Scientific data dissemination a data catalogue to assist research organizations

Eduardo Batista de Moraes Barbosa
Mestre em computação aplicada pelo Inpe de São José dos Campos.
E-mail: eduardo@cptec.inpe.br
Galeno de Sena
Doutor em ciências de informática pela Pontifícia Universidade Católica do Rio de Janeiro - PUC/RJ; pós-doutorado no National Institute of Multimedia Education, Chiba-shi, Japão.
E-mail: gsena@feg.unesp.br

Abstract

The constant scientific production in the universities and in the research centers makes these organizations produce and acquire a great amount of data in a short period of time. Due to the big quantity of data, the research organizations become potentially vulnerable to the impacts on information booms that may cause a chaos as far as information management is concerned. In this context, the development of data catalogues comes up as one possible solution to the problems such as (I) the organization and (II) the data management. In the scientific scope, the data catalogues are implemented with the standard for digital and geospatial metadata and are broadly utilized in the process of producing a catalogue of scientific information. The aim of this work is to present the characteristics of access and storage of metadata in databank systems in order to improve the description and dissemination of scientific data. Relevant aspects will be considered and they should be analyzed during the stage of planning, once they can determine the success of implementation. The use of data catalogues by research organizations may be a way to promote and facilitate the dissemination of scientific data, avoid the repetition of efforts while being executed, as well as incentivate the use of collected, processed an also stored data.

Keywords

Data Catalogue. Data dissemination. Scientific metadata. Z39.50.

Disseminação de dados científicos - um catálogo de dados para ajudar as organizações de pesquisa

Resumo

A constante produção científica nas universidades e nos centros de pesquisa faz com que essas organizações produzam e adquiram grande quantidade de dados em um curto período de tempo. Devido ao volume de dados, as organizações de pesquisa se tornam potencialmente vulneráveis aos impactos da explosão de informação, que pode causar um caos na gestão da informação. Nesse contexto, o desenvolvimento de catálogos de dados aparece como uma solução para os problemas tais como (I) a organização e (II) a gestão de dados. No âmbito científico, os catálogos de dados são implementados com o padrão para metadados digitais geoespaciais, amplamente utilizado na catalogação da informação científica. O objetivo desse trabalho é apresentar as características de acesso e armazenamento de metadados em sistemas de banco de dados para melhorar a descrição e a disseminação de dados científicos. Serão abordados aspectos relevantes que devem ser considerados durante a fase de planejamento, uma vez que eles podem determinar o êxito da implementação. O uso dos catálogos de dados por organizações de pesquisa pode ser uma maneira de promover e facilitar a disseminação de dados científicos, evitar a duplicação de esforços em sua consecução, bem como estimular o uso dos dados coletados, processados e armazenados.

Palavras-chave

Catálogos de dados. Disseminação de dados. Metadados científicos. Z39.50.

INTRODUCTION

To large extent of the research Organizations the production and the acquisition of scientific data are considered delayed and expensive tasks, due the financial investments and the time demand. However, the reuse of data in researches is rare, due to lack of appropriates documentation and dissemination of what already was produced.

The constant scientific production in the universities and research centers makes these organizations produces and acquires great amount of data in short time. The growth of the database occurs in such way from the aggregation of new data to the system, as for the generation of analyses and/or maintenance of what already it exists.

For research organizations, data are the basic resources and therefore they must be of easy access to the users. In general, the necessity of localization and access to the specific data inside of great datasets is common, becoming relevant the documentation and organization of the databases.

A solution that comes being adopted by some organizations is the data catalogues that assists the users in the datasets localization and analysis (CALLAHAN AND JOHNSON, 1995). Data catalogues contain descriptive information about data as, for example, its content and quality. The development of these systems has gotten success in the scientific information management, due the easiness in allowing to the users the analysis of data without the necessity of acquire it.

The emphasis of this paper is in the use of metadata (data about the data) standard to make possible a common terminology to the data description. The purpose of using standards is to prevent that the same data is described in different ways by the organizations, what it could vary widely for another one. Thus, the metadata makes possible an interface between the data producer and users, becoming possible the common agreement of the data.

The objective of this study is to present the characteristics of tools for access and storage scientific metadata in database systems. These tools are widely used by organizations that work with scientific data, therefore allow its dissemination,

preventing the duplication in its attainment and making possible the knowledge of the data in the organization. The approach will include a discussion on metadata in scientific context and will present the Standard for Digital Geospatial Metadata, widely used in the scientific information cataloging. Following, will be presented some aspects that must be considered during the planning phase of the data catalogues development.

METADATA IN THE SCIENTIFIC CONTEXTO

The use of metadata is pointed by the scientific community as an efficient solution to the information description (Moura and Campos, 2002). Metadata can be defined simply as data about data, however this is not a consensus and others definitions can be found (IKEMATU, 2001).

In the scientific context, metadata contains information to describe the content, the quality, the condition and others characteristics of data (CALLAHAN and JOHNSON, 1996; HART and PHILLIPS, 1998).

There are different metadata standards with the varieties purposes to the necessities of resources description. The Dublin Core Metadata Initiative (DCMI) contains a set of expressions for electronic resources description from the Internet. However, in the scientific context the elements supplied by the DCMI are considered limited, because they can not cover specific characteristics from scientific data. An example of metadata standard with specific purpose is the Government Information Locator Service (GILS), whose purpose is to catalogue (United States) governmental information. An other example is the digital libraries standard Bibliographic-1 (BIB-1) used for to register bibliographic data knowledge. In the scientific context, the standard supplied by Federal Geographic Data Committee (FGDC) is sufficiently complete and specific.

THE STANDARD FOR DIGITAL GEOSPATIAL METADATA

The Standard for Digital Geospatial Metadata (GEO) was the joint developed 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 will 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 to it development. In 1994, GEO was approved by the FGDC as standard for data documentation in the United States. The version developed by the ASTM was incorporated to 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 others standards (GILS and BIB-1). However, it has its proper set of elements that cannot be mapped for others standards. The elements numbered between 1 and 1999 had been inherited from BIB-1, elements between 2000 and 2999 had been inherited from GILS and the too much elements, numbered between 3000 and 3999, are specific of GEO. The standard has three classrooms of appointed elements, that is, (i) relation elements, to allow the relationship between the searched term and its position in the metadata, (ii) structure elements, to specify that it has left of the metadata will be searched and (iii) truncation elements, used to truncate the words in the text.

The structure of GEO is divided in seven groups, where only the first (Identification Information) and the last one (Metadata Reference Information) are needed. The objective of each one is presented in summary as to follow.

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);
   - Restrictions access.

2. Data Quality Information
   This group contains general information on dataset quality.

3. Spatial Data Organization Information
   This group contains information of which mechanisms had 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 user to describe the dataset information.

6. Distribution Information
   This group contains information on options to data supply. The supplier corresponds to the point of contact (person/organization) listed in the Identification Information

group. Some information includes the ways of access to the dataset as, for example, ftp, email, etc.

7. Metadata Reference Information
   This group contains information on the last metadata update.

DATA CATALOGUE OVERVIEW

PLANNING PHASE

Data catalogues (DC) are defined as systems to describe datasets and to indicate its localization for use. The key factor that favors its use is to make possible the users to determine the data relevance and the quality, without acquire it for detailed analysis. However, the effective use of the DC depends strongly on the way as the information description. Callahan and the Johnson (1995) list six key factors that must be considered during the planning process:

- Completeness;
- Easiness;
- Coherence;
- Precision;
- Availability;
- Publicity.

The DC will be strongly used if the organizations really register all datasets of its databases. For example, if queries fail, the users will be certain that the dataset do not exists in the organization. However, if the DC will be incomplete, the users never will be certain if the datasets exists or not.

One of the objectives of these systems is to make possible fast access and easy localization to the datasets. Therefore it does not have the necessity of promoting extensive training to use.

Enough information must be considered to allow the users determine if the datasets will be used. However, determine which information must be firstly considered is a difficult task. Each dataset has different types of information and, consequently, the DC must be flexible to this variation. The content of the DC must be carefully determined by data classifiers. The data utility depends on the relevance of the information that are returned by search.

Many users access DC for tests of the stored information, therefore, the DC must prevent incomplete descriptions that can cause low credibility.

The easiness of access from the computer networks must allow any user into the organization have access to the DC. The people must know that these systems exists, to understand that they must be used and to apply them in benefit of the development of its works.

THE DATA DESCRIPTION

The data description and its publication over the Internet is not only a technological question therefore, in general, it involves changes of concepts, reorganization, learning and planning in the organizations (BARBOSA and SENA, 2002 and 2006).

Normally, applications are developed with friendly interfaces, whose purpose is to minimize the efforts used to data description. Unlikely, this is the point where the efforts cease. Some reasons used to prevent the data documentation are presented to follow (CALLAHAN and JOHNSON, 1996):

- It is a tedious task;
- The data localization and the quality are already well known;
- This task keeps the scientist far from its real work (makes science);
- A phone call can facilitate the data localization;
- The recognition is small close to the time expense;
- Use of the data without credits to the "owner" of these.

As result is observed that considerable resources are expenses, but the initial interest of the researchers in participating of the process is wasted quickly. In this context, the challenge of the organizations is to motivate them to register his data and to keep it up to date.

It is too observed that the biggest investments must be directed not only in technological resources, but mainly in the human resources. These last ones will carry through tasks that hardly will be automated (metadata description). The description process requires of the responsible person for the task knowledge on some characteristics, which who was only involved in its attainment can know (BARBOSA and SENA, 2002 and 2006).

DATA CATALOGUE IMPLEMENTATION

The DC implementation was accomplished in the Center of Forecast and Climate Studies (CPTEC), from the National Institute for Spatial Research (Inpe). During this study there were different tools to data search from the Internet. We chose the Isite program due the following characteristics: (i) applications development using the GEO standard; (ii) availability of tools for the applications development through the Internet; (iii) free technical support; (iv) used by hundreds organizations and (v) accomplishment of queries in transparent way to the users (in the local and remote databases).

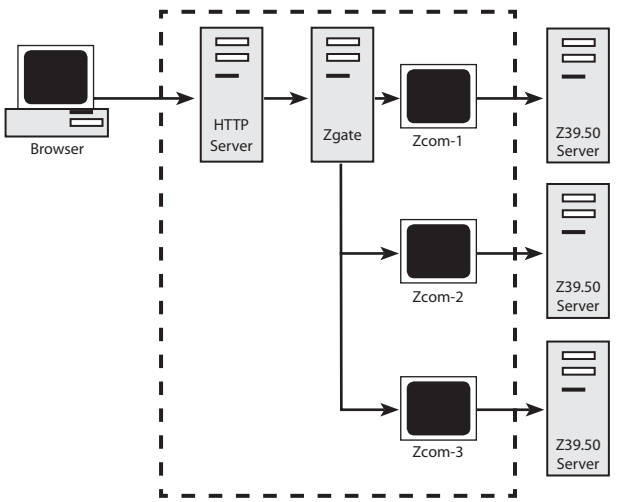
Isite can be defined as a complete information system for Internet, therefore it has the characteristics to integrate

database systems with other systems and protocols as WWW and Z39.50 (Gamiel, 1998). The package, developed by the Center for Networked Information Discovery and Retrieval (CNIDR), represents the investments results of the National Science Foundation to promote and to implement the integration tools for information recovery through the Internet.

The Isite includes a data index tool, called Iindex, that can be configured for diverse profiles as GEO, BIB-1, GILS, etc. It also includes the Z39.50 server, called Zserver, to answer queries submitted through Z39.50 protocol and a set of applications destined to the accomplishment search, called Zgate, Zcon and Isearch. Using Web pages, in a transparent way to the users, is possible to use the Isite tools to data retry in remote databases. An example of that architecture is presented in Figure 1.

Figure 1

The Isite architecture



Z39.50 PROTOCOL

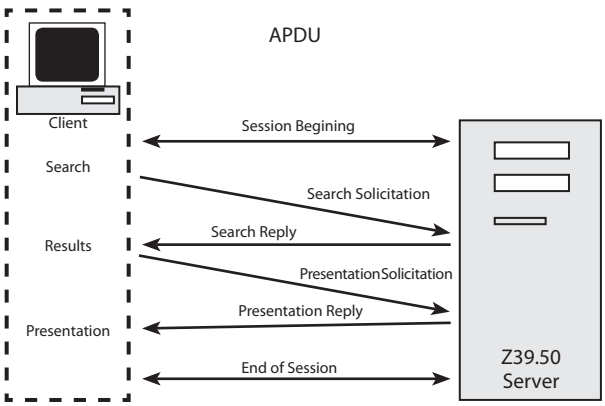
The communication protocol Z39.50, defined as Information Retrieval (Z39.50): Application Service Definition and Protocol Specification, was developed by National Information Standard Organization (NISO) committee, to specify rules and procedures for computers communication and data search in distributed way (ANSI/NISO Z39.50-1995, 1995).

The Z39.50 standard was proposed in 1984 for the specific use with bibliographical information. Its first version was approved in 1988 and, during 1989 its administration and technical development became accomplished by Library of Congress.

From the technical point of view, Z39.50 applications must be qualified to allow the data exchange between client and server computers. The client begins dialogues with the server, with the purpose to data retry according to the search conditions requested to the DC (Lynch, 1997). The basic functioning of a Z39.50 client (Figure 2) can be summarized by the session establishment (through a beginning solicitation to the server), its maintenance and the closing session, from end of session solicitation.

Figure 2

Basic schematic diagram of Z39.50



During the session accomplishment, the search solicitations, requested by clients, are sent to the server through technical messages defined as Application Protocol Data Units (APDUs) specified according to Abstract Syntax Notation One (ASN.1). After the serach solicitation, the server becomes responsible for its accomplishment until the reply to the client. The client receives the reply contends the query total results and makes solicitation for results presentation. In a transparent way to the users, the data are formatted and presented to the client.

The resources of the Z39.50 are widely used by the library community therefore they present great versatility for applications development to information dissemination. One of its advantages is to allow the DC access with the most diverse types of information (bibliographic, governmental, meteorological, etc.). Implementations using the Z39.50 resources do not have to be complex, due that offers opportunity to the information dissemination through distributed systems in transparent way to the users. Examples of libraries with on-line access to DC are easily found, mainly, in the universities and data public access systems in countries as Australia, Canada and United States.

If its development were associated with data management applications, the Z39.50 reveals very dynamic and rich in functionalities. The resources of this standard must be seen as an open door to discovery and distributed exploration of the DC. Its use is extremely advantageous for applications that have as objective data dissemination through the Internet, therefore allows the use of only an interface to facilitate the search accomplishment in such a way remote and local databases.

THE DATABASE MODELING

The Unified Modeling Language (UML), standardized by the Object Management Group (OMG) in 2003, is being extensively used in software projects, due the facilities to represent diverse aspects from these projects. The UML is defined as a language to the specification, construction, visualization and documentation of software systems (Booch et al., 2006). In general, UML diagrams are used by the developers to represent the system characteristics (static and dynamic) and to provide the communication between different people involved in the development process.

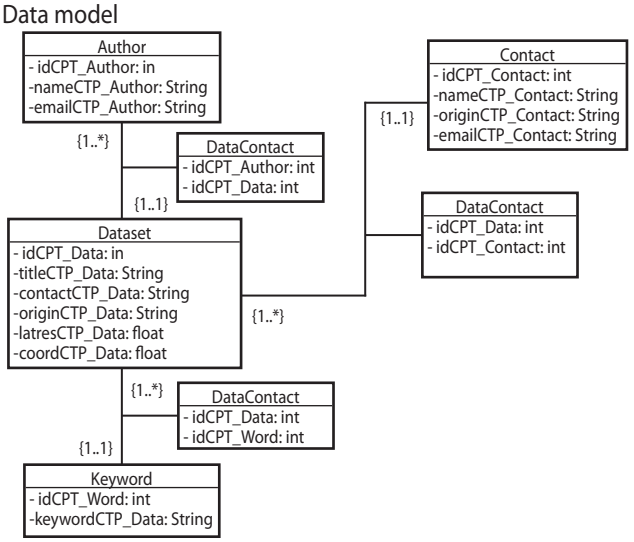
The data model elements for the DC (Figure 3) are presented using class diagrams from UML, that represents an alternative way to the entity-relationship diagram (Chen, 1976). In these diagrams, a class is presented through a box with two sections, where in the upper section is the object name (or entity name) and its attributes name and data types are showed in the lower section.

In this study, the database system is used to store and to facilitate tasks related to the metadata 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 agility and flexibility for data access.

The data model (Figure 3) is composed for the following tables: Author, Dataset, Contact and Keyword, whose attributes, in its majority, are associates to the described groups from GEO (Section 2.1). The Author table stores information to the person for the metadata generation (group Identification Information). The information about the contact person with the data users are registered in the Contact table (group Distribution Information). Dataset, contains information to the datasets and its attributes (summarized showed) having been extracted of the diverse groups described from GEO. The Keyword table stores keyword used in the content to the dataset description.

Figure 3

Data model



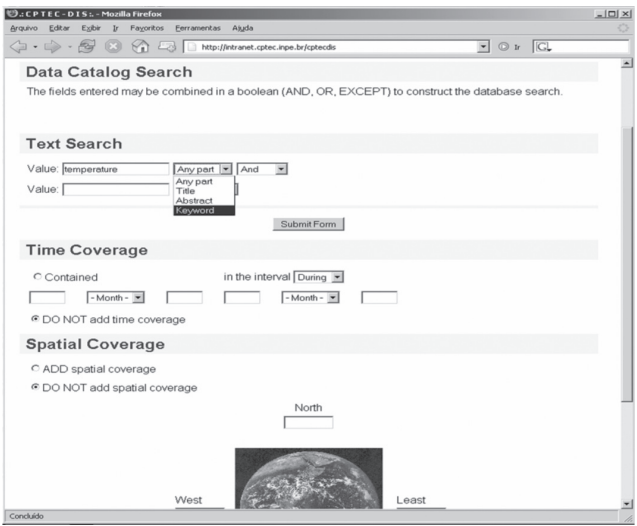
THE EXAMPLE OF THE DATA CATALOGUE APPLICATION

The application developed, called CPTEC – Data Information System (CPTEC-DIS), uses friendly interfaces for the accomplishment metadata search. The standards adopted were Hypertext Markup Language (HTML) and Common Gateway Interface (CGI) that make possible the dynamic Web pages, generally, with results of database queries. The database system integration with the Internet was carried through DBI module, a generic interface from the Perl language.

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

Figure 4

Web page showing 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 boxes located to the right of page (Figure 4). There is possibility to combine terms using boolean operators: "and", "or" or "except".

The Web page with the search results (Figures 5) to the term "temperature" located in the keywords, presents firstly the summarized information to the data. In the begin of page, is showed a table contends referring information to the search (database name, search status and the results). Below, is showed a list contends the metadata title, followed by the option "Details…" (Web link format), that makes possible the metadata access in its complete form, for a detailed analysis (Figure 6).

Figure 5

Web page showing search results

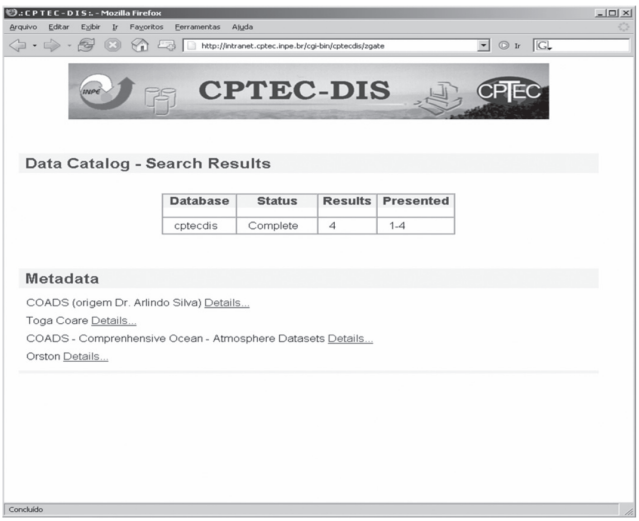
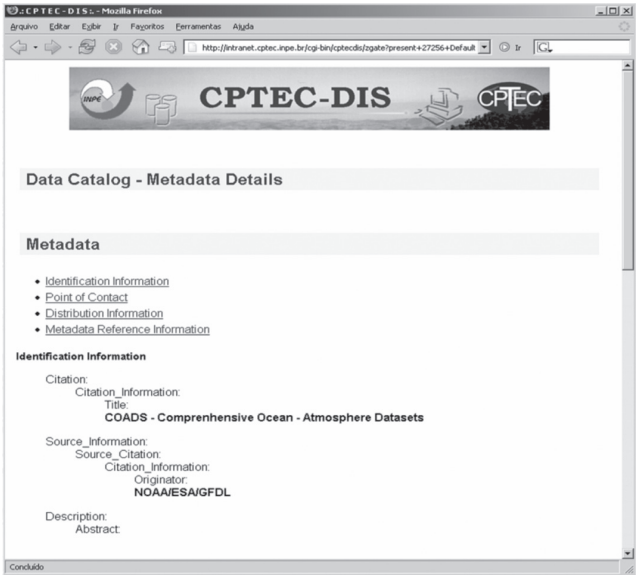


Figure 6

Web page showing metadata details



CONCLUSION

The DC has been used as solution for problems as data organization and management in the organizations. During the planning phase, the metadata description process must receive special attention. Frequently, these applications fail, due to lack of organizational culture and, mainly, research interest to participate of this process. In this context, a challenge to the organizations is to motivate people to register and keep up to date the data documentation. It is necessary the recognition from the organization to the process relevance that can determine the system success (BARBOSA and SENA, 2006).

To a large extent of the research organizations the paper publication constitutes a key information to the professional recognition, being also of relevance for promotions. The metadata publication in the DC is an important process for these organizations and, therefore, the recognition of this kind of production must receive similar consideration to the paper publication (CALLAHAN and JOHNSON, 1995 and 1996).

The GEO standard can be used to create a common grammar for the data description. This legalized conceptualization certainly will provide a potential integration between research organizations with the purpose to promote the interchange for scientific data exchange.

The development of friendly electronic interfaces should stimulate the DC use. From these interfaces, it can be carried through, in easy and agile ways, tasks as search and data analysis. The application CPTECDIS, during the initial phase, will be available only to the internal CPTEC/Inpe users through an Intranet. However, the use of dynamic Web pages has the purpose to become it available, in a close future, for access by researchers from others organizations.

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

Artigo submetido em 23/10/2007
e aceito para publicação em 22/08/2008

REFERENCES

ANSI/NISO. Information retrieval (Z39.50): application service definition and protocol specification. 1995. Disponível em: <http://lcweb.loc.gov/z3950/agency>. Acesso em: 10 jan. 2007.

BARBOSA, E. B. M. Uma ferramenta para disseminação de dados científicos do CPTEC/Inpe através de um banco de metadados. 2002. 62 f. Trabalho de Conclusão de Curso (Especialização em Informática Empresarial)- Faculdade de Engenharia de Guaratinguetá, Universidade Estadual de São Paulo, 2002.

_____; SENA, G. J. Um banco de metadados para auxiliar a disseminação de dados científicos em instituições de pesquisas. In: INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGY MANAGEMENT, 3.; WORLD CONTINUOUS AUDITING, 11., São Paulo. Proceedings... São Paulo, 2006.

BOOCK, G.; RUMBAUGH, J.; JACOBSON, I. UML: guia do usuário. [S. l.]: Campus, 2006.

CALLAHAN, S. D., JONHSON, B. D. Scientific data set catalogues. In: AGSO FORUM ON GIS IN THE GEOSCIENCES, 2., 1995, Canberra. Proceedings… Canberra, 1995. p. 29-31.

_____. Dataset publishing: a means to motivate metadata entry. In: IEEE METADATA CONFERENCE, 1., 1996, Maryland. Proceedings… Maryland: Silver Springs, 1996.

CHEN, P. P. The entity-relationship model: toward a unified view of data. ACM Transactions on Database Systems, p. 9-36, 1976.

DATE, C. J. Introduction to database systems. 7th ed. [S. l.]: Addison Wesley Professional, 2000.

GAMIEL, K. The ISite information system: version 2.00: the clearinghouse for networked information discovery and retrieval. [S. l.]: Center for Networked Information Discovery and Retrieval, 1998. No. NCR-9216963.

HART, D., PHILLIPS, H. Metadata primer: how to guide on metadata implementation. recurso. 1998. Disponível em: <http://www.lic.wisc.edu/metadata/metaprim.htm>. Acesso em: 10 jan. 2007.

IKEMATU, R. S. Gestão de metadados: sua evolução na tecnologia da informação. DataGramaZero, v. 2, n. 6, 2001.

LYNCH, C. A. The Z39.50 Information retrieval standard, Part I: a strategic view of is past, present and future. D-Lib Magazine, 1997. Disponível em: <http://www.dlib.org/ dlib/april97/04lynch.html>. Acesso em: 10 jan. 2007.

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, v. 1, n. 8, p. 16-31, 2002.

NEBERT, D. D. Z39.50: application profile for geospatial metadata. v. 2.2. 2000. Disponível em: <http://www.blueangeltech.com/standards/GeoProfile/geo22.htm>. Acesso em: 10 jan. 2007.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. Database systems concepts. 5. ed. [S. l.]: McGraw-Hill, 2005.