Earth Sci Inform (2009) 2:235–247
DOI 10.1007/s12145-009-0039-2

RESEARCH ARTICLE

Searching for metadata using knowledge bases and topic maps in Spatial Data Infrastructures

Odilon Corrêa da Silva · Jugurta Lisboa-Filho ·
José Luís Braga · Karla A. V. Borges

Received: 1 June 2009 / Accepted: 6 October 2009 / Published online: 27 November 2009

Abstract Concern for environmental issues has become a reality in all sectors of society, mainly among researchers and professionals who work directly with environmental status. In this context, several studies have been undertaken on sustainable development of the Brazilian Amazon, generating a large amount of data and information. Environmental area characterization involves the knowledge about their natural, economic and social resources, as well as understanding the interaction and correlation among them. Such interdisciplinary character requires new solutions for knowledge representation. This study proposes to minimize metadata recovery problems in Spatial Data Infrastructures by using Topic Maps and Thesaurus. This approach applied to an interface aims to allow users to visually recover information from metadata catalogs.

Keywords Information retrieval systems · Metadata · Spatial data infrastructure · Thesauri · Topic maps

Communicated by: H.A. Babaie

SI: Spatial data infrastructures for the Amazon: a first step towards a Global Forest Information System

O. C. da Silva (✉) · J. Lisboa-Filho · J. L. Braga
Departamento de Informática,
Universidade Federal de Viçosa (UFV),
CEP 36570-000 Viçosa, MG, Brazil
e-mail: odilon.correa@gmail.com

J. Lisboa-Filho
e-mail: jugurta@ufv.br

J. L. Braga
e-mail: zeluis@dpi.ufv.br

K. A. V. Borges
Prodabel, Empresa de Informática e Informação do Município de
Belo Horizonte, Av. Pres. Carlos Luz, 1275,
CEP 31230-000 Belo Horizonte, MG, Brazil
e-mail: karla@pbh.gov.br

Introduction

The Amazon region extends over eight countries in northern South America. In Brazil, the Legal Amazon encompasses the states of Acre, Amazonas, Amapá, Rondônia, Roraima, Pará and Tocantins, all in the northern region of the country, also including the states of Mato Grosso in the Midwestern region and the state of the Maranhão in the northeastern region. In total, the Amazon comprises approximately 5 million km^2, representing 58% of the Brazilian territory (IBGE).[1]

The Amazon ecosystem is a major reservoir of biodiversity on the planet, with great potential still untapped, and huge quantities of minerals, agricultural land and many other resources. It is within this context that one seeks a path toward sustainable development in the region, which could establish a relationship between economic growth and reduction of environmental degradation. The Brundtland Commission (OECD 1996) defined Sustainable Development as that which "meets the needs of the present without impairing the ability of future generations to meet their own needs", which constitutes one of the greatest challenges facing humanity in this century.

The sustainable planning of any activity such as urban planning, reforestation and agricultural projects requires, firstly, knowledge on the environment in which this activity will be inserted (Burrough and McDonnell, 1998). Environmental area characterization requires knowledge on natural resources and understanding the interactions and

[1] Brazilian Institute of Geography and Statistics—http://www.ibge.gov.br

correlations among them (Resende et al. 1995). Geographical Information Systems (GIS) along with remote sensing techniques and products can help in knowledge accumulation by continual managing and updating of spatial data (Resende et al. 1995; Burrough and McDonnell 1998). Knowledge generated by a community that deals with computational representation of space has to be available to researchers, allowing them to analyze the characteristics of these elements and their interactions.

The need to share information and the rapid Internet growth have made it a preferred medium for data dissemination. Internet capillarity triggered the development of a new class of information systems with a different architecture from its predecessors. This movement has extended to spatial data. The main GIS suppliers have alternatives for accessing spatial data through the global network (Davis and Alves 2005).

Dissemination and sharing of spatial data sets from different backgrounds have a huge expansion if supported by a computational environment in which the data are freely shared in an integrated way. However, this is not the current reality. According to Davis and Alves (2005), organizations that aim to share data often face significant issues such as encoding formats, storage, quality standards, content limitations, map projection parameters and even data structures. Maguire and Longley (2005) suggested the use of Spatial Data Infrastructures (SDI) to minimize this problem, enabling cooperation and sharing of data.

The open and distributed nature of SDIs is nevertheless, an obstacle to retrieval and sharing of geospatial information (Athanasis et al. 2008). In general, the search on a SDI is mainly based on keywords, spatial coordinates, temporal or thematic classification. With the search result, the user can access the source to view or download data, analyze metadata or start a new search. This approach has a number of difficulties, particularly for inexperienced users, who may not know which keywords to use, how to fill properly query forms or define the number of criteria to use (Hochmair 2005). Such limitations in the recovery of geospatial information may hinder search in complex contexts such as the environment.

Given these issues, we sought a new approach to the recovery and sharing of geospatial information, which together with the questions arising from the interaction among SDI components, is a central theme in the Geospatial Semantic Web (Egenhofer 2002).

Thesaurus is one of the various forms of information representation used by Information Recovery Systems. The flexibility to establish new relationships between terms, hierarchies and cross-references gives the tool a variety of uses, including processes that range from indexing to the effective retrieval of information. This work proposes to minimize metadata recovery problems in SDI using knowledge bases and topic maps. This approach aims to allow users to visually recover metadata stored in SDI.

In "Spatial Data Infrastructure", the article presents the concept of SDI. "Knowledge representation" presents forms of knowledge representation and its utilization in information retrieval. "An architecture for Information Retrieval Systems in metadata catalog" presents a proposal of architecture for a system of information retrieval in a SDI metadata catalog. "Case study" presents a case study that illustrates how an SDI has been improved with the proposed architecture. Finally, "Conclusions" presents the conclusions of the work.

Spatial Data Infrastructure

The term Spatial Data Infrastructure was proposed in 1993 by the Mapping Sciences Committee of the U.S. National Research Council and initially used to describe the provision of standardized access to spatial information (Maguire and Longley 2005).

Davis and Alves (2005) discusses that the focus on the SDI concept is a natural consequence of the evolution of the Web and its architecture, given the guidelines that have been drawn up by its regulatory body, the World Wide Web Consortium (W3C). The main WEB change is that the presentation is no longer the central focus, with the content sharing its place, which allows the semantic structuring of data. This change of focus was consolidated in the concept of the Semantic Web, "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in co-operation" (Berners-Lee et al. 2001). Davis and Alves (2005) proposed an architecture for SDI development in which data are provided by different information services, forming then a service oriented infrastructure (Fig. 1) that was used in developing the case study in "Case study".

Nebert (2004) states that the term SDI is used to describe a set of technologies, policies, standards and human resources required (especially map-producing organizations) to promote availability and access to spatial data. The paper describes a set of principles, norms and protocols, so that adjacent SDI exchange information. It describes the aspects required to begin an SDI construction, which include spatial data, metadata, catalogs, online data visualization, data access, geoservices, training and public policy.

In environments where there is great data heterogeneity, metadata allow users to relate data using a classification system common among various types of data available in order to improve data integration and sharing.

Currently, there are spatial metadata standards available to document collections of geospatial data. Among them, the experiences of the U.S. government with the Content

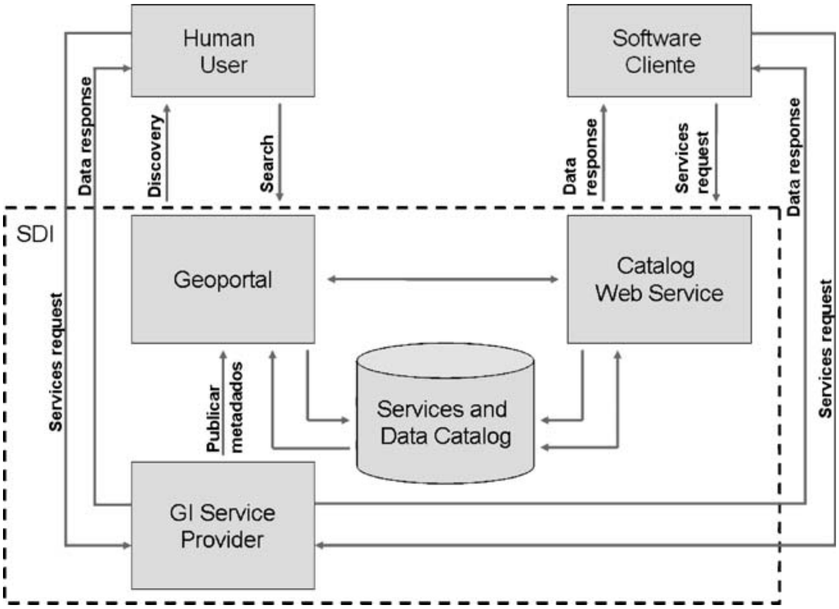


Fig. 1 Geoportals and SDI (Davis and Alves (2005)

Standards for Digital Geospatial Metadata (CGSDM) of the Federal Geographic Data Committee—FGDC (FGDC 1998) and the ISO/TC 211 (ISO standard 19115) are widely used. The ISO/TC211 standard was developed by the ISO Committee aiming to establish a structured set of standards for information directly or indirectly associated with spatial location. These standards are usually based on elements of quality defined by the ICA (Guptill and Morrisson 1997) and provide information on existing data, framing of data in certain applications and conditions for access and transfer to users. The use of a reference model of metadata for geographic data allows the description of their properties according to different descriptive approaches and this reference model should be flexible enough to fit easily to new application requirements of users.

Nogueras-Iso et al. (2005) discuss some issues related to SDI development, emphasizing the role played by metadata. The authors emphasize the importance of metadata, addressing issues related to the interoperability among the standards ISO, FGDC and Dublin Core, and benefits of using thesauri and ontologies for information retrieval in metadata catalogs. Another emphasized issue is the little or total lack of relationship between Digital Libraries and SDI. The authors warn that experience in this area could enhance the development of concepts in many aspects of a SDI. In this sense, they suggest that digital libraries could be considered as part of SDI.

Knowledge representation

According to Davenport and Prusak (1998), providing knowledge in a format that is accessible to the user requires the definition of an encoding. A key aspect to efficiently achieve these results is how the knowledge is represented. Garshol (2004) states that one of the concerns of an information architect is to create sites in which users can actually find what they are looking for. Information architects use several techniques and technologies for site creation and organization, but their main instruments are related with organizing information, which have their origin in other subjects. Some of these techniques come from the library management, such as controlled vocabularies, taxonomies and thesauri.

This section discusses particularly thesauri, ontologies and topic maps, exploring their specific contributions to both representation and retrieval of information.

Thesauri

The word thesaurus, according to Dodebei (2002), comes from the Greek word thesaurós meaning treasure or repository of words. The word became popularized with the publishing of the analog dictionary "Thesaurus of English Words and Phrases" by Peter Mark Roget, in London (1852 apud Dodebei; Dodebei 2002). Roget coined the term thesaurus to describe his catalogue of words, because the term also designates vocabulary, dictionary or lexicon. Roget's dictionary is different from the others because the words have been arranged according to their meaning and not by alphabetical order and had the merit of establishing the denomination of vocabularies that relate its terms through some kind of meaning.

Foskett (1972 apud Dodebei; Dodebei 2002) points out that thesaurus is a tool for terminological control, and in agreement with Lancaster (1972 apud Dodebei; Dodebei

2002), the author lists as the main functions the control of synonyms and near-synonyms, homograph distinction and provision of facilitators to search for related terms and references, improving the consistency of indexing and transforming the search language into indexing language, reducing time and increasing efficiency in indexing and information retrieval tasks.

W3C establishes the SKOS standard (Simple Knowledge Organization Systems), which is based on RDF (Resource Description Framework), to represent thesauri and other similar types of systems for knowledge organization (SKOS 2009). The specifications for constructing a thesaurus are defined by the standards ISO 2788 (ISO 1986) and ISO 5964 (ISO 1985). Figure 2 illustrates a UML Class diagram (Booch et al. 2005) for representation of a thesaurus component.

Figure 2 shows that this model of representation has multilanguage support, meeting the specifications of ISO 5964 (ISO 1985). The elements of the model are described below:

- Scope Note: defines, explains or limits the meaning of the descriptor for indexing purposes;
- Term: A term of the thesaurus;
- Translation: Provides support for multilingual thesaurus;
- Top Term: indicates the top term of the thesaurus;
- Preferred Term: indicates the synonyms or non-descriptors, not valid for indexing;
- Broader Term: marks the broader term to which the descriptor belongs;
- Narrower Term: indicates the specific terms (types or classes) involved;
- Related Term: indicates terms semantically associated with the descriptor.

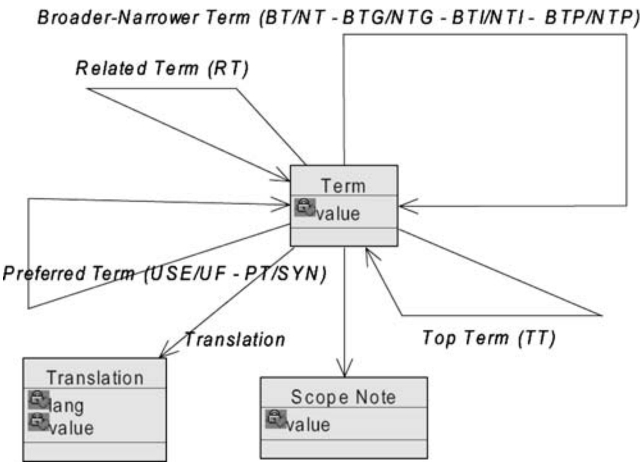

Broader-Narrower Term (BT/NT - BTG/NTG - BTI/NTI - BTP/NTP)
Related Term (RT)
Term
value
Preferred Term (USE/UF - PT/SYN)
Translation
Top Term (TT)
Translation
lang
value
Scope Note
value

Fig. 2 Model for a thesaurus representation (Miguel 2009)

There are several thesauri created and maintained by non-governmental organizations available for consultation and access to content on the Internet. GEMET (General Multilingual Environmental Thesaurus) is a thesaurus of environment terms developed by the European Topic Center and currently maintained by the European Environment Information and Observation Network. GEMT is publicly available for consultation via the web portal and through Web Services, in addition to providing their databases in various formats.

Dodebei (2002) considers that the determination of the conceptual universe can be obtained by both prior domain knowledge and observation of the conceptual field, recognizing, in the methodology field, induction and deduction as choice processes for the composition of this universe.

In the inductive process, which was denominated “Empirical” by Lancaster (1972 apud Dodebei; Dodebei 2002) and “Analytical” by the American National Standard Institute (ANSI), the terminology is obtained through the identification of terms collected in the literature. However, in the conceptual process, the terminology is obtained by consensus among experts in the subject. The method is also called Gestalt, as opposed to “Analytical”, or Committee Approach as opposed to “Empirical” (Dodebei 2002). The author indicates that the two procedures derive from two principles governing term survey, which are the Literary Warrant or Bibliographic Warrant, by Hulmer (1950 apud Dodebei; Dodebei 2002) and the User Warrant or Personal Warrant, by Lancaster (1972 apud Dodebei; Dodebei 2002). These methods are used in the theoretical foundation and construction of the semantic basis described in “An architecture for Information Retrieval Systems in metadata catalog”.

Ontologies

Gruber’s (1995) definition of ontology is frequently cited in the literature: “An Ontology is a formal, explicit specification of a shared conceptualization”. The most important consideration in this statement is the notion of a conceptualization. A conceptualization can be developed with terminologies and vocabularies, establishing properties and allowing the knowledge to be reused, avoiding rework or rediscovery of equivalent terminology (Guarino 1998).

Ontology, a term with a long history in philosophy, has been incorporated into the vocabulary of Artificial Intelligence, Knowledge Engineering and Knowledge Representation, among other fields of Computer Science. Over the last two decades, research in Artificial Intelligence has been using ontologies for the formal description of things in the world, which is essential for intelligent systems to act and ponder upon the world in which they will operate (Welty and Guarino 2001).

It is a fact that there is still no consensus as regards the definitions and applications of ontologies. Maedche and Staab (2001) consider that the process of building ontologies is complex and lengthy, involving a survey of concepts and terms of interest in a particular domain, according to the aim one wants to achieve. Therefore, since the thesaurus is a very important component for a knowledge retrieval system, besides reaching a better consensus on its application than ontologies, these facts, together with aspects of implementation, supported the decision to adopt the thesaurus as an instrument of knowledge representation in SDI.

Topic maps

The digital environment that has been set up in recent decades has the collections of digital objects growing with regard to both typology and complexity. In this scenario, text, images, sounds, videos, websites and various other digital objects require different types of treatment and representation for efficient information retrieval (Burke 1999).

According to Sigel (2000), the knowledge management seeks to organize and optimize knowledge repositories so that the user can effectively retrieve information. Rath (2003) believes that talking about knowledge structure is talking about topic maps. The ISO 13250 (ISO 2002) defines topic maps as "a unified international standard to describe knowledge structures and formalize their association with information resources". Topic maps can be seen as a paradigm that can organize and retrieve information stored in SDI. Choosing topic maps allows for, e.g., the creation of an information retrieval system in catalogs of metadata. Topic maps are similar, in many ways, to semantic networks, conceptual maps and mind maps, although only topic maps are standardized, with structure formalized by ISO 13250 (ISO 2002), which allows future implementation of interoperability.

The forms of knowledge representation through thesauri and ontologies described in this section influenced the creation of the Topic Maps Standard (ISO 2002). It is therefore appropriate to say that topic maps is a map that represents the knowledge found in a domain, besides enabling to present relevant concepts and relationships among them, as it is done in the thesaurus.

Topic maps consist of three concepts (Fig. 3): Topics, Associations and Occurrences, represented by the acronym TAO (Pepper 2000). Despite the simplicity, these elements have high descriptive power. These elements are described below:

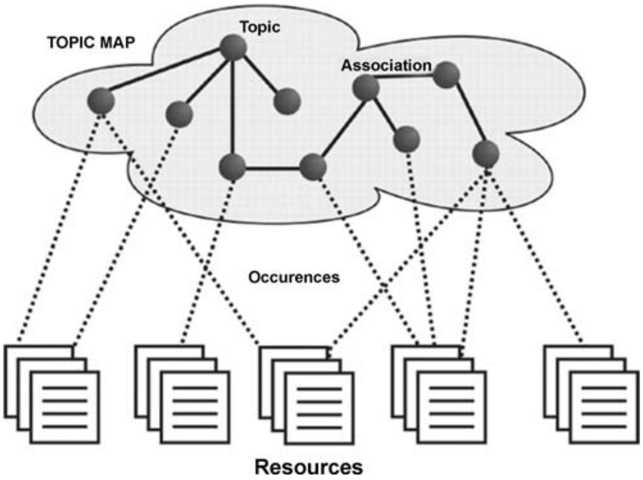


Fig. 3 Elements of topic maps standard (adapted from: Garshol 2002)

- **Topics:** can represent a theme in an application domain as a person, an entity, a concept, etc. Strictly speaking, the term "topic" refers to an element of the topic maps representing a theme. The topics have relationship such as class/subclass, for example, "Institution" is a supertype of "University", which is a supertype of "Federal University of Viçosa";
- **Associations:** can describe relationships between topics. An association is a link element that establishes a relationship between two or more topics. An example of association is the topics "Institution" and "Event" are linked by the association "To organize" which is represented by the role "Organize" and "Organized-by";
- **Occurrences:** a topic can be linked to one or more information resources which are relevant to the topic in question in some way. Resources are termed topic occurrences. An occurrence is a link that connects a topic with a resource. As example, consider the resources website (http://www.ufv.br) and year of foundation (1922) as occurrences of the topic "Federal University of Viçosa".

Ahmed (2009) presents a design pattern for topic maps based on ontologies. This method can be used to represent thesaurus in topic maps following a few principles proposed by Garshol (2004). The design standard proposed by Ahmed (2009) and adapted by Garshol (2004) can be used to represent the thesaurus maintained by SDI users in topic maps. The topic maps structure can relate topics and resources, and can be applied to relate thesaurus terms with SDI metadata. The creation of a knowledge bases and implementation of an interface to retrieve information in metadata catalogs of a SDI using topic maps can help their users.

Information Retrieval Systems

The term Information Retrieval attributed to computer systems is highly debatable, and many authors prefer the term Document Retrieval and Text Retrieval (Ferneda 2003). Note that the systems do not recover “information”, but references or documents whose content is relevant to the needs of the user. Thus, the designation Information Retrieval (IR) will be used in this work, it being understood that this is “information” contained in the documents and texts retrieved.

The basic principles of the probabilistic model for Information Retrieval Systems (IRS) were launched by Maron and Kuhns (1960 apud Ferneda; Ferneda 2003) and formally defined by Robertson and Jones (1976 apud Ferneda; Ferneda 2003). The first IRS were based on the frequency of words in the text. Although the use of more sophisticated methods have arisen as a result of a natural evolution of mathematical models in search of a deeper semantic processing of the text, research using statistical models continued to generate new models and improving old ideas. That is the case of the Boolean model, which was proposed in 1854, and several other models that have been updated for use in information retrieval on the Web (Ferneda 2003).

The presence of Boolean mechanisms in these systems has not resulted necessarily in greater ease of use. This fact is demonstrated in studies of Savoy and Picard (1998) and Jansen et al. (2000), which point out the low use of Boolean operators by users of Web search engines. AND is the most used among the operators (Table 1).

The process of information retrieval consists of identifying among the documents of a system, which meet the information needs of the user (Burke 1999). Grand and Soto (2002) state that visual search techniques enhance the user’s perception of the environment perception. The authors think that navigation resource is essential, since it helps users to build their own representation of the environment. The growing need to manage and provide an increasing amount of data led the development of several techniques for visualization of information. Card et al. (1999) point out that the use of visual and interactive representation and supported by computer can broaden the user’s perception of the involved knowledge.

Table 1 Use of Boolean operators and modifiers in search expressions (N expressions = 51.473)

Operator or modifier	Number of expressions	% of all expressions
AND	4094	8
OR	177	0,34
AND NOT	105	0,20
( )	273	0,53
“ ”	3282	6

Adapted from Jansen et al. (2000, p.217)

The hyperbolic tree is an example of technology for visualization and navigation based on the focus + context technique (Lamping et al. 1995). This technique can be used for graphical representation of a topic maps, which enables the User to more easily view and navigate the information content without the need to know exactly the wanted term. By using the hyperbolic tree to represent a topic maps, the topic positioned by the user at the focus appears larger. During navigation, the appearance of peripheral topics grows exponentially as the topic maps is moved. Because of these characteristics, the hyperbolic navigation together with thesauri, topic maps and IR methods provide a good alternative for viewing and retrieval of large data sets, for example, metadata catalogs of an SDI.

In the context of SDIs, the researchers Aditya and Kraak discuss and propose the development of interfaces and mechanisms of IR in metadata catalogs (Aditya and Kraak 2007; Aditya 2007). For the authors, it is essential the development of search interfaces for SDIs and propose the use of maps and graphs to support the discovery of geospatial resources. In another study, Miguel (2009) analyzes the main problems affecting IRS in SDI related to semantic structures. The main objective of this research is the integration of terminology models in SDI, aiming at simplifying the resource classification and improving IR. However, the author fails to address a way to present the proposed solutions to users of a SDI, for example, an interface for consultation and research.

An architecture for Information Retrieval Systems in metadata catalog

This section presents an architecture for Information Retrieval Systems in metadata catalogs available in SDIs. The system aims to provide an alternative solution to IR problems in a metadata catalog of a SDI. It is thus expected not only to minimize some of these problems, but also to bring more semantics to query operations in metadata catalogs. The specification of the system is directed to the needs of an SDI of an environmental project. However, this specification and the development of the system have been carried out in a way that it can be applied to other SDIs.

In this section, diagrams of the Unified Modeling Language (UML) are used to describe the proposed architecture. The UML diagrams used are: use case, sequence, activity, component and class (Booch et al. 2005). The use cases illustrated in Fig. 4 defines the services provided by the information retrieval system in metadata catalogs. The services available to the User are based on indexing, generating, consulting and evaluating a database.

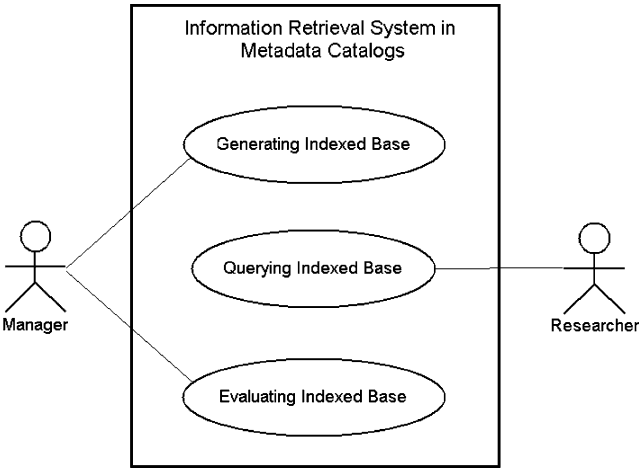


Fig. 4 Use cases for Information Retrieval System in metadata catalogs

The use case Generating Indexed Base must be able to extract terms and relationships from the Semantic Base (Thesaurus), resources from the Resource Base (Metadata Catalog) and generate the Indexed Base (Topic Maps) with terms, relationships and occurrence between terms and resources. Figure 5 illustrates the sequence diagram of this use case.

The use case Querying Indexed Base must be able to extract occurrences from resources of the Indexed Database that meet the query criteria and make the resources retrieved in the Resource Base available to the User. Figure 6 shows the sequence diagram of this use case.

The use case Evaluating Indexed Base must be able to extract resources from the Resource Base, terms from the Semantic Base and provide the User a list of terms and resources that were not indexed in the Indexed Database. Since the focus of the work is the generation and query to Indexed Database the sequence diagram of this use case is not presented. The following sections describe the proposed architecture.

System's components

In order to implement the use cases defined in the previous section, components that make up the architecture of the information retrieval system in metadata catalog were defined. The system consists of three modules that provide their services through interface classes. Two applications (one for management and another for query), a module controlling persistence and three data repositories are also part of the system architecture. Figure 7 shows the components and their relationships, which are described in details below.

Component description:

- Resource Base: This repository stores the resources extracted from the catalog metadata of the SDI. The resource is made up by the identification of metadata and part of their content. The content consists of elements (eg, abstract, keyword) extracted from the structure of metadata which are usually stored in XML;
- Semantic Base: This repository stores the semantic content of a thesaurus, with a framework that allows storing the identification of the thesaurus, its terms, concepts and relationships. The framework was developed following ISO 2788 (ISO 1986) and ISO 5964 (ISO 1985) specifications for thesaurus construction;
- Indexed Base: This repository stores the result of the indexing process between the Resource Base and Semantic Base. It stores the topics (terms of the thesaurus), their associations (relationships of terms) and occurrences (relationship between the term of the thesaurus and metadata identification). The result of the

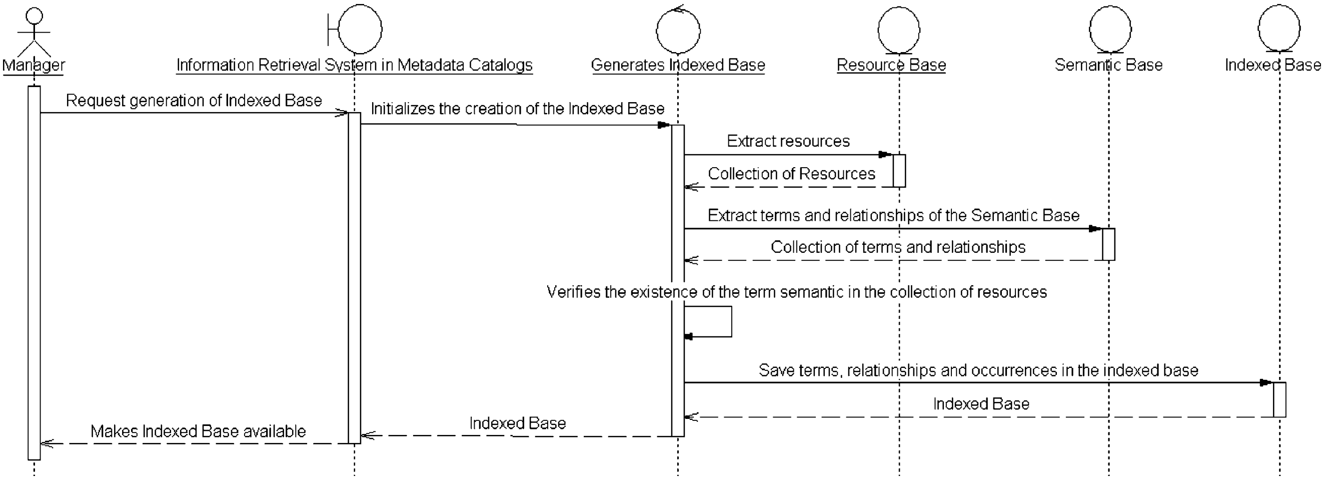

Manager
Information Retrieval System in Metadata Catalogs
Generates Indexed Base
Resource Base
Semantic Base
Indexed Base
Request generation of Indexed Base
Initializes the creation of the Indexed Base
Extract resources
Collection of Resources
Extract terms and relationships of the Semantic Base
Collection of terms and relationships
Verifies the existence of the term semantic in the collection of resources
Save terms, relationships and occurrences in the indexed base
Indexed Base
Indexed Base
Makes Indexed Base available

Fig. 5 Sequence diagram of the use case generating indexed base

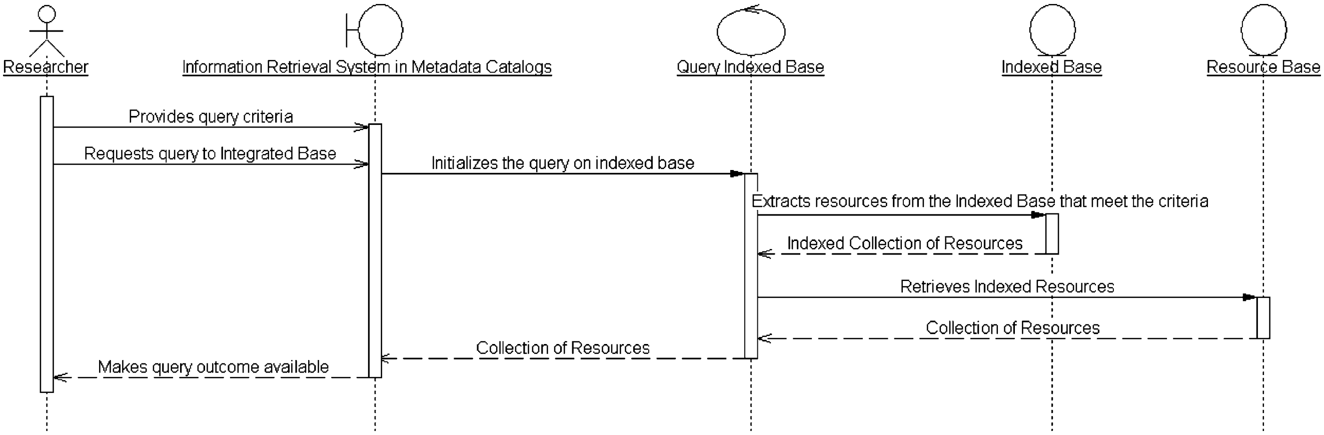

Researcher
Information Retrieval System in Metadata Catalogs
Query Indexed Base
Indexed Base
Resource Base
Provides query criteria
Requests query to Integrated Base
Initializes the query on indexed base
Extracts resources from the Indexed Base that meet the criteria
Indexed Collection of Resources
Retrieves Indexed Resources
Collection of Resources
Collection of Resources
Makes query outcome available

Fig. 6 Sequence diagram of the use case querying indexed base

indexing process follows ISO 13250 specifications (ISO 2002) for construction of topic maps;

- Indexed Base Generation Module: This module is responsible for implementing an interface with the application Management and managing the generation of Indexed Database as a whole. The process of generating the Indexed Database is described in the next section;
- Query to Indexed Base Module: This module is responsible for implementing the interface with the application Query and managing the process of query to Indexed Database;
- Indexed Base Evaluation Module: This module is responsible for implementing an interface with the application Management and managing the evaluation of Indexed Base. This process analyzes the Integrated Base in relation to the Resource Base and Semantic Base, generating a list of which metadata and terms were not indexed.
- Persistence Module: This module is responsible for monitoring the persistence and retrieval of repository data;
- Management Application: This application is responsible for translating the features of the modules for Generation and Evaluation of Indexed Base to something that the User can understand and therefore interact with the system;
- Query Application: This application is responsible for translating the features of the module Query to the Indexed Base to something that the User can understand and therefore interact with the system. This application displays the Indexed Base through a hyperbolic tree, in which the User can browse, select terms of interest and finally process the search.

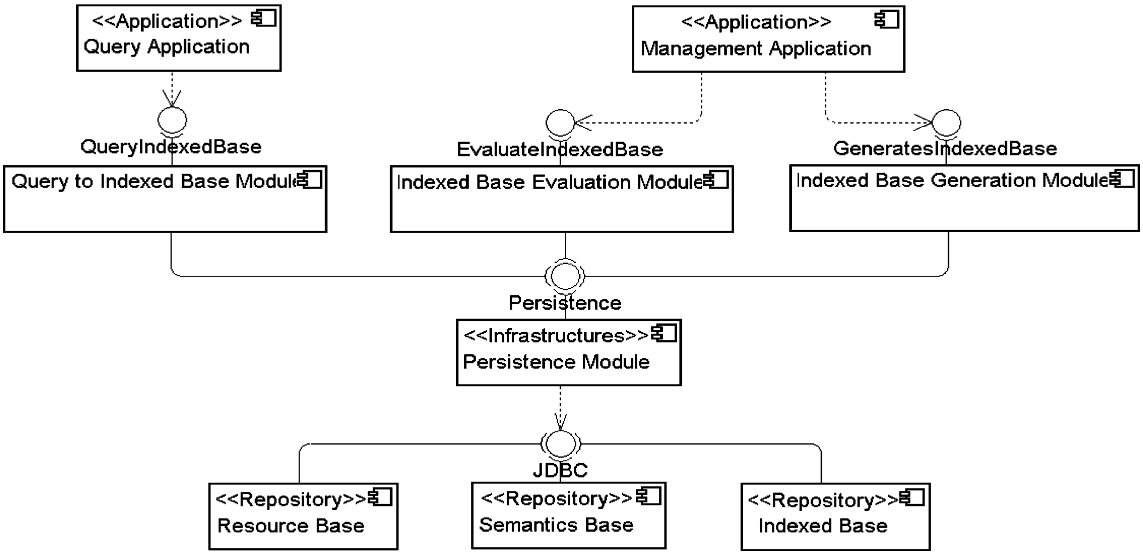


Fig. 7 Diagram of system components

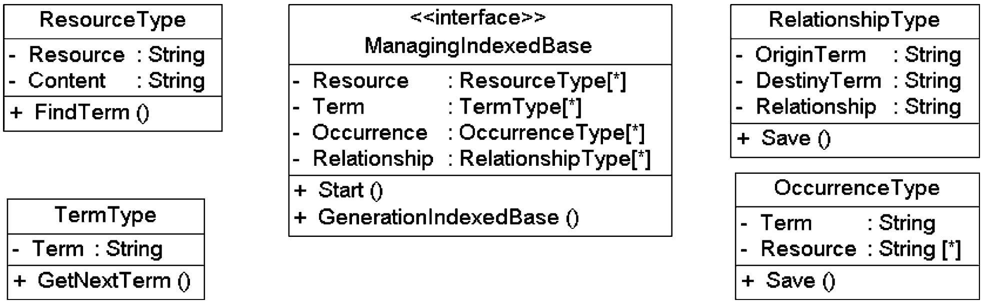


Fig. 8 Class diagram used in generation of indexed base module

Generation of the indexed base

Figure 8 shows the set of classes used in the Generation of Indexed Base Module. The features of the module is defined by *ManagingIndexedBase* class and the attribute's type is defined by the others.

The *ResourceType* class defines the resources structure extracted from the Resource Base. The *TermType* class defines the terms structure extracted from the Semantic Base. The *RelationshipType* class defines the relationships structure extracted from the Semantic Base and will persist in the Indexed Base. The *OccurrenceType* class defines the resources structure that will persist in the Indexed Base. Finally, the *ManagingIndexedBase* class is responsible for the whole process control in Generation of Indexed Base.

When an user requests the generation of an indexed base, an object of the class *ManagingIndexedBase* is instantiated and the method *Start*() is invoked. *Start*() is responsible for updating the attributes *Resource*, *Term* and *Relationships* with collections retrieved from the *Resource* and *Semantic bases*. Then the method *GenerationIndexedBase*() will be invoked initiating. The method *GenerationIndexedBase*() manages the indexing process from collections that have been updated and running the flow of activities as in Fig. 9.

Case study

This section describes the development of a SDI, the process of creation and evolution of a semantic base with SDI users and the implementation of IRS in metadata catalog, as it was proposed in the previous section. The case study was developed in the ambit of the Project Terrestrial Ecosystems in Antarctica (TERRANTAR[2]). However, the specifications of the system have been carried out in a way that it can be applied to other SDIs. Because of the similarity of the environmental context, this case study serves to illustrate the potential of the proposed architecture in the context of projects aimed at the issue of the Amazon.

The main objective of the TERRANTAR Project is to manage and share the results of research on the Antarctic continent. Data generated are recorded and stored in different formats, such as vector and/or raster geographic data, theses, articles, spreadsheets and photos. The first data collection that has been cataloged and made available in TERRANTAR's SDI is the result of projects developed by researchers in the Soil Science Department of the Federal University of Viçosa.

TERRANTAR's Spatial Data Infrastructure

The main objective of the TERRANTAR's SDI is to give support to geospatial data sharing. However, technological compatibility is always a concern. One of the latest initiatives to promote interoperability was coordinated by the Open Geospatial Consortium (OGC), which has proposed standards for the use of services in geographic information sharing.

Considering the initial requirements of the project, a technical feasibility study was conducted on existent standards, methods, concepts and technologies that could underlie its development. The focus of the study was the search for computational tools to optimize and facilitate SDI development. Figure 10 illustrates the SDI framework of the TERRANTAR Project improved by IRS metadata developed with the proposed architecture. The figure also shows the open source tools used in the implementation of the SDI and its relationship with elements of this framework. Finally, it can be seen that spatial data can be stored in several formats.

Cataloging and metadata access are provided by Geonetwork.[3] The metadata repository is maintained by PostgreSQL.[4] Implementation of Web Services is provided by GeoServer.[5] This tool provides geographic data in various formats, providing interoperability for users. Finally, OpenLayer[6] allows users to view and analyze spatial

[2] TERRANTAR—Antarctica Terrestrial Ecosystems—http://www.terrantar.com.br

[3] Geonetwork—http://geonetwork-opensource.org

[4] PostgreSQL—http://www.postgresql.org

[5] GeoServer—http://www.geoserver.org

[6] OpenLayer—http://www.openlayers.org

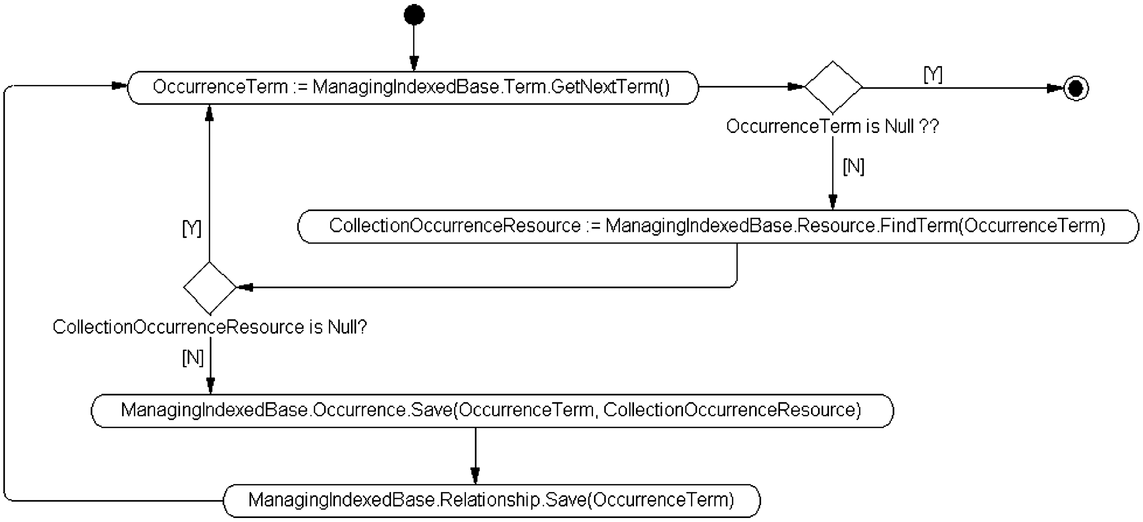

OccurrenceTerm := ManagingIndexedBase.Term.GetNextTerm()
[Y]
OccurrenceTerm is Null ??
[N]
CollectionOccurrenceResource := ManagingIndexedBase.Resource.FindTerm(OccurrenceTerm)
[Y]
CollectionOccurrenceResource is Null?
[N]
ManagingIndexedBase.Occurrence.Save(OccurrenceTerm, CollectionOccurrenceResource)
ManagingIndexedBase.Relationship.Save(OccurrenceTerm)

Fig. 9 Activity diagram of the method GenerationIndexedBase

data in interactive maps without necessarily having to download them.

Geonetwork provides a solution to the portal of communication with users. The portal has an interface that follows a traditional approach for metadata recovery. Users can fill out keyword fields, spatial coordinates or time ratings. This approach has a very high research cost, since metadata are usually stored in semi-structured formats (e.g. XML). Another drawback is the user's experience and knowledge to combine the criteria to be used.

Creating the semantics base

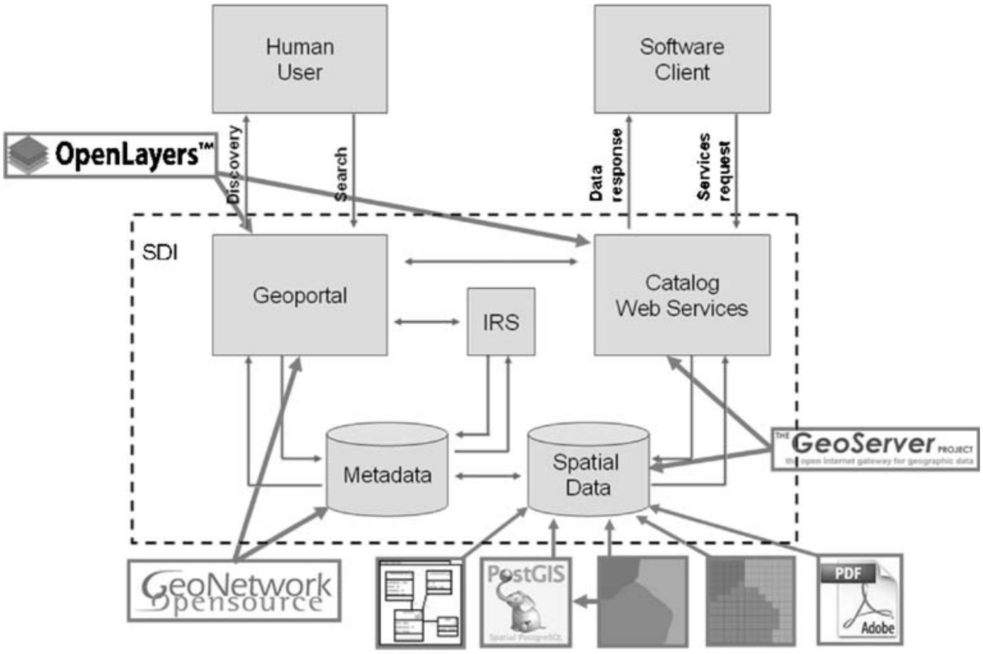
This section presents the method specification for the Semantics Base generation to be used in a SDI. It also describes the process of creation of this knowledge database that must be fed by a user community connected to a specific domain. In the focus of this case study, the base is fed by people involved with research related on the Antarctic Ecosystem.

The formation and creation of semantics are supported by a management system for indexing languages (MSIL) and its evolution is influenced by users' interaction. The production, collection, evaluation and information sharing are roles played by SDI users. New terms for the thesaurus are collected by the community and logged into MSIL, becoming available to a user group responsible for its evaluation and ultimately available for viewing and consumption of the SDI. In the case of a community working with geospatial data, MSIL can be understood as a component of SDI, which can improve metadata production, following the observations that accumulated experiences in the field of Digital Libraries can enhance the development of SDI aspects (Nogueras-Iso et al. 2005).

The creation of a thesaurus in collaboration with a community follows the User Endorsement method by Lancaster (1972 apud Dodebei; Dodebei 2002). But a literature review must be carried out to define the main themes and categories of the thesaurus specialization. The step of Literary guarantee used by Hulmer (1950 apud Dodebei; Dodebei 2002) had its aim to extract themes, categories and terms of the GEMET thesaurus that met the initial requirements of the SDI. Generation of the thesaurus is supported in the premises of collective intelligence, together with a MSIL added to an SDI. The method consists of three steps, which are described below:

- **Formation of the SDI community:** the basic elements of a community are people (eg, members, leaders, and collaborators), a context and its objectives. In this step the user is informed about rules and norms to be followed. Users are registered and provided with a login and password depending on their roles (Administrator, Content Reviewer, Editor, Registered User) in the community;
- **Spatial data and metadata publishing:** using a SDI, the community Editors can publish their spatial data, jointly with metadata, which must contain terms identified in the SDI thesaurus. Data and metadata are validated by the Content Reviewer, who will review all information posted by the Editors. Finally, the material released by the Content Reviewer is available to the community in the SDI. A diagram of the publishing step is shown in Fig. 11;
- **Thesaurus publishing:** the published thesaurus can be accessed by SDI users who search semantic knowledge of context.

Fig. 10 Components and SDI of the TERRANTAR project



Metadata retrieval using semantic base and topic maps

The current process of testing the IRS module in a metadata catalog is not fully automated (Fig. 12). In the first step, metadata and thesaurus are imported and structured in RDF-SKOS, for the Resource and Semantic Bases, respectively. The routine Generation of Indexed Base is then run. After that, the application Query to Indexed Base can be used.

The developed query module is focused on User browsing behavior, ie, usability research is prioritized. Usability involved in this case is to be able to follow the reasoning of the research supported by a graph representing the context of information available. The User can select the wanted terms and relationships and process the query. Figure 13 illustrates the interface of the query module, which is available on GeoPortal of the SDI.

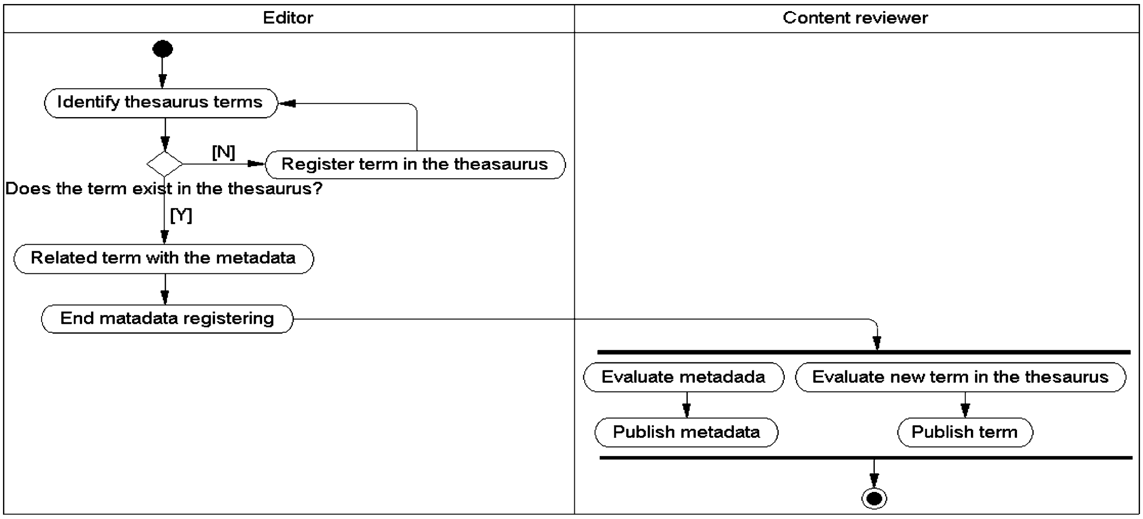

Editor
Content reviewer
Identify thesaurus terms
[N]
Register term in the theasaurus
Does the term exist in the thesaurus?
[Y]
Related term with the metadata
End matadata registering
Evaluate metadada
Evaluate new term in the thesaurus
Publish metadata
Publish term

Fig. 11 Activity diagram sequencing publishing activities

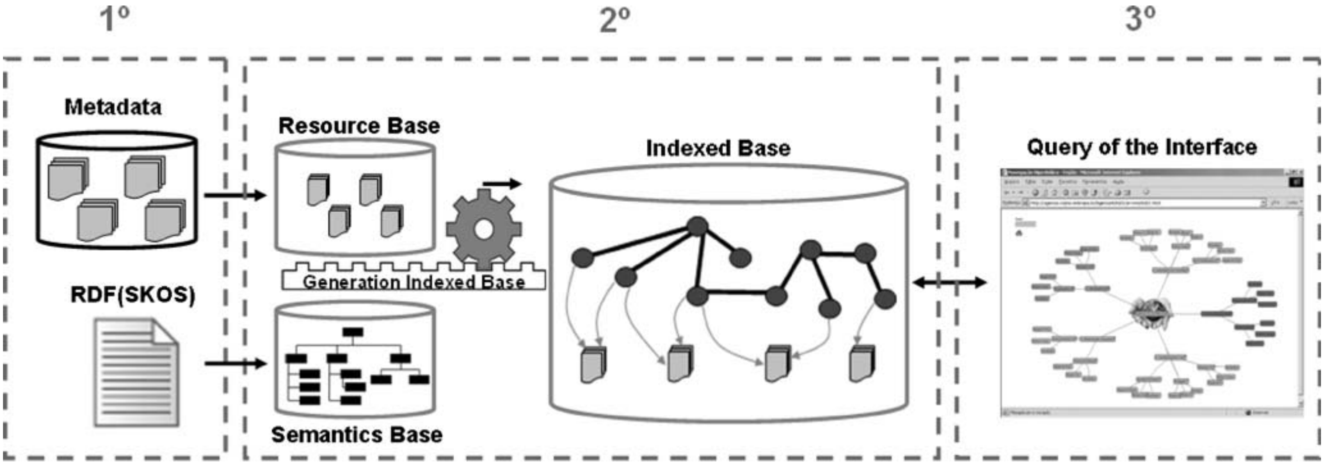


Fig. 12 Steps to test IRS metadata

Conclusions

This paper described the use of the thesaurus and topic maps to aid in the retrieval process in SDI metadata catalogs. This work's main contributions include: (1) proposal of a method for creating a thesaurus in a SDI; (2) definition of a system to retrieve metadata in SDI supported by thesaurus and topic maps; (3) implementation of an interface based on visual search techniques.

The interface proposed in this work offer a viable alternative for recovery of visual metadata in SDI. These techniques can be easily implemented, without overloading the environment and still produce good results. This approach has the advantage of providing the users with a SDI knowledge bases of the context of SDI.

The method, architecture and interface described in this work are currently being validated by the TERRANTAR Project. Once they have been validated, they can be applied

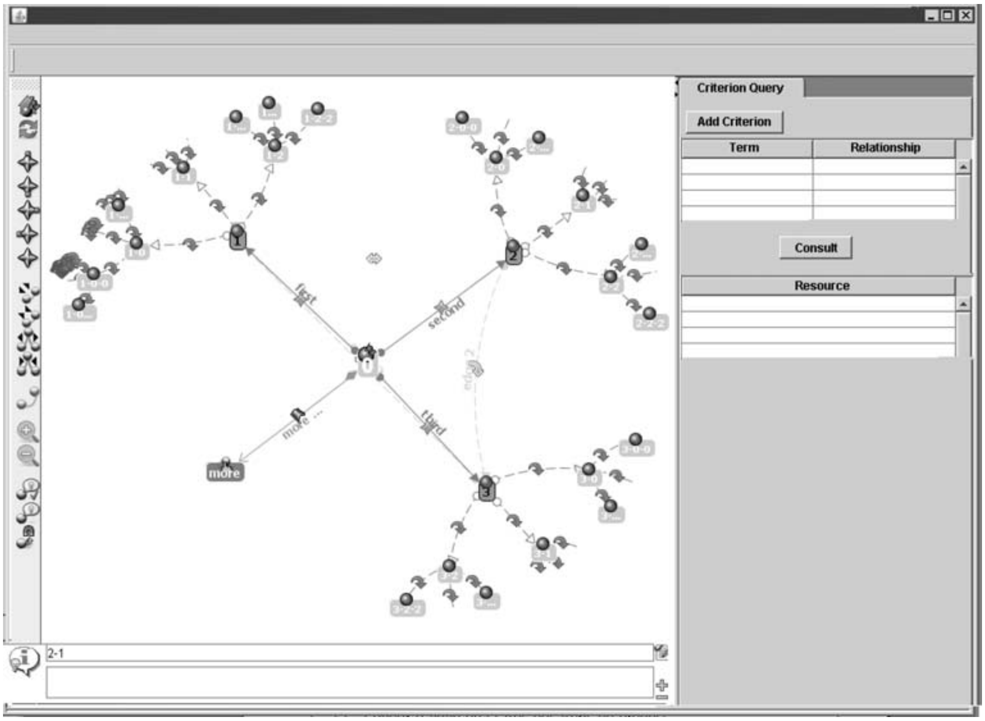


Fig. 13 Search interface of the query module

to SDI in the Amazon context, such as the Amazon Protection System (SIPAM[7]). The SIPAM base incorporates updated information on the Brazilian Legal Amazon. The use of this information in projects developed by SIPAM and third party agencies generates knowledge that assists in the articulation, planning and coordination of government global actions, seeking the protection, social inclusion and sustainable development of the region.

Acknowledgments This project was partially financed by the Research Support Foundation of the Minas Gerais State (FAPEMIG) and by the Brazilian National Research Council (CNPq / MCT / CT-Info).

References

Aditya T (2007) The national atlas as a metaphor for improved use of a national geospatial data infrastructure. PhD thesis, Enschede: ITC

Aditya T, Kraak M-J (2007) A search interface for an SDI: implementation and evaluation of metadata visualization strategies. Trans GIS 11(3):413–435

Ahmed K (2009) Topic map design patterns for information architecture. XML 2003, http://www.techquila.com/tmsinia.html. Accessed March 2009

Athanasis N, Kalabokidis K, Vaitis M, Soulakellis N (2008) Towards a semantics-based approach in the development of geographic portals. Comput Geosci 35(2):301–308. doi:10.1016/j.cageo.2008.01.014

Berners-Lee T, Hendler J, Lassila O (2001) The semantic Web. Sci Am 284(5):34–43

Booch G, Rumbaugh J, Jacobson I (2005) The unified modeling language user guide, Addison Wesley Longman Publishing Co., Inc., Redwood City, CA

Burke MA (1999) Organization of multimedia resources: principle and practice of information retrieval. Gower, Aldershot

Burrough PA, McDonnell RA (1998) Principles of geographical information systems. Oxford University Press, Oxford, p 333p

Card SK, Mackinlay JD, Shneiderman B (1999) Readings in information visualization: using vision to think. Kaufmann, San Francisco 686p

Davenport TH, Prusak L (1998) Conhecimento empresarial: como as organizações gerenciam seu capital intelectual. Campus, Rio de Janeiro (in Portuguese)

Davis Jr CA, Alves LL (2005) Local Spatial Data Infrastructures based on a service-oriented architecture. In Proc. of Brazilian Symposium on Geoinformatics, 30–45

Dodebei VLD (2002) Tesauro: linguagem de representação da memória documentária. Niterói: Intertexto; Rio de Janeiro: Interciência (in Portuguese)

Egenhofer M (2002) Toward the semantic geospatial web. Tenth ACM international symposium on Advances in geographic information systems. McLean, Virginia, ACM

Ferneda E (2003) Recuperação da informação: análise sobre a contribuição da ciência da computação para a ciência da informação. USP, São Paulo, 147p. Tese (Ciências da Comunicação); Escola de Comunicação e Arte da Universidade de São Paulo. (in Portuguese)

FGDC. Federal Geographic Data Committee. FGDC-STD-001-1998 (1998) Content standard for digital geospatial metadata. Federal Geographic Data Committee, Washington 78p

Garshol L (2002) What are topic maps? http://www.xml.com/pub/a/2002/09/11/topicmaps.html. Accessed March 2009

Garshol LM (2004) Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. J Inf Sci 30(4):378–391

Grand BL, Soto M (2002) Visualization of the semantic web: topic maps visualization. Sixth International Conference on Information Visualization (IV'02), pp 344–349

Gruber TR (1995) Towards principles for the design of ontologies used for knowledge sharing. Int J Human-Comput Stud 43(5–6):907–928

Guarino N (1998) Formal ontology and information systems. In Proc. of the First Int. Conference on Formal Ontology in Information Systems, Trento, Italy

Guptill SC, Morrisson JL (eds) (1997) Elements of spatial data quality. ICA Commission on Spatial Data Quality, Pergamon 197p

Hochmair H (2005) Ontology matching for spatial data retrieval from Internet portals. In Proceedings of Geospatial Semantics. Lecture Notes in Computer Science 3799, Mexico City, Mexico, pp 166–182

ISO 13250 (2002) Topic maps. Second Edition. International Organization for Standardization (ISO)

ISO 2788 (1986) Guidelines for the establishment and development of monolingual thesauri, International Organization for Standardization (ISO)

ISO 5964 (1985) Guidelines for the establishment and development of multilingual thesauri, International Organization for Standardization (ISO)

Jansen BJ, Spink A, Saracevic T (2000) Real life, real users, and real needs: a study and analysis of user queries on the web. Inf Process Manag 36(2):207–227

Lamping J, Rao R, Pirolli P (1995) A focus+context technique based on hyperbolic geometry for visualizing large hierarchies, human factors in computing systems, CHI '95 Conference Proceedings, ACM, pp 401–408

Maedche A, Staab S (2001) Ontology learning for the semantic web. IEEE Intelligent Systems March/April (2001), pp. 72–79

Maguire DJ, Longley PA (2005) The emergence of geoportals and their role in Spatial Data Infrastructures. Comput Environ Urban Syst 29:3–14

Miguel JL (2009) Contributions to the problem of knowledge management in Spatial Data Infrastructures. Zaragoza. Phd Dissertation; Computer Science and Systems Engineering Department of University of Zaragoza

Nebert DD (ed) (2004) Developing Spatial Data Infrastructures: the SDI cookbook, Version 2.0 (GSDI-Technical Working Group)

Nogueras-Iso J, Zarazaga-Soria FJ, Muro-Medrano PR (2005) Geographic information metadata for Spatial Data Infrastructures—resources, interoperability and information retrieval, Springer

OECD. Organisation for Economic Cooperation and Development (1996) Pollution prevention and control: environmental criteria for sustainable transport. Organisation for economic cooperation and development, Paris

Pepper S (2000) The TAO of topic maps—finding the way in the age of infoglut. Ontopia. http://www.ontopia.net/topicmaps/materials/tao.html. Accessed March 2009

Rath HH (2003) The topic maps handbook. Empolis, Gütersloh

Resende M, Curi N, Rezende SBD, Corrêa GF (1995) Pedologia: base para distinção de ambientes. NEPUT, Viçosa 304p. (in Portuguese)

Savoy J, Picard J (1998) Retrieval effectiveness on the web. Inf Process Manag 37(4):543–569

Sigel A (2000) Towards knowledge organization with topic maps. Proceedings of XML Europe 2000. Graphic Communications Association, Alexandria, pp 603–611

SKOS. Simple Knowledge Organization System. Julho, 2009. Disponível em: http://www.w3c.org/2004/02/skos/

Welty C, Guarino N (2001) Supporting ontological analysis of taxonomic relationships. Data Knowl Eng 39:51–74

[7] http://www.sipam.gov.br/