Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira

Carlos Henrique Marcondes
Doutor em ciência da informação, DEP-IBICT/UFRJ. Consultor do projeto BDB/IBICT. Universidade Federal Fluminense, Departamento de Ciência da Informação.
marcondes@alternex.com.br.

Luís Fernando Sayão
Doutor em ciência da informação, DEP-IBICT/UFRJ. Consultor do projeto BDB/IBICT. Comissão Nacional de Energia Nuclear. Centro de Informações Nucleares
lsayao@cnen.gov.br

Resumo

Descreve as opções tecnológicas e metodológicas para atingir a interoperabilidade no acesso a recursos informacionais eletrônicos, disponíveis na Internet, no âmbito do projeto da Biblioteca Digital Brasileira em Ciência e Tecnologia, desenvolvido pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Destaca o impacto da Internet sobre as formas de publicação e comunicação em C&T e sobre os sistemas de informação e bibliotecas. São explicitados os objetivos do projeto da BDB de fomentar mecanismos de publicação pela comunidade brasileira de C&T, de textos completos diretamente na Internet, sob a forma teses, artigos de periódicos, trabalhos em congressos, literatura "cinzenta", ampliando sua visibilidade e acessibilidade nacional e internacional, e também de possibilitar a interoperabilidade entre estes recursos informacionais brasileiros em C&T, heterogêneos e distribuídos, através de acesso unificado via um portal, sem a necessidade de o usuário navegar e consultar cada recurso individualmente.

Palavras-chave

Bibliotecas digitais; Publicações eletrônicas; Arquivos abertos; Interoperabilidade; Metadados; Padrões; Tecnologia da informação; Informação em ciência e tecnologia; Comunicação científica; Acesso à informação.

Integration and interoperability in accessing electronic information resources in science and technology: the proposal of Brazilian Digital Library

Abstract

This paper describes technological and methodological options to achieve interoperability in accessing electronic information resources, available in Internet, in the scope of Brazilian Digital Library in Science and Technology Project - BDL, developed by Brazilian Institute for Scientific and Technical Information - IBICT. It stresses the impact of the Web on publishing and communication in science and technology and also on information systems and libraries. The work stresses the two main objectives of BDL project: promoting electronic publishing of different full text materials - theses, journal articles, papers in events, "grey" literature - by Brazilian scientific community, so amplifying their nationally and internationally visibility; and achieving, through a gateway, interoperability among those heterogeneous electronic information resources available in the Web, thus avoiding a user to navigate and query those resources one by one separately.

Keywords

Digital libraries; Electronic publishing; Open archives; Interoperability; Metadata; Standards; Information technology; Science and technology information; Scientific communication; Information access.

INTRODUÇÃO

"*Um laboratório sem uma biblioteca é como se fosse um animal descorticado: as atividades motoras continuam a funcionar, mas falta a coordenação da memória e da vontade*" (Zilman, 1979, p. 115)

Há muito tempo está constatada a relação entre desenvolvimento econômico e social e o estágio de desenvolvimento da ciência e tecnologia de um país. À ciência e tecnologia está reservado um papel fundamental na luta pelo desenvolvimento da sociedade brasileira. Informação é insumo fundamental para o desenvolvimento da ciência. É em torno do problema da otimização dos fluxos e da transferência de informação científica que surge, na segunda metade do século XX, a ciência da informação (Pinheiro, 1995).

A convergência e o uso integrado das tecnologias de comunicação, de computação e de conteúdos em formato digital, cujo paradigma é a Internet, tem contribuído nos anos recentes para criar um novo ambiente de acesso, disseminação, cooperação e promoção do conhecimento em uma escala global. Estamos em meio a este processo, cujas conseqüências ainda não podemos avaliar completamente. Novos suportes de conhecimento, que não guardam similares com os materiais impressos em papel, estão sendo inventados a cada dia.

Do ponto de vista da informação como subsídio às atividades acadêmicas e em C&T, a Internet vem proporcionar facilidades que extrapolam o conceito tradicional de informação bibliográfica baseada em documentos, como artigos de periódico, trabalhos em congressos, teses etc. Novos recursos informacionais estão à disposição da comunidade de pesquisa além desses tradicionais, agora em versão eletrônica, como documentos multimídia, listas de discussão, fóruns eletrônicos, conferências em linha, imagens (de satélites, de microscópios, em tempo real), modelos animados, bancos de *preprints* eletrônicos, os *e-prints* etc. Estes recursos tanto servem de subsídio à pesquisa quanto de canais de comunicação e publicação dos resultados e de garantia de primado e originalidade intelectuais dos mesmos.

Mais que somente recursos informacionais, os novos recursos disponíveis via Internet são acima de tudo novas ferramentas cognitivas, no sentido emprestado a elas por Pierre Lévy (1993), capazes de abrir novas possibilidades cognitivas e intelectuais que extrapolam em muito aquelas oferecidas por documentos em papel, de leitura linear. Para muitos autores, a Internet representa, neste sentido, uma mudança de paradigma comparável à invenção da imprensa por Gutemberg. Esta mudança de paradigma se faz sentir também no aspecto da comunicação científica. A Internet é um mecanismo de comunicação de alcance mundial, instantâneo, interativo e multidirecional: qualquer um pode publicar nela, o que foi publicado é imediatamente acessível, o autor pode receber um retorno e avaliação imediatos sobre o que publicou, de qualquer lugar. Um autor acadêmico almeja a máxima divulgação para seus trabalhos, para que os resultados de sua pesquisa tenham o maior impacto possível sobre as pesquisas de seus pares e sobre outras publicações. Estudos recentes confirmam que as publicações eletrônicas são muito mais citadas que as publicações em papel: "*The mean number of citations to offline articles is 2,74, and the mean number of citations of online articles is 7.03, an increase of 157%"* (Lawrence). Desenvolver mecanismos de publicação eletrônica para a comunidade acadêmica brasileira, aumentando sua visibilidade, torna-se, portanto, uma questão essencial para o desenvolvimento e maturidade da pesquisa científica brasileira.

Dessa forma, os paradigmas de comunicação científica, tendo por base o periódico científico em papel, com seu esquema de revisão por pares e o monopólio das grande editoras científicas, vêm sofrendo grande impacto com o surgimento da Internet e grande questionamento por parte da comunidade científica de todo o mundo. Desde o surgimento do primeiro arquivo eletrônico de *preprints*, ou *eprints*, o ArXiv, no Los Alamos National Laboratory, criado em 1991 pelo físico Paul Ginsparg (Ginsparg, 1996), que a própria comunidade científica internacional oferece uma alternativa prática para a publicação de seus trabalhos. Uma lista que permite dimensionar a amplitude dos arquivos eletrônicos por todo o mundo pode ser encontrada em http://www.osti.gov/eprints/ppnbrowse.html A alternativa dos *e-prints* vem também se articulando na chamada OpenArchives Initiative (http://www.openarchivesinitiative.org)

Estas transformações têm exercido profunda influência sobre a concepção e funcionamento dos sistemas de informação automatizados, especialmente aqueles voltados para as atividades de pesquisa. O rompimento de barreiras tecnológicas importantes, experimentadas na última década, permitiram o surgimento de um novo patamar para esses sistemas: antes orientados basicamente para recuperação de referências bibliográficas em bases de dados isoladas e textos em papel, voltam-se hoje para a recuperação distribuída de objetos digitais – textos completos, imagens em movimento, som etc. –, estabelecendo como palavras de ordem a publicação na Internet e a interoperabilidade entre fontes de informação heterogêneas e globalmente distribuídas.

Com o projeto da Biblioteca Digital Brasileira, o IBICT quer abrir a possibilidade, fomentar e fornecer meios para que a comunidade brasileira de C&T possa *publicar seus trabalhos de forma rotineira, diretamente na rede,* aumentando com isso sua visibilidade nacional e internacional, otimizando o fluxo da comunicação científica e reduzindo o ciclo de geração de novos conhecimentos.

Por outro lado, somente a disponibilidade de textos brasileiros em C&T *on-line* não teria grande impacto sobre a comunicação científica e a ciência no país sem a existência de serviços de informação que viabilizem o acesso de forma fácil a estes conteúdos. O país também tem acumulado experiências bastante significativas, embora isoladas, na criação de bibliotecas digitais e repositórios de informações na rede. À medida que experiências brasileiras neste sentido se multiplicam, como o Prossiga, o Scielo, o repositório de teses da USP, o arquivo de *e-prints* do Impa (http://www.preprint.impa.br/indexEngl.html) etc., disponibilizando de forma crescente recursos informacionais em texto completo na Web, fica patente, para as organizações brasileiras que trabalham com sistemas de informação para C&T, a importância da questão da interoperabilidade entre bibliotecas digitais e outros recursos informacionais digitais: *como consultar, de uma única vez, todas estas fontes de forma integrada e transparente, com o mínimo de esforço, com a máxima rapidez, e obter resultados consolidados?*

São assim dois os objetivos fundamentais do projeto BDB: a) *fomentar a publicação de recursos informacionais de interesse para C&T na rede*, propiciando à comunidade científica brasileira meios para publicar diretamente na Web, dando maior visibilidade à produção brasileira em C&T, tanto nacionalmente, quanto internacionalmente; b) *viabilizar o acesso rápido e integrado a estes recursos*, facilitando a descoberta na Internet de recursos informacionais brasileiros de interesse para a ciência e tecnologia, de forma integrada e, dessa forma, encurtando o ciclo de comunicação científica entre pares nas comunidades brasileiras de C&T.

O PROBLEMA DA INTEROPERABILIDADE

Um aspecto problemático da cultura de nosso tempo é o assim chamado fenômeno da explosão informacional, a grande quantidade de informações produzidas e disponibilizadas por diferentes atividades sociais, dificultando sua identificação, acesso e utilização. Na emergência da sociedade da informação, o valor desta como insumo para qualquer atividade, seja ela uma decisão econômica, um processo cultural ou de ensino/aprendizagem, uma pesquisa científica ou tecnológica, está relacionado diretamente ao seu potencial de orientar de forma econômica o dispêndio de energia para a realização desta atividade. Para que possa realizar todo este potencial, a informação relevante para um dado problema deve estar disponível no tempo certo. De nada adianta a informação existir, se quem dela necessita não sabe da sua existência ou se ela não puder ser encontrada.

Esta situação assume proporções alarmantes com o surgimento da Internet. Uma notícia divulgada no *Boletim Edupage* em português, de 05/04/98, publicado pela RNP, levanta o problema da busca de informações na Internet e comenta os resultados de um estudo sobre o desempenho dos assim chamados "mecanismos de busca":

"*ACHANDO UMA AGULHA (OU 7.079 PÁGINAS EM UMA AGULHA) NA WEB*

Um estudo realizado pelo NEC Research Institute afirma que a Internet explodiu para mais de 320 milhões de páginas na Web, uma estimativa que não inclui milhões de páginas com acesso protegidas por senhas ou "muros de pesquisa" que bloqueiam acesso a browsers ou mecanismos de busca. O estudo indica que a pesquisa do mecanismo de busca HotBot tem o índice mais abrangente da Web, mas, ainda assim, cobre apenas 34% das páginas indexáveis. A cobertura de alguns dos outros mecanismos de busca inclui: AltaVista (28%); Northern Light (20%); Excite (14%); Lycos (3%)."

Uma novidade em termos de mecanismos de busca que parece alentadora são os projetos CLEVER e GOOGLE (http://www.google.com), com suas propostas de ordenamento e priorização (*ranking*) dos resultados de uma busca, tendo por base os *sites* mais referenciados por *links* a partir de outros (Clever, 1999).

A enorme quantidade de informação armazenada e disponibilizada via Internet torna cada vez mais crítico o problema da *identificação* de informação relevante, assim chamada *information discovery*. Diferentes estratégias para fazer frente à explosão informacional trazida pela Internet podem hoje ser divisadas, como os mecanismos de busca gerais (AltaVista, Excite, Lycos, Infoseek, Yahoo e outros), os localizadores de informações especializados, como o GILS (http://www.usgs.gov/gils/) ou portais temáticos como o SIGNPOST (http://www.signpost.org) americano, o OMNI (http://www.omni.ac.uk) e o SOSIG (http://www.sosig.ac.uk)ingleses, o PROSSIGA – Comunicação e Informação para a Pesquisa – (http://www.prossiga.br) ou LIS – Localizador de informações em Saúde – (http://www.bireme.br) no Brasil.

Ambas as alternativas, os mecanismos de busca gerais e os portais temáticos oferecem soluções parciais para a localização de informações na Internet, *principalmente as de interesse para C&T*. As deficiências dos mecanismos de busca são já bastante conhecidas e discutidas na literatura (Sneiderman,1997). Entre as principais, pode-se citar as seguintes: baixa qualidade da indexação, por ser feita automaticamente, que resulta em grande quantidade de informações recuperadas, a maioria sem relevância (em termos de recuperação de informação, oferecem alta revocação, mas baixa precisão); cobertura parcial da Internet; as ferramentas de busca não são especializadas; indexam páginas HTML isoladas, e não recursos; além disto, grande quantidade de informações disponíveis na Internet estão sob a forma de registros contidos em bases de dados, que ficam assim "escondidas"; estes registros são acessados somente por meio das interfaces destas bases de dados, o que pressupõe uma interação entre um usuário humano com a base de dados e, portanto, ficam inacessíveis aos programas robôs.

Por sua vez, as bibliotecas digitais e os portais temáticos isolados resolvem somente em parte o problema do acesso a recursos informacionais de interesse para C&T publicados na rede: continuam limitadas ao "seu" acervo. Descobrir, avaliar, tratar e indexar estes recursos por profissionais de informação é caro e lento. Estes recursos estão sendo criados em número crescente, armazenados em diferentes servidores isolados, operados por interfaces de busca diferentes, o que obriga um usuário a uma dispendiosa busca, *site* a *site*, para encontrar informações relevantes.

Do ponto de vista de um usuário acadêmico ou pesquisador, o interessante e confortável seria poder submeter sua necessidade de informação e interagir com uma *única interface* e ter retornadas informações de diferentes fontes, de forma consolidada. Este é um tema que, sob diferentes denominações, está sendo cada vez mais discutido: *digital libraries federation* e *distributed archives* (Liu, 2001), *confederated digital libraries* (Leiner, 1998), *distributed subject gateways* (IMesh, 1999), *networked digital library* (Davis, 1995), *multiple information sources* (Paepcke, 2000) *cross-searching*, *heterogeneous distributed databases*, *metasearches* (Gravano, 1996) etc.

A questão começa a ser levantada na *An Intenational research agenda for digital libraries*, de 1998 – um agenda de pesquisa conjunta da NSF (EUA) e União Européia – em três grupos de trabalho temáticos: *global resource discovery*, *interoperability* e *metadata*. Hoje é endereçada diretamente por diferentes iniciativas de pesquisa, como a Joint NSF – JISC International Digital Libraries Research Programme, como o consórcio Imesh (http://www.desire.org/html/subjectgateways/community/imesh/), o Scout Project, e por iniciativas práticas como OpenArchives Initiative, Arc - Cross Archive Searching Service, NCSTRL (Univ. Cornell, EUA), NDLTD (Virginia Tech, University, EUA), Digital Library Federation (consórcio de bibliotecas digitais americanas), ROADS (UKOLN JISC, Inglaterra). As diferentes denominações sob as quais o tema aparece na literatura convergem para os conceitos de integração e interoperabilidade entre bibliotecas digitais, que consistiria na *possibilidade de um usuário realizar buscas a recursos informacionais heterogêneos, armazenados em diferentes servidores na rede, utilizando-se de uma interface única sem tomar conhecimento de onde nem como estes recursos estão armazenados*.

Hoje, no cenário mundial, identificam-se várias alternativas de interoperabilidade e acesso integrado a recursos informacionais heterogêneos publicados na rede. Estas podem ser agrupadas basicamente em duas alternativas, embora ainda não tenha se fixado uma nomenclatura amplamente aceita: *buscas distribuída a diferentes servidores e busca em uma base de metadados centralizada*. Em ambas as alternativas, *o usuário interage com uma única interface Web, de onde é submetida a busca.*

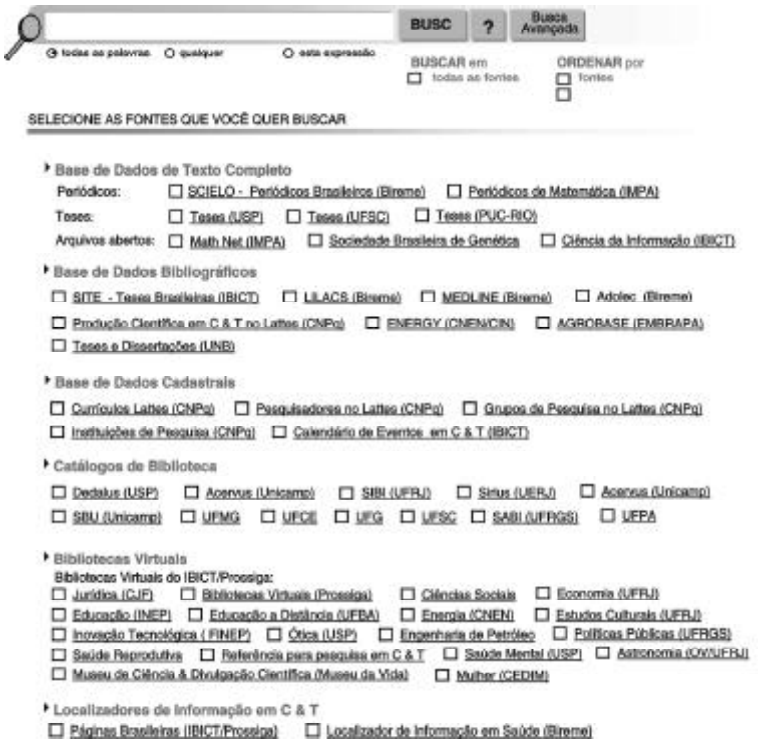
Na primeira alternativa, a interface de busca distribui a consulta (*broadcast search*) a diferentes *sites*, segundo um protocolo padrão, identificados pela interface como capazes de fornecer respostas satisfatórias, e os resultados são consolidados e integrados. Exemplo típico desta alternativa é o conhecido protocolo Z39.50, usado para proporcionar interoperabilidade entre catálogos automatizados de bibliotecas. Esta alternativa apresenta as seguintes vantagens: novos provedores de dados podem ser acrescentados, desde que sejam aderentes ao padrão ou padrões utilizados, com reconfiguração mínima. Como desvantagens, pode-se apontar que provedores de dados precisam rodar *software* servidor do protocolo padrão para serem consultáveis. Alguns destes *softwares* consomem muitos recursos por parte dos provedores de dados, como no caso do Z39.50 (Troll, 2001). Em alguns casos, são necessários servidores especializados, como os servidores de índices, que roteiam as consultas para os servidores capazes de atendê-las. Alguns dos padrões tecnológicos utilizados são os seguintes: Z39.50 (ISO/NISO), Whois + +, LDAP, CIP, SDLIP, DIENST. Esta alternativa é utilizada nos seguintes sistemas: NCSTRL (University of. Cornell, EUA), NDLTD – Networked Digital Library of Theses and Dissertations - *federated search* (Powel, 1998), California Digital Library (http://www.cdlib.org/), Berkeley Environmental Digital Library, EUA, ROADS, ISAAC/SCOUT Project (University of Stanford, EUA).

Na segunda alternativa, metadados referentes a documentos eletrônicos são coletados periodicamente, alimentando uma base comum de metadados sobre a qual são realizadas as buscas. Este esquema é bastante conhecido da colaboração/cooperação entre as instituições participantes para manutenção do Catálogo Coletivo/base de metadados centralizada. Dentro desta alternativa variam os esquemas de centralização destes metadados. O esquema do envio de metadados por parte das instituições cooperantes é mais tradicional e largamente conhecido pela comunidade de informação, inclusive a brasileira, em sistemas/bases de dados como LILACS/BIREME, SITE/IBICT, INIS/CIN. Este esquema apresenta as seguintes vantagens: novos provedores de dados podem ser acrescentados, desde que sejam aderentes ao padrão ou padrões utilizados; melhor desempenho em função da consulta ser na base de metadados local do provedor de serviços. Como desvantagens este esquema apresenta: manutenção pelo provedor de dados da base comum de metadados; grande ônus administrativo e gerencial por parte do provedor de serviços para sincronizar o envio dos dados por parte dos provedores de dados e processá-los para incluí-los na base comum de metadados; necessidade de sincronização entre os dados armazenados nos provedores de dados e os metadados coletados pelo provedor de serviços.

O esquema de coleta automática de metadados (*harvesting*) é mais recente: metadados de diversos provedores de informação tornam-se ¨visíveis¨ através de protocolos padronizados e são coletados automaticamente de forma periódica e armazenado em um *data warehousing*, ou base centralizada de metadados, onde são efetuadas as buscas de forma integrada. Vantagens deste esquema são as seguintes: novos provedores de dados podem ser acrescentados, desde que sejam aderentes ao padrão utilizado; melhor desempenho em função de a consulta ser na base de metadados local do provedor de serviços. Desvantagens: manutenção pelo provedor de dados da base comum de metadados; necessidade de sincronização entre os dados armazenados nos provedores de dados e os metadados coletados pelo provedor de serviços (Liu, 2001). Padrões utilizados: OpenArchives Harvest Protocol, Open Archives Metadata Set, Dublin Core, XML. Exemplos são as experiências do sistema MARIAN

(Virginia Technical University), do portal da NDLTD (Suleman, 2001), da OpenArchives Initiative, que, através do protocolo OAI harvest protocol, permite colheita (*harvesting*) de metadados, do Arc – Cross Archive Searching Service, primeiro serviço a proporcionar acesso integrado a diversos arquivos eletrônicos (http://www.arc.cs.odu.edu).

FIGURA1
Proposta de interface de busca heterogênea para a BDB



MODELO DE INTEROPERABILIDADE DA BDB

O projeto de implantação da BDB no país pressupõe forte ação de integração, liderada pelo IBICT, dos mais importantes provedores de conteúdos e de serviços de informação para C&T do país. Esta integração se dará em torno das questões prioritárias para a BDB, que são a publicação e a disponibilidade de textos completos e outros objetos digitais na Internet e a interoperabilidade entre os diversos sistemas/serviços de informação participantes através de um portal único de acesso, preservando-se a independência e peculiaridades de cada sistema/serviço participante. Para conseguir cooperação dos eventuais provedores de dados, o conjunto de metadados, a configuração, os padrões e procedimentos por parte dos provedores de dados para garantir interoperabilidade com a BDB deverão ser os mais simples e menos onerosos para os provedores de dados, garantindo sua máxima independência.

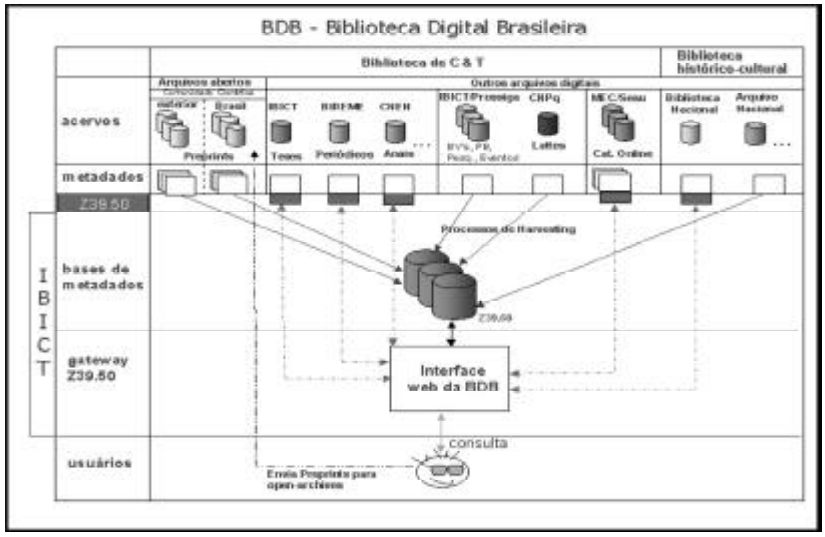
Apesar do avanço acelerado das tecnologias Web e das tecnologias de informação e comunicação, o projeto da BDB prevê a utilização de tecnologias consolidadas, cujo grau de estabilidade e confiabilidade tenham sido comprovados no país e no exterior, e que, prioritariamente, tenham sido aplicados nas principais experiências internacionais análogas à da proposta da BDB. Além do mais, essas tecnologias devem ser passíveis de serem implantadas, mantidas e, quando necessário, alteradas pelo corpo técnico do IBICT. Sempre que possível, serão adotadas tecnologias abertas, não proprietárias, que permitam garantir o grau de interoperabilidade desejável entre os diversos partícipes da BDB e que possam facilmente ser repassadas aos parceiros do Projeto. Esta opção tem como objetivo a disseminação no país de um corpo de protocolos e padrões que possam ser adotados pelos futuros integrantes da BDB e por outros sistemas que queiram aderir a sistemas/redes internacionais.

No intuito de disseminar mais facilmente as tecnologias de publicação na Internet entre as diversas comunidades de conhecimento, sempre que possível serão adotados *software* de domínio público, preservada a qualidade, documentação e mantenabilidade dos mesmos. Isto se deve à constatação de que há uma diversidade surpreendente de *software* livres, confiáveis e de qualidade que estão sendo adotados por instituições importantes na área de informação.

O modelo de interoperabilidade proposto para a BDB aproxima-se bastante dos modelos do portal da NDLTD (Suleman, 2001) e do Arc – Cross Archive Searching Service (http://www.arc.cs.odu.edu). Ambos os sistemas fazem *harvesting* de metadados de provedores de dados, alimentando uma base de dados central de metadados. O portal da BDB na Internet será a materialização da Biblioteca Digital Brasileira em C&T. Trata-se de um *site* que, através de diferentes mecanismos de interoperabilidade, possibilitará ao pesquisador acesso unificado e integrado a diferentes recursos infor-macionais brasileiros de interesse para C&T, heterogêneos e distribuídos, sem a necessidade de navegar e consultar cada recurso individualmente. A figura 1 dá uma idéia da proposta do portal da BDB.

Entre estes recursos informacionais se incluirão, já na versão inicial do portal, periódicos eletrônicos brasileiros que fazem parte do portal Scielo, mantido pela Bireme, anais eletrônicos de eventos brasileiros em C&T a serem disponibilizados pela CNEN/CIN, bancos de teses eletrônicas hoje já existentes na USP, Unicamp, UFSC, PUC-Rio, ENSP/Fiocruz, repositórios de *e-prints* brasileiros que começam a ser dissemindos na Internet, como o do Impa. Na figura 2, é mostrado o modelo geral de interoperabilidade da BDB.

FIGURA 2



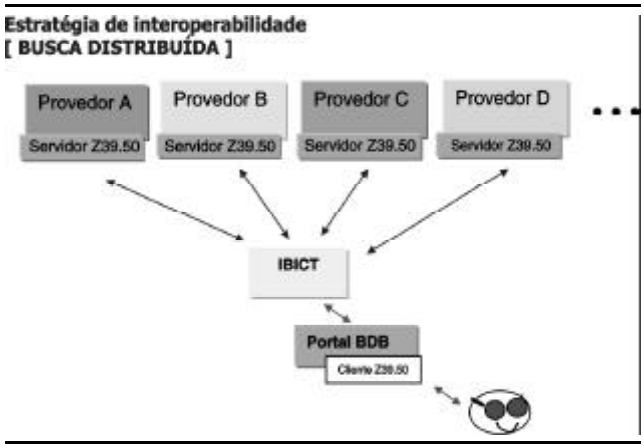
Em um ambiente de publicação e acesso a documentos publicados na rede como o da BDB, a OpenArchives Initiative reconhece dois atores institucionais fundamentais: provedores de dados e provedores de serviço. Estas definições serão utilizadas aqui para explicitar a solução de interoperabilidade adotada, com base no que é estabelecido na OpenArchives Initiative:

Provedores de dados: de uma forma ampla, seriam todas as instituições brasileiras que possuem *site* Internet que disponibiliza documentos eletrônicos em C&T. Eventualmente, este *site* abriga também um ambiente de submissão/publicação de documentos em texto completo; um autor registra seu documento no *site* através de um conjunto de metadados e, opcionalmente, armazena aí seu documento em formato eletrônico. O *site* do provedor de dados provê facilidades de busca para acesso aos documentos nele armazenados e, caso seja aderente ao padrão OpenArchives Harvest Protocol, permitirá também que os metadados dos documentos do seu acervo sejam visíveis a um programa de *harvest*. Exemplos típicos seriam o SCIELO, os diversos arquivos eletrônicos existentes no mundo, como o CogPrints (http://cogprints.soton.ac.uk/), ou o arquivo aberto do IMPA (http://www.preprint.impa.br/indexEngl.html).

Provedor de serviços: instituições que provêem serviços de valor agregado sobre documentos eletrônicos disponibilizados por um ou mais provedor de dados. Exemplos destes serviços seriam a montagem de bases de dados qualificadas, o acesso unificado a documentos armazenados em diferentes provedores de dados, revisão e avaliação de documentos publicados em um ou mais provedores de dados, *linkagem* de recursos informacionais. Exemplos típicos seriam o serviço Arc (http://arc.cs.odu.edu/) que provê acesso unificado a diferentes arquivos abertos e o objeto deste projeto, a própria BDB.

Dado o papel integrador da proposta da BDB, seu modelo de interoperabilidade se baseia em dois elementos: mecanismos de submissão de consultas, a partir da interface única do portal, aos diferentes recursos informacionais que comporão a BDB e conjunto de metadados que descreverão e fornecerão uma visão unificada dos diferentes conjuntos de documentos.

FIGURA 3



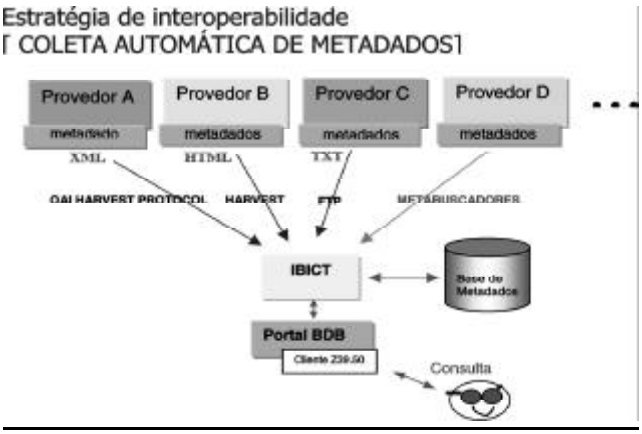
Com relação aos mecanismos de submissão de consultas, a BDB deverá incorporar as principais alternativas tecnológicas discutidas anteriormente, buscas distribuídas e busca em uma base centralizada de metadados obtida mediante coleta automática (*harvesting*). Para isso, o portal da BDB disporá de um programa cliente Z39.50, que lhe permitirá acesso integrado por meio da distribuição de consultas aos catálogos na Internet das principais bibliotecas universitárias do país e estrangeiras servidas pelo protocolo Z39.50. Qualquer outro recurso informacional na Internet que seja servido por este protocolo também poderá ser acessado do portal da BDB, como, por exemplo, um arquivo de *e-prints* ou o Scielo, o qual planeja implementar um servidor Z39.50 para sua base. Esta opção é ilustrada na figura 3.

Além de integrar, via consultas distribuídas, os recursos informacionais servidos pelo protocolo Z39.50, a BDB manterá em seu *site* uma base comum de metadados, obtida pelo processo de *harvesting* dos metadados dos recursos/serviços de informação que não tiverem servidor Z39.50. Estes recursos/serviços serão objeto de coleta automática periódica, usando o OAI Harvesting protocol dos provedores de dados compatíveis com este protocolo. Outras soluções foram pensadas de modo a não onerar tecnicamente provedores de dados não compatíveis com este protocolo, como coleta de metadados via FTP em arquivos HTML ou arquivos texto. Os formatos de arquivos passíveis de coleta automática são ilustrados na figura 4.

FIGURA 4
Solução Técnica
[PROCESSOS DE HARVESTING]

Estes metadados serão processados e armazenados na base comum de metadados, mantida no *site* da BDB; esta base, por sua vez, contará também com um servidor protocolo Z39.50, que a tornará acessível a partir do portal da BDB como qualquer outro recurso servido por este protocolo. Arquivos acadêmicos de *preprints* eletrônicos que poderão ser implantados em departamentos de instituições de ensino superior, institutos de pesquisa, sociedades científicas, periódicos eletrônicos, como o próprio *Ciência da Informação* do IBICT, ou projetos específicos interinstitucionais, como Projeto Genoma, poderão se integrar à BDB segundo esta opção. Ela é ilustrada na figura 5.

FIGURA 5
Estratégia de interoperabilidade
[COLETA AUTOMÁTICA DE METADADOS]



O segundo elemento do esquema de interoperabilidade da BDB é o conjunto de metadados. Ele deverá contemplar e integrar, em uma descrição unificada, as diferentes tipologias de documentos originários dos diferentes recursos informacionais que comporão a BDB. A referência emergente nesta área, avalizada pela comunidade de informação, é Dublin Core Element Set. É resultado de intenso trabalho de discussão e padronização em nível internacional, mantida por um ativo grupo e fórum internacionais, a Dublin Core Metadata Initiative, que já realizou diversos encontros; é usado em diferentes sistemas, inclusive na OpenArchives Initiative e se encontra em pleno desenvolvimento.

O conjunto de metadados Dublin Core é composto de 13 elementos descritivos (Dublin Core, 1999), suporta qualificadores para especificar o significado de um elemento (Dublin Core, 2000) e pode ser codificado em formatos como HTML e XML (Cox, 2000), (Beckett, 2000), (Beckett. 2001). Como é explicitado na própria proposta Dublin Core, o conjunto de metadados deve ser tão simples e intuitivo a ponto de permitir que o próprio autor descreva seu trabalho. É exatamente assim que funcionam os ambientes de submissão de trabalhos: ao submeter seu trabalho, o autor preenche um formulário com os metadados pertinentes.

Dada a sua característica integradora de diferentes recursos informacionais heterogêneos, que armazenam diferentes tipos de documentos, a BDB usará o conjunto Dublin Core,

expandido com qualificadores para suportar características especiais de alguns tipos de documentos. Detalhes da estrutura de metadados da BDB serão objeto de um trabalho posterior.

CONCLUSÕES

As experiências internacionais em torno da questão das publicações na rede e da interoperabilidade entre de bibliotecas digitais são bastante recentes, contemporâneas mesmo; as soluções para estes problemas são hoje o foco das maiores atenções por parte dos pesquisadores de ciência da informação e dos diversos sistemas de informação em C&T. Apesar de toda esta ebulição, está patente que já existe um conjunto de padrões e tecnologias maduros e consolidados que são bases para vários sistemas de informação importantes já em operação pelo mundo. As questões de interoperabilidade endereçadas pelo projeto da BDB são ainda bastante restritas, limitadas a um enfoque tecnológico. A interoperabilidade entre recursos informacionais heterogêneos na Internet tem várias outras dimensões – semântica, política/humana, entre comunidades, internacional, interlingüística (Powell, 1998) –, como alerta Miller.

O sucesso do projeto em uma perspectiva a médio e longo prazo vai depender de o IBICT adotar uma nova postura institucional com relação ao acompanhamento das tendências e padrões tecnológicos envolvidos, à pesquisa, ao desenvolvimento e à adaptação de tecnologias. Este quadro é tão amplo que seria impossível ao IBICT acompanhá-lo sozinho. Para fazer frente a este desafio, o IBICT deve assumir o papel de articulador entre diferentes parceiros nacionais e a comunidade acadêmica para um trabalho conjunto em torno de uma agenda de pesquisas que inclua questões de interesse do projeto da BDB. Para isso, o projeto sugere:

"Acompanhamento e articulação com os fóruns internacionais que discutem questões como metadados, interoperabilidade, publicações na rede, comunicação científica via rede, preservação de documentos eletrônicos, direitos autorais em documentos eletrônicos, *linkagem* de recursos informacionais etc." como World Wide Web Consortiun, Dublin Core Metadata Initiative, Open Archives Initiative, Digtal Library Federation. Com o objetivo de acompanhar o desenvolvimento das tecnologias associadas a bibliotecas digitais e o desenrolar das controvérsias sobre estes assuntos e seus desdobramentos, é necessária uma aproximação com as organizações e/ou redes internacionais que operam e/ou desenvolvem experiências com bibliotecas digitais heterogêneas distribuídas. É também de grande importância a participação do IBICT nos principais fóruns internacionais que discutem as questões relacionadas ao tema. Sugere-se:

– propor uma agenda incluindo os temas de pesquisa citados anteriormente à comunidade de pesquisas brasileira e uma linha de fomento correspondente, envolvendo universidades, institutos de pesquisa e instituições parceiras;

– propor um fórum nacional sobre o tema amplo de bibliotecas digitais que sirva para discussão e troca de experiências;

– propor um programa de formação de quadros" (IBICT, 2001, p.16).

O projeto da BDB, embora tenha um compromisso pragmático com a prestação de serviços à comunidade acadêmica brasileira, coloca também questões relativas ao planejamento de ICT no país. Desde fins da década da de 80, com a Ação Programa de Informação em Ciência e Tecnologia (Brasil, 1984) e com os PADCTs (Brasil, 1985), não surgiram documentos abrangentes de planejamento de ICT no Brasil. Os atuais Livros Verdes, o da Sociedade da Informação (2000) e o de Ciência, Tecnologia e Inovação (2001), não contemplam questões relativas à ICT.

Hoje, no entanto, a comunicação científica é cada vez mais fortemente dependente das tecnologias de informação. O projeto da BDB menciona itens que constituiriam a infra-estrutura necessária para viabilizar um ambiente de informação integrado. Planejar e implantar esta infra-estrutura seria claramente papel do IBICT. Um ambiente como este garantiria aos usuários "a informação na ponta dos dedos" e simultaneamente ampla visibilidade à produção acadêmica brasileira. Componentes desta infra-estrutura seriam: ambientes para submissão e armazenamento de documentos eletrônicos, mecanismos de *linkagem* de documentos eletrônicos entre si, de modo que um usuário pudesse ter acesso imediatamente a referências e fontes citadas em um documento eletrônico, endereços eletrônicos persistentes, sem os problemas de *links* inválidos, base de autoridades, linguagens de descrição, esquemas de classificação temática e sistemas de metadados para interoperabilidade entre sistemas e descoberta de informações, armazenamento e preservação de documentos eletrônicos por longo tempo.

Um empreendimento amplo e com um caráter integrador como a proposta da BDB deve ser gerido por um comitê dirigente que inclua as parcerias estratégicas do IBICT neste projeto e representantes da comunidade de C&T. Este Comitê Dirigente deve se reunir periodicamente para analisar e aprovar o Relatório de Atividades da BDB e seu Plano de Trabalho para o próximo período. Entre as reuniões do comitê dirigente, a BDB será gerida por um comitê executivo.

Embora os objetivos da BDB sejam ambiciosos, mostram-se também plenamente viáveis em termos tecnológicos; a implantação da BDB pode se iniciar com um investimento baixo, beneficiando-se da experiência e das metodologias já desenvolvidas nos principais centros internacionais. Estes objetivos também são necessários para que a comunidade acadêmica brasileira possa dispor de uma infra-estrutura informacional compatível com os padrões internacionais. Isto permitirá ao país se inserir plenamente nos fóruns científicos internacionais dentro do paradigma atual da comunicação científica. E permitirá principalmente que os sistemas de informação brasileiros se integrem ao fluxo mundial de informações, dando maior visibilidade à produção brasileira em C&T. O interesse do pesquisador é conseguir a maior visibilidade possível para sua produção acadêmica. Os serviços de informação em C&T, entre eles as bibliotecas digitais como a BDB, devem ajudar os autores dos documentos neles armazenados a obter a máxima visibilidade da sua produção, adotando mecanismos que maximizem a integração e interoperabilidade amplas entre serviços de informação.

REFERÊNCIAS BIBLIOGRÁFICAS E FONTES DE INFORMAÇÃO

ALLEN, J.; MEALLING, M. The architecture of the Common Indexing Protocol (CIP). IETF Internet draft. Disponível em: <http://search.ietf.org/internet-drafts/draft-ietf-cip-arch-02.txt>. Acesso em: 4 out. 2001.

BECKETT, Dave; MILLER, Eric; BRICKLEY, Dan. An XML encoded of simple Dublin core metadata. [S. l.] : Dublin Core Metadata Initiative, 2001. Disponível em: <http://dublincore.org/documents/2001/04/11/dcmes-xml>. Acesso em: 6 jun. 2001.

BECKETT, Dave; MILLER, Eric; BRICKLEY, Dan.. Using Dublin core in XML. [S. l.] : Dublin Core Metadata Initiative, 2000. Disponível em: <http://www.purl.org/dc/documents/wd/dcmes-xml-20000714.htm>. Acesso em: 1 jun. 2001.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The semantic web. *Scientific American*, New York, n. 5, May 2001. Disponível em: <http://www.scian.com/2001/0501issue/0501berners-lee.html>. Acesso em: 24 maio 2001.

BRASIL. Ministério da Ciência e Tecnologia. PADCT: documento base [e] documentos sínteses dos subprogramas. Brasília : CNPq, 1985. 97 p.

BRASIL. Presidência da Republica. Secretaria de Planejamento. III PBDCT: III Plano Básico de Desenvolvimento Científico e Tecnológico: informação em ciência e tecnologia. Brasília : CNPq, 1984, 69 p. (Ação Programada em Ciência e Tecnologia).

BRYAN, Martin. An introduction to Extensible Markup Language (XML). [S. l.] : SGML Center, 1997. Disponível em: <http://www.personal.u-net.com/~sgml/xmlintro.html>. Acesso em: 21 out. 2000.

CIÊNCIA, tecnologia e inovação: desafio para a sociedade brasileira: livro verde. Brasília : Ministério da Ciência e Tecnologia, Academia Brasileira de Ciências, 2001.

CLEVER PROJECT. Hypersearching the web. *Scientific American*, New York, n. 6, 1999. Disponível em: <http://www.sciam.com/1999/0699issue/0699raghavan.html>. Acesso em: 31 maio 1999.

COX, Simon; MILLER, Eric; POWELL, Andy. Recording qualified Dublin core metadata in HTML meta elements. [S. l.] : Dublin Core Initiative, [2001?]. Disponível em: <http://dublincore.org/documents/2000/08/15/dcq-html/>. Acesso em: 6 jul. 2001.

DAVIS, James R. Creating a networked computer science technical report library. *D-Lib Magazine*, Sept. 1995. Disponível em: <http://www.dlib.org/dlib/september95/09davis.html>. Acesso em :16 jul. 2001.

DIENST architecture summary description. Ithaca : Cornell University, 2000. Disponível em: <http://www.cs.cornell.edu/cdlrg/dienst/architecture/architetucture.htm>. Acesso em: 5 out. 2001.

DIENST protocol version 4.1. Draft. Ithaca : Cornell University, 1998. Disponível em: <http://www.cs.cornell.edu/NCSTRL/protocol.htm>. Acesso em: 5 out. 2001.

DIENST software summary description. Ithaca : Cornell University, 2000. Disponível em: <http://www.cs.cornell.edu/cdlrg/dienst/software/DienstSoftware.htm>. Acesso em: 5 out. 2001.

DUBLIN core metadata elements set, version 1.1: reference description. [S. l.] : Dublin Core Initiative, 1999. Disponível em: <http://dublincore.org/documents/1999/07/02/dces/>. Acesso em: 6 jun. 2001.

DUBLIN core qualifiers. [S. l.] : Dublin Core Initiative, 2000. Disponível em: <http://purl.org/dc/documents/rec/dcmes-qualifiers-20000711.htm>. Acesso em: 5 out. 2001.

EXTENSIBLE markup language (XML). [S. l.] : World Wide Web Consortium, 2000. Disponível em: <http://www.w3.org/XML/>. Acesso em: 30 jan. 2001.

GINSPARG, P. Winners and losers in the global research village. In: CONFERENCE ON ELECTRONIC PUBLISHING IN SCIENCE, 1996, Paris. Proceedings... Disponível em: <http://xxx.lanl.gov/blurb/pg96unesco.html>. Acesso em: 5 out. 2001.

GONÇALVES, Marcos Andre; FRANCE, Robert K.; FOX, Edward A. MARIAN: flexible interoperability for federated digital libraries. In: ECDL-01: EUROPEAN CONFERENCE ON RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, 5., 2001, Darmstadt. Proceedings... Disponível em: <http://www.dlib.vt.edu/reports/ecdl2001_8.pdf>. Acesso em: 5 out. 2001.

GREENSTEIN, Daniel. The arts and humanities data service: three years' on. *D-Lib Magazine*, Dec. 1998. Disponível em: <http://www.dlib.org/dlib/december98/greensteisn/greenstein.html>. Acesso em: 1 jul. 2001.

THE IMESH toolkit: an architecture and toolkit for distributed subject gateways. *International Digital Libraries*, n. 19, Jan. 1999. NSF 99-6.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA (Brasília, DF). *Projeto técnico da biblioteca digital brasileira em C&T*. Brasília, 2001. 40 p.

LAWRENCE, Steve. Free online availability substantially increases a paper's impact. Disponível em: http://wwwnature.com/nature/debates/e-access/Articles/lawrence.html. Acesso em: 10 jun. 2001.

LEVAN, Ralf. Dublin core and Z39.50. [S. l.] : Dublin Core Metadata Initiative, 1998. Disponível em: <http://dublincore.org/documents/1998/02/02/dc-z3950/>. Acesso em: 9 mar. 2001.

LIU, Xiaoming *et al*. Arc: an OAI service provider for digital library federation. *D-Lib Magazine*, v. 7, n. 4, Apr. 2001. Disponível em: <http://www.dlib.org/dlib/april01/liu/04liu.html>. Acesso em: 25 jul. 2001.

LYNCH, Clifford. The Z39.50 information retrieval standard. *D-Lib Magazine*, v. 7, n. 4, Apr. 2001. Disponível em: <http://www.dlib.org/dlib/april97/liu/04lynch.html>. Acesso em: 16 mar. 2001.

MEDEIROS, Norm. XML and the resource description framework: the great web hope. *Online*, v. 24, n. 5, Sept. 2000. Disponível em: <http://www.onlineinc.com/onlinemag/OL2000/medeiros9.html>. Acesso em: 9 jan. 2001.

MILLER, Paul. UK interoperability focus: Bath, UK, UKOLN. Disponível em: <http://www.ukoln.ac.uk/interop-focus/about/>. Acesso em: 3 out. 2001.

NETWORK WORKING GROUP. Architecture of the whois + + service. RFC 1835.. Disponível em: <http://src.doc.ic.ac.uk/computing/internet/rfc/rfc1835.txt>. Acesso em: 4 out. 2001.

PAEPKE, Andreas *et al*. Search middleware and the simple digital library interoperability protocol . *D-Lib Magazine*, v. 6, n. 3, Mar. 2000. Disponível em: <http://www.dlib.org/dlib/march00/paepcke/03paepcke.html>. Acesso em: 26 jul. 2001.

PAYETTE, Sandra; RIEGER, Oya Y. Z39.50: the user's perspective. *D-Lib Magazine*, Apr. 1997. Disponível em: <http://www.dlib.org/dlib/april97/cornell/04payette.html>. Acesso em: 8 set. 2000.

PINHEIRO, Lena Vânia Ribeiro; LOUREIRO, José Mauro Matheus. Traçados e limites da ciência da informação. *Ciência da Informação*, Brasília, v. 24, n. 1, p. 42-53, jan./abr. 1995.

POWELL, James. Multilingual federated searching across heterogeneous collections. *D-Lib Magazine*, Sept. 1998. Disponível em: <http://www.dlib.org/dlib/september98/power//09power.html>. Acesso em: 25 jul. 2001.

RIFKIN, Adam. A look at XML. WebDeveloper.com. Disponível em: <http://webdeveloper.internet.com/xml/xml_a_look_at_xml.html>. Acesso em: 5 set. 2000.

SHNEIDERMAN, Ben; BYRD, Don; CROFT, W. Bruce. Clarifying search: a user-interface framework for text searches. *D-Lib Magazine*, Jan. 1997. Disponível em: <http://www.dlib.org/dlib/january97/retrieval/01sneiderman.html>. Acesso em: 5 out. 2001

SOCIEDADE da informação no Brasil : livro verde. Brasília : Ministério da Ciência e Tecnologia, 2000. (organizado por Tadao Takahashi.).

SOMPEL, Herbert van de; LAGOZE, Carl. The Santa Fe convention of the open archives initiative. *D-Lib Magazine*, v. 6, n. 2, Feb. 2000. Disponível em : <http://www.dlib.org/dlib/february00/vandesompel-oai/vandesompel-oai.html>. Acesso em: 5 out. 2001

SULEMAN, Hussein, *et al.* Networked digital library of theses and dissertations: bringing the gap for global access: part 1: mission and progress. *D-Lib Magazine*, v. 7, n. 9, Sept. 2001. Disponível em: <http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html>. Acesso em: 19 set. 2001.

SULEMAN, Hussein, *et al*. Networked digital library of theses and dissertations: bringing the gap for global access: part 2: services and research. *D-Lib Magazine*, v. 7, n. 9, Sept. 2001. Disponível em: <http://www.dlib.org/dlib/september01/suleman/09suleman-pt2.html>. Acesso em: 19 set. 2001.

TROLL, Denise; MOEN, Bill. Report to the DLF on the Z39.50 Implementers' Group. DLF : 2001. Disponível em: <http://www.diglib.org/architectures/zig0012.htm>. Acesso em: 22 ago. 2001.

XML schema part 0: primer. W3C Working Draft, 7 April 2000. Disponível em: <http://www.w3.org/TR/2000/WD-xmlschema-0-20000407/>. Acesso em: 5 set. 2000.

Z39.50 profiles. Disponível em: <http://locweb.loc.gov/z3950/agency/profiles/profiles.html>. Acesso em: 30 mar. 2001.

Artigo recebido em 10/11/2001.