JISTEM Revista de Gestão da Tecnologia e Sistemas de Informação
Journal of Information Systems and Technology Management
Vol. 8, No. 1, 2011, *p. 73-86*
ISSN online: 1807-1775
DOI: 10.4301/S1807-17752011000100004

DATA INFORMATION SYSTEM TO PROMOTE THE ORGANIZATION DATA OF COLLECTIONS – MODELING CONSIDERATIONS BY THE UNIFIED MODELIGN LANGUAGE (UML)

Eduardo Batista de Moraes Barbosa
Centro de Previsão de Tempo e Estudos Climáticos - CPTEC
Instituto Nacional de Pesquisas Espaciais - INPE
Galeno José de Sena
Universidade Estadual Paulista - UNESP

ABSTRACT

It can be argued that technological developments (e.g., measuring instruments like software, satellite and computers, as well as, the cheapening of storage media) allow organizations to produce and acquire a great amount of data in a short time. Due to the data volume, research organizations become potentially vulnerable to the information explosion impacts. An adopted solution is the use of information system tools to assist data documentation, retrieval and analysis. In the scientific scope, these tools are developed to store different metadata (data about data) patterns. During the development process of these tools, the adoption of standards such as the Unified Modeling Language (UML) stands out, whose diagrams assist the different scopes of software modeling. The objective of this study is to present an information system tool that assists organizations in the data documentation through the use of metadata and that highlights the software modeling process, through the UML. The Standard for Digital Geospatial Metadata will be approached, widely used to the dataset cataloging by scientific organizations around the world, and the dynamic and static UML diagrams like use cases, sequence and classes. The development of the information system tools can be a way to promote the scientific data organization and dissemination. However, the modeling process requires special attention during the development of interfaces that will stimulate the use of the information system tools.

Keywords: *Information systems; Scientific data; Data dissemination; Modeling process; UML*

Recebido em/*Manuscript first received*: 23/06/2008 *Revised version received:* 31/12/2010 Aprovado em/*Manuscript accepted*: 10/02/2011

Endereço para correspondência/ *Address for correspondence*

Eduardo Batista de Moraes Barbosa, Centro de Previsão de Tempo e Estudos Climáticos – CPTEC, Instituto Nacional de Pesquisas Espaciais – INPE Rod. Presidente Dutra, Km. 39 - Cachoeira Paulista - SP - 12630-000 E-mail: eduardo.barbosa@cptec.inpe.br

Galeno José de Sena, Universidade Estadual Paulista – UNESP, Faculdade de Engenharia de Guaratinguetá – FEG Av. Dr. Ariberto Pereira da Cunha, 333 - Guaratinguetá - SP - 12516-410 E-mail: gsena@feg.unesp.br

ISSN online: 1807-1775
Publicado por/*Published by*: TECSI FEA USP – 2011

RESUMO

Pode-se afirmar que a evolução tecnológica (desenvolvimento de novos instrumentos de medição como, softwares, satélites e computadores, bem como, o barateamento das mídias de armazenamento) permite às Organizações produzirem e adquirirem grande quantidade de dados em curto espaço de tempo. Devido ao volume de dados, Organizações de pesquisa se tornam potencialmente vulneráveis aos impactos da explosão de informações. Uma solução adotada por algumas Organizações é a utilização de ferramentas de sistemas de informação para auxiliar na documentação, recuperação e análise dos dados. No âmbito científico, essas ferramentas são desenvolvidas para armazenar diferentes padrões de metadados (dados sobre dados). Durante o processo de desenvolvimento destas ferramentas, destaca-se a adoção de padrões como a Linguagem Unificada de Modelagem (UML, do Inglês Unified Modeling Language), cujos diagramas auxiliam na modelagem de diferentes aspectos do software. O objetivo deste estudo é apresentar uma ferramenta de sistemas de informação para auxiliar na documentação dos dados das Organizações por meio de metadados e destacar o processo de modelagem de software, por meio da UML. Será abordado o Padrão de Metadados Digitais Geoespaciais, amplamente utilizado na catalogação de dados por Organizações científicas de todo mundo, e os diagramas dinâmicos e estáticos da UML como casos de uso, sequências e classes. O desenvolvimento das ferramentas de sistemas de informação pode ser uma forma de promover a organização e a divulgação de dados científicos. No entanto, o processo de modelagem requer especial atenção para o desenvolvimento de interfaces que estimularão o uso das ferramentas de sistemas de informação.

Palavras-chave: *Sistemas de informação; Dados científicos; Disseminação de Dados; Processo de modelagem; UML*

1. INTRODUCTION

In the last decades, the technological development of new materials and advanced computing devices have allowed universities and research centers to produce and acquire great amount of data in a short space of time. In general, the growth of the database occurs from the aggregation of new data to the system, as well as of the data resulting from the generation of analyses and/or the maintenance of what already exists in database.

Data should be considered as the main assets for research organizations and therefore they must be easily accessed by their users. The necessity of locating and accessing specific data inside great datasets is common, which makes data collection documentation and organization releveant. As a consequence, organizations that do not document their data are subject to the overlapping of efforts in data collecting and maintaining, as well as vulnerable to problems like inconsistencies. 12 55

A solution that has been adopted by some organizations is the development of information system (IS) tools to assist users in the datasets documentation, location and analysis. In the scientific scope, the IS tools contain descriptive information about the dataset. The development of these data IS (DIS) has obtained success in the scientific information management, due to the easiness in allowing the users to analyze the data without the necessity of acquiring them. As examples of very successful implementations that use these concepts, one can mention the Inter-American Institute for Global Change Research with the IAI-DIS (URL:

https://iaibr3.iai.int/twiki/bin/view/DIS), the Large Scale Biosphere-Atmosphere Experiment in Amazonia in the LBA-DIS (URL: http://lba.cptec.inpe.br/beija-flor) and the Geonetwork tool used by the Brazilian Institute of Geography and Statistics as the data access interface (URL: http://www.metadados.geo.ibge.gov.br/geonetwork).

The objective of this study is to present an information system tool that assists organizations in the data documentation through the use of metadata (data about data) and to make some considerations about the IS modeling process, according to the Unified Modeling Language (UML). The Standard for Digital Geospatial Metadata, widely used in the data cataloging by scientific organizations around the world, will be approached. The use of metadata standards can prenvent the same data from being described in different ways by organizations, what could vary widely from one to another. In this sense the metadata technology makes it possible for the conception of an interface between data producers and users, allowing for the common agreement about the data (Barbosa & Sena, 2008). The process of the IS tool modeling will be highlighted in this work through the use of the UML. The UML, standard by the Object Management Group (OMG, 2003), uses a standardized graphical notation to create abstract models of a system, through several types of diagrams that can increase the understanding of an application under development.

2. INFORMATION SYSTEMS OVERVIEW

The data volume generated by organizations is high and therefore they should seek for efficient means for data management, in order to prevent the loss or repurchase of data already owned.

The database technology is described as one of the fields which is rapidly developing in the computer and IS areas (Elmasri, 2003). Database systems have been expanded to several areas and specific features were added to improve information dissemination, mainly through the Internet.

In the scientific scope, the IS should store information – using database technology – to assist the dataset description (content, scope, geographical coordinates, quality, etc.) and to allow a preliminary data analysis (Barbosa & Sena, 2006; Callahan & Johnson, 1995). That is, its primary feature is to enable users to make preliminary data analysis, without the need to acquire them. In this way, the IS should promote and assist the data collection organization process and encourage its dissemination among organizations. As a result, the wide information dissemination should minimize duplication efforts and promote dataset knowledge in research organizations.

2.1 METADATA STANDARDS

Metadata, the layer of data abstraction (Shankaranarayanan & Even, 2006), is among the most enigmatic elements in IS, since people have vague and sometimes conflicting views of its role and value. However, recent studies (Even et. al, 2006; Fisher, et. al, 2003) have explored the effects of metadata on decision-making efficiency,

and pointed that its use may promote the data quality evaluation both objectively and subjectively. Moreover, the use of metadata is pointed by the scientific community as an efficient solution to the information description (Moura & Campos, 2002).

In the earth sciences scope, data can take different representations – e.g., numerical measurement from instruments (balloons, radars, sensors) and/or from simulations (mathematical models for weather and climate forecasting). In that context, metadata must contain information to describe the content, quality, condition, geographic coordinates and others characteristics to identify the data (Callahan & Johnson, 1996; Hart & Phillips, 1998).

There are, actually, several metadata standards in order to meet the needs of resources description. The Dublin Core Metadata Initiative (DCMI) (URL: http://dublincore.org/) contains a set of terms to describe the electronic resources from the Internet. Its original intent was to provide descriptions for networked resources, but it may be applied to other resources, depending on how closely their metadata resembles typical Internet documents and on what purpose the metadata is intended to serve. An example of specific metadata standard is the Government Information Locator Service (GILS) (URL: http://www.gils.net), whose purpose is to identify and describe information resources throughout the Federal government and to provide assistance in obtaining the information. Another example is the Machine-Readable Cataloguing (MARC 21) (URL: http://www.loc.gov/marc), used for storing and exchanging bibliographic records and related information in machine-readable form. In the scientific scope, the Content Standard for Digital Geospatial Metadata (GEO), provided by the Federal Geographic Data Committee (FGDC) (URL: http://www.fgdc.gov/metadata), is quite complete and specific for digital geospatial data description.

2.2 THE CONTENT STANDARD FOR DIGITAL GEOSPATIAL METADATA

The Content Standard for Digital Geospatial Metadata (GEO) was developed jointly by the FGDC and the American Society for Testing Materials (ASTM). Its purpose is to supply elements for digital spatial data information. These elements should specify labels that would be used by geo-processing programs, aiming to facilitate the search, the attainment of results and the presentation of geospatial data.

The development of GEO was initiated by the ASTM in 1990 and, in 1992, the FGDC joined its development. In 1994, GEO was approved by the FGDC as a standard for data documentation in the United States. The version developed by the ASTM was incorporated into the GEO through the alphanumeric and numerical markers that must be used in the data research and presentation (Nelbert, 2000).

The GEO has 334 elements that, in some cases, are inherited from other standards (GILS and MARC 21). However, it has its own set of elements that cannot be mapped by other standards. The elements numbered between 1 and 1999 have been inherited from MARC 21, elements between 2000 and 2999 have been inherited from GILS and the several other elements, numbered between 3000 and 3999, are GEO specific. The elements in the standard are classified into three categories, namely: (i) relation elements, to allow the relationship between the searched term and its position in

the metadata; (ii) structure elements, to specify the metadata that will be searched; and (iii) truncation elements, used to truncate the words in the text.

The structure of GEO is divided into seven groups, where only the first group (Identification Information) and the last one (Metadata Reference Information) are essential for a minimum metadata description. The objective of each group is summarized below.

1. Identification Information

This group contains the basic goal-information on the dataset as, for example:

- Textual description;
- Time period information;
- Spatial reference;
- Keywords;
- Point of contacts (person and organization); and
- Access restrictions.

2. Data Quality Information

This group contains general information on the dataset quality.

3. Spatial Data Organization Information

This group contains information about what mechanisms that have been used to represent the dataset spatial information.

4. Spatial Reference Information

This group contains information on the system projection coordinates.

5. Entity Attribute Information

This group allows the user to describe the dataset information.

6. Distribution Information

This group contains information on options for data supply. The supplier corresponds to the same point of contact (person/organization) listed in the Identification Information group. Some information includes the ways to access the dataset (e.g., ftp, email, etc.).

7. Metadata Reference Information

This group contains information on the last metadata update.

3. THE UNIFIED MODELING LANGUAGE – UML

The UML is being extensively used in software projects, due to the facilities to represent diverse aspects from the projects. The UML is defined as a language for the specification, construction, visualization and documentation of the system artifacts (Booch *et al.*, 2006). The language includes thirteen diagrams divided into three categories: (i) *behavior diagrams* (which depict behavioral features of a system or business process), (ii) *interaction diagrams* (which represent different aspects of the

system's interactions) and (iii) *structure diagrams* (which show the static structure of the system being modeled) (OMG, 2009) (URL: http://www.omg.org). Each diagram provides a partial representation of the system, semantically correct, with the aim of assisting the understanding of its architecture.

Generally, the UML diagrams are used to represent the system characteristics (dynamic and static) and to provide the communication between different people involved in the development process.

In the next few sections, the modeling process will be presented using the three diagrams categories (behavior, interaction and structural) considered during the implementation of the Data Information System (DIS), as well as an example of the IS application.

3.1 FUNCTIONAL REQUERIMENTS MODELING

The functional requirements of the DIS will be presented through a use-case (UC) diagram from the UML (Fig. 1). The goal of a use-case diagram is to describe *what* the system does from the standpoint of an external observer. UC diagrams can be considered as the basic concepts for specifying functional requirements of the IS (Langlands & Edwards, 2008). In summary, they can be used to improve the high-level communication of the functions of a system as well as of its scope (Bell, 2003) during the development phase.

UC diagrams are composed by actors (who will interact with the system) and scenarios (examples of what happens when an actor interacts with the system), connected by relationships. Initially, it has been identified the following functions, or UC, for the DIS: metadata maintenance (insert, update and delete) and metadata search.

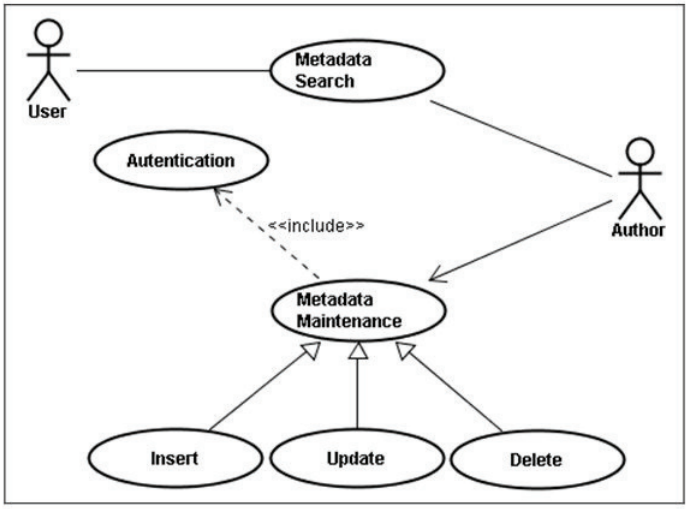

Figure 1 – DIS use-cases

To use the system (Fig. 1), an actor should be either a researcher, who carries out tasks related to the metadata maintenance and search, or a simple user, who may make only metadata searches. However, the metadata maintenance tasks can be carried out only by researchers previously registered in the DIS authentication system.

Table 1 – Use case (metadata search) specification

NAME:	Metadata search
DESCRIPTION:	**Search for dataset previously registered in the dis**
Actor (s):	Simple user or Author
PRIMARY SCENARIO	
1. **The system presents a form with optional fields to be filled**	
2. **The user should fill at least one form field and identify its localization in the text**	
3. **The user can combine the filling of more than one field of a form**	
4. **The user clicks on the search button (*submit*)**	
5. **The system sends the query**	
6. **The user gets the response to the query in a friendly format**	
7. **The user can access the metadata details by clicking on the link details**	
ALTERNATIVE SCENARIO	
5a. **Before sending the query the system checks whether the user has filled at least one field. If not (none filled), the consultation should not be sent and a warning page should be presented**	

The UC, usually, must be followed by the *descriptive scenario* (Table 1), whose goal is to specify the steps for the UC completion. The scenarios should contain fields to provide better understanding of a UC, such as its name, purpose, actors (who will activate the process) and steps (primary and alternative scenarios). In this way, each UC (metadata maintenance and search) should contain a descriptive scenario associated with it. Table 1 shows the steps needed for any user (simple user or author) to make a metadata search. These steps can be graphically represented by means of the UML *interaction diagrams* (Fig. 2).

The interaction diagrams are used to present the interactions between system objects. These diagrams are classified as sequence, communication, temporal and overview, and share the common properties of the other system objects, like name and

content.

Sequence diagrams (Fig. 2) are based on a bi-dimensional graph, where the *Y* axis includes the messages exchanged between objects in time and the *X* axis, the objects (aligned on the top). From Figure 2, one can observe the interactions between system objects to perform a metadata search, specified previously in Table 1. In this example, the Web Form is an interface between an actor (user/author) and the system classes (keyword and dataset), where the metadata can be showed in a friendly format. The messages are calls to the services (operations) provided by the system objects.

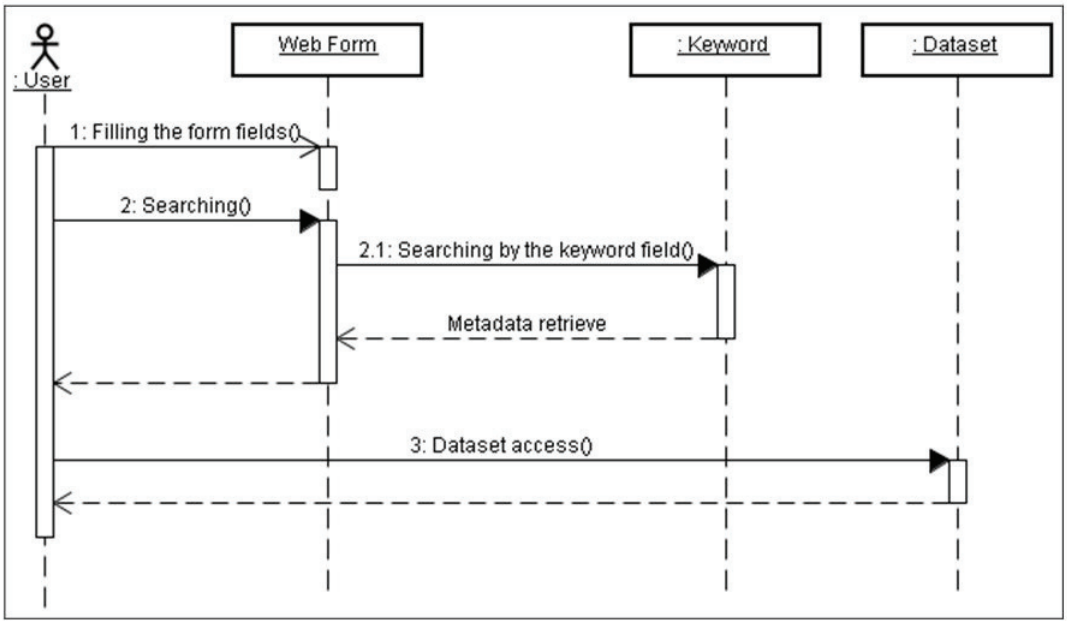

Figure 2 – DIS metadata searches

3.2 STRUCTURE MODELING

The data model elements of the DIS (Fig. 3) will be presented using class diagrams from the UML, which represent an alternative way to the entity-relationship diagram (Chen, 1976). In these diagrams, a class is presented through a box with two sections, with the object name (or entity name) included in the upper section and its attribute's name and data types shown in the lower section.

In this project, the database system is used to store the metadata and to facilitate the tasks related to their management (query, insert, update and remove). The data stored in tables, projected in compliance with the relational model restrictions (Date, 2000; Silberschatz *et al.*, 2005), guarantees both agility and flexibility to the data access.

The data model (Fig. 3) is composed of the following relations (or tables): Author, Dataset, Contact and Keyword, whose attributes, most of them, are associated with the described groups from GEO (Section 2.2). The Author table stores information about the person responsible for the metadata generation (group Identification Information). Information about the contact person for the data users is recorded in the

Contact table (group Distribution Information). Dataset contains information concerning the datasets and its attributes have been extracted from the diverse described groups from GEO. The Keyword table stores keywords used to the dataset description.

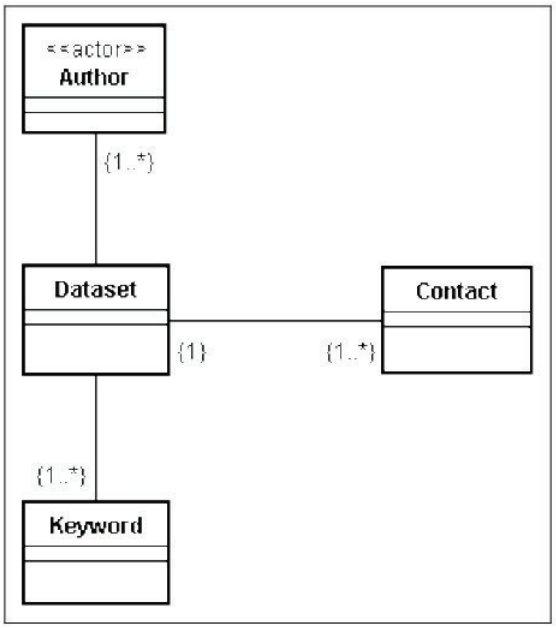

Figure 3 – DIS class diagram

4. A DIS BASED ON THE UML MODELING

10:19 This section presents some of the screens designed to query a scientific metadata DIS, whose implementation was based on the UML modeling (Section 3). For a more complete description of the DIS, we refer the reader to the works of Barbosa and Sena (Barbosa & Sena, 2008; Barbosa & Sena, 2006).

The DIS uses friendly interfaces for the accomplishment of metadata search. The standards adopted were Hypertext Markup Language (HTML) and Common Gateway Interface (CGI) that make it possible for the presentation of dynamic Web pages, generally, with the results of database queries. The database system integration with the Internet was carried through a DBI module, a generic interface from the Perl language (URL: http://www.perl.org).

To carry through metadata search, the users must define the terms that will be used, as well as, identify their location in the text, the data time variation or the area covered by the data (Fig. 4).

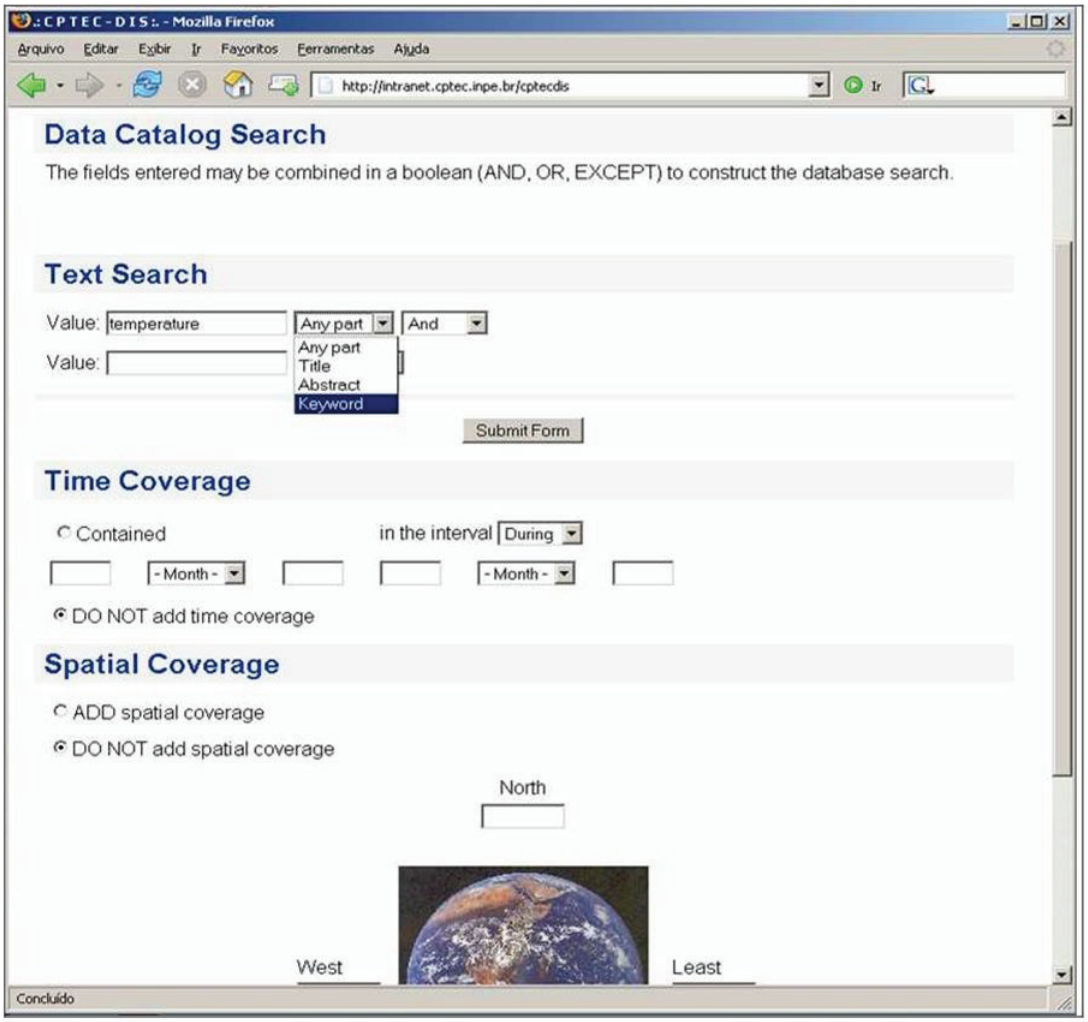

Figure 4 – DIS search form

The available options to consult terms in the text are: (i) any part, (ii) title, (iii) abstract and (iv) keyword. The users must choose one of these options using combo boxes located to the right of the page (Fig. 4). There is the possibility to combine terms using Boolean operators: "and", "or" or "except". In addition to the text search fields, one can see in the form fields for the time and the spatial coverage of the searched metadata.

The web page with the search results (Fig. 5) to the term "temperature", located in the keywords, presents initially summarized information about the data. At the beginning of the page, a table contents is showed referring to the information of the search (database name, search status and the results).

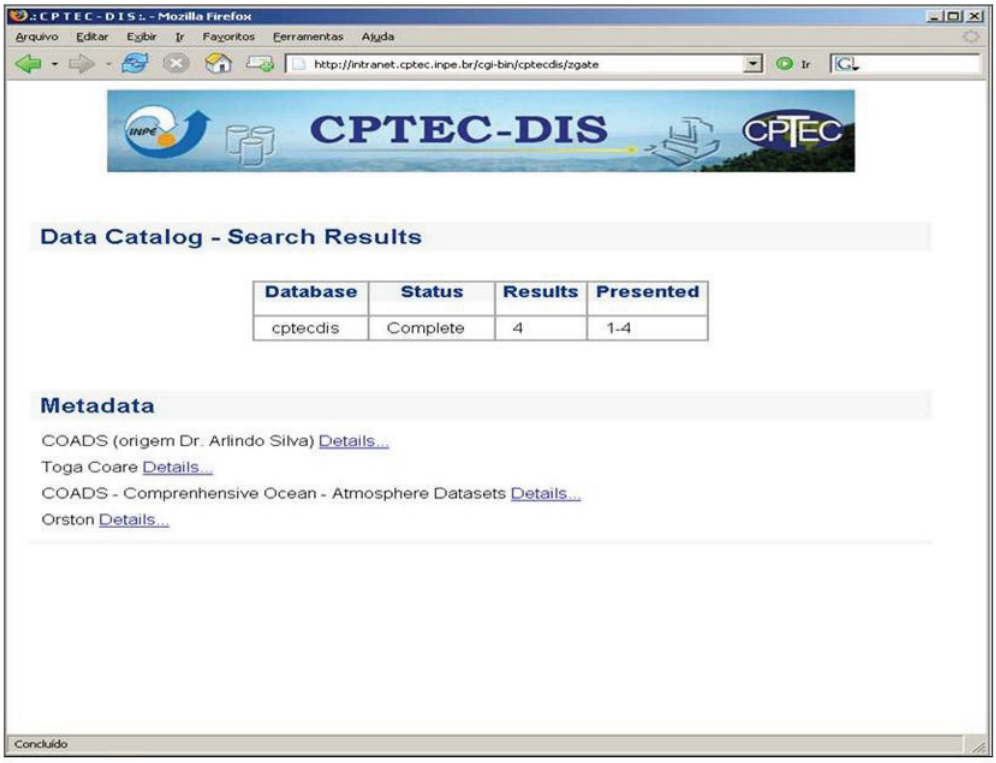

Figure 5 – Metadata search results

(a) **(b)**

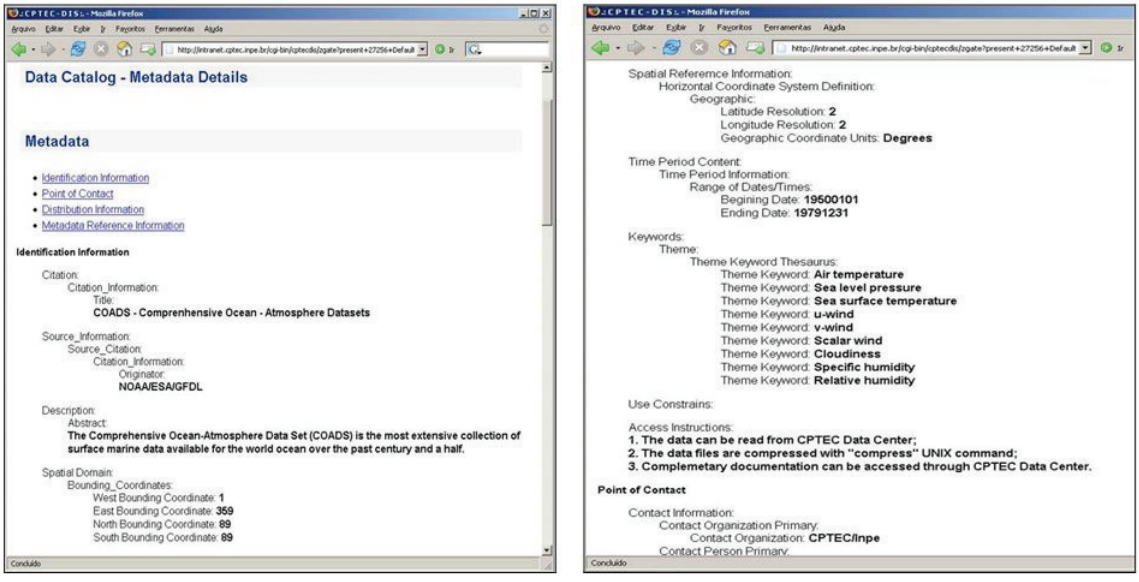

Figure 6 – Dataset details (The metadata identification information view)

Above, a list containing the metadata title is showed, followed by the option

"Details…" (web link format), that makes it possible the metadata access in its complete form, for a detailed analysis (Fig. 6).

The screens above presented, which structure a query (Fig. 4) and metadata access (Fig. 5 and 6), were developed in accordance with the UML model, in the following sense: the UC diagram (Fig. 1) and the descriptive scenario (Table 1) support the specifications of the steps to the query, while the sequence diagram (Fig. 2) defines the sequence of objects that should be accessed during a metadata query.

5. FINAL CONSIDERATIONS

The metadata standards may be used to create a common grammar for the data description. This conceptualization must provide a potential integration between organizations in order to promote the interchange for the scientific data exchange. This paper presented a DIS intended to query scientific metadata based on the GEO standard. The work emphasized the system modeling (their specifications and interactions) through the use of the UML.

With the use of the UML, as a standard graphical language, it is possible to abstract the essential details of the underlying problem, in order to better characterize the IS modeling process. Yet, with respect to the IS modeling, this paper presents the three types of diagrams provided by the UML (behavior, interactions and structure) and emphasizes their use in the implementation of the system. These diagrams were used to assist the development of friendly electronic interfaces that will stimulate the DIS use. Through the interfaces, tasks such as search and data analysis, as well as metadata maintenance, can be carried through in an easy and agile way. Through the UML diagrams it is possible to develop modular systems, easily adaptable to the needs of users.

The development of the DIS by research organizations can be a way to promote and to facilitate the scientific data dissemination, to prevent the duplication of efforts in their attainment, as well as to stimulate the reuse of the collected, already processed and stored data.

It should be noted that the modeling techniques of IS described in this work can be easily adapted to the development of other systems, with other metadata standards, regardless of the platform used in the implementation of the systems.

REFERENCES

Barbosa, E.B.M. & Sena, G.J. (2008). Scientific Data Dissemination – A Data Catalogue to Assist Research Organizations. *Ciência da Informação***.** *37,* 1.

Barbosa, E.B.M.& Sena, G.J. (2006). Um Banco de Metadados para Auxiliar a Disseminação de Dados Científicos em Instituições de Pesquisas. In: *3rd CONTECSI International Conference on Information Systems and Technology Management and 11th World Continuous Auditing*, São Paulo. Anais.

Bell, D. UML basics (2008): An introduction to the Unified Modeling Language. URL: <http://www.ibm.com/developerworks/rational/library/769.html>, 2003.Boock, G., Rumbaugh, J. & Jacobson, I. (2006). *UML Guia do Usuário***.** Campus.

Callahan, S.D. & Jonhson, B.D. (1995) Scientific Data Set Catalogues. In: *Proceedings of Second AGSO Forum on GIS in the Geosciences*, Canberra, (29-31).

Callahan, S.D. & Jonhson, B.D. (1996). Dataset Publishing – A Means to Motivate Metadata Entry. In: *First IEEE Metadata Conference*, Silver Springs, Maryland.

Chen, P.P. (1976). The Entity-Relationship Model – Toward a Unified View of Data. *ACM Transactions on Database Systems*, (9-36)

Date, C.J. (2000). *Introduction to Database Systems,*. 7th. Ed. Addison Wesley Professional.

Elmasri, R., Navathe (2003), S.B. Fundamentals of Database Systems. 4th Ed. Addison Wesley.

Even, A., Shankaranarayanan, G. & Watts, S. (2006). Enhancing decision making with process metadata: Theoretical framework, research tool, and exploratory examination. In: *Proceedings of the 39th Hawaii International Conference on System Sciences (HICSS-39)*. IEEE Press, Los Alamitos, CA.

Fisher, C., Chengalur-Smith, I., & Ballou, D. (2003). The impact of experience and time on the use of data quality information in decision making. *Information Systems Research. 14*, 2, (170-188).

Hart, D., Phillips, H. Metadata Primer – How to Guide on Metadata. Implementation. URL: http://www.lic.wisc.edu/metadata/metaprim.htm ,1998. Accessed: Jan/2007. Langlands M., Edwards, C. Business vs System Use Cases – Part One.

URL: http://www.requirementsnetwork.com , 2008. Accessed: May/2008.

Moura, A.M.C. & Campos, M.L.M. A Metadata Approach to Manage and Organize Electronic Documents and Collections on the Web. *Journal of the Brazilian Computer Society, 1*, 8, (16-31).

Nebert, D.D. Z39.50 Application Profile for Geospatial Metadata, 2.2

URL: <http:www.fgdc.gov/standards/projects/GeoProfile>, 2000. Accessed: Jan/2007.

OMG, Object Management Group, Inc. Introduction To OMG's Unified Modeling Language (UML). URL: <http://www.omg.org/gettingstarted/what_is_uml.htm>, 2009.

Accessed: Jun/2010.

Shankaranarayanan, G. & Even, A. (2006). The Metadata Egnima. *Communications of the ACM*, *49*, 2.

Silberschatz, A., Korth, H.F. & Sudarshan, S. (2005). Database Systems Concepts. 5th. Ed. McGraw-Hill.