UM BANCO DE METADADOS PARA AUXILIAR A DISSEMINAÇÃO DE DADOS CIENTÍFICOS EM INSTITUIÇÕES DE PESQUISAS

A METADATA DATABASE TO ASSIST THE DISSEMINATION OF SCIENTIFIC DATA IN RESEARCH ORGANIZATIONS

Eduardo Batista de Moraes Barbosa (CPTEC – Instituto Nacional de Pesquisas Espaciais, São Paulo, Brasil) eduardo@cptec.inpe.br
Galeno José de Sena (FEG – Universidade Estadual Paulista "Júlio de Mesquita Filho", São Paulo, Brasil) gsena@feg.unesp.br

ABSTRACT

The continuous scientific production in universities and research centers makes these organizations produce and acquire great amount of data in short time. Due to the data volume generated, research organizations become potentially vulnerable to the impacts of the information explosion, which can cause a chaos in the information management. In this context, the development of data catalogues appears as solution to problems such as: (i) organization and (ii) management of data in the organizations. In the scientific scope, the data catalogues are implemented in conjunction with the digital standard for geographic metadata, widely used in the in the cataloging of scientific information. The objective of this paper is to present the characteristics of access and storage of scientific metadata in database systems to improve the description and the dissemination of scientific data. It will be approached important aspects that must be considered during the development phase, since they can determine the success of the implementation. From a data flow diagram, the phases considered since receiving the data until its publication and subsequent liberation for queries against the data catalogue will be illustrated. The use of data catalogues in research organizations can be a way to promote and to facilitate the dissemination of the scientific data, to prevent the duplication of efforts in its attainment, as well as, to stimulate the reuse of the data collected, processed and stored.

Keywords: Data Catalogue, Data Dissemination, Scientific Metadata, Z39.50

RESUMO

A constante produção científica em universidades e centros de pesquisa faz com que estas instituições produzam e adquiram grande quantidade de dados em curto espaço de tempo. Devido ao volume de dados gerados, instituições de pesquisas tornam-se potencialmente vulneráveis aos impactos da explosão de informações, que pode acarretar um caos no gerenciamento das mesmas. Diante deste contexto, o desenvolvimento dos catálogos de dados (CD) aparece como solução para problemas como: (i) organização e (ii) gerenciamento dos dados nas instituições. No âmbito científico, os CD são implementados em conjunto com o padrão digital para metadados geográficos, largamente utilizado na catalogação de informações científicas. O objetivo deste artigo é apresentar as características de acesso e armazenamento de metadados científicos em sistemas de banco de dados para facilitar a descrição e a disseminação de dados científicos. Serão abordados aspectos relevantes que devem ser considerados durante a fase de desenvolvimento, pois

podem determinar o sucesso da implementação. A partir de um diagrama de fluxo de dados, serão ilustradas as fases consideradas desde o recebimento dos dados até sua publicação e conseqüente liberação para consultas no CD. A utilização dos CD em instituições de pesquisa pode ser uma maneira de promover e facilitar a disseminação dos dados científicos, evitar a duplicação de esforços em sua obtenção, bem como, estimular o reuso dos dados já coletados, processados e armazenados.

Palavras-chave: Catálogo de dados, Disseminação de dados, Metadados científicos, Z39.50

1. INTRODUÇÃO

Em grande parte das instituições de pesquisa a produção e a aquisição de dados científicos são consideradas tarefas demoradas e custosas, uma vez que demandam investimentos financeiros e tempo. No entanto, a reutilização de dados em pesquisas é rara, devido à falta de documentação e divulgação apropriadas daquilo que já foi produzido.

A constante produção científica em universidades e centros de pesquisas faz com estas instituições produzam e adquiram grande quantidade de dados em curto espaço de tempo. O crescimento dos acervos ocorre tanto a partir da agregação de novos dados ao sistema, como pela geração de análises e/ou manutenção daquilo que já existe.

Para instituições de pesquisa, os dados são recursos básicos, e por isso devem ser de fácil acesso aos usuários. Em geral, a necessidade de localização e acesso rápido a dados específicos, dentro de grandes conjuntos de dados, é comum, tornando relevante a documentação e a organização dos acervos.

Uma solução que vem sendo adotada por algumas instituições é o desenvolvimento de catálogos de dados, que auxiliam os usuários na localização e análise preliminar de conjuntos de dados (Callahan e Johnson, 1995). Catálogos de dados são sistemas de armazenamento que contêm informações descritivas sobre os dados como, por exemplo, seu conteúdo, abrangência temporal e geográfica, e qualidade. O desenvolvimento desses sistemas tem obtido sucesso no gerenciamento das informações científicas. Um fator chave que reforça sua utilização é a facilidade em permitir aos usuários a análise de um dado sem a necessidade de adquiri-lo.

A tecnologia de bancos de dados tem se expandido para diversas áreas de aplicações e, com a utilização de recursos específicos, agregam funcionalidades que viabilizam a disseminação de informações, principalmente, com o auxílio da Internet. Com a possibilidade de disseminação de informações mundialmente, diversas instituições têm se preocupado em padronizar o conteúdo daquilo que será disponibilizado.

A ênfase neste artigo está na utilização de padrões de metadados (informações sobre os dados). A utilização destes padrões viabiliza a definição de uma terminologia comum para descrever um dado. O propósito em definir-se um padrão é evitar que a mesma informação seja descrita de maneira diferente por instituições diferentes, o que poderia variar amplamente de uma instituição para outra. Assim, a tecnologia de metadados possibilita uma interface entre o produtor do dado e quem irá utilizá-lo, tornando possível o entendimento comum do dado.

A utilização do termo metadados relacionado ao armazenamento e, conseqüentemente, à descrição de conjuntos de dados científicos em sistemas de bancos de dados, intuitivamente leva ao emprego do termo metabanco de dados. Entretanto, um metabanco de dados se refere ao nível de abstração mais baixo de um sistema de banco de

dados (Figura 1). Por exemplo, ao executar a maior parte de suas funções internas, um sistema gerenciador de banco de dados (SGBD) acessa o metabanco de dados para encontrar informações pertinentes aos esquemas por ele implementados. Kerhervé e Gerbé (1997) sugerem que o metabanco de dados pode ser considerado o conjunto de metadados necessários para garantir eficiência aos processos internos do SGBD. Diante desse contexto, optou-se por empregar o termo catálogo de dados, julgando ser este mais apropriado ao armazenamento e à descrição de metadados científicos em SGBD.

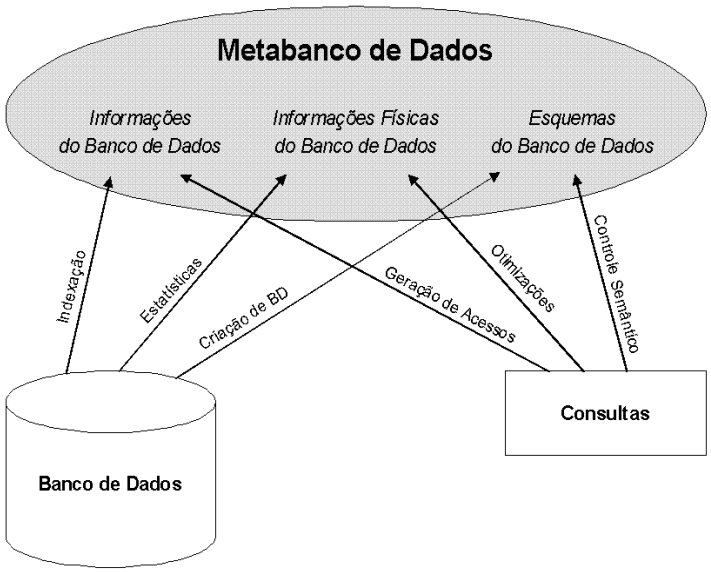

Figura 1 – Abstração do metabanco de dados em um SGBD. (Fonte: Kerhervé e Gerbé, 1997)

O presente trabalho tem por objetivo apresentar as características de ferramentas para acesso e armazenamento de metadados científicos em sistemas de bancos de dados. Estas ferramentas são amplamente utilizadas em instituições que trabalham com dados científicos, pois permitem que estes sejam disseminados, evitando a duplicação de esforços em sua obtenção e possibilitando o conhecimento do acervo de dados da própria instituição. A abordagem incluirá uma discussão sobre metadados no contexto científico e, na seqüência, será apresentado o padrão digital de metadados geográficos, largamente utilizado na catalogação de informações científicas. A seguir, serão apresentados os catálogos de dados e salientados alguns aspectos relevantes que devem ser considerados durante a fase de desenvolvimento, uma vez que podem determinar o sucesso da implementação. A partir de um diagrama de fluxo de dados serão ilustradas as fases consideradas desde o recebimento dos dados até sua publicação e conseqüente liberação para consultas no catálogo de dados.

2. METADADOS NO CONTEXTO CIENTÍFICO

A utilização de metadados é apontada pela comunidade científica como uma solução eficiente para a descrição de informações (Moura e Campos, 2002) de interesse desta comunidade como, por exemplo, resultados de pesquisas científicas. Metadados podem ser definidos simplesmente como dados sobre dados. Entretanto, esta definição não é um consenso (Ikematu, 2001) e outras podem ser encontradas:

- Metadados fornecem o contexto para entender os dados através do tempo;
- Metadados estão associados a objetos que auxiliam usuários potenciais a ter vantagem completa do conhecimento da existência ou característica dos dados;
- Metadados são componentes de informações ou instruções de alto nível que descrevem o conteúdo, qualidade, estrutura e acessibilidade de um conjunto específico de dados;
- Metadados são instrumentais para transformar dados brutos em conhecimento.

No contexto científico, os metadados contêm informações para descrever o conteúdo, a qualidade, a condição e outras características relevantes dos dados (Callahan e Johnson, 1996; Hart e Phillips, 1998).

Diversos padrões de metadados têm sido criados com a finalidade de atender as necessidades de descrição de recursos específicos. O padrão desenvolvido pelo *Dublin Core Metadata Initiative* (DCMI) contém um conjunto especializado de expressões para descrição dos recursos eletrônicos a partir da Internet. Porém, quando se trata da catalogação de metadados específicos, os elementos fornecidos pelo DCMI são considerados limitados, uma vez que os dados podem possuir características particulares não cobertas pelo mesmo. Um exemplo de padrão de metadados com finalidade específica é o *Government Information Locator Service* (GILS), cuja finalidade é catalogar especificamente informações governamentais. O padrão utilizado por bibliotecas digitais, para cadastrar informações em catálogos de dados, é o *Bibliographic*-1 (BIB-1). No contexto de dados científicos, o padrão fornecido pelo *Federal Geographic Data Commite* (FGDC) é bastante completo e específico.

2.1 PADRÃO DE METADADOS GEO-ESPACIAIS DIGITAIS

O padrão para metadados geo-espaciais digitais (GEO), formalmente definido *Content Standard for Digital Geospatial Metadata* (Padrão de Conteúdo para Metadados Geo-Espaciais Digitais), foi desenvolvido pelo *Federal Geographic Data Commite* (FGDC) e pela *American Society for Testing Materials* (ASTM), com a finalidade de fornecer um conjunto de elementos referentes a dados digitais ou a informações espaciais geo-referenciadas. Estes elementos têm o objetivo de especificar rótulos que serão utilizados em programas de geo-processamento, visando facilitar a consulta, a obtenção de resultados e a apresentação de dados geo-espaciais.

O desenvolvimento do GEO foi iniciado pela ASTM em 1990 e, a partir de um fórum sobre metadados geo-espaciais, em 1992, o FGDC juntou-se ao desenvolvimento. Em 1994, o GEO foi aprovado pelo FGDC e oficializado como padrão de documentação de dados pelo governo americano. A versão desenvolvida pela ASTM foi incorporada ao GEO através dos marcadores alfanuméricos e numéricos para cada elemento ou grupo de elementos do padrão. Os marcadores são específicos e devem ser utilizados na pesquisa e apresentação de dados (Nelbert, 2000).

O padrão GEO possui um conjunto de 334 elementos que, em alguns casos são herdados de outros padrões como o GILS e o BIB-1. No entanto, possui seu próprio conjunto de elementos que não pode ser mapeado para os demais padrões. Os elementos numerados entre 1 e 1999 foram herdados do BIB-1, elementos entre 2000 e 2999 foram

herdados do GILS e os demais elementos, numerados entre 3000 e 3999, são específicos do GEO. O padrão possui ainda três classes de elementos designados de (i) elementos de relação, que permitem o relacionamento entre o termo pesquisado e sua posição nos metadados, (ii) elementos de estrutura, para especificar que parte dos metadados será pesquisada e (iii) elementos de truncamento, utilizados para truncar as palavras no texto, quando necessário.

A estrutura do GEO se divide em sete grupos, onde somente o primeiro (*Identification Information*) e o último (*Metadata Reference Information*) são obrigatórios, sendo os demais opcionais. O propósito de cada um destes grupos é apresentado de forma resumida a seguir. Mantêm-se, na apresentação, as designações dos grupos em sua forma original (em Inglês).

1. *Identification Information* (Informações de Identificação)

Este grupo contém a meta-informação básica sobre o conjunto de dados como, por exemplo:

- Descrição textual;
- Informação sobre o período de tempo;
- Referência espacial;
- Palavras chaves;
- Pessoa e instituição de contato para maiores informações sobre o conjunto de dados;
- Restrições de acesso.

2. *Data Quality Information* (Informações sobre a Qualidade dos Dados)

Este grupo contém informações gerais sobre a qualidade do conjunto de dados.

3. *Spatial Data Organization Information* (Informações sobre a Organização Espacial dos Dados)

Este grupo contém informações de quais mecanismos foram utilizados para representar a informação espacial do conjunto de dados.

4. *Spatial Reference Information* (Informações sobre a Referência Espacial)

Este grupo contém informações sobre o sistema de projeção e sistemas de coordenadas utilizadas.

5. *Entity Attribute Information* (Informações sobre os Atributos das Entidades)

Este grupo permite ao usuário descrever as informações contidas no conjunto de dados.

6. *Distribuition Information* (Informações de Distribuição)

Este grupo contém informações sobre quem irá fornecer o dado e sobre as opções para obtê-lo. O fornecedor, em geral, corresponde à pessoa de contato/instituição listado no grupo *Identification Information*. Algumas informações incluem os meios de acesso ao conjunto de dados como, por exemplo, ftp, e-mail, etc.

7. *Metadata Reference Information* (Informações de Referência sobre os Metadados)

Este grupo contém informações sobre a última atualização dos metadados.

3. VISÃO GERAL SOBRE CATÁLOGOS DE DADOS

3.1 ASPECTOS RELEVANTES CONSIDERADOS NA FASE DE DESENVOLVIMENTO

Os Catálogos de Dados (CD) são definidos como sistemas para descrever um conjunto de dados e indicar a sua localização para uso. O fator chave que favorece sua utilização é possibilitar aos usuários determinar a relevância e a qualidade dos dados para um propósito específico, sem a necessidade de ter que adquiri-los para uma análise mais detalhada. No entanto, o uso efetivo dos CD depende fortemente da maneira como as informações são descritas no servidor. De acordo com Callahan e Johnson (1995), seis fatores chaves devem ser considerados durante o processo de desenvolvimento desses sistemas (CD):

- Completude;
- Facilidade de utilização;
- Coerência das informações;
- Precisão;
- Disponibilidade;
- Serem Públicos.

Os CD serão fortemente utilizados se forem completos, isto é, se as instituições realmente documentarem todos os conjuntos de dados de seus acervos. Por exemplo, se consultas a determinados conjuntos de dados falharem, os usuários estarão certos que os mesmos não existem na instituição. No entanto, se o CD for incompleto, os usuários jamais estarão certos se os conjuntos de dados existem ou não.

Um dos objetivos destes sistemas é viabilizar o acesso e a localização dos conjuntos de dados de maneiras rápida e fácil. Por isso não há a necessidade de se promover treinamentos extensivos para a utilização de um sistema. Dados científicos, geralmente, possuem componentes espaciais, que também devem ser consideradas como uma forma de busca.

Informações suficientes devem ser consideradas para permitir aos usuários determinarem se os conjuntos de dados serão utilizados ou não para um propósito específico. Determinar quais as informações devem ser primeiramente consideradas é uma tarefa difícil. Cada conjunto de dados possui potencialmente diferentes tipos de informações e, como conseqüência, os CD devem ser flexíveis para acomodar esta variação. O conteúdo dos CD deve ser determinado criteriosamente por quem classifica os dados, uma vez que sua utilidade depende da relevância das informações que são retornadas pelas consultas.

Muitos usuários acessam um CD para teste de veracidade das informações armazenadas. Por isso, os CD devem ser precisos e evitarem descrições incompletas, que podem acarretar baixa credibilidade.

A facilidade de acesso a partir das redes de computadores deve permitir que qualquer usuário dentro das instituições possa ter acesso às informações dos CD. As pessoas devem saber que estes sistemas existem, entender que devem ser utilizados e aplicá-los em benefício do desenvolvimento de seus trabalhos.

3.2 DESCRIÇÃO DOS DADOS

O processo de descrição ou catalogação de dados e, conseqüentemente, a sua disponibilização através da Internet não é apenas uma questão tecnológica, pois, em geral, envolve mudanças de conceitos, reestruturação organizacional, aprendizagem e planejamento nas instituições.

Normalmente, aplicações computacionais são desenvolvidas com interfaces gráficas amigáveis, cuja finalidade é minimizar os esforços empregados neste processo. Infelizmente, este é o ponto onde os esforços cessam. Algumas razões comumente utilizadas para evitar a documentação dos dados são apresentadas a seguir (Callahan e Johnson, 1996):

- Trata-se de uma tarefa tediosa;
- A localização e a qualidade dos dados já são bem conhecidas;
- Esta tarefa mantém o cientista longe de seu trabalho real (fazer ciência);
- Um telefonema pode facilitar a localização dos dados;
- O reconhecimento é pequeno perto do tempo gasto;
- Receio de tornar os dados públicos;
- Utilização dos dados sem o devido crédito ao "dono" destes.

Como resultado observa-se que recursos consideráveis são gastos, mas o interesse inicial dos pesquisadores em participar do processo dissipa-se rapidamente. Nesse sentido, o desafio das instituições é motivá-los a documentar e manter atualizadas as informações de seus conjuntos de dados.

Diante desse contexto, observa-se que durante a implantação do processo, os maiores investimentos devem ser direcionados não aos recursos tecnológicos, mas principalmente aos recursos humanos, uma vez que estes últimos realizarão tarefas que dificilmente serão automatizadas, no que se refere à catalogação de metadados.

No que diz respeito a dados científicos, o processo de catalogação requer da pessoa responsável pela tarefa um conhecimento mais preciso sobre algumas características, que só quem esteve envolvido em sua obtenção pode conhecer. Isto reforça a afirmação de que, um fator relevante para a utilização de metadados na catalogação de informações, particularmente no caso de dados científicos, é conscientizar as pessoas envolvidas em sua obtenção da importância de documentá-los.

4. IMPLEMENTAÇÃO DO CATÁLOGO DE DADOS

O presente estudo foi desenvolvido no Centro de Previsão de Tempo e Estudos Climáticos (CPTEC), do Instituto Nacional de Pesquisas Espaciais (INPE), em Cachoeira Paulista.

Na etapa inicial, foram realizadas diversas entrevistas com os administradores dos sistemas IAI-DIS e LBA-DIS por se tratar, no Brasil, de referências de sucesso no contexto de catalogação e disseminação de informações científicas a partir da Internet. Após algumas reuniões, chegou-se à conclusão da adoção de um programa chamado *Isite*, devido a algumas características encontradas, destacando-se (i) possibilidade de desenvolvimento de aplicações seguindo o padrão geo-espacial; (ii) disponibilidade de ferramentas para o

desenvolvimento de aplicações de busca para a Internet; (iii) suporte técnico gratuito; (iv) utilização por centenas de instituições pelo mundo e (v) possibilidade da realização de consultas de maneira transparente aos usuários, tanto na base de dados local quanto em bases de dados remotas.

O sistema *Isite* corresponde a um conjunto de programas integrados fornecidos gratuitamente pelo FGDC, para diversas plataformas de sistemas operacionais, com a finalidade de viabilizar o desenvolvimento de ferramentas de pesquisa e disseminação de informações personalizadas através da Internet.

O *Isite* pode ser definido como um sistema completo de informações para Internet, pois tem a característica de integrar sistemas de bancos de dados com outros sistemas e protocolos como *World Wide Web* (WWW) e, principalmente, o Z39.50 (Gamiel, 1998). O pacote foi desenvolvido inicialmente pelo *Center for Networked Information Discovery and Retrieval* (CNIDR), responsável por sua manutenção, e representa o resultado dos investimentos da *National Science Foundation* para promover e implementar a integração entre as ferramentas para recuperação de informações distribuídas através da Internet.

A ferramenta utilizada para indexar os dados é chamada *Iindex* e pode ser configurada para diversos perfis como, por exemplo, o GEO, o BIB-1, o GILS, etc. O *Isite* inclui também a aplicação *Zserver*, que corresponde a um servidor Z39.50 utilizado para responder às consultas submetidas por clientes Z39.50 à base de dados. No entanto, para funcionar corretamente é necessário configurar alguns parâmetros como, por exemplo, a localização do arquivo de *logs*, o nome do banco de dados, o número máximo de sessões consecutivas que serão aceitas, a porta de comunicação para permitir acesso remoto ao servidor e o tipo de servidor.

Faz parte do *Isite* um conjunto de aplicativos destinados à realização de pesquisas a partir da Internet. Esses aplicativos, chamados *Zgate* e *Zcon*, integram o *Isearch* módulo *Commom Gateway Interface* (CGI). Após a execução do servidor *Zserver* e com a utilização de páginas da Internet devidamente configuradas, é possível realizar consultas a diversos bancos de dados remotos, de maneira transparente ao usuário. Um exemplo da arquitetura de funcionamento do *Isearch* módulo CGI pode ser observado na Figura 2.

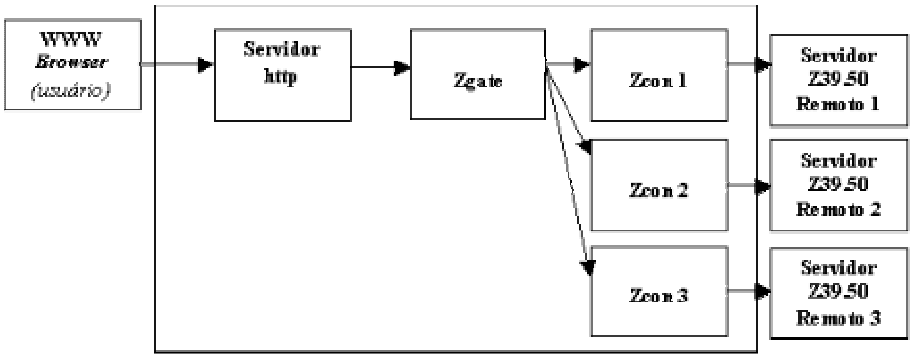

Figura 2 – Arquitetura de funcionamento do *Isearch* módulo CGI.

4.1 PROTOCOLO Z39.50

O protocolo de comunicação Z39.50, formalmente definido *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification* (Recuperação de

Informações (Z39.50): Definição do Serviço da Aplicação e Especificação do Protocolo), foi desenvolvido com o objetivo de especificar regras e procedimentos para comunicação entre dois sistemas de computação, com o propósito de consultar e recuperar informações (dados bibliográficos, imagens, textos, etc.) de maneira distribuída (ANSI/NISO Z39.50-1995, 1995).

O padrão Z39.50 foi originalmente proposto em 1984, para a utilização específica com informações bibliográficas. Sua primeira versão foi aprovada em 1988, desenvolvida por um comitê designado pela *National Information Standard Organization* (NISO). A partir de 1989, foi estabelecido que a administração do padrão Z39.50 seria realizada pela *Library of Congress* (Biblioteca do Congresso Americano), que se tornou responsável pela coordenação técnica do desenvolvimento do protocolo, registro de novas implementações e trabalho editorial do mesmo.

Sob o ponto de vista técnico, uma aplicação que utiliza o Z39.50 deve estar habilitada a permitir a troca de informações entre um cliente e um servidor. O componente cliente é definido como o responsável pelo início dos diálogos junto ao servidor, tendo como objetivo final recuperar os registros que satisfaçam as condições da consulta requisitada no catálogo de dados do servidor (Lynch, 1997). O principio básico de funcionamento de um cliente Z39.50 (Figura 3) resume-se ao estabelecimento da sessão através de uma requisição de início, sua manutenção e, por fim, do fechamento da sessão a partir da requisição de final da sessão.

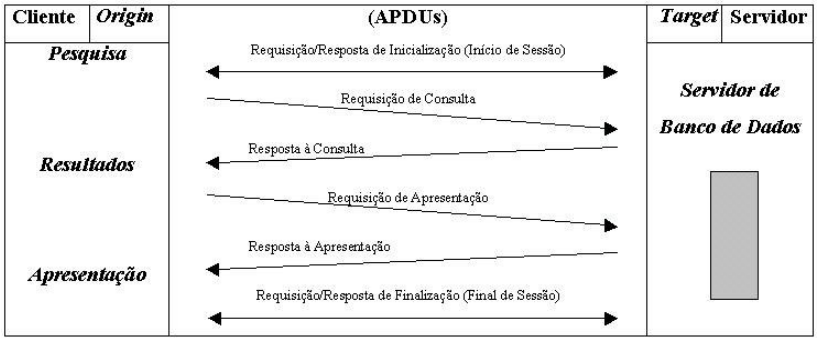

Figura 3 – Princípio de funcionamento básico do Z39.50.

Durante a realização de uma sessão as requisições de consultas, solicitadas pelo cliente, são enviadas ao servidor através de mensagens definidas tecnicamente como *Application Protocol Data Units* (APDUs) (Unidades de Dados do Protocolo da Aplicação) que são especificadas segundo o *Abstract Syntax Notation One* (ASN.1) (Notação Sintática Abstrata), que é um padrão de sintático utilizado para descrever as APDUs. Após o envio da requisição de consulta, o servidor torna-se responsável por sua realização até a devolução da resposta ao cliente. O cliente recebe a resposta do servidor, contendo o total de resultados obtidos pela consulta e faz uma requisição para apresentação destes resultados. De maneira transparente ao usuário, após o recebimento da requisição de apresentação, os dados são formatados adequadamente e apresentados ao cliente.

Os recursos do Z39.50 são largamente utilizados pela comunidade bibliotecária, pois apresentam grande versatilidade para o desenvolvimento de aplicações objetivando disseminar este tipo de informação. Uma de suas vantagens é permitir, de maneira transparente aos usuários, o acesso aos CD com os mais diversos tipos de informações. Implementações que utilizam os recursos do Z39.50 não devem ser complexas, uma vez que oferecem a oportunidade de disseminar informações e realizar consultas em sistemas

distribuídos, sem a necessidade de um entendimento detalhado dos procedimentos utilizados para o acesso, por ser este realizado de maneira transparente. Exemplos de bibliotecas com acesso *on-line* a catálogos de dados são facilmente encontrados na Internet, principalmente, em universidades e sistemas públicos de acesso a dados de países como Austrália, Canadá e Estados Unidos.

Quando associado a aplicações como criação e gerenciamento de CD, o Z39.50 mostra-se muito dinâmico e rico em funcionalidades. Os recursos desse padrão devem ser vistos como uma porta aberta ao descobrimento e à exploração dos CD distribuídos. Sua utilização é extremamente vantajosa para aplicações que tenham como objetivo disseminar dados através da Internet, pois permite a utilização de uma interface única objetivando facilitar a realização de consultas tanto locais quanto remotas.

4.2 DEFINIÇÃO DO BANCO DE DADOS

O banco de dados (BD) foi modelado com a finalidade de armazenar os metadados em tabelas e facilitar a realização de tarefas como: atualização, remoção e consulta a qualquer informação cadastrada. O armazenamento de informações em tabelas, projetadas em conformidade com as restrições (regras) do modelo relacional (Date, 2000; Silberschatz et al., 2005), em especial com relação ao projeto normalizado de BD, garante maior agilidade e flexibilidade para acesso aos dados. Com a utilização das funcionalidades oferecidas pelo SBGD MySQL (Axmark et al., 2001) tem-se a garantia de consistência, integridade e de segurança no acesso aos dados.

A definição dos metadados necessários, com vistas à identificação dos atributos (campos) das tabelas do BD, foi efetuada mediante um consenso, após a realização de reuniões junto aos de pesquisadores do CPTEC/INPE.

O modelo de Entidades e Relacionamentos para BD (Chen, 1976; Date, 2000; Silberschatz et al., 2005) é apresentado na Figura 4. O diagrama apresentado foi construído com o aplicativo DBDesigner (http://www.fabforce.net). No diagrama pode-se verificar, de maneira objetiva, os relacionamentos entre as tabelas (entidades) incluídas no BD.

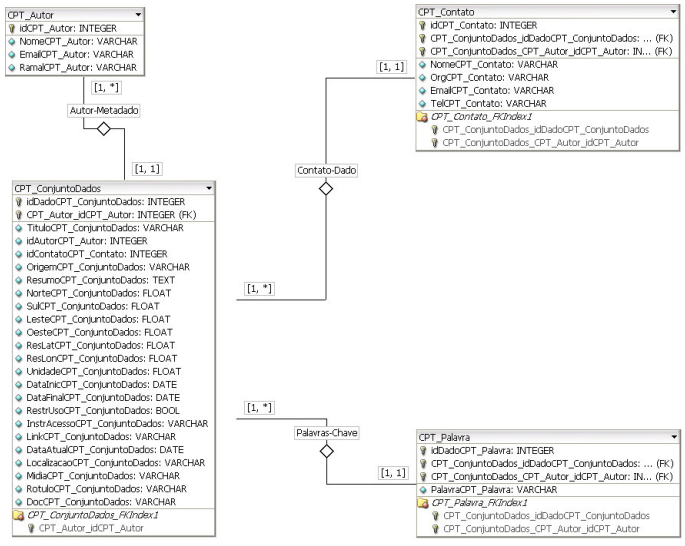

Figura 4 – Diagrama de entidades e relacionamentos representando a estrutura do banco de dados.

4.3 INTERFACE DE CONSULTA

O desenvolvimento de interfaces amigáveis para a realização das consultas em um CD e para a conseqüente apresentação dos resultados visa facilitar as análises dos conjuntos de dados. O padrão adotado na implementação das interfaces gráficas foi o *Hypertext Markup Language* (HTML) que, em conjunto com a tecnologia CGI, viabiliza a confecção de páginas de Internet dinâmicas, em geral obtidas a partir dos resultados de consultas ao BD. A integração do BD MySQL com a Internet foi realizada a partir de uma interface genérica da linguagem *Pratical Extration and Report Language* (Perl), chamada *DataBase Interface* (DBI).

Para realizar consultas no CD, os usuários devem definir os termos que serão utilizados, bem como, identificar sua localização no texto, sua variação temporal e/ou a área espacial coberta pelo dado. A página utilizada para as consultas ao CD (Figura 5) fornece as seguintes opções:

- Termos de Pesquisa;
- Cobertura Temporal;
- Cobertura Espacial.

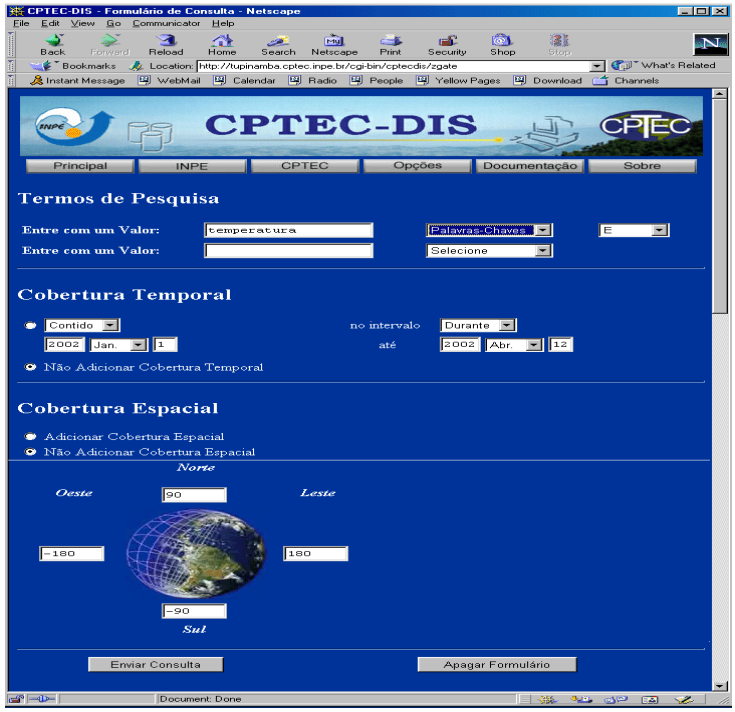

Figura 5 – Página para a realização das consultas.

As opções disponíveis para consultar um termo no texto são: (i) título, (ii) palavras-chave, (iii) resumo e (iv) qualquer parte do texto. Como ilustrado na Figura 5, o usuário deve escolher uma destas opções em uma das caixas de listagem localizadas à direita da área da página destinada ao preenchimento de termos. Nos dois campos destinados ao preenchimento de termos (valores), são permitidas combinações utilizando um dos operadores booleanos: "E", "OU" ou "EXCETO".

A página apresentada na Figura 6 é confeccionada dinamicamente de acordo com os resultados que satisfazem uma consulta. Na primeira página de respostas (Figura 6), são apresentadas informações resumidas referentes aos registros que satisfazem as condições de consulta. A página de respostas apresentada refere-se à consulta do termo "temperatura", selecionando-se a opção “palavras-chaves” para busca nos metadados (Figura 5). Na parte superior da página, pode-se observar uma tabela contendo informações referentes à consulta como, o nome do BD, o *status* da consulta, a quantidade total de registros recuperados e que registros estão sendo apresentados na página. Abaixo, encontra-se uma lista contendo os títulos dos registros recuperados, seguidos da opção "*Show*…" em formato de link, opção esta que possibilita o acesso ao registro desejado em sua forma completa, para uma análise mais detalhada.

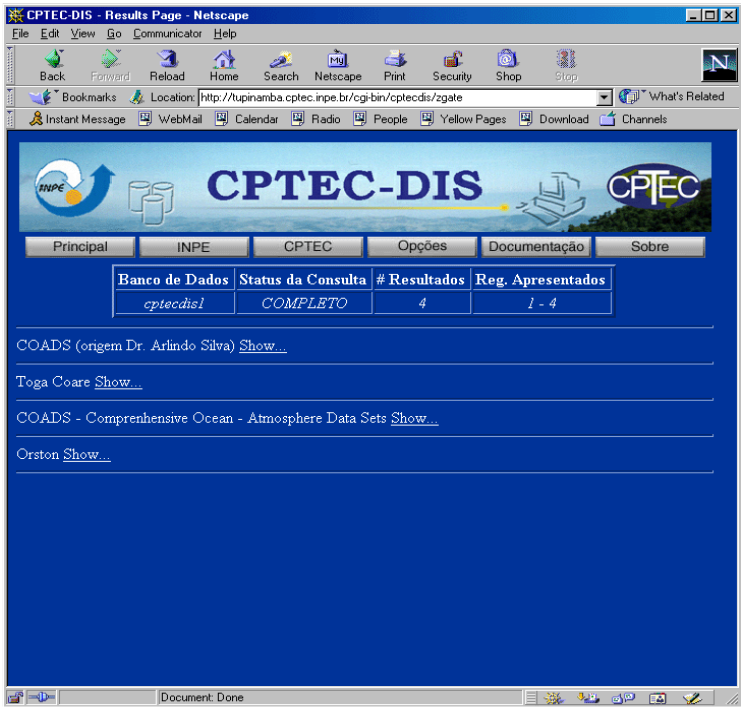

Figura 6 – Página de respostas à consulta realizada no catálogo de dados.

A partir da segunda página de respostas (Figura 7), pode-se ter acesso a informações detalhadas sobre cada conjunto de dados, o que fornece melhores condições para a análise dos resultados da consulta pelo usuário.

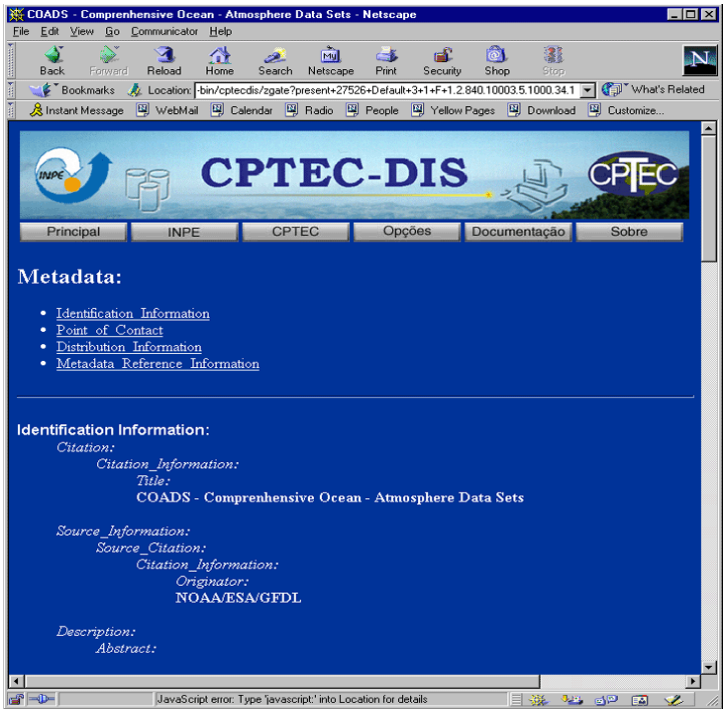

Figura 7 – Página com os detalhes da consulta.

4.4 DIGRAMA DE FLUXO DE DADOS GENERALIZADO

A figura abaixo (Figura 8) apresenta um possível diagrama seguido por instituições de pesquisas (Barbosa, 2002). Pode-se observar, no diagrama de fluxo de dados (Rumbaugh et al., 1991), que os dados chegam ao acervo a partir de diversas fontes como, por exemplo:

- Outras instituições (universidades, centros de pesquisas, etc.);
- Sensores;
- Modelos computacionais;
- Continuação de pesquisas;
- Trabalhos de campo, etc.

Deve-se ressaltar, com relação ao diagrama abaixo, a necessidade de tratamento de pré-armazenamento, exigida para alguns conjuntos de dados como procedimento anterior ao seu armazenamento no acervo da instituição. A partir deste ponto, os dados armazenados podem ser acessados e o processo de catalogação deve ser iniciado.

Uma vez catalogados, os metadados são, então, armazenados nos CD científicos e disponibilizados para a realização de consultas.

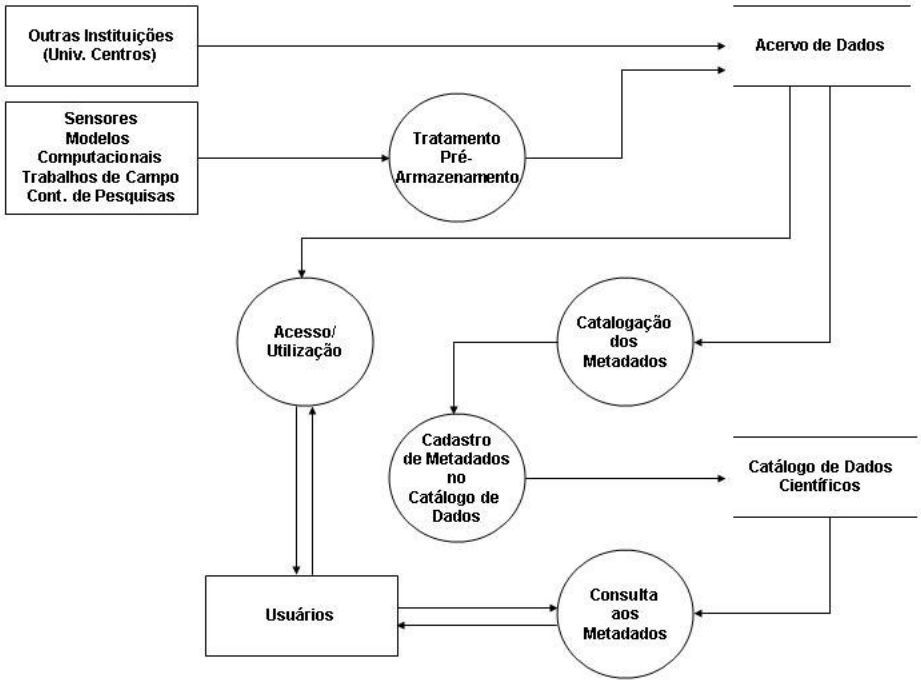

Figura 8 – Diagrama de fluxo de dados generalizado.
(Fonte: Barbosa, 2002)

5. CONCLUSÕES

Os CD têm sido utilizados como a solução para problemas como (i) a organização dos acervos e (ii) o gerenciamento dos dados nas instituições. Dentre os aspectos considerados em seu desenvolvimento, o processo de catalogação dos metadados deve receber especial atenção. Freqüentemente a implantação do sistema costuma falhar, devido à falta de cultura organizacional e de interesse, em geral, dos pesquisadores, em participar deste processo. Nesse contexto, observa-se um desafio chave para as instituições, qual seja motivar as pessoas a documentar e manter atualizada a documentação de seus dados. É necessária uma conscientização institucional quanto ao reconhecimento da relevância deste processo, conscientização esta que pode determinar o sucesso do sistema.

Em grande parte das instituições de pesquisa o número de artigos publicados constitui uma informação chave no processo de reconhecimento profissional, sendo também de relevância para promoções dentro da própria instituição. As publicações de metadados em CD têm uma função importante para estas instituições e, portanto, o reconhecimento deste tipo de produção deve receber consideração semelhante ao da publicação de artigos (Callahan e Johnson, 1995 e 1996).

O padrão para metadados geo-espaciais digitais pode ser utilizado na criação de uma gramática comum para a descrição dos conjuntos de dados, ou seja, uma ontologia. Esta conceituação formalizada certamente proporcionará uma integração potencial entre instituições de pesquisa com a finalidade de facilitar o intercâmbio para troca de dados científicos.

O desenvolvimento de interfaces eletrônicas amigáveis tem a finalidade de viabilizar e estimular a utilização dos CD. A partir destas interfaces, pode-se realizar, de maneira fácil e ágil, tarefas como consulta e análise preliminar de informações científicas.

No período inicial, o CD aqui descrito estará disponível apenas aos usuários internos ao CPTEC/INPE através de uma intranet. Entretanto, a utilização de páginas *Web* dinâmicas tem a finalidade de tornar a aplicação disponível, num futuro próximo, para acesso e realização de consultas por pesquisadores de outras instituições, a partir da Internet.

A partir do diagrama de fluxo de dados pôde-se apresentar, de maneira generalizada, uma seqüência de processos, considerando desde a obtenção e/ou aquisição de conjuntos de dados pelas instituições até a sua publicação para consulta via CD. Foram destacadas neste diagrama fontes heterogêneas de obtenção e/ou aquisição de dados científicos, o que torna necessário a realização de atividades de pré-armazenamento no acervo da instituição. Observou-se, também, que a catalogação dos metadados deve ocorrer após o armazenamento dos dados no acervo. Como conseqüência, o processo de catalogação deve ser realizado de forma cautelosa, de modo a garantir informações consistentes e relevantes sobre os diversos conjuntos de dados.

A utilização dos CD em instituições de pesquisa pode ser uma maneira de promover e facilitar a disseminação dos dados científicos, evitar a duplicação de esforços em sua obtenção, bem como, estimular a reutilização dos dados já coletados, processados e devidamente armazenados.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ANSI/NISO Z39.50-1995, 1995 : **INFORMATION RETRIEVAL (Z39.50): APPLICATION SERVICE DEFINITION AND PROTOCOL SPECIFICATION.** Recurso *on-line* disponível na Internet. URL: *http://lcweb.loc.gov/z3950/agency*.

Axmark, D., Widenius, M., Cole, J., Lentz, A., DuBois, P., 2001 : **MYSQL REFERENCE MANUAL.** Recurso on-line disponível na Internet. URL: *http://www.mysql.com*.

Barbosa, E.B.M., 2002 : **UMA FERRAMENTA PARA DISSEMINAÇÃO DE DADOS CIENTÍFICOS DO CPTEC/INPE ATRAVÉS DE UM BANCO DE METADADOS.** Monografia de Especialização em Informática Empresarial, Faculdade de Engenharia de Guaratinguetá (FEG/UNESP), 62 p.

Callahan, S.D., Jonhson, B.D., 1995 : **SCIENTIFIC DATA SET CATALOGUES.** Proceedings of Second AGSO Forum on GIS in the Geosciences, Canberra, ACT, 29-31pp.

Callahan, S.D., Jonhson, B.D., 1996 : **DATASET PUBLISHING - A MEANS TO MOTIVATE METADATA ENTRY.** Paper presented at the 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 pp.

Date, C.J., 2000 : **INTRODUCTION TO DATABASE SYSTEMS. 7th. Edition.** Addison Wesley Professional.

Gamiel, K., 1998 : **THE ISITE INFORMATION SYSTEM - VERSION 2.00. THE CLEARINGHOUSE FOR NETWORKED INFORMATION DISCOVERY AND RETRIEVAL.** Center for Networked Information Discovery and Retrieval (CNIDR). This material is based in work sponsored by the National Science Foundation under Cooperative Agreement No. NCR-9216963. 20 pp.

Hart, D., Phillips, H., 1998 : **METADATA PRIMER – HOW TO GUIDE ON METADATA IMPLEMENTATION.** Recurso *on-line* disponível na Internet. URL: *http://www.lic.wisc.edu/metadata/metaprim.htm*.

Ikematu, R.S., 2001 : **GESTÃO DE METADADOS: SUA EVOLUÇÃO NA TECNOLOGIA DA INFORMAÇÃO.** DataGramaZero - Revista de Ciência da Informação - V.2, N.6.

Kerhervé, B, Gerbé, O., 1997 : **MODELS FOR METADATA OR METAMODELS FOR DATA?** Proceedings of Second IEEE Metadata Conference.

Lynch, C.A., 1997 : **THE Z39.50 INFORMATION RETRIEVAL STANDARD, PART I. A STRATEGIC VIEW OF IS PAST, PRESENT AND FUTURE.** D-Lib Magazine. Recurso *on-line* disponível na Internet. URL: *http://www.dlib.org/dlib/april97/cornell/04lynch.html*.

Moura, A.M.C., Campos, M.L.M., 2002 : **A METADATA APPROACH TO MANAGE AND ORGANIZE ELECTRONIC DOCUMENTS AND COLLECTIONS ON THE WEB.** Journal of the Brazilian Computer Society. V.1, N.8, 16-31 pp.

Nebert, D.D., 2000 : **Z39.50 APPLICATION PROFILE FOR GEOSPATIAL METADATA. V.2.2.** Recurso *on-line* disponível na Internet. URL: *http://www.blueangeltech.com/standards/GeoProfile/geo22.htm*.

Rumbaugh, J., Blaha, M., Premeriani, W., Eddy, F., Lorensen, W., 1991 : **OBJECT-ORIENTED MODELING AND DESIGN.** Prentice Hall.

Silberschatz, A., Korth, H.F., Sudarshan, S., 2005 : **DATABASE SYSTEMS CONCEPTS. 5th. Edition.** McGraw-Hill.