding a blank node to store the disambiguation probability (liability of disambiguation) and by adding the major and minor inverse relations. An example of SKOS-Mapping can be seen in figure 5. There, the concept *3154 (fen)* of GEMET is correctly mapped to the WordNet concept *8763104 (marsh, marshland, fen, fenland)* with a probability of 91.08755%. Also an unrelated minor mapping is found, but it is given a low probability (8.912453%).

All the components in the application layer are packed into a black-box component called *Ontology Manager*, which is only accessible through the Application Programming Interface (API) called *WebOntologyManagerAPI* (i.e., the *Ontology Manager* component applies a *Facade* design pattern). This API includes the methods to allow other components to access the ontologies managed by WOS. These methods, displayed in figure 6, can be classified in two categories: query and administration.

- With respect to query methods, *query* and *getRelatedConcepts* methods allow users to browse through the relations between concepts and to search concepts by their label in different languages. The *query* method uses the disambiguation mechanism described before to expand the results returned, providing equivalent terms from the same or different ontologies.
- As regards to administration methods, they allow users to create a new ontology given its metadata, modify its metadata, delete it, and import or

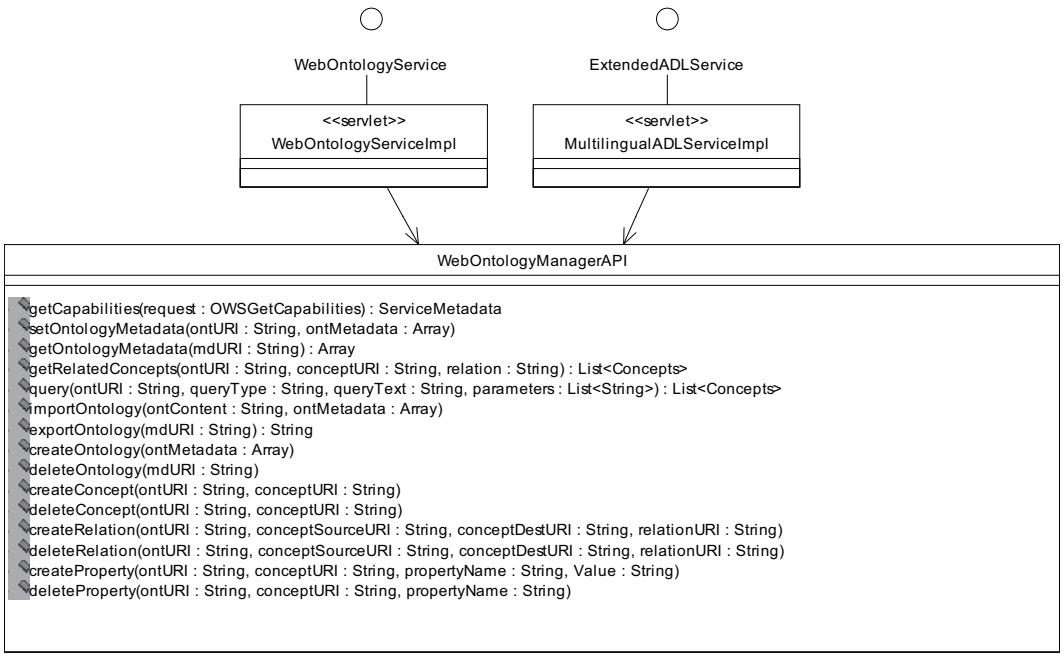

Figure 6. Web Ontology Service Implementation

export it in SKOS format. Additionally, the API includes methods to update concept properties and relations between concepts from different ontologies.

4.2 External interface

The service layer at the top of the layered architecture in figure 3 provides access to the WOS service through the HTTP protocol. Using the core functionality accessible through the *WebOntologyManagerAPI*, two web services have been built to provide compliance with well-established architecture specifications: the OGC Web Services Architecture (WSA) [43], and the service architecture proposed in the Alexandria Digital Library (ADL) project [44].

With respect to the compliance with OGC, this community aims at facilitating the adoption of open, spatially enabled reference architectures in enterprise environments worldwide. Therefore, OGC WSA has specified an Application Programming Interface each OGC Web service of an SDI should conform to. The objective is to promote interoperability among OGC service specifications by increasing commonality and discouraging non-essential differences. According to this API, every OGC service inherits from a general service whose unique operation is *getCapabilities* [43]. The *getCapabilities* operation provides a description of the service, its operations, parameters and data types. It is used for clients to identify whether a service provides the needed functionality and how to access it. Although OGC has developed numerous specifications for SDI web services, they have not created a specification for a service to manage ontolo-

gies yet. The WOS service can fulfil this gap. It simplifies the management of ontologies, and thanks to the compliance with the general OGC architecture, it can be integrated with the rest of OGC services in an SDI. Therefore, this work proposes the creation of a new OGC service called *WebOntologyService*, which extends the standard *OGC_WebService* interface (as other services in the OGC WSA do) with methods that provide the functionality to manage lexical ontologies. The top part of figure 7 shows the integration of *WebOntologyService* with the rest of OGC services. It must be noted that this new service interface does not include update methods for concepts and relations because the intention in this external access is to consider each ontology as a whole, managing their changes as different versions of the whole ontology. As depicted in figure 6, the *WebOntologyService* is implemented by the *WebOntologyServiceImpl*, which provides the bridge to the *WebOntologyManagerAPI*.

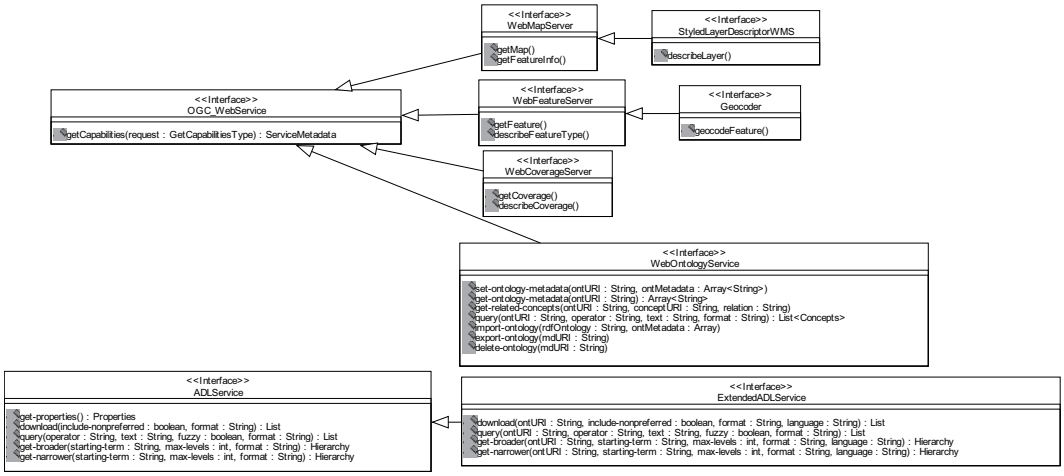

Figure 7. External interfaces of the WOS service

Concerning the ADL compliance, the WOS service also supports the ADL Thesaurus protocol [45], a protocol designed for the distribution of thesauri through the Web. The *ADLService* interface represents this protocol in figure 7. Additionally, it is worth noting that we have proposed and extension called *ExtendedADLService* to provide access to multiple thesauri, being those thesauri able to support properties in multiple languages. As depicted in figure 6, the *ExtendedADLService* is implemented by the *MultilingualServiceImpl*, which provides the bridge to the *WebOntologyManagerAPI*.

5 Applicability of WOS in SDI discovery

The objective behind the incorporation of WOS into SDI discovery is to enhance its capabilities, moving from data retrieval strategies to information retrieval strategies. Data retrieval consists mainly in determining which records in an SDI catalog system contain the words specified in the user query, but

very frequently this is not enough to satisfy the user information need [46]. On the opposite, information retrieval is more concerned with retrieving information about a subject than retrieving the data which satisfies exactly a given query. Usually, there is discordance between the query terms typed by casual users and the keywords inserted in metadata records. It seems sensible to think that discovery in metadata catalogs should not be implemented just as a simple word matching between user queries and metadata records. Thus, the integration of selected information retrieval techniques into metadata catalogs helps to understand the sense of user vocabularies and to link this meaning to the underlying concepts expressed in metadata records.

An information retrieval model can be defined as the specification for the documents (in our case, metadata records), queries and the comparison algorithm to retrieve the relevant documents. Next subsections are devoted to present a proposal for a retrieval model where WOS plays a fundamental role for the creation of metadata and for query expansion alternatives. The next subsection describes the process of metadata creation. Then subsection 5.2 describes the information retrieval model. Finally, section 5.3 shows the results of applying the proposed method for the retrieval of a metadata collection.

5.1 Metadata creation

SDIs are characterized by integrating information from many different sources, which may range from individuals (e.g., concerned citizens or graduate students in geography) and non-profit institutions (e.g., universities or non- governmental organizations for humanitarian help) to large remote sensing companies or governmental institutions (e.g., national mapping agencies, cadasters or environmental agencies). This great variety of sources implies a consequent heterogeneity in the metadata creation process, both in the wide choice of metadata standards and in the different expertise of metadata creators. According to the different resources and organizational procedures of the institutions contributing to the SDI, metadata may be created by scientific spatial data producers, by library cataloguers, or by administrative staff. Therefore, it is important to provide users with metadata edition tools that facilitate the content creation, i.e. generating those metadata elements that can be automated and guiding in the edition of descriptive elements that must be typed manually. Moreover, given that typing errors in metadata creation can imply not finding a resource, control of content quality is even more important. Being homogeneous in the selection of the terms used to describe a resource is another important issue. If two resources have similar characteristics, they should be described with the same terms. Otherwise, a query system will only return a subset of the records it should return. The use of controlled vocabulary for the most relevant elements of metadata can help to reduce the time of

creation, the number and impact of human errors, and increase the homogeneity. In order to reuse these vocabularies in different SDI services, it becomes essential to manage them uniformly by means of services such as our WOS proposal.

For instance, the OGC catalogue service specification [47] (standardizing the interface of discovery systems in SDIs) recommends the use of ISO19119 [11] for service description and ISO19115 [9] or Dublin Core [10] for geographic information description. All these standards define a big number of metadata elements, and many of them must or may contain terms from controlled vocabularies. Some examples in ISO19115 are the *descriptive keywords*, the *topic category*, the *distribution format* or the *spatial representation type*. As already mentioned, values for these elements could be facilitated through a WOS instance integrated with a metadata edition tool, reducing in that way the cost of creation and improving its quality and homogeneity.

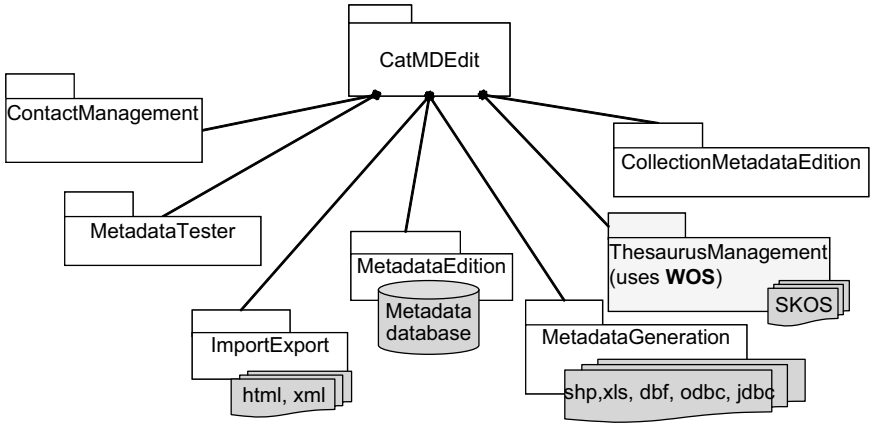

Figure 8. Integration of WOS with CatMDEdit

In order to test the WOS functionality in this direction, WOS technology has been fully integrated with the last version of the CatMDEdit Open Source metadata edition tool [10] [48]. Figure 8 shows the architecture of CatMDEdit. Among the components used to edit metadata in different schemes, the *Thesaurus Management* component uses the *WOSOntologyManagerAPI* to provide access to thesauri stored as lexical ontologies. Moreover, it must be noted that thanks to this integration CatMDEdit not only facilitates the selection of terms in different languages, but also gives access to their definitions, synonyms, narrower-broader concepts and related concepts (from the same lexical ontology or from a different ontology connected through the concept core).

[10] http://catmdedit.sourceforge.net/

5.2 Information retrieval model

5.2.1 General Context

An information retrieval process implies a series of typical operations such as text processing, indexing of documents, query processing, searching and ranking of retrieved documents. Figure 9 shows a schema of these operation interactions based on the model proposed by [46], but customized to the special characteristics of metadata management. Additionally, the figure remarks the interaction with the WOS component for query processing (see section 5.2.2 for further details).

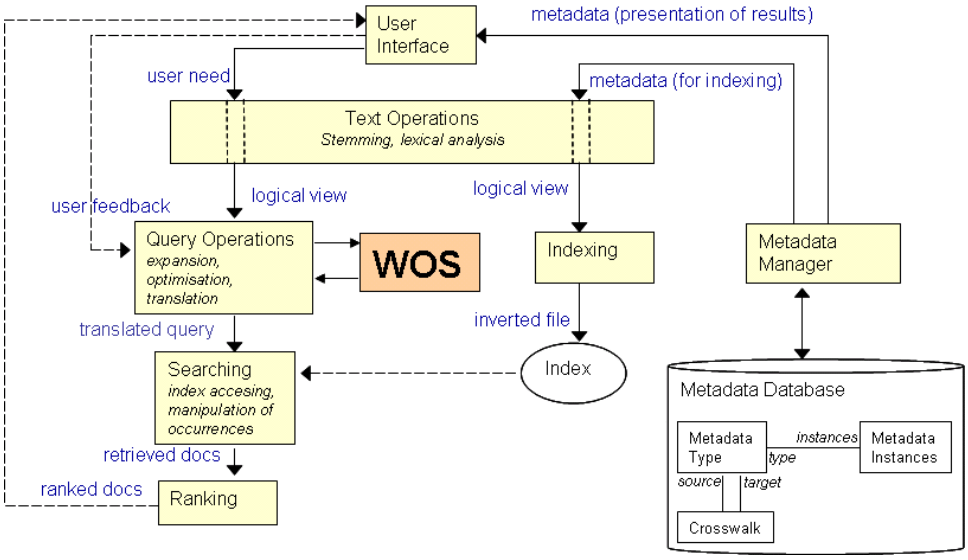

Figure 9. Structure of an information retrieval system (IRS) [46]

As regards the specific decisions taken in the operations involved in this information retrieval process, this work proposes the use of CatServer. This catalog system, described in [49], provides a functional kernel for catalog services handling XML-encoded metadata. With respect to the information retrieval model applied, CatServer is based on the Extended Boolean Model [46], i.e. it combines the simplicity of the Simple Boolean Model with the slightly more sophisticated ranking of results supplied by the Extended Model. Additionally, it is worth noting that this catalog system fulfills two main requirements. On the one hand, the system is independent from the metadata standards or schemas followed by the metadata inserted in the catalog. The idea behind this requirement is to use CatServer as a basis for the implementation of different metadata-driven services such as geographical data catalogs, service catalogs, or even Web Feature Servers (including its gazetteer variant). On the other hand, CatServer is able to manage large amounts of metadata records and be efficient enough in response time.

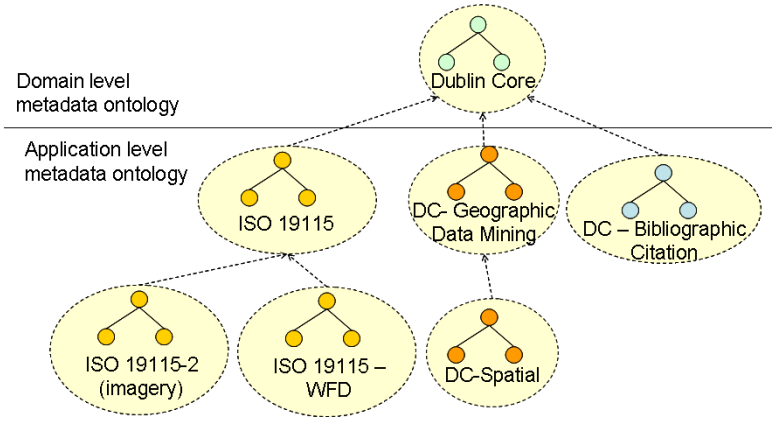

Figure 10. A hierarchy of metadata ontologies

In order to be independent from metadata standards, two design decisions have been taken in the development of CatServer:

- Firstly, metadata are directly stored in XML at CatServer. This modus operandi is significantly different from other catalogs which convert the XML into a persistent object model. The great advantages of the adopted approach are its retrieving speed (since it only has to retrieve the XML) and its independence from metadata standards. Otherwise, as it happens with the persistent object model approach, the inclusion of new standards involves code rewriting.
- Secondly, apart from the storage in XML format, the independence from metadata standards is fulfilled thanks to the fact that the different metadata schemas share a common core [38]. This common core is needed if the system wants to provide the user with the functionality of querying all the metadata instances stored, independently of the metadata schema used (e.g. we need a common set of queryable properties). As depicted in figure 10, the only prerequisite of the standards supported by our system is to provide their XML Schema and their mapping to the common core of Dublin Core. That is to say, as it is shown in figure 9, the metadata database maintains a knowledge base of the supported *metadata types* (schemas) and the *crosswalks* between them (at least a crosswalk towards the Dublin Core common core).

With respect to the second requirement related to the efficiency and the management of huge amounts of metadata records, it must be noted that the Inverted Index structure[46] was chosen and adapted to speed up queries. This structure could be defined as a sequence of *(key, pointer)* pairs where each pointer refers to a record in a database which contains the key value in some particular field. The index is sorted by the key values to provide fast searching for a particular key value (e.g. using binary search). The index is "inverted" in the sense that the key value is used to find the record rather than the other way around. For catalog systems enabling searches with filters on more than

one database field, multiple indexes (sorted by those keys) may be created.

The index structure of CatServer is slightly different. It consists of a pair *(key, array)* where the key has the same meaning, but there is an array instead of a pointer to a register. The array is a metadata identifier array which represents those metadata records that contain the word in a specific XML metadata element tag. The index structure has been implemented by means of a relational database table. The usual way of working is to build an Inverted Index for every XML metadata element tag for which the clients need to search. Figure 11 (left) shows two Inverted Indexes built over the Dublin Core elements *title* and *subject* (the examples uses an excerpt of metadata describing the *Natura 2000 sites* dataset, a set of areas of special interest for biodiversity protection across Europe).

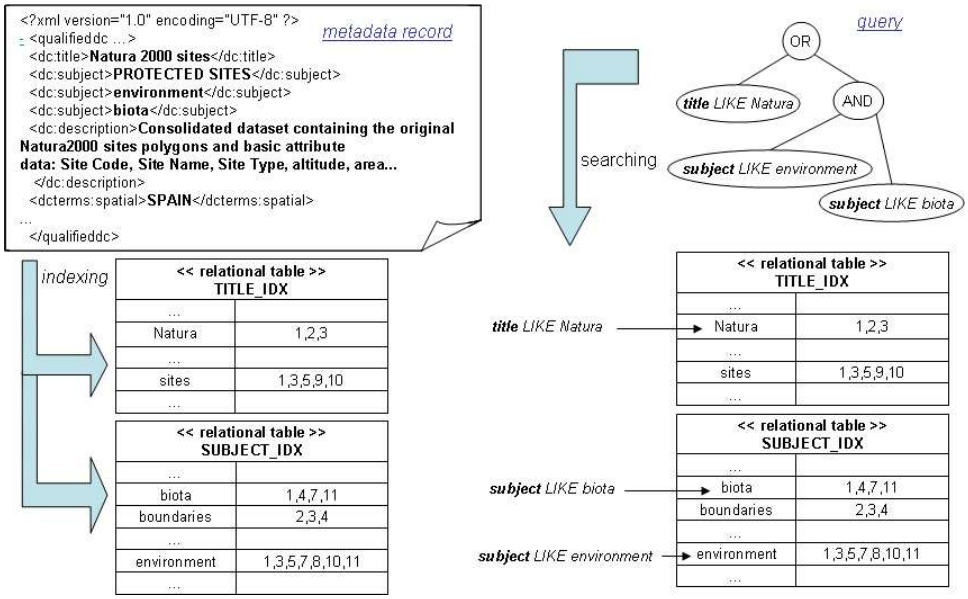

Figure 11. Retrieval example: XML tags and Inverted Index implementation correspondence (left); querying process (right)

Once the indexes are built, the system can retrieve the information with only the tag name, which determines the index to examine, and the key. For instance, let us consider the query represented in figure 11 (right). This query aims at retrieving those metadata records whose *title* contains *Natura* or whose *subject* contains *biota* and *environment* (*title LIKE '%Natura%' OR (subject LIKE '%biota%' AND subject LIKE '%environment%')*). Thus, CatServer would obtain three arrays of metadata identifiers: one for *Natura*, one for *biota*, and another for *environment*. The next step in the process is to combine these arrays as sets of metadata records. The *AND* implies an intersection operation between the *biota* array and the *environment* array. The *OR* implies a union operation between the *Natura* array and the subset obtained in the previous step.

Evidently, not all the results are equally important. As mentioned at the beginning of this subsection, the ranking process is based on the Extended Boolean Model. Therefore, the subset of metadata is in fact a list of metadata records ordered by relevance. Following with the example, metadata records satisfying both operands of the *OR* logic expressions are more relevant than those which only satisfy one of them, i.e. they appear before in the ranked list.

5.2.2 Query expansion

As stated in [5], SDIs aim at being a basic infrastructure for all kind of users and providers of spatial data within all levels of government, the commercial sector, the non-profit sector, academia and citizens in general. Therefore, in many situations SDI users (and applications built on top of these SDIs) do not have a clear understanding about which keywords they should introduce in their queries. Sometimes the users are professionals with a high level of expertise, but other times it is also usual to find citizens and novice users just exploring for the first time the possibilities offered by SDI services. Thus, the keywords used to express the concepts behind the user queries may differ from the keywords used by metadata creators. This is partially solved by offering search interfaces that guide the user through a thesaurus or other type of linguistic/terminological ontology that contain the more appropriate terms, ideally the same terms also used by metadata creators. However, the ideal situation of having created metadata by selecting terms from a unique lexical ontology does not occur very frequently. Quite the opposite, an SDI project that implies the cooperation of different institutions usually derives in a collection of metadata records using a wide range of thesauri and other classification schemes. Content creators from different organizations and application domains apply their own criteria for the classification of resources, generating very diverse terminology even for the description of similar resources. Moreover, this situation is even more problematic when the catalog system stores metadata records written in different languages. In that case, the terminological differences between users and metadata creators become a really difficult barrier for information retrieval.

Therefore, despite guiding the user in the construction of queries by means of a lexical ontology, retrieval may be of low quality due to the heterogeneity of metadata content and the great variety of SDI users' expertise. Thus, we propose query expansion as a well-known technique to improve the initial query formulation [46, Chap.5]. In particular, we propose to expand user queries by making profit of the knowledge behind the lexical ontologies managed by WOS. This query expansion is similar to works like [50] or [51], which present systems where thesauri are used as the basis for discovery services, and the thesaurus hierarchical structure helps to find resources either directly related

to the "concepts" found in user queries or "closely" related to "the user's concepts of interest".

As depicted in figure 9 we propose to extend the basic functionality provided by CatServer with a module that processes the terms included in the user query in order optimize and expand them with related terms obtained through a WOS service. Assuming that the user is guided by an initial terminological ontology, the *Query Operations* module will expand the user queries in two directions:

- *Expansion through the initial lexical ontology.* Firstly, the concepts selected by the user through an initial lexical ontology (and displayed in a particular language) are expanded with all the existing alternative labels in the different languages supported by this initial lexical ontology. By alternative labels of a concept we mean the preferred labels of this concept in all the languages supported by the ontology and all the synonym labels of this same underlying concept in those languages.
- *Expansion through disambiguation (related lexical ontologies).* Secondly, the *Query Operations* module tries to expand the query with the labels corresponding to related concepts in other lexical ontologies managed by WOS. Using the disambiguation component, described in section 4, it is possible to interrelate ontologies thanks to the connection with an upper-level ontology. If the user selects a very specific concept in the initial terminological ontology, this strategy will not probably find similar concepts in other ontologies. But in the case of searching more general concepts, this strategy will help to find synomyms or translations existing in related ontologies, which may have a richer vocabulary or support more languages than the initial one.

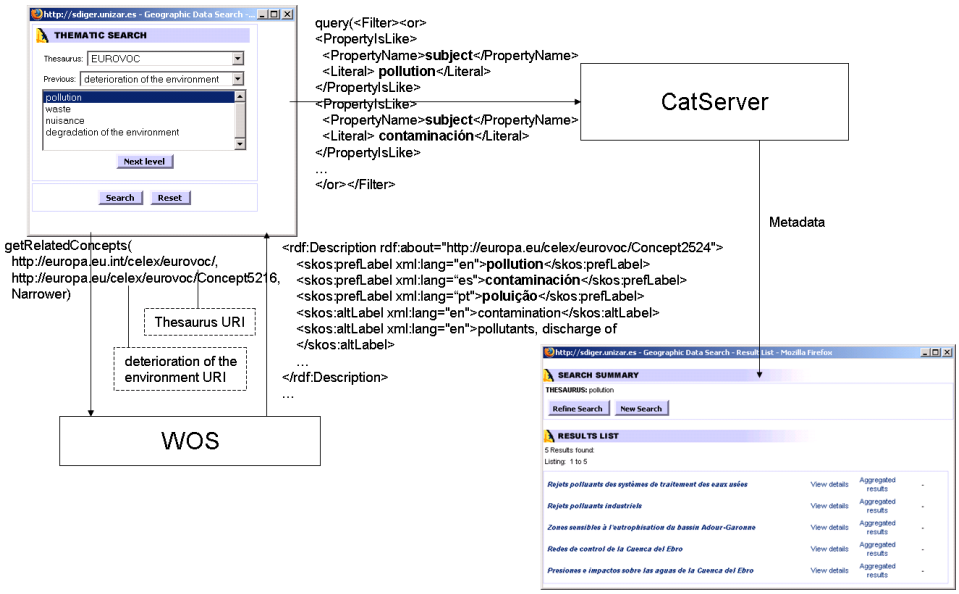

Figure 12. Example of query expansion for a thematic catalog

Figure 12 shows an example of the first type of query expansion (*expansion through the initial lexical ontology*). This example can be described in a sequence of three main steps:

- Firstly, a thematic search interface allows the user to browse the concepts contained in a lexical ontology (left side of figure 12). Although the search interface only shows the preferred labels in the language the user selected for human-computer interaction, we assume that the lexical ontology is multilingual, i.e. it gives support for several languages [11] . For instance, whenever the user browses the narrower concepts of a first concept (e.g., the concept *deterioration of the environment* identified by the URI *http://europa.eu/eurovoc/Concept5216* in the EUROVOC lexical ontology [52]), the thematic search interface interacts with the WOS service to retrieve all the preferred and alternative labels of the narrower concepts and in all the available languages (e.g., *pollution* in English but also *contaminación* in Spanish). Figure 12 shows an excerpt of the *getRelatedConcepts* request sent to the WOS service and the response in SKOS format returned by WOS.
- Secondly, a click on the *search* button represents that the user has stopped browsing the lexical ontology and has decided the final concepts to be included in the query. At this moment the search interface constructs the query that will be sent to the catalog system (CatServer). This query is compliant with the OGC Filter encoding specification [53] and contains an expression that includes all the possible alternatives of preferred and alternative labels in different languages obtained from the WOS service.
- And thirdly, the CatServer system launches the searching and ranking processes to obtain the metadata records that satisfy the expanded user query. Thanks to the fact that WOS provides preferred terms in different languages, the returned metadata records may have been written in multiple languages. For instance, the results shown on the right side of figure 12 include records in French (*Rejets pollutants des systèmes ...*) and Spanish (*Presiones e impactos sobre ....*).

With respect to the second strategy for query expansion (*expansion through disambiguation*), the *Query Operations* module applies a basic routine to estimate the reliability of expanding an original set of keywords with a new term belonging to a new different ontology, not used in the original set. This basic routine consists of four steps:

- The first step is the collection of all the major mappings of the concepts in the original query with respect to the upper-level ontology used by the WOS service. From now on we will use the name *synset* for these major mappings

[11] This search interface belongs to the set of search services offered by the SDIGER project (see section 5.3 for more details).

because this is the name given to the concepts in Wordnet, which is the upper-level ontology used for the disambiguation functionality described in section 4.1. As a result of this first step, we obtain for each concept in the query an initial collection of *synsets*.

- Secondly, we will also collect the *synsets* corresponding to a concept from a different ontology, which may be a candidate for query expansion. Initially, all the concepts of the ontologies stored in the WOS are considered as candidates.
- Thirdly, we will compute the reliability of a new candidate concept as the number of *synset* coincidences with the *synsets* of the original query concepts divided by the number of *synsets* of the new concept and multiplied by 99:

$$reliability = \frac{|synset\ matches\ of\ new\ concept|}{|synsets\ of\ new\ concept|} \times 99 \quad (1)$$

The reason to use a final factor of 99 and not 100 in equation 1 is to obtain a maximum reliability percentage of 99 for automatically expanded concepts, reserving uniquely a 100-reliability percentage for the concepts which were originally in the query.

- Finally, the reliability of a new candidate concept is compared with a *threshold* reliability. If the reliability percentage is greater than a *threshold* reliability, the query is expanded with this new concept. This means that the query expression will include as alternatives the preferred and alternative labels of this new concept in the different languages available. A *threshold* of 50% is considered as an appropriate value to detect related concepts to the initial set.

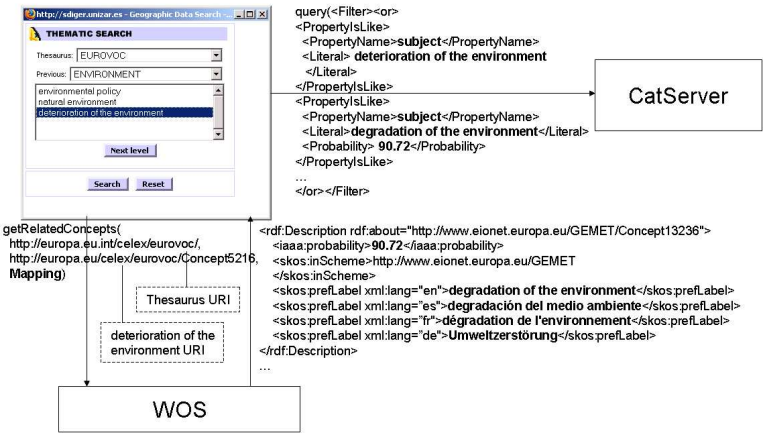

Figure 13. Expansion through disambiguation

It is worth noting that this expansion technique is integrated in the WOS service. The disambiguation functionality is provided through the *getRelatedConcepts* operation, using *Mapping* as relation type. Figure 13 shows an example of this type of query expansion. The concept *deterioration of the environment*

belonging to the EUROVOC vocabulary is expanded with the concept *degradation of the environment* of GEMET. This new concept of GEMET has been mapped to the original concept of EUROVOC with a reliability of 90.72%.

5.3 Testing the retrieval model

In order to quantify the retrieval effectiveness of an information retrieval system, performance measures such as precision (number of relevant hits divided by the number of hits) and recall (number of relevant hits divided by the number of relevant documents) must be computed upon the results obtained from evaluation experiments, which are conducted under controlled conditions. This requires a testbed comprising a fixed number of documents, a standard set of queries, and relevant and irrelevant documents in the testbed for each query.

For the case of testing the retrieval model presented in previous section and verifying the influence of WOS in the improvement of information retrieval performance, this model has been applied within the context of the SDIGER project [54]. SDIGER is a pilot project on the implementation of the Infrastructure for Spatial Information in Europe (INSPIRE) to support access to geographic information resources concerned with the European Water Framework Directive. This project includes a thematic catalog searcher (see left side of figure 12) that makes use of a WOS instance to access multilingual thesauri and to help in the construction of user queries, which are automatically expanded with cross-language terminology by means of the strategies explained in section 5.2.2. The multilingual thesauri managed by the WOS instance integrated within the SDIGER project have been the Multilingual Agricultural Thesaurus (AGROVOC) [28], the European Vocabulary Thesaurus (EUROVOC) [52], the GEneral Multilingual Environmental Thesaurus (GEMET), and the UNESCO Thesaurus [55]. All of them have been defined by well-known organizations and give support to several European languages, at least the three ones required for the project: English, French and Spanish.

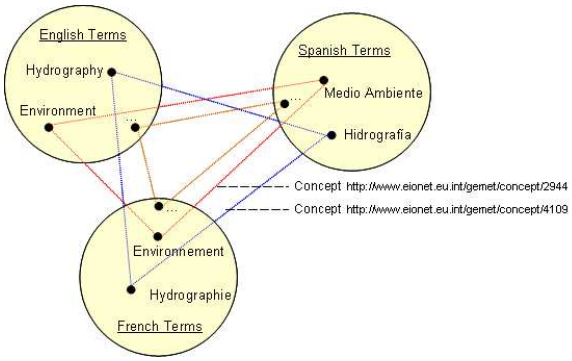

Figure 14. Mapping between terms in different languages

The SDIGER metadata corpus consists of around 26,000 metadata records in Spanish, English and French, which contain about 350 different keywords (in Spanish, English and French) to describe their associated data. Many keywords in such metadata records have been extracted from different thesauri but others have been randomly typed by metadata creators. In addition, each metadata record is written only in one language and this includes the terms used as keywords. Therefore, this was an appropriate corpus to analyze the impact of multilingual dispersion in the information retrieval performance.

Previous to the analysis of performance, it was necessary to obtain a series of topics and their relevance with respect to metadata records. This way, it would be possible to compare different retrieval (and query expansion) strategies. The topics were selected upon an analysis of the concepts behind the 350 different keywords found in the metadata records. After mapping terms in different languages, identifying synonyms, and eliminating redundancies introduced by plurals and other derived lexical forms, 204 different concepts were obtained. This concept extraction process was semi-automatic. Firstly, the language of each of the 350 keywords was identified by means of the *language* descriptor found in the metadata records containing these keywords. Besides, this language classification was verified with a multilingual dictionary. Secondly, as shown in figure 14, a manual mapping between terms in different languages was applied to identify the concepts that would be used later as topics for the experiments. Thirdly, the identification of synonyms and elimination of related lexical forms was applied as well with the aid of a multilingual dictionary. And at last, spatial data experts from the institutions contributing to the SDIGER project assigned manually the relevance of metadata records with respect to each topic.

For the sake of facilitating topic relevance assignment, experts were provided with the *Inverted Indexes* automatically created by CatServer and an initial pre-assignment of relevance according to the following rule: *"a metadata record is relevant to a topic if it contains one of the possible terms (labels) that represent the concept in that topic"*. The experts only had to revise this initial pre-assignment for possible mistakes due to word-sense ambiguity. However, in most cases the initial pre-assignment was accurate. In contrast to text information retrieval, where full documents are indexed, in this case we are indexing metadata records, which are short summary texts created by experts. This has two important advantages in comparison with classical text information retrieval. On the one hand, the texts are short and there are few noise words, reducing the possibilities of mistaking a noise word for a label representing a real topic of the resource described. And on the other hand, most terms found in metadata are quite specific, reducing the possibilities of polysemy. In fact, a search system just based on word-matching of topic terms would yield a high precision. The main problem that affects the performance of search systems over this metadata corpus is the problem of detecting the

correspondence among translation of terms and some synonymy issues. That is to say, a simple word-matching strategy for retrieval yields a low recall.

Once the corpus was fully established, a series of experiments were conducted using CatServer in order to compare different alternatives for query expansion. These experiments can be classified into three categories according to the query expansion strategy applied:

- *No query expansion.* The first three experiments consisted in selecting a particular language (e.g., English, French or Spanish) and sending queries to CatServer using topic terms in that particular language without applying any strategy for query expansion. In other words, these three experiments were oriented to study the three original languages separately (used in metadata records) and the problems derived from the multilingual dispersion.
- *Expansion through the initial ontology.* A second series of three experiments was oriented to analyze the effect of expanding queries thanks to the knowledge stored in the lexical ontologies. This strategy matches with the first heuristic described in section 5.2.2 for query expansion. In these experiments, it is assumed that the user is browsing a lexical ontology (GEMET, AGROVOC, or UNESCO) for the definition of user queries. When the user decides the final concepts to include in the query, the user query is automatically expanded with all the existing terms in different languages for those user selected concepts.
- *Complete query expansion.* A final experiment is devoted to analyze the effect of applying both two heuristics for query expansion described in section 5.2.2. This experiment assumes that the user is using the GEMET lexical ontology for the definition of the query. As regards to the query expansion, apart from extending the topic concepts to the all possible terms, the expansion also considers related concepts in the lexical ontologies of AGROVOC and UNESCO. Using the strategy called *expansion through disambiguation*, based on the disambiguation mechanism (see section 4.1) and the reliability formula (explained in section 5.2.2), the concepts of GEMET were connected to related concepts in UNESCO and AGROVOC.

With respect to the performance measures obtained upon these experiments, it is worth mentioning that we have focused on the analysis of recall. Given the characteristics of the metadata corpus, the comparison of precisions for each experiment is not relevant. As stated before, the results obtained in experiments not using query expansion always get a high precision because the metadata collection contains very specific concepts, which are rarely affected by polysemy conflicts. Additionally, the topics used for the queries correspond to concepts extracted from the own keywords contained in metadata records. This can be also extrapolated to the other two series of experiments using query expansion. Again, thanks to the lack of polysemy and the specificity of the topics used in the experiments, the automatic expansion is supposed

to be precise. On the one hand, the translations of terms derived from the use of a lexical ontology are inherently accurate (the lexical ontology has been constructed by experts with knowledge in different languages). On the other hand, the expansions due to the mappings of concepts between different ontologies are also accurate because of the specificity of the topics.

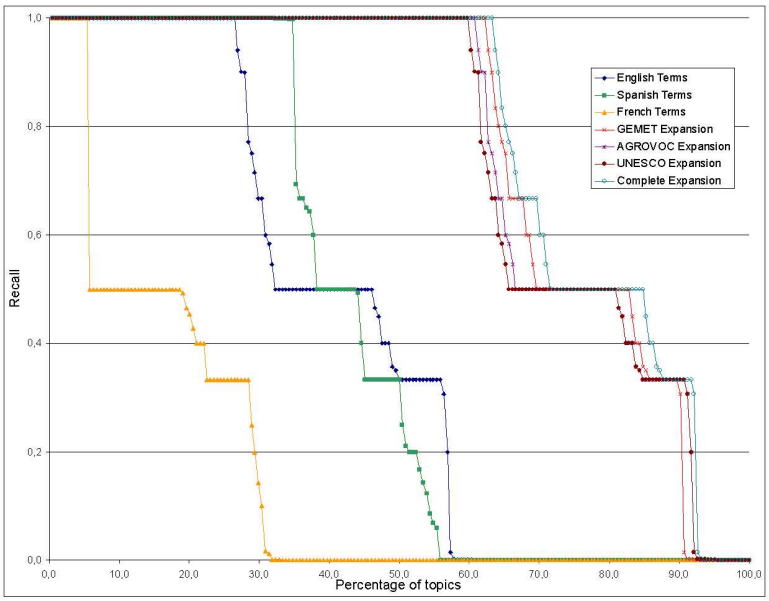

Figure 15. Comparison of recall using different query expansion alternatives

Figure 15 shows the recall curves obtained in each of the aforementioned experiments. The topics in the *x-axis* of each recall curve are ordered by the recall obtained in the experiment strategy. This fact does not allow the comparison of recall for a particular topic in two experiments. However, the main purpose of the figure is to provide a general idea of the average recall in each experiment. The area covered in the polygons bounded by the recall curves and the positive sides of both *x-axis* and *y-axis* denotes the recall improvements in each experiment.

As a result of the experiments, it can be observed that query expansion strategies based on lexical ontologies (i.e., use of translations and synonyms) imply an important recall improvement. Without query expansion, only 6% of topics using French terms have a full recall. This is slightly improved in the case of experiments using English and Spanish terminology: 26% and 32% of topics with full recall respectively. But anyway, it can be verified that this strategy without query expansion produces low recall measures in such a multilingual corpus. Quite the opposite, the experiments guided by the use of a lexical ontology such as GEMET, AGROVOC and UNESCO obtain a high recall for most of the topics: 60% of topics have a full recall, and 80% of topics have a recall higher than 50%. At last, the experiment using complete query expansion provides a small increase in recall with respect to the use of a single ontology.

Theoretically, the experiment with complete query expansion should have been obtained a perfect recall. However, there are still a small number of concepts that are not contained in the lexical ontologies used for the experiments. It must be taken into account that topics are derived from the keywords found in metadata records, but these keywords may not have been selected from a lexical ontology. Additionally, the labels (terms in multiple languages) used for a concept in a lexical ontology may not necessarily match with the terms manually mapped for the extraction of topic concepts.

Finally, it must be noted that independently of the query expansion method and the lexical ontology used, the results obtained are similar. This is caused by the fact that the ontologies selected for the experiment are thematically related to the metadata collection and contain a subset of concepts which are similar to the keywords contained in metadata records.

6 Conclusions

This work has presented a Web Ontology Service, called WOS, compliant with the OGC Web Services Architecture specification and whose purpose is to facilitate the management and use of ontologies in an SDI. Designed as a centralized service, the architecture of this service aims at reducing the cost of creation of a new ontology, improving reusability and avoiding duplicities and inconsistencies. It is planned to submit the specification of this Web Ontology Service as a new OGC Web Service specification that could be integrated in the future with the rest of Web Service specifications already issued by the Open Geospatial Consortium, to at least obtain the required feedback to improve, if necessary, the functionality offered by this service.

In addition, focusing on the objective of resource classification and improvement of information retrieval, this work has analyzed the potential benefits that this service may provide to the discovery components of an SDI. On the one hand, WOS can be used to facilitate the creation of metadata content, since it provides access to terminological ontologies (concepts, properties, definitions and relations between concepts and other ontologies) recommended by metadata standards. On the other hand, it has been proven that a WOS service can be easily integrated within an information retrieval system to facilitate the construction of user queries and improve the recall of such systems. This work has proposed an automatic approach for the expansion of user query concepts. It has been shown how the WOS service can be used to exploit the knowledge of lexical ontologies to expand the original concepts with translations, synonyms and related concepts in similar lexical ontologies.

As future work, we plan to improve the strategies for query expansion. Cur-

rently, we are assuming in our query expansion approach that metadata records may include any of the concepts contained in the lexical ontologies managed by WOS. However, in most cases the collection of metadata records accessible through a catalog system only includes a small subset of the concepts in a lexical ontology. To avoid this problem and make the query expansion strategies more efficient, a next step will be to prune the expanded concepts with a thematic topic map extracted off-line from a collection of resources. A first approach for the extraction of these topic maps has been already proposed in [56]. This initial approach analyzes the metadata records, extracts the concepts belonging to a lexical ontology, and uses those concepts to prune the lexical ontology and obtain a thematic topic map with the concepts actually used in the metadata collection.

Finally, it is also planned to explore new possibilities for the applicability of WOS in other SDI operational scenarios such as resource visualization or resource access and further processing. The objective is to extend WOS functionality to give support to non-lexical ontologies expressed in formal ontology languages such as OWL. This would not imply a complete redesign of the WOS architecture because both SKOS and OWL are based on RDF. In fact, it would be possible to redefine SKOS resources, properties and relation types in terms of OWL constructs.

References

[1] T. Gruber, A translation approach to portable ontology specifications, Technical Report KSL 92-71, Knowledge Systems Laboratory, Standford University, Stanford, CA (1992).

[2] P. C. Smits, A. F. Christensen, Resource discovery in a european spatial data infrastructure, IEEE Transactions on Knowledge and Data Engineering 19 (1) (2007) 85–95.

[3] F. T. Fonseca, M. J. Egenhofer, C. A. Davis, K. A. V. Borges, Ontologies and knowledge sharing in urban GIS, Computers, Environment and Urban Systems 24 (2000) 251–271.

[4] W. Kuhn, Geospatial Semantics: Why, of What, and How?, Journal on Data Semantics III, Special Issue on Semantic-based Geographical Information Systems, Lecture Notes in Computer Science 3534 (2005) 1–24.

[5] D. Nebert (Ed.), Developing Spatial Data Infrastructures: The SDI Cookbook v.2.0, Global Spatial Data Infrastructure (GSDI), http://www.gsdi.org, 2004.

[6] J. Nogueras-Iso, M. A. Latre, P. R. Muro-Medrano, F. J. Zarazaga-Soria, Building eGovernment services over Spatial Data Infrastructures, 3rd International Conference on Electronic Government - EGOV'04, Lecture Notes in Computer Science 3183 (2004) 387–391.

[7] European Committee for Standardization (CEN), Geographic information - standards, specifications, technical reports and guidelines, required to implement spatial data infrastructures, CEN/TR 15449 (2006).

[8] J. Nowak, J. Nogueras-Iso, S. Peedell, Issues Of Multilinguality In Creating A European SDI - The Perspective For Spatial Data Interoperability, in: Proceedings of the 11th EC GI & GIS Workshop, ESDI Setting the Framework, Alghero, Italy, 2005.

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

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

[11] International Organization for Standardization (ISO), Geographic information - Services, ISO ISO19119:2005 (November 2005).

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

[13] B. Heath, D. McArthur, R. Vetter, Metadata lessons from the iLumina digital library, Communications of the ACM 48 (7) (2005) 68–74.

[14] J. Lieberman (Ed.), OpenGIS Web Services Architecture, v0.3, Open Geospatial Consortium Inc, OGC 03-025, 2003.

[15] J. Nogueras-Iso, F. J. Zarazaga-Soria, P. R. Muro-Medrano, Geographic Information Metadata for Spatial Data Infrastructures - Resources, Interoperability and Information Retrieval, Springer Verlag, 2005.

[16] J. Nogueras-Iso, F. J. Zarazaga-Soria, J. Lacasta, R. Béjar, P. R. Muro-Medrano, Metadata Standard Interoperability: Application in the Geographic Information Domain, Computers, Environment and Urban Systems 28 (6) (2004) 611–634.

[17] L. Bermudez, M. Piasecki, Metadata Community Profiles for the Semantic Web, Geoinformatica 10 (2006) 159–176.

[18] N. Weißenberg, R. Gartmann, Ontology Architecture for Semantic GeoServices for Olympia 2008, in: Geodaten- und Geodienste-Infrastrukturen - von der Forschung zur praktischen Anwendung. Beiträge zu den Münsteraner GI-Tagen 26./27. Juni 2003, Vol. 18, IFGIprints, Münster, Germany, 2003, pp. 267–283.

[19] J. Nogueras-Iso, J. Lacasta, J. A. Bañares, P. R. Muro-Medrano, F. J. Zarazaga-Soria, Exploiting disambiguated thesauri for information retrieval in metadata catalogs, Lecture Notes on Artificial Intelligence (LNAI) 3040 (2004) 322–333.

[20] M. Gatius, M. Bertran, H. Rodríguez, Multilingual and Multimedia Information Retrieval from Web Documents, in: 4th International Workshop on Natural Language and Information Systems (NLIS'04) (DEXA'04 workshop), IEEE Computer Society, Zagaroza, Spain, 2004.

[21] J. Nogueras-Iso, F. J. Zarazaga-Soria, J. Lacasta, R. Tolosana, P. R. Muro-Medrano, Improving multilingual catalog search services by means of multilingual thesaurus disambiguation, in: 10th European Commission GI&GIS Workshop, ESDI: The State of the Art, Warsaw, Poland, 2004.

[22] N. Ostländer, S. Tegtmeyer, T. Foerster, Developing an SDI for time-variant and multi-lingual information dissemination and data distribution, in: Proc. of 11th EC GI&GIS Workshop, ESDI: Setting the Framework, Alghero, Italy, 2005.

[23] International Organization for Standardization (ISO), Geographic information – Rules for application schema, ISO 19109:2005 (2005).

[24] International Organization for Standardization (ISO), Geographic information – Methodology for feature cataloguing, ISO 19110:2005 (2005).

[25] International Organization for Standardization (ISO), Geographic information - Profiles for feature data dictionary registers and feature catalogue registers , ISO DRAFT ISO/TC 211 / SC N 1561 (2004).

[26] M. Lutz, E. Klien, Ontology-Based Retrieval of Geographic Information, Journal of Geographical Information Science 20 (3) (2006) 233–260.

[27] O. Shafiq, I. Toma, R. Krummenacher, T. Strang, D. Fensel, Using Triple Space computing for communication and coordination in Semantic Grid, in: 3rd Semantic Grid Workshop in conj. with the 16th Global Grid Forum, Athens, Greece, 2006, pp. 13–16.

[28] B. Lauser, M. Sini, G. Salokhe, J. Keizer, S. Katz, Agrovoc Web Services: Improved, real-time access to an agricultural thesaurus, Quarterly Bulletin of the International Association of Agricultural Information Specialists (IAALD) 1019-9926 (2) (2006) 79–81.

[29] European Environment Agency, GEneral Multilingual Environmental Thesaurus (GEMET). Version 2.0, European Topic Centre on Catalogue of Data Sources (ETC/CDS), http://www.eionet.europa.eu/gemet (2004).

[30] L. Hill, Alexandria Digital Library Feature Type Thesaurus, Version of July 3, 2002, University of California at Santa Barbara, Alexandria Digital Library, http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/index.htm (2002).

[31] Canadian GeoSpatial Data Infrastructure (CGDI), Homepage of the Canadian GeoSpatial Data Infrastructure (CGDI), http://www.geoconnections.org/CGDI.cfm (2005).

[32] A. Miles, D. Brickley (Eds.), SKOS Core Vocabulary Specification, W3C, W3C Working Draft 10 May 2005, 2005, http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20050510.

[33] Semantic Web Advanced Development for Europe (SWAD-Europe), Homepage of Semantic Web Advanced Development for Europe Thesaurus Activity, http://www.w3.org/2001/sw/Europe/reports/thes/ (2001).

[34] F. Manola, E. Miller (Eds.), RDF Primer, W3C, W3C Recommendation 10 February 2004, 2004, http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

[35] R. Volz, D. Oberle, B. Motik, S. Staab, KAON SERVER - A Semantic Web Management System, in: Proceedings of the 12th World Wide Web, Alternate Tracks - Practice and Experience, Hungary, Budapest, 2003.

[36] A. Farquhar, R. Fikes, J. Rice, The Ontolingua Server: A Tool for Collaborative Ontology Construction, Technical Report KSL 96-26, Stanford University, Knowledge Systems Laboratory (1996).

[37] S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, L. A. Stein, OWL Web Ontology Language Reference, W3C, W3C Recommendation 10 February 2004, 2004, http://www.w3.org/TR/2004/REC-owl-ref-20040210/.

[38] R. Tolosana-Calasanz, J. Nogueras-Iso, R. Béjar, P. R. Muro-Medrano, F. J. Zarazaga-Soria, Semantic interoperability based on Dublin Core hierarchical one-to-one mappings, International Journal of Metadata, Semantics and Ontologies (IJMS&O) 1 (3) (2006) 183–188.

[39] R. Heery, P. Johnston, D. Beckett, N. Rogers, JISC metadata schema registry, in: 5th ACM/IEEE-CS joint conference on Digital libraries, 2005.

[40] G. A. Miller, WordNet: An on-line lexical database, Int. J. Lexicography 3 (1990) 235–312.

[41] P. Vossen, Introduction to EuroWordNet, Computers and the Humanities (Special Issue on EuroWordNet) 32 (2-3) (1998) 73–89.

[42] A. Miles, D. Brickley (Eds.), SKOS Mapping Vocabulary Specification, W3C, 2004, http://www.w3.org/2004/02/skos/mapping/spec/.

[43] A. Whiteside (Ed.), OGC Web Service Common Specification, v1.0, Open Geospatial Consortium Inc, OGC 05-008, 2005.

[44] G. Janée, J. Frew, The ADEPT Digital Library Architecture, in: Proc. of the second ACM/IEEE-CS joint conference on Digital libraries, Portland, Oregon, USA, 2002, pp. 342 – 350.

[45] G. Janée, S. Ikeda, L. L. Hill, The ADL Thesaurus Protocol, http://www.alexandria.ucsb.edu/∼gjanee/thesaurus/specification.html (2003).

[46] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, New York, ACM Press, Addison Wesley, 1999.

[47] D. Nebert, A. Whiteside (Eds.), OpenGIS - Catalogue Services Specification (version: 2.0), Open Geospatial Consortium Inc, OpenGIS Project Document 04-021r2, 2004.

[48] F. J. Zarazaga-Soria, J. Lacasta, J. Nogueras-Iso, M. P. Torres, P. R. Muro-Medrano, A Java Tool for Creating ISO/FGDC Geographic Metadata, in: Geodaten- und Geodienste-Infrastrukturen - von der Forschung zur praktischen Anwendung. Beiträge zu den Münsteraner GI-Tagen 26./27. Juni 2003, Vol. 18, IFGIprints, Münster, Germany, 2003, pp. 17–30.

[49] R. Tolosana-Calasanz, D. Portolés-Rodríguez, J. Nogueras-Iso, P. R. Muro-Medrano, F. Zarazaga-Soria, CatServer: a server of GATOS, in: Proceedings of AGILE Conference 2005, Estoril, Portugal, 2005, pp. 359–366.

[50] D. Tudhope, C. Binding, D. Blocks, D. Cunliffe, Query expansion via conceptual distance in thesaurus indexed collections, Journal of Documentation 62 (4) (2006) 509–533.

[51] P. Clark, J. Thompson, H. Holmback, L. Duncan, Exploiting a thesaurus-based semantic net for knowledge-based search, in: Proc 12th Conf on Innovative Application of AI (AAAI/IAAI'00), 2000, pp. 988–995.

[52] European Union Publication Office, European Vocabulary (EUROVOC), http://eurovoc.europa.eu (2005).

[53] P. Vretanos(Eds), Filter Encoding Implementation Specification, Version 1.1, OpenGIS project document OGC 04-095, OpenGIS Consortium Inc (3 May 2005).

[54] F. J. Zarazaga-Soria, J. Nogueras-Iso, M. A. Latre, A. Rodríguez, E. López, P. Vivas, P. Muro-Medrano, Research and Theory in Advancing Spatial Data Infrastructure Concepts, ESRI Press, 2007, Ch. Providing SDI Services in a Cross-Border Scenario: the SDIGER Project Use Case, pp. 113–126.

[55] United Nations Educational, Scientific and Cultural Organization (UNESCO), UNESCO Thesaurus: A Structured List of Descriptors for Indexing and Retrieving Literature in the Fields of Education, Science, Social and Human Science, Culture, Communication and Information, UNESCO Publishing, Paris, 1995, http://www.ulcc.ac.uk/unesco/.

[56] J. Lacasta, J. Nogueras-Iso, R. Tolosana, F. Lopez, F. Zarazaga-Soria, Automating the thematic characterization of geographic resource collections by means of topic maps, in: Proceedings of 9th AGILE International Conference on Geographic Information Science: Shaping the future of Geographic Information Science in Europe, Visegrád, Hungary, 2006, pp. 81–127.