Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

UM AMBIENTE BASEADO EM ARQUIVOS ABERTOS PARA INTEGRAÇÃO DE DADOS ECOLÓGICOS

Dissertação apresentada ao Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

EVANDRINO GOMES BARROS

Belo Horizonte
Agosto de 2005

Resumo

Por uma década, o Brasil tem realizado pesquisas ecológicas de longa duração, envolvendo dados com aspectos geográficos e históricos, coletados a partir de uma extensa rede de sítios do Programa PELD (Pesquisas Ecológicas de Longa Duração). Entretanto, essa rede não possui sistemas para coletar ou compartilhar esses dados, bem como para analisá-los apropriadamente. Além disso, na maioria dos sítios, esses dados estão armazenados em fontes textuais e não estruturadas. Como solução para esse problema, esta dissertação propõe um ambiente, baseado em bibliotecas digitais, para integração de dados dos vários sítios da rede. Esse ambiente se utiliza de padrões abertos de interoperabilidade, como o protocolo OAI-PMH, para coleta de dados dos diversos sítios e sua inclusão no repositório de uma biblioteca digital, denominada BDiG-PELD (Biblioteca Digital Georreferenciada do PELD), que oferece serviços de busca e navegação, combinados com facilidades de georreferenciamento. Esse ambiente inclui ainda uma interface de entrada de dados descentralizada, por meio da qual os pesquisadores preenchem seus dados de coleta de campo em arquivos textuais, que são então processados por sistemas locais e transformados em metadados EML *(Ecological Metadata Language)* para posterior colheita pela BDiG-PELD. Para demonstrar a efetividade desse ambiente, foi construído um protótipo da BDiG-PELD usando dados de um dos sítios da rede PELD. A partir do processo de carga de dados, foi então realizada uma avaliação experimental, do ponto de vista de usabilidade e qualidade dos dados, da interface de entrada de dados, sendo seus resultados analisados. Os resultados obtidos foram bastantes satisfatórios e confirmam que o ambiente da BDiG-PELD é uma solução prática e econômica para integração de dados dos sítios de pesquisa do Programa PELD.

Abstract

For a decade, Brazil has been maintaining a long-term ecological research program, involving heterogeneous data with important historic and geographic aspects, collected from a network of ecological sites of the PELD (Pesquisa Ecológicas de Longa Duração) program. However, this network has no system for collecting or sharing such data as well as for supporting data analyses. Moreover, in most of these sites, such data are stored in textual and unstructured sources. As a solution to this problem, this dissertation proposes a digital-library-based environment for integrating data from this network of sites using open interoperability standards, such as OAI-PMH, for harvesting data from the several network sites and inserting them into the repository of a digital library, named BDiG-PELD, which offers searching and browsing services, combined with georeferencing facilities. This environment also includes a decentralized data input interface, through which PELD researchers fill their field sample data in text files that are processed by a local system and transformed into EML (Ecological Metadata Language) metadata which are then made available for harvesting by BDiG-PELD. To demonstrate the effectiveness of this environment, we have built a BDiG-PELD prototype using data from one of the PELD network sites. Based on the data input process, we have also conducted an experimental evaluation, focused on usability and data quality aspects, of the data input interface and analyzed its results. The results have been quite satisfactory and confirm that the BDiG-PELD environment is a practical and economical solution for integrating data from the PELD research sites.

Dedico este trabalho à minha esposa Fernanda, por tanto carinho e paciência.

Agradecimentos

Numa jornada de tantas aulas, tarefas, seminários e horas de trabalho a fio, pude perceber que outras lições, muito especiais, foram a mim proporcionadas, de uma maneira mais informal, mas muito sólida. Por isso tenho tanto a agradecer.

Ao meu orientador Prof. Alberto Henrique Frade Laender por sua dedicação e por estar sempre pronto a educar. Agradeço também ao Prof. Marcos André Gonçalves por colaborações imprescindíveis.

Aos amigos do Laboratório de Bancos de Dados que além de uma convivência prazerosa e espirituosa, muito me ensinaram sobre determinação, respeito à diversidade, amizade e trabalho em equipe. Em especial à Karla e Ricardo, por tanto apoio.

Também, aos pesquisadores do Instituto de Ciências Biológicas da UFMG, integrantes do Programa de Pesquisas Ecológicas de Longa Duração, por tantas lições de interdisciplinaridade, sempre pautadas pela seriedade e colaboração.

Aos professores do Departamento de Ciência da Computação da UFMG, por tantos incentivos e ensinamentos sobre perseverança e disciplina. Por tanto conhecimento. Agradeço também aos funcionários do DCC-UFMG, sempre solícitos e cordiais.

Ainda agradeço a aqueles que mesmo não vinculados à vida acadêmica, ainda me proporcionam muito aprendizado.

Aos meus pais e irmãos, por uma formação que me deu autonomia emocional e intelectual e por tantas lembraças carinhosas.

Também, aos meus tios Eliseu e Luciana e primos Eliseu e Ana Cláudia, pais e irmãos mineiros.

Aos meus amigos e parentes, por tanto estímulo, confiança e apreço.

E, muito especialmente, à minha esposa Fernanda por tantas lições de paciência e dedicação. Por tanto carinho e confiança. Por tanto amor.

Paradoxalmente, ou não, saio de um mestrado em Ciência da Computação, mais humano e disposto a me educar, cada vez mais.

A todos vocês, agradeço muito.

Sumário

1 Introdução **1**
1.1 Motivação . . . 1
1.2 Caracterização do Problema . . . 2
1.3 Abordagem Proposta . . . 4
1.4 Contribuições . . . 5
1.5 Trabalhos Relacionados . . . 5
1.6 Organização da Dissertação . . . 7

2 Conceitos, Padrões e Arquiteturas para Interoperação de Dados **8**
2.1 Bibliotecas Digitais . . . 8
2.2 Arquivos Abertos . . . 9
2.3 Metadados Ecológicos . . . 12
2.4 Georreferencimento em Bibliotecas Digitais . . . 13

3 Ambiente BDiG-PELD **16**
3.1 Arquitetura do Ambiente . . . 16
3.2 Provedor de Dados . . . 18
3.2.1 Banco de Dados Local . . . 18
3.2.2 Interface de Entrada de Dados . . . 20
3.3 Provedor de Serviços . . . 22
3.3.1 Protocolo OAI-PMH . . . 22
3.3.2 Infraestrutura ODL . . . 22
3.4 Implementação da BDiG-PELD . . . 25

4 Serviços da BDiG-PELD **27**
4.1 Especificação de Requisitos . . . 27
4.1.1 *Societies* . . . 27
4.1.2 *Structures* . . . 28
4.1.3 *Scenarios* . . . 29
4.1.4 *Spaces* . . . 30
4.1.5 *Streams* . . . 30
4.2 Serviço de Busca . . . 30
4.3 Serviço de Navegação . . . 32

4.3.1 Módulo de Navegação 32
4.3.2 Interface de Navegação 33
4.4 Facilidades de Georreferenciamento 37
4.4.1 Locus 37
4.4.2 Interface de Georreferenciamento 38

5 Avaliação Experimental **44**
5.1 Descrição do Experimento 44
5.2 Resultados 46
5.2.1 Aspectos de Usabilidade 46
5.2.2 Qualidade dos Dados 47
5.3 Discussão 48

6 Conclusões e Trabalhos Futuros **49**

Referências Bibliográficas **52**

Lista de Figuras

2.1 Abordagens de interoperação 11
2.2 Principais elementos da EML 14
2.3 Trecho de documento com metadados EML *(Ecological Metadata Language)* . . . 15

3.1 Arquitetura do ambiente BDiG-PELD 17
3.2 Principais tabelas do conjunto de tabelas básicas 19
3.3 Tabelas básicas e tabelas específicas 20
3.4 Interface do Sistema de Informação para o Sítio 4 21
3.5 Arquivo *csv* com dados de nutrientes de uma lagoa 23
3.6 Trecho de um documento OAI-PMH 23
3.7 Página Web para submissão de arquivos *csv* 24

4.1 Pagina Inicial da interface da BDiG-PELD com formulário de busca 31
4.2 Segundo formulário de busca da interface da BDiG-PELD 32
4.3 Página com resultados de uma busca 33
4.4 Dados gerais de uma coleta 34
4.5 Dados da classificação taxonômica de uma coleta 35
4.6 Dados do projeto de pesquisa 36
4.7 Dados sobre atributos de coleta 37
4.8 Dados originais de uma coleta 38
4.9 Dados originais de uma coleta recuperados em formato *csv* 39
4.10 Esquema relacional do *data mart* do serviço de navegação 39
4.11 Página inicial do serviço de navegação 40
4.12 Navegação com várias subcategorias de taxonomia 40
4.13 Filtro a partir de uma instância da subcategoria espécie 41
4.14 Navegação com a combinação das categorias Taxonomia e Local 41
4.15 Navegação com a combinação das categorias Taxonomia, Local e Data 42
4.16 Lista de amostras obtidas através do Serviço de Navegação 42
4.17 Arquitetura do módulo Locus 43
4.18 Interfaces de busca e georreferenciamento 43

Lista de Tabelas

1.1 Sítios do Programa PELD 3

2.1 Verbos do protocolo OAI-PMH 10
2.2 Descrição do principais elementos da linguagem EML 13

3.1 Componentes ODL utilizados no ambiente da arquitetura 24
3.2 Principais programas da interface de acesso da BDiG-PELD 25
3.3 Quantidade de linhas de código 26

4.1 Exemplos de uso das perspectivas 5S 28

5.1 Resultados do experimento de entrada de dados 46
5.2 Avaliação da qualidade dos dados 48

Capítulo 1

Introdução

1.1 Motivação

Com o uso crescente da Web, várias comunidades de usuários passaram a interagir e promover o intercâmbio de conhecimento de maneira mais democrática e em maior escala. Algumas dessas comunidades apoiam-se na Web para o desenvolvimento de pesquisas científicas através da publicação de trabalhos científicos e dos resultados de seus experimentos [29, 37]. No entanto, tais comunidades poderiam ter maior interação se possuíssem sistemas de informação que facilitassem a publicação e acesso a dados científicos através de serviços especializados, atendendo a suas preferências e necessidades. Uma dessas comunidades é a que estuda a biodiversidade brasileira e que necessita de sistemas de informação na Web para integrar dados de várias fontes e acessá-los através de serviços de busca e navegação.

No entanto, essa comunidade, primeiro, precisa ter seus sistemas de informação de biodiversidade locais devidamente estruturados para que, posteriormente, possam fazer parte de uma rede integrada de pesquisa. Seguindo o curso da integração dos dados, é necessária uma infraestrutura computacional aberta e que respeite as particularidades de cada fonte de dados no tocante à diversidade dos dados e formas de armazenamento locais. Finalmente, já integrados, os dados devem ser acessados publicamente através de serviços que facilitem o acesso a eles.

Quanto à natureza, sistemas de informação em biodiversidade envolvem vários tipos de dados heterogêneos que incluem características ecológicas e geográficas. Entretanto, os sistemas desse tipo atualmente disponíveis oferecem um suporte limitado para gerenciamento integrado desses dados [20], não gerenciando, por exemplo, dados de localização juntamente com informações de biodiversidade. Isso muitas vezes leva os usuários à utilização alternada de sistemas de informação geográficos e sistemas de informação em biodiversidade, para, de alguma forma, combinar as informações extraídas deles. Sistemas de informação em biodiversidade com serviços que permitem a busca e interpretação, tanto textual quanto geográfica, podem tornar mais fácil as atividades dos pesquisadores da área.

Dentro desse contexto surgiu o Programa PELD-Pesquisas Ecológicas de Longa Duração[1],

[1] www.icb.ufmg.br/∼peld

um programa financiado pelo CNPq e que tem como objetivo promover a organização e consolidação do conhecimento existente sobre a composição e o funcionamento dos ecossistemas brasileiros, gerando informação e ferramentas para avaliar sua diversidade biológica por meio de pesquisas de longa duração. O Programa possui uma agenda comum cujos temas centrais são [44]:

1. Conservação da biodiversidade;
2. Padrões e controle de produtividade primária e secundária;
3. Dinâmica de populações e organização de comunidades e ecossistemas;
4. Dinâmica (fluxos) de nutrientes;
5. Efeitos de perturbações naturais e impactos de atividades antrópicas.

Atualmente o Programa PELD é composto de uma rede de pesquisa com 12 sítios brasileiros, listados na Tabela 1.1, cuja premissa é a integração dos dados de biodiversidade e sua disseminação para a comunidade científica. Um sítio de pesquisa do Programa PELD representa um importante ecossistema brasileiro, onde são feitos levantamentos específicos de sua biodiversidade através de coletas de campo e experimentos, sob as dimensões temporal e geofísica. Um exemplo de coleta é a diversidade e densidade de espécies de zooplâncton em um lago. Com o Programa PELD é esperado o crescimento da pesquisa ecológica voltada para a conservação da biodiversidade e que ajude a formular políticas públicas de planejamento ambiental e de desenvolvimento sustentável.

No plano internacional, o Programa PELD está inserido no programa ILTER[2] *(International Long Term Ecological Research)*, uma rede internacional que conta com 21 países participando ativamente e trocando experiências, da qual o Brasil é membro ativo desde 1998, tendo, inclusive, participação efetiva no comitê organizador [44].

1.2 Caracterização do Problema

Para atender as premissas do Programa PELD, deve ser oferecido aos seus usuários um sistema de informação que integre todos os dados do programa. No entanto, isso traz vários desafios e questões complexas que se devem não só à diversidade dos dados manipulados e de seu conteúdo, mas também ao grande volume de dados existentes. Um fator mais agravante é a falta de fontes de dados estruturadas, ou seja, aquelas que não possuem bancos de dados estruturados e, conseqüentemente, sistemas de informação para acesso aos dados. Essas fontes podem ser arquivos textuais, planilhas eletrônicas sem esquemas padronizados e, até mesmo, anotações manuscritas. Por exemplo, o sítio do Parque Estadual do Rio Doce em Minas Gerais possui 10 anos de coleta sobre os parâmetros físicos e químicos de lagos da Bacia do Rio Doce, cujos registros são mantidos em planilhas de dados particulares ou mesmo em

[2] www.ilternet.edu

Número	*Sítio*	*Áreas Estudadas*	*Instituição*
1	Floresta Tropical Úmida - Manaus	fragmentos florestais, pastagens e florestas secundárias	INPA (Instituto Nacional de Pesquisas da Amazônia)
2	Pantanal Sul	cerrado e pantanais	EMBRAPA (Empresa Brasileira de Pesquisas Agropecuárias)
3	Cerrado Centro-Oeste	cerrado	UnB (Univ. Nacional de Brasília)
4	Mata Atlântica e Sistema Lacustre do Médio Rio Doce	Mata Atlântica, lagos, rios e plantações de *Eucalyptus spp*	UFMG (Univ. Federal de Minas Gerais)
5	Restingas e Lagoas Costeiras do Norte Fluminense	Mata Atlântica Costeira, restingas e manguezais	UFRJ (Univ. Fed. do Rio de Janeiro)
6	Planície de Inundação do Alto Rio Paraná	Rio Paraná, tributários, lagos/planícies de inundação	Univ. Estadual de Marigá
7	Sistema Hidrológico Banhado Taim	banhados/dunas, praias arenosas, lagos e gramíneas	UFRGS (Univ. Fed. do Rio Grande do Sul)
8	Estuário Lagoa dos Patos e Costa Adjacentes	estuário e águas costeiras	Fundação UFRG (Univ. Fed. do Rio Grande)
9	Floresta Ombrófila Mista e Transições	Mata de Araucária e Florestas Plantadas Mistas	PUC-PR (Pontifícia Univ. Católica) do Paraná
10	Biodiversidade de Fragmentação de Ecosistemas nos Cerrados Marginais do Nordeste	Chapadinha(MA) - Nordeste do Maranhão e Parque Nacional das Setes Cidades-PI	UFPI (Univ. Fed. do Piauí)
11	Caatinga	Bacia do Rio Taperoá e Bacias do Seridó a Oeste	UFPB (Univ. Fed. da Paraíba)
12	Pantanal Norte	Parte Norte do Pantanal do Mato Grosso	UFMT (Univ. Fed. do Mato Grasso)

Tabela 1.1: Sítios do Programa PELD

anotações manuais. Além desses dados, ainda são mantidos, da mesma maneira, dados sobre as áreas de diversidades genética, vegetal e faunística do Parque.

Um sistema de informação em biodiversidade deve manipular dados heterogêneos e ser capaz de gerenciar grandes bancos de dados referentes às espécies (quando e onde elas foram observadas, por quem e como), incluindo também os dados geográficos que caracterizam o ecossistema onde cada espécie foi observada e a sua distribuição espacial [20]. Com a posição geofísica das coletas é possível estabelecer correlações espaciais relevantes no estudo da biodiversidade. Exemplos dessas correlações são questões como "Em que locais uma determinada espécie ocorre?" ou "Quais coletas foram realizadas em lagos próximos ao Parque Estadual

do Rio Doce?". No entanto, dentro do Programa PELD, para responder tais perguntas, os pesquisadores precisam utilizar os dados de biodiversidade, na forma hoje existente, e um sistema de informação geográfico separadamente, o que torna as análises demoradas e, às vezes, incompletas. Correlacionar dados ecológicos, mantidos em sistemas de informação em biodiversidade e sistemas de informação geográficos a partir de um único ambiente envolve questões relacionadas à interoperação.

1.3 Abordagem Proposta

Esta dissertação propõe um ambiente, baseado em bibliotecas digitais, para integração de dados dos vários sítios do Programa PELD. Esse ambiente se utiliza do protocolo OAI-PMH para coleta de dados dos diversos sítios PELD e sua posterior inclusão no repositório de uma biblioteca digital, denominada BDiG-PELD (Biblioteca Digital Georreferenciada do PELD).

Esse ambiente se utiliza ainda de uma interface de entrada de dados descentralizada, a partir da qual os dados de coleta de campo são submetidos ao banco de dados local do respectivo sítio e publicados na forma de dados EML[3] *(Ecological Metadata Language)*, linguagem própria para descrever dados ecológicos. Esses dados são então colhidos e integrados ao repositório central da BDiG-PELD. A interface de entrada de dados combina arquivos textuais, bancos de dados e repositórios XML no padrão OAI[4] *(Open Archives Initiative)*, como abordagem para permitir que os sítios PELD possam participar de um ambiente de interoperação.

A BDiG-PELD oferece a partir do seu repositório central, serviços de busca textual e navegação combinados com facilidades de georreferenciamento por meio de uma interface única. Parte desses serviços foi implementada com o uso da infraestrutura ODL *(Open Digital Libraries)* [47, 48], cujos componentes, baseados no padrão OAI, oferecem facilidades para interoperação, busca e navegação, e do Locus [21, 22], um localizador geográfico desenvolvido no Laboratório de Bancos de Dados da UFMG.

Para demonstrar a efetividade do ambiente proposto, uma versão-piloto da BDiG-PELD foi criada com base nos dados do sítio do Programa PELD sediado no Parque Estadual do Rio Doce, a partir da qual foi realizada uma avaliação experimental criteriosa, com usuários reais, sob os aspectos de usabilidade da interface de entrada de dados e qualidade dos dados, sendo os seus resultados analisados e discutidos.

Além disso, especificamos os requisititos da versão-piloto da BDiG-PELD de acordo com a abordagem 5S *(Streams, Structures, Spaces, Scenarios and Societies)* [27] que define um modelo para descrição e construção de bibliotecas digitais. A partir dessa especificação, foi possível obter um melhor entendimento dos requisitos da comunidade usuária da BDiG-PELD.

[3] knb.ecoinformatics.org
[4] www.openarchives.org

1.4 Contribuições

São duas as principais contribuições desta dissertação. A primeira é a proposta efetiva de um ambiente para interoperação centrado em uma biblioteca digital que permite integrar dados ecológicos de várias fontes e que agrega serviços que combinam busca textual e navegação com facilidades de georreferenciamento [17, 18]. A arquitetura desse ambiente contempla tecnologias combinadas para atender vários requisitos, tais como, entrada de dados descentralizada, interoperação de dados, publicação através de um repositório central e serviços especializados para seus usuários. Para ajudar no entendimento desses requisitos e na especificação da biblioteca digital ainda utilizamos a modelagem 5S. A segunda é a avaliação, sob os aspectos da usabilidade e qualidade dos dados, da abordagem utilizada para viabilizar a participação de fontes de dados pouco estruturadas ou informatizadas nesse ambiente [17]. Realizada a partir do sítio do Parque Estadual do Rio Doce, cuja diversidade de amostras é ampla e inclui um número considerável de pesquisadores, a avaliação permitiu discutir todo o processo de construção de provedores de dados pouco estruturados e avaliar a sua viabilidade quando aplicado a outros sítios de pesquisa do Programa a serem integrados.

Esta dissertação é resultado do envolvimento de várias disciplinas que fundamentam as pesquisas em bibliotecas digitais, tais como, bancos de dados, recuperação de informação, engenharia de software e interfaces homem-máquina, o que o torna um estudo de caso relevante nessa área.

1.5 Trabalhos Relacionados

Muitos trabalhos tratam da integração de dados originados de bancos de dados distintos, sejam eles ecológicos ou não, para prover sistemas de informação com uma diferente gama de serviços. Parte desses trabalhos utiliza abordagens baseadas em bancos de dados centralizados, com um esquema único [33, 34] e com alguns serviços agregados. Outros tratam a questão como uma rede de integração de provedores de dados, com dados descritos através de uma linguagem de metadados padrão com grande poder de descrição [20, 43], e oferecem serviços de maior valor agregado. A seguir, apresentamos uma breve descrição de alguns desses trabalhos, fazendo uma comparação dos mesmos com a abordagem proposta nesta dissertação.

O Metcat *(Metadata Catalog)* [34], desenvolvido pela Universidade da Califórnia, através do NCEAS *(National Center of Ecological Analisys and Synthesis)*, é um *servlet* Java que recebe e direciona os dados ecológicos submetidos para subsistemas específicos. Ele integra 180 estações da organização americana OBFS *(Organization of Biological Field Stations)* e mais de 25 sítios de pesquisa da rede norte-americana ILTER *(International Long Term Ecological Research)*. Nessa solução, cada pesquisador edita seus dados ecológicos através de uma aplicação cliente própria (Morpho) e pode consultá-los pela Web. Seus principais subsistemas são armazenamento, replicação, consulta e validação. Toda comunicação é feita sob HTTP com conteúdo EML. No entanto, ao contrário do ambiente proposto neste trabalho, o Metacat não oferece facilidades de georreferenciamento nem permite a inclusão de novos serviços baseados

em componentes, embora tenha uma arquitetura independente de padrões particulares de metadados.

O sistema de informação SINBIOTA[5] (Sistema de Informação Ambiental do Biota) mantém uma base taxonômica e georreferenciada da bacia hidrográfica do Rio Mogi-Guaçu, São Paulo. Os serviços oferecidos baseiam-se em análises espaciais a partir de mapas sobre conservação, uso do solo e hidrografia [33]. O SINBIOTA recebe diretamente os dados de campo através de formulários Web e, portanto, não configura uma arquitetura de interoperação. Seu esquema de banco de dados é proprietário e implementado por meio de um SGBD relacional. Assim como na BDiG-PELD, não há serviços de personalização ou colaboração para os usuários. O SINBIOTA possui serviços de georreferenciamento bem integrados embora utilize uma infraestrutura proprietária, diferentemente da nossa que possui um arcabouço mais genérico podendo ser aplicada a outras áreas do conhecimento.

A ETANA-DL[6] *(Eletronic Tools and Ancient Near Eastern Archives - Digital Library)* é uma biblioteca digital que mantém dados sobre objetos arqueológicos de um consórcio mundial de sítios de arqueologia. A interoperação entre provedores de dados e a ETANA-DL é realizada através do protocolo OAI-PMH e da infraestrutura ODL. Nossa abordagem é semelhante à da ETANA-DL no que se refere a infraestrutura, entretanto a ETANA-DL não possui facilidades para georreferenciamento e sua linguagem de metadados é mais simples, embora envolva serviços de colaboração.

A *Alexandria Digital Library*-ADL[7] é uma biblioteca digital com coleções de objetos georreferenciados. A ADL utiliza um *gazetteer* que permite o georreferenciamento de locais. A ADL oferece ainda acesso público a um acervo de mais de 15.000 itens incluindo mapas, imagens e dados espaciais. Mapas digitais são acessados na ADL com interfaces de consulta por coordenadas ou nomes. Sua arquitetura segue o modelo de três camadas: servidores gerenciam as coleções de dados espaciais, um *middleware* implementa os serviços de acesso às coleções via protocolo HTTP e clientes são utilizados pelos usuários para consulta e navegação pelas coleções. Com filosofia semelhante, a BDiG-PELD implementa um localizador geográfico de seus objetos ecológicos através do Locus [21], que permite localizar espécies e sítios PELD, mas que também possibilita a combinação de serviços de busca textual e navegação com facilidades de georreferenciamento.

O trabalho que mais se aproxima da nossa abordagem é o apresentado em [20], que propõe uma arquitetura genérica, também baseada em componentes ODL, para gerenciamento integrado de dados heterogêneos sobre seres vivos e seus ecossistemas, combinando buscas textuais e consultas por conteúdo gráfica. Nessa arquitetura é proposto um novo componente ODL para consultas por conteúdo de imagem. Nossa abordagem se difere desse trabalho por possibilitar consultas textuais e navegações cujas respostas podem ser georreferenciadas, como discutido na Seção 4.4.

Em relação aos trabalhos descritos acima, a abordagem apresentada neste trabalho oferece

[5] sinbiota.cria.org.br/info/
[6] feathers.dlib.vt.edu
[7] www.alexandria.ucsb.edu

alguns diferenciais, que são:

1. Disponibiliza uma interface de entrada de arquivos textuais com dados de campo. Tal interface recebe os arquivos pela Web e transforma seus dados em metadados ecológicos para que sejam coletados pela biblioteca central. A interface proposta não interfere no gerenciamento de dados de sítios que já possuam sistemas de informação próprios, bem como permite a participação de sítios ecológicos que ainda não possuam nenhum sistema de informação local.

2. Utiliza padrões abertos para interoperação entre os sítios da rede e o repositório central, cuja implementação envolve componentes abertos de software, ou seja, sem qualquer custo de licenciamento.

3. Oferece, além dos recursos básicos de busca, a possibilidade de explorar os dados armazenados através de serviços de navegação e recuperá-los em sua versão original de entrada, bem como de visualizá-los em mapas através de facilidades de georreferenciamento.

4. Envolve uma avaliação criteriosa de usabilidade do módulo de entrada de dados e de qualidade dos dados armazenados, que contou com a participação dos usuários finais.

1.6 Organização da Dissertação

O restante desta dissertação está organizado da seguinte forma. O Capítulo 2 apresenta uma visão geral dos principais conceitos, padrões e arquiteturas para interoperação de dados. No Capítulo 3 é descrita, detalhadamente, a arquitetura proposta, assim como seus principais módulos. O Capítulo 4 descreve a avaliação experimental realizada e discute seus resultados. Finalmente, o Capítulo 5 apresenta as principais conclusões do trabalho e as perspectivas para seu andamento futuro.

Capítulo 2

Conceitos, Padrões e Arquiteturas para Interoperação de Dados

Neste capítulo apresentamos os principais conceitos, padrões e arquiteturas para interoperação de dados, foco deste trabalho. O capítulo está dividido em seções da seguinte forma. Na Seção 2.1 apresentamos o conceito de bibliotecas digitais. A Seção 2.2 descreve arquivos abertos, alternativa utilizada para interoperação de dados. Na Seção 2.3 detalhamos a linguagem de metadados ecológicos EML, utilizada no processo de interoperação. Finalmente, a Seção 2.4 caracteriza os serviços de georreferenciamento usualmente encontrados em bibliotecas digitais.

2.1 Bibliotecas Digitais

Suleman [47] define as bibliotecas digitais como sistemas de informação dedicados a resolver as necessidades de busca e interoperação de seus usuários. Conforme [15], uma biblioteca digital é uma combinação envolvendo uma coleção de objetos digitais, um conjunto de usuários e sistemas que oferecem uma variedade de serviços, tais como indexação, catalogação, busca, navegação, recuperação, recomendação e preservação de dados.

Pela abrangência dessas definições constatamos que as bibliotecas digitais são sistemas de informação complexos e com caráter interdisciplinar. Tal complexidade e interdisciplinaridade podem ser verificadas claramente na abordagem proposta neste trabalho, pois nela tratamos de vários assuntos, tais como, processo de interoperação, repositório de documentos XML, especificação de requisitos, serviços de busca e navegação, e serviços especializados, como, por exemplo, o de georreferenciamento.

Por essas características, bibliotecas digitais necessitam de um modelo formal ou teórico para especificar melhor as complexas interações entre seus vários assuntos e os requisitos de seus usuários. No entanto, segundo [27], poucas iniciativas são encontradas nesse campo. Uma delas é abordagem 5S [26, 27], um modelo formal para descrição e construção de bibliotecas digitais, descrita mais detalhadamente na Subseção 4.1, onde apresentamos o modelo da BDiG-PELD segundo essa abordagem.

Apesar da carência de modelos formais para especificação de bibliotecas digitais, existem muitos avanços quanto às iniciativas para melhorar o compartilhamento de dados entre bibliotecas digitais [47]. Uma dessas iniciativas é a criação e uso de arquivos abertos ou padrão OAI *(Open Archives Initiative)*, cujos detalhes são apresentados na Subseção 2.2. Outros avanços também são encontrados em pesquisas que buscam formular arquiteturas reutilizáveis para construção de bibliotecas digitais, mesmo para comunidades com necessidades bem específicas. Um desses casos é a infraestrutura ODL [48], apresentada, também, na Subseção 2.2.

Finalmente, outro aspecto muito pouco explorado é relacionado à capacidade de georreferenciamento em bibliotecas digitais, o qual é considerado ainda exclusivo dos sistemas de informação geográficos [20, 31]. Entretanto, algumas comunidades de usuários seriam muito beneficiadas com essa característica, como, por exemplo, a de biodiversidade. Esse tema também é objeto deste trabalho e por isso será melhor discutido na Seção 4.4.

Um exemplo de biblioteca digital é a BDBComp, que foi desenvolvida no Laboratório de Banco de Dados da UFMG com o propósito de arquivar, preservar, indexar e disseminar a produção científica da comunidade brasileira de Ciência da Computação [35]. A BDBComp possui um repositório de metadados no formato *Dublin Core*[1] e oferece serviços de busca e navegação, além do serviço de interoperação baseado no padrão OAI. Todos esses serviços estão disponíveis por meio de interfaces específicas. Em breve a BDBComp oferecerá, também, serviços de recomendação e auto-arquivamento.

Outra biblioteca digital, também concebida sob o padrão OAI, é a NDLTD (Netwoked Digital Library of Theses and Dissertations) [25], uma iniciativa para reunir teses e dissertações em nível mundial. Seu catálogo central colhe dados entre as instituições participantes, como, por exemplo, universidades, que publicam seus dados no formato ETDMS *(Electronic Thesis and Dissertation Metadata Set)*, uma extensão do *Dublin Core* para descrever teses e dissertações, e oferece serviços tais como busca, navegação, personalização e recomendação.

2.2 Arquivos Abertos

Em geral, interoperação entre sistemas refere-se à habilidade de um sistema trabalhar cooperativamente com outros sistemas com o propósito de oferecer melhores serviços aos usuários [47]. Uma das abordagens para se alcançar a interoperação no contexto de bibliotecas digitais é proposta pela iniciativa OAI, uma organização formada por pesquisadores, bibliotecários, editores e arquivistas, cujo objetivo principal é criar padrões para possibilitar a interoperação entre sistemas. O padrão estabelecido por essa iniciativa é o protocolo OAI-PMH, o qual específica como dois repositórios de dados podem intercambiar uma seqüência de registros estruturados [47]. Repositórios de dados que atendem a esse protocolo são chamados arquivos abertos, sendo o termo "arquivos" originado da comunidade de *EPrints* [6], na qual é aceito como sinônimo para repositórios de artigos científicos. Contudo, com a iniciativa OAI esse termo é ampliado para repositórios de dados. Já o termo "abertos" não significa acesso livre

[1] dublincore.org

ou ilimitado, e sim a disponibilidade de interfaces que facilitam a coleta de dados armazenados em repositórios baseados no protocolo OAI-PMH.

A iniciativa OAI surgiu em outubro 1999 [46] e logo recebeu a atenção da comunidade de bibliotecas digitais, bem como adesões. Uma das primeiras foi a biblioteca digital de apoio ao ensino em Ciência da Computação denominada *Computer Science Teaching Center* [2].

Outras iniciativas para interoperação, simultaneamente à iniciativa OAI, surgiram, por exemplo, na comunidade de B2B *(Business-to-Business)* que tentou estender a funcionalidade da tecnologia EDI *(Eletronic Data Interchange)* com uso da linguagem XML [7]. O protocolo SOAP *(Simple Object Access Protocol)* [36], elaborado mais recentemente pela academia e indústria, também é uma dessas iniciativas e tenta viabilizar a interoperação entre sistemas ao propor mecanismos para troca de dados sob a linguagem XML [19], em ambientes computacionais distribuídos com a utilização de chamadas RPC *(Remote Procedure Call). Web Services* caracterizam uma alternativa a mais para interoperação, sendo uma de suas recentes iniciativas a linguagem WSDL *(Web Services Description Language)* [10] que permite descrever a localização de serviços Web, seus parâmetros, operações e resultados esperados.

Iniciativas como SOAP e WSDL, entre outras, procuram tornar o processo de interoperação na Web mais automatizado embora trabalhem em um nível mais sintático ao contrário do protocolo OAI-PMH que se concentra mais na semântica de interoperação, pois estabelece uma série de comandos ou verbos (Tabela 2.1) que auxiliam nesse processo, e também atua como uma camada de auto-nível para construção de bibliotecas digitais [48], conforme demonstrado na Subseção 3.3.1.

Verbo	Resposta
Identify	Descrição do provedor de dados.
ListMetadataFormat	Padrões de metadados disponíveis em um provedor de dados.
ListSets	Conjuntos de dados (hierárquicos) disponíveis no repositório.
ListIdentifiers	Identificadores dos registros de um repositório.
ListRecords	Registros de um repositório.
GetRecord	Um registro individual de dados de um repositório em um formato específico.

Tabela 2.1: Verbos do protocolo OAI-PMH

Uma importante decisão da iniciativa OAI foi a forma de estabelecer o processo de interoperação entres repositório de dados. Entre duas abordagens possíveis, federação e colheita, foi escolhida a de colheita *(harvesting)* pois torna as barreiras para a interoperação menores, na maioria dos casos, segundo [47]. Para esclarecer melhor, na abordagem de federação os dados requisitados por um usuário em uma biblioteca local são a combinação de buscas em múltiplos repositórios remotos. Essa abordagem é considerada mais dispendiosa pois acarreta na queda de desempenho da rede, caso o volume recuperado seja grande, e exige buscas em tempo real, bem como a disponibilidade integral dos nós remotos. Já a abordagem de colheita requer que cada repositório de dados transfira periodicamente uma seleção de metadados para a biblioteca digital central, a partir da qual o usuário recupera os dados pretendidos mais rapidamente

e com maior disponibilidade de acesso. Naturalmente, o esforço para coletar, armazenar, indexar e oferecer mecanismos de busca robustos é maior na biblioteca central, cujo papel na iniciativa OAI é definido como o de provedor de serviços. Outro papel importante, na iniciativa OAI, é o de provedor de dados, desempenhado por cada repositório remoto de dados. A Figura 2.1 faz uma comparação entre as duas abordagens.

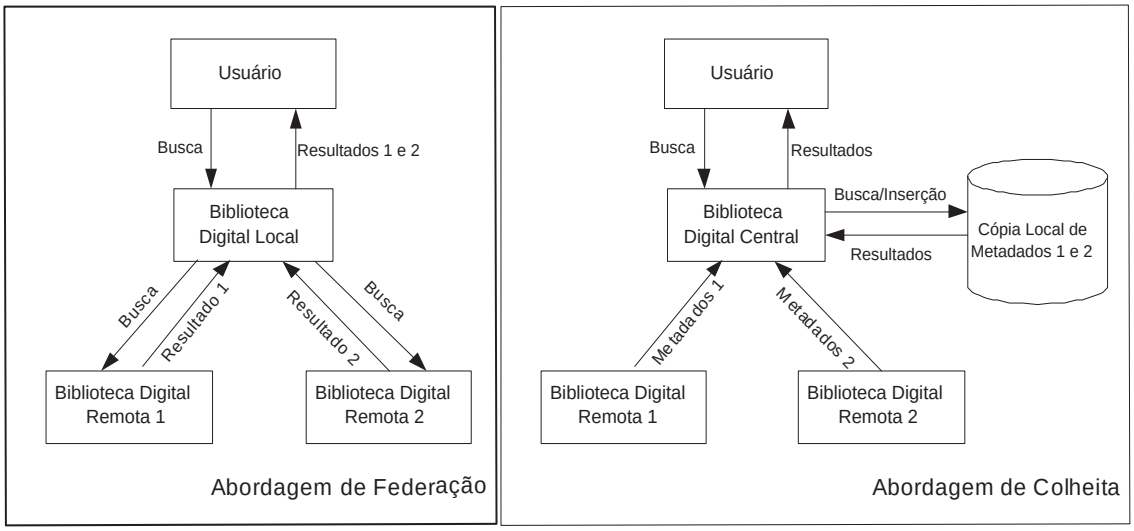


Figura 2.1: Abordagens de interoperação

A partir dessa exposição podemos entender a seguinte afirmação de Suleman [47]: "Bibliotecas digitais podem ser modeladas como redes de arquivos abertos, nas quais cada arquivo aberto é um provedor de dados ou um provedor de serviços".

Enquanto as iniciativas para melhora a interoperação alcançam seus objetivos, outras iniciativas tentam estabelecer padrões para construção de arquiteturas para as bibliotecas digitais, geralmente baseados em componentes de software que permitem atender necessidades específicas dos usuários, mas que possam ser aplicados a diferentes domínios ou assuntos sem deixar de incorporar os avanços obtidos pelas iniciativas para troca de dados entre sistemas.

Entre essas iniciativas, destacamos o projeto Fedora *(Flexible and Extensible Digital Object and Repository Architecture)* [40], um arcabouço baseado em código aberto para gerenciamento, armazenamento e disseminação de objetos complexos, cujos relacionamentos são definidos através da linguagem RDF *(Resource Description Framework)* [8]. Esse arcabouço possui uma arquitetura baseada em *Web Services* que provê funções para gerenciamento de objetos complexos através de interfaces SOAP.

Outra iniciativa é a infra-estrutura ODL *(Open Digital Library)*, proposta por Suleman [47], que consiste de um conjunto de componentes para construção de bibliotecas digitais, cujos serviços utilizam, tanto interna ou externamente, a simplicidade da semântica do protocolo OAI-PMH sob HTTP, como meio de comunicação. Seus componentes são reutilizáveis e oferecem serviços de interoperação, busca, navegação e colaboração. Tal infra-estrutura pode ser ampliada, também, com a inclusão de novos serviços, já que utiliza padrões abertos.

A infra-estrutura ODL por sua praticidade e efetividade foi a iniciativa adotada neste

trabalho para tratar de questões inerentes à sua construção, tais como, serviços aos usuários e interoperação. Retornaremos à infra-estrutura ODL na Subseção 3.3.1, onde trataremos de sua implementação.

2.3 Metadados Ecológicos

Uma questão importante no processo de interoperação OAI é estabelecer se o conteúdo da transferência pode conter somente metadados ou se deve incluir também os dados propriamente ditos. Outro fator, também importante, é a definição do padrão de metadados a ser utilizado. A iniciativa OAI estabelece que devem ser tratados, inicialmente, somente os metadados, mas que entre eles exista algum que indique "ponteiros" para o objeto com dados originais. Seu padrão de metadados é o *Dublin Core*, cujos 15 campos permitem descrever objetos digitais disponíveis na Internet, como, por exemplo, teses, dissertações, etc.

Apesar da iniciativa OAI adotar o padrão *Dublin Core*, nosso ambiente requer o uso de um padrão mais adequado para representar os complexos objetos ecológicos gerados pelo Programa PELD. Encontramos na literatura, inicialmente, três padrões de metadados utilizados em importantes redes de pesquisa em biodiversidade, os quais são apresentados a seguir.

O padrão DwC2 *(Darwin Core 2)*, composto de metadados com 41 campos, foi inicialmente adotado pelo projeto GBIF[2] *(Global Biodiversity Information Facility)*, do qual participam 50 países e organismos internacionais, com o objetivo de estabelecer uma infraestrutura distribuída de informações primárias em biodiversidade, cujo foco principal são espécies e dados sobre suas observações somente, mas com interligações com as áreas de genética e ecologia [3].

Por sua vez, o padrão ABCD *(Access to Biological Collection Data)* foi adotado pela rede BioCASE *(Biodiversity Collection Access Service for Europe)* [1], formada por sistemas de informação biológicos do continente Europeu e Israel. Desenvolvidos pelo grupo TDWG *(Taxonomical Databases Working Group)* [9], responsável por propor estruturas de dados e de interoperação, ambos os padrões ABCD e DwC2 são complementares e possuem algumas diferenças básicas. O padrão DwC2 possui 44 elementos sem a possibilidade de definir estruturas hierárquicas ou repetitivas, ao contrário do padrão ABCD com seus 700 elementos aproximadamente. A semelhança entre eles é quanto ao foco da descrição, pois ambos tratam de coleções de espécies e dados sobre suas observações. A partir de 2002, o grupo GBIF adotou o padrão ABCD o que possibilitou sua integração com a rede BioCASE, também facilitada pelo uso comum da arquitetura aberta de interoperação denominada DiGIR *(Distributed Generic Information Retrieval)* [4, 24], basead em um ambiente cliente/servidor com SOAP para recuperação de informações distribuídas, sob o protocolo HTTP.

Já o padrão EML *(Ecological Metadata Language)* foi conjuntamente criado por pesquisadores da área de ecologia, pelo NCEAS *(National Center for Ecological Analisis and Synthesis)* e pela rede internacional ILTER *(International Long Term Ecological Research)*, da qual a rede brasileira faz parte. O padrão ou linguagem EML promove a catalogação histórica de dados de natureza ecológica com ênfase em seus aspectos essenciais que são: geografia, data

[2] www.gbif.org

e hora, taxonomia, metodologia, dados gerais e específicos [13]. Suas origens já remontam os inícios dos anos 80 quando a rede internacional ILTER foi criada com 8 universidades norte-americanas e com o financiamento da fundação norte-americana NSF (National Science Fundation). Os primeiros intercâmbios de dados já foram incorporados à rede por meio de arquivos textuais. Em 1994, o governo americano propôs a publicação dos dados de forma *on-line*, mas somente em 1996 surgiu o padrão *ILTER Metadata*, o primeiro rascunho da EML, que já incorporava padrões de outras áreas para atender uma rede cada vez mais abrangente e diversa [42].

Para ilustrar a abrangência da linguagem EML, apresentamos um pouco da sua estrutura. Um documento EML possui o elemento raiz (`<eml>`). Diretamente abaixo dele pode vir elementos, descritos na Tabela 2.2, que representam as principais seções de um documento EML. Os elementos são: `<dataset>`, `<citation>`, `<software>` ou `<protocol>`, e opcionalmente `<additionalMetadata>`. A Figura 2.2 apresenta esses elementos através de um trecho de seu esquema em XSD *(XML Schema Definition)* [12], representado graficamente com o uso da ferramenta XMLSpy [11].

Elemento	Descrição
`<dataset>`	Descreve um conjunto de dados que pode incluir uma ou mais tabelas bem como imagens espaciais em modo *raster* ou vetorial. Também permite incluir os dados propriamente ditos das coletas.
`<citation>`	Apresenta as referências bibliográficas utilizadas no processo de coleta.
`<software>`	Descreve o software ou utilitário que pode acessar ou manipular dados de coleta do documento.
`<protocol>`	Descreve protocolo científico ou métodos e procedimentos utilizados para coleta ou catalogação dos dados do documento.
`<additionalMetadata>`	Permite especificar, incluir e preencher metadados de outros padrões, como, por exemplo, metadados CSDGM-FGDC.

Tabela 2.2: Descrição do principais elementos da linguagem EML

Como exemplo de um documento EML, apresentamos a Figura 2.3, na qual podemos destacar a inclusão dos dados de coletas de campo (marcação `<inline>` entre linhas 23 e 27) assim como do seu esquema (marcação `<attributeList>` entre as linhas 18 a 22), localização geofísica de uma coleta (marcação `<geographicCoverage>` entre as linhas 9 e 17) e taxonomias das espécies envolvidas (marcação `<taxomicCoverage>`) entre as linhas 28 e 31). A EML é bastante extensa, de modo que somente as marcações mais representativas foram utilizadas para descrever os dados do Programa PELD.

2.4 Georreferencimento em Bibliotecas Digitais

Hill [31] define georreferenciamento como a capacidade de relacionar informações, como, por exemplo, documentos, conjuntos de dados, informações biológicas e espécies de seres vivos, a locais geográficos através de nome de lugares ou de coordenadas de latitude e longitude. A

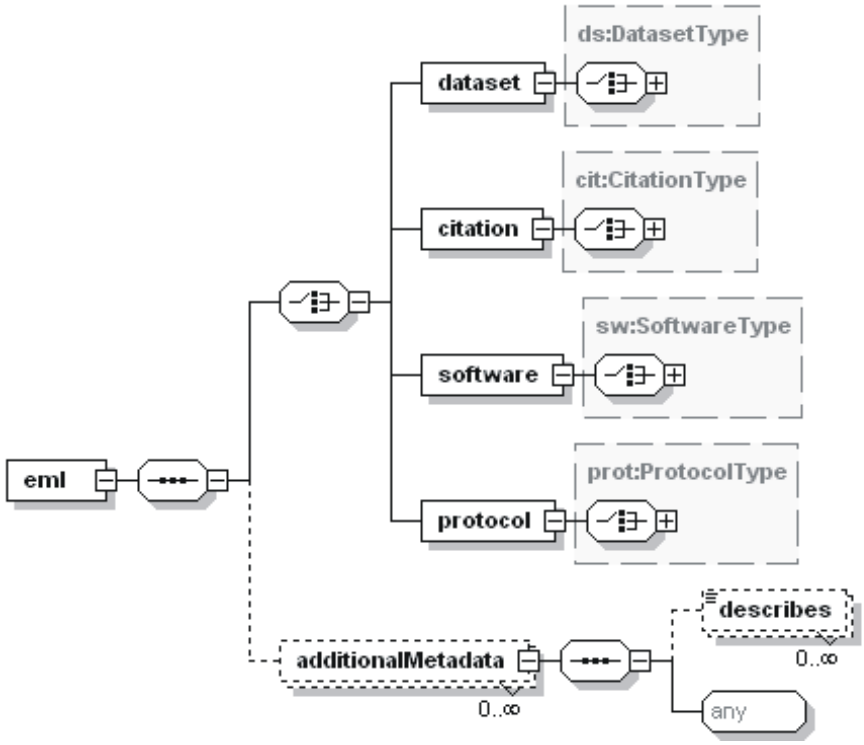


Figura 2.2: Principais elementos da EML

incorporação dessa característica às bibliotecas digitais representa uma nova classe de serviços frente aos tradicionais, geralmente de busca e navegação textuais.

É uma possibilidade poderosa para os usuários georreferenciar objetos que possam ser localizados por um nome de lugar *(placename)* sem que seja informado qualquer coordenada geográfica. Outra possibilidade, também interessante, é o fato de se poderem realizar buscas por meio de relações espaciais indiretas, como, por exemplo, "coletas localizadas perto do Parque Estadual do Rio Doce". No entanto, para se chegar a bibliotecas digitais georreferenciadas, sete questões devem ser tratadas, segundo [32]. A primeira delas e mais complexa é a descoberta de recursos passíveis de georreferenciamento. Relacionados a ela estão as questões de integração dos dados descobertos em *gazetteers* [30], dicionários que traduzem nomes de locais em coordenadas geográficas, e escalonamento de resultados de buscas *(ranking)*. As outras questões são: estrutura de dados robusta, escalabilidade, interfaces e interoperação entre recursos geoespaciais.

Tratamentos padronizados para essas questões ainda estão em aberto, mas não impediram o surgimento de iniciativas já bem consolidadas, como, por exemplo, a biblioteca digital ADL *(Alexandria Digital Library)* [28]. Outra iniciativa, que utilizamos para viabilizar as facilidades de georreferenciamento propostas neste trabalho, é o localizador geográfico Locus desenvolvido pelo Laboratório de Bancos de Dados da UFMG [21, 22]. Em sua utilização algumas dessas questões também foram trabalhadas, cujos detalhes são apresentados na Seção 4.4.

Outra funcionalidade pouco explorada em bibliotecas digitais georreferenciadas é a dimensão temporal, útil em análises em áreas, como, por exemplo, geopolítica e biodiversidade. Um

```
1      <eml>
2         <dataset>
3            <title>Composição da comunidade zooplanctônica de rios e lagos do trecho médio
4              da bacia do Rio  Doce-MG - Lagoa Águas Claras - 1/8/04 14h:30m -  Coleta Mensal
5            </title>
6            <creator><individualName><givenName>Francisco</givenName>
7                                  <surName>Barbosa</surName></individualName>
8            </creator> ...
9            <geographicCoverage>
10              <geographicDescription> Parque Estadual do Rio Doce - Lagoa Águas Claras
11              </geographicDescription>
12              <boundingCoordinates><westBoundingCoordinate>49°20'22"W </westBoundingCoordinate>
13                                   <eastBoundingCoordinate>49°20'22"W </eastBoundingCoordinate>
14                                  <northBoundingCoordinate>40°1'11"S </northBoundingCoordinate>
15                                  <southBoundingCoordinate>40°1'11"S </southBoundingCoordinate>
16              </boundingCoordinates>
17           </geographicCoverage>...
18           <attributeList>
19              <attribute><attributeName>TAXONOMIA</attributeName> ... </attribute>
20              <attribute><attributeName>PROFUNDIDADE</attributeName> ... </attribute>
21              <attribute><attributeName>DENSIDADE</attributeName>... </attribute>
22           </attributeList> ...
23           <inline>
24              Chlamydomonas sp.; 4;126,43
25              Chlorella sp.; 0 ;2907,857143
26              Closteriopsis;sp. 1; 1 ;42,14285714
27           </inline> ...
28           <taxonomicCoverage>
29            <taxonRankName>GENERO</taxonRankName><taxonRankValue>Chlamydomonas</taxonRankValue>
30            <taxonRankName>ESPECIE</taxonRankName><taxonRankValue>sp.</taxonRankValue> ...
31           </taxonomicCoverage> ...
32        </dataset> ...
33      </eml>
```

Figura 2.3: Trecho de documento com metadados EML *(Ecological Metadata Language)*

dos trabalhos que abordam essa questão no âmbito de localizadores geográficos é apresentado em [41].

Capítulo 3

Ambiente BDiG-PELD

Apresentamos neste capítulo os detalhes do ambiente BDiG-PELD. O capítulo está dividido em seções da seguinte forma. Na Seção 3.1 mostramos os principais componentes da arquitetura do ambiente, divididos entre provedores de dados e serviços. Na Seção 3.2, detalhamos os componentes de um provedor de dados e seu funcionamento. Em seguida, na Seção 3.3, apresentamos o funcionamento e os componenente de um provedor de serviços. Finalmente, na Seção 4.3, descrevemos o processo e as principais características para a implementação do ambiente.

3.1 Arquitetura do Ambiente

Conforme já discutido à Subseção 2.2, a iniciativa OAI propõe a abordagem de colheita para o processo de interoperação com dois papéis básicos em sua arquitetura, o de provedor de dados e o de provedor de serviços. Os provedores de dados publicam dados para serem colhidos pelos provedores de serviços. Os provedores de serviços agregam valor aos dados na forma de serviços oferecidos aos usuários. A partir dessa definição, a BDiG-PELD é um provedor de serviços e cada sítio PELD é um provedor de dados.

A interoperação entre provedores de dados e de serviços atende ao paradigma cliente-servidor pelo uso do protocolo de colheita OAI-PMH que permite ao provedor de serviços colher dados dos provedores através de requisições HTTP periódicas e seletivas. O protocolo OAI-PMH é utilizado nessa arquitetura através da infraestrutura ODL [47] desenvolvida na Virginia Tech[1] e composta de componentes para interoperação em bibliotecas digitais e implementação de serviços de busca, navegação, etc.

A Figura 3.1 apresenta a arquitetura idealizada para esse ambiente, no qual a BDiG-PELD atua como provedor de serviços. A BDiG-PELD possui um repositório central de dados, a infraestrutura ODL, uma interface para usuários e o localizador geográfico Locus. Cada provedor de dados possui um banco de dados relacional local, um repositório de dados XML, uma interface para entrada de dados e a camada OAI, através da qual se dá a interoperação

[1] Virginia Polytechnic Institute and State University

entre a BDiG-PELD e os provedores de dados. Essa arquitetura é descrita no decorrer deste capítulo.

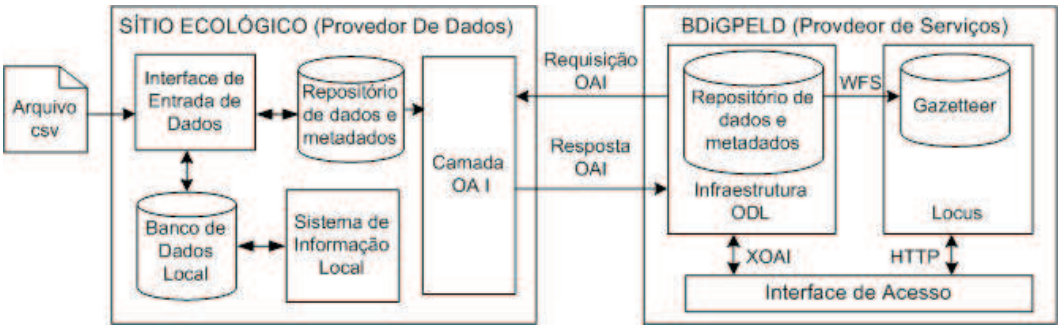


Figura 3.1: Arquitetura do ambiente BDiG-PELD

Dados ecológicos, em arquivos *csv (comma-separated value)*, são carregados no provedor de dados através da interface de entrada de dados que consiste e valida esses dados. A interface de entrada insere dados no banco de dados local do provedor de dados para serem acessados por um sistema de informação. A interface também gera e publica os metadados do arquivo recebido em um repositório de dados XML. Através da camada OAI, esse repositório disponibiliza dados para o provedor de serviços. A colheita é iniciada pela BDiG-PELD por meio de uma requisição OAI. Ao receber a requisição, a camada OAI do provedor de dados a processa e devolve o resultado através de uma resposta OAI. O sistema de informação local permite acesso e manipulação aos dados específicos do sítio, tais como, usuários e projetos de pesquisa, bem como aos dados das coletas, cujas cópias são também mantidas localmente.

No provedor de serviços temos ainda a infraestrutura ODL, que possui módulos para iniciar requisições, receber respostas e processá-las. O processamento realizado diz respeito ao armazenamento de documentos XML em um repositório central e à construção de índices para o serviço de busca. Um serviço da BDiG-PELD identifica dados do repositório associados a locais geofísicos e insere suas coordenadas no banco de dados geográfico, ou *gazetteer*, do Locus. A interoperação entre a infraestrutura ODL e o Locus é realizada através do protocolo WFS *(Web Feature Service)*, apropriado para acesso a objetos espaciais, pelo qual podem ser transportados coordenadas geográficas, nome de locais e suas descrições.

Os usuários interessados em dados ecológicos utilizam a interface de consulta da BDiG-PELD[2] através dos serviços de busca e navegação para exibir e recuperar dados de campo, bem como georreferenciar as ocorrências de seu interesse. Para oferecer esse serviço, a interface de consulta interage com a infraestrutura ODL por meio do protocolo XOAI *(Extended OAI)* [47] e com o Locus por meio de *servlets* Java, via protocolo HTTP. O protocolo XOAI é uma extensão do protocolo OAI para interoperação entre os próprios componentes da infraestrutura ODL ou com componentes proprietários no ambiente. A seguir apresentamos a descrição dos módulos que compõem a arquitetura do ambiente BDiG-PELD.

[2]www.lbd.dcc.ufmg.br/cgi-bin/bdigpeld/index.pl

3.2 Provedor de Dados

Pelo padrão OAI, cada provedor de dados é responsável pela publicação de seus dados em um ambiente de interoperação. Para provedores de dados pouco estruturados, como é o caso da maioria dos sítios PELD do programa PELD, é necessário estabelecer uma infraestrutura mínima que atenda a esse padrão. A arquitetura proposta neste trabalho adota, para o provedor de dados, uma interface de entrada de dados que permite o armazenamento dos dados em bancos de dados locais e sua publicação em um repositório XML. Os dados ecológicos são submetidos à interface de entrada de dados através de arquivos *csv*, preenchidos em qualquer planilha eletrônica, de acordo com o tipo de coleta a que se refere. Os componentes de um provedor de dados são descritos a seguir.

3.2.1 Banco de Dados Local

Boa parte dos sítios do Programa PELD não possui um banco de dado local, por isso tivemos que construí-lo para armazenar dados básicos cadastrais de um sítio e dados das amostras coletadas em campo ou geradas em experimentos de laboratório. Sua especificação foi feita a partir de várias reuniões envolvendo as principais áreas de pesquisa do Programa PELD. Embora essas reuniões tenham sido feitas a partir do sítio ecológico do Parque Estadual do Parque do Rio Doce somente, concentramos o foco da modelagem de dados no fato de que todos os sítios do Programa possuem dois conjuntos de dados distintos.

O primeiro conjunto é relativo aos dados básicos das coletas, comuns a todos os sítios de pesquisa do Programa. Ou seja, qualquer coleta do programa possui dados, como por, exemplo, hora, data, local, responsável pela coleta, projeto de pesquisa, área de pesquisa, coordenadas latitudinal e longitudinal, etc. Esse conjunto compõe as tabelas básicas cadastrais que permitem a normalização dos dados da tabela AMOSTRAGEM, cujas instâncias representam coletas. A Figura 3.2 apresenta a definição das principais tabelas básicas de acordo com a notação gráfica da ferramenta DBDesigner 4 [5].

O outro conjunto de dados diz respeito aos dados específicos das coletas de campo realizadas por diferentes subprojetos de pesquisa. Para cada tipo de coleta foram criadas tabelas específicas que detalham seus dados básicos, mantidos na tabela básica AMOSTRAGEM. Com essa abordagem, novas coletas podem ser incorporadas facilmente ao banco de dados por meio da criação de tabelas específicas.

Na Figura 3.3 podemos observar que a tabela AMOSTRAGEM pode representar coletas sobre nutrientes em lagos (tabela AMOSTRAGEM_NUTRIENTES) ou sobre densidade de espécies de seres vivos encontradas em lagos (tabela AMOSTRAGEM_DENSIDADE). Essa figura também utiliza a notação gráfica da ferramenta DBDesigner 4.

A tabela TAXONOMIA dessa figura também faz parte das tabelas básicas e permite armazenar todas as classificações taxonômicas de um provedor de dados, muito importantes em qualquer estudo de biodiversidade, e por isso merece alguns comentários.

Tais comentários dizem respeito ao cadastramento de taxonomias, geralmente uma tarefa difícil de gerenciar, pois um pesquisador pode atribuir a um mesmo ser vivo uma classificação

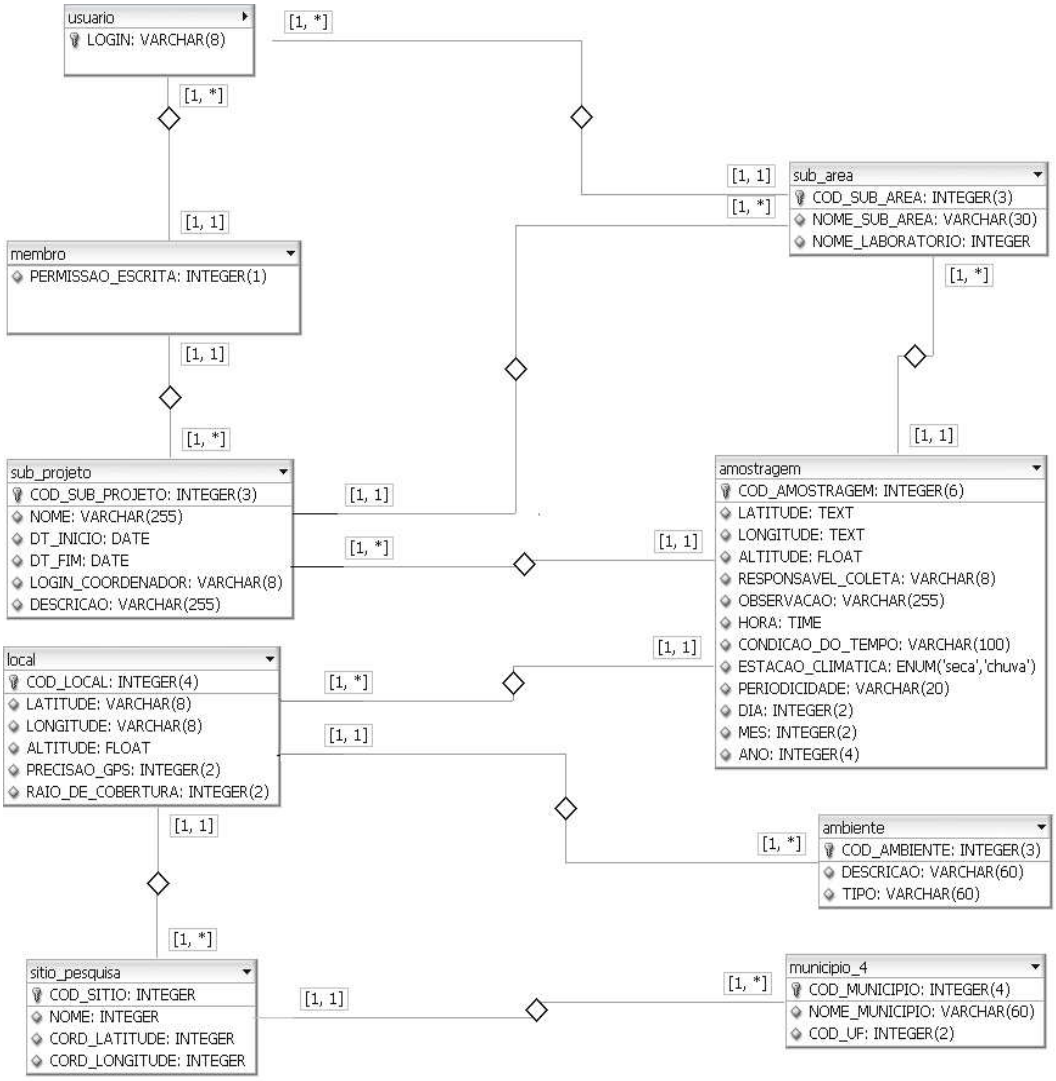


Figura 3.2: Principais tabelas do conjunto de tabelas básicas

taxonômica mais completa se comparada a uma realizada por outro pesquisador ou mesmo pode cometer erros ortográficos ao efetuar uma classificação. Para esses casos e outros, estabelecemos um procedimento para gerenciamento de taxonomias. Antes de qualquer entrada de dados de coleta no banco de dados do provedor, o pesquisador deve submeter sua lista de taxonomias a um especialista que tentará cadastrar as taxonomias no banco de dados local do sítio, garantindo a integridade de dados. Atualmente, um pesquisador pode consultar quais taxonomias já estão cadastradas no banco de dados local através do sistema de informação local. Contudo, o gerenciamento de taxonomias ainda é praticamente todo manual, mas poderá incorporar, em trabalhos futuros, um fluxo colaborativo automatizado.

Não obstante o esforço para construir um banco de dados para um provedor de dados, ainda tivemos que construir um sistema de informação Web para cadastramento dos dados básicos, ou melhor, para povoar as tabelas básicas do provedor. A Figura 3.4 apresenta

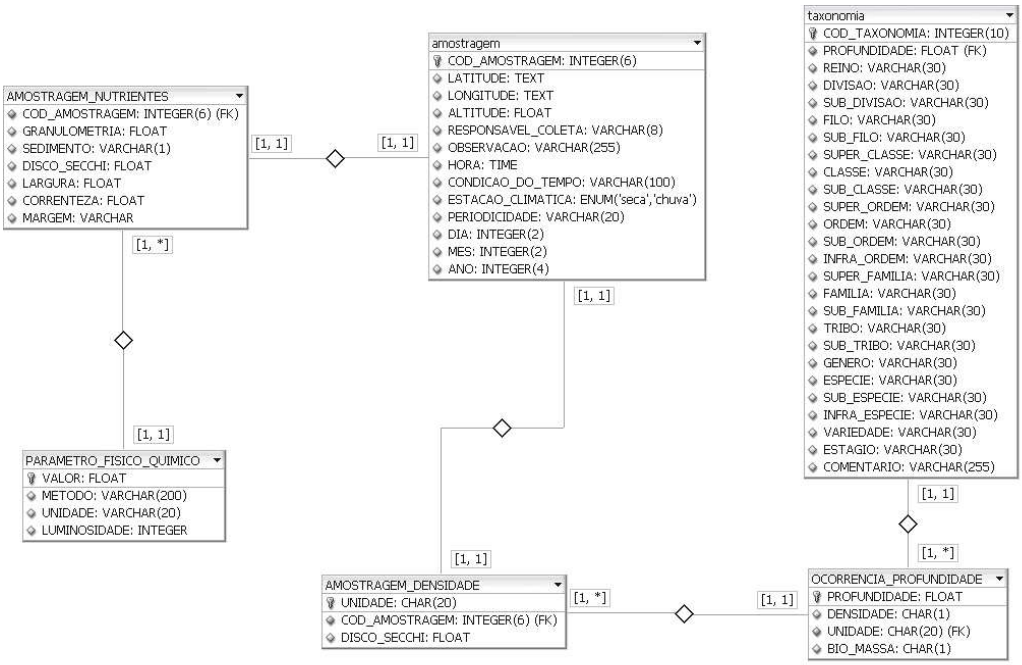


Figura 3.3: Tabelas básicas e tabelas específicas

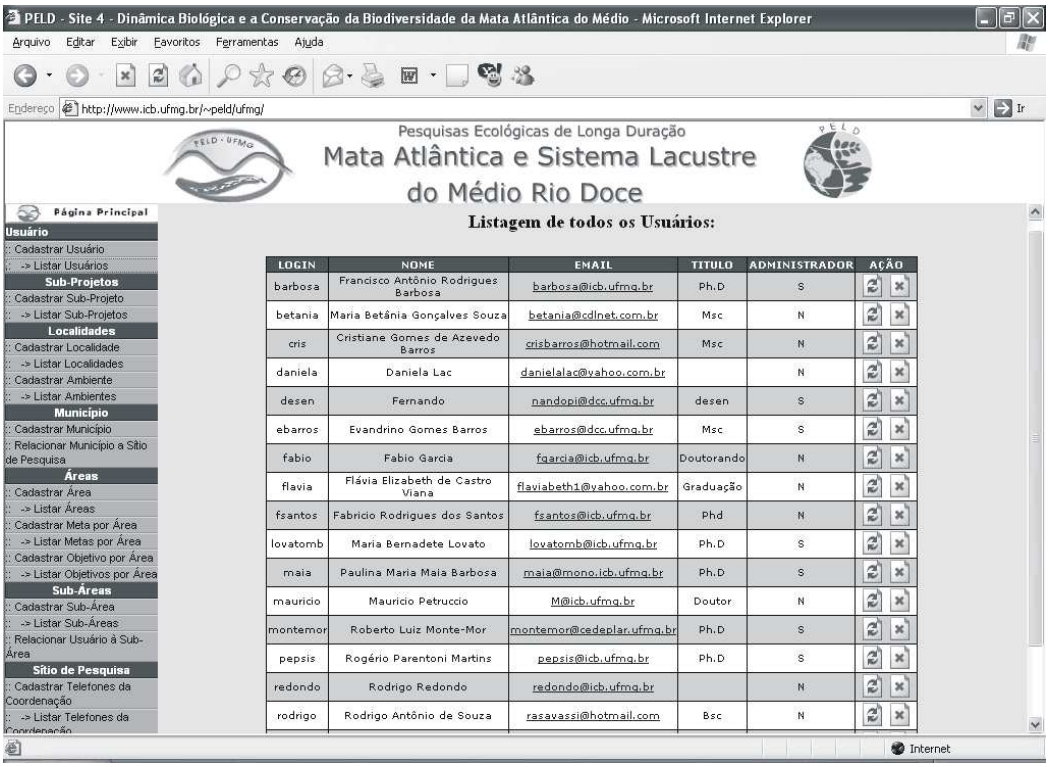
a interface desse sistema (www.icb.ufmg.br/∼peld/ufmg) através da página de cadastro de usuários e pesquisadores. Nela, à sua esquerda, encontramos uma barra de opções para acesso às funcionalidades do sistema e, disposta no centro, a lista de usuários com dados de *login*, nome, correio eletrônico, titulação e se o usuário é administrador do sistema ou não.

As tabelas específicas foram povoadas através da interface de entrada de dados discutida na subseção a seguir.

3.2.2 Interface de Entrada de Dados

Através de arquivos *csv* submetidos à uma interface de entrada de dados, atualizamos as tabelas específicas e também geramos, simultaneamente, documentos XML para serem colhidos pelo repositório central. Essa forma de operação evitou que gerássemos uma interface, no sistema de informação, para cada tipo de coleta, o que tomaria muito tempo, além de termos que desenvolver um módulo específico para gerar os documentos XML. Outra vantagem é a adaptabilidade a qualquer tipo de amostragem pois depende exclusivamente das tabelas específicas relacionadas ao tipo de arquivo *csv* submetido. Essa interface pode ser utilizada por qualquer sítio do Programa PELD para receber dados de coleta e publicá-los, através de metadados EML. Dessa forma, simplificamos o processo para tornar um sítio ecológico do Programa PELD em um provedor de dados.

A interface de entrada de dados funciona da seguinte forma. Inicialmente um arquivo *csv* com dados, como, por exemplo, o apresentado na Figura 3.5, é submetido à interface de entrada de dados por meio da página Web apresentada na Figura 3.7. Ao receber um arquivo *csv*, a interface de entrada de dados identifica qual o seu tipo de coleta e preenche as tabelas do banco de dados relacionadas à coleta. O banco de dados local do provedor também funciona como um catálogo para os dados básicos presentes no arquivo *csv*. Por exemplo, o campo do arquivo *csv* que indica o responsável pela coleta só pode ser preenchido com valores já cadastrados no banco de dados através do sistema de informação que o mantém, caso contrário o arquivo *csv* é rejeitado. O mesmo acontece com a classificação taxonômica de espécies, entre outros dados básicos. Essa interface também gera e publica, automaticamente, documentos XML no padrão EML a partir do arquivo *csv* submetido. A Figura 2.3 apresenta um documento XML gerado pela interface de entrada a partir da submissão de um arquivo *csv*.



LOGIN	NOME	EMAIL	TITULO	ADMINISTRADOR	AÇÃO
barbosa	Francisco Antônio Rodrigues Barbosa	barbosa@icb.ufmg.br	Ph.D	S	
betania	Maria Betânia Gonçalves Souza	betania@cdlnet.com.br	Msc	N	
cris	Cristiane Gomes de Azevedo Barros	crisbarros@hotmail.com	Msc	N	
daniela	Daniela Lac	danielalac@yahoo.com.br		N	
desen	Fernando	nandopi@dcc.ufmg.br	desen	S	
ebarros	Evandrino Gomes Barros	ebarros@dcc.ufmg.br	Msc	S	
fabio	Fabio Garcia	fgarcia@icb.ufmg.br	Doutorando	N	
flavia	Flávia Elizabeth de Castro Viana	flaviabeth1@yahoo.com.br	Graduação	N	
fsantos	Fabricio Rodrigues dos Santos	fsantos@icb.ufmg.br	Phd	N	
lovatomb	Maria Bernadete Lovato	lovatomb@icb.ufmg.br	Ph.D	S	
maia	Paulina Maria Maia Barbosa	maia@mono.icb.ufmg.br	Ph.D	S	
mauricio	Mauricio Petruccio	M@icb.ufmg.br	Doutor	N	
montemor	Roberto Luiz Monte-Mor	montemor@cedeplar.ufmg.br	Ph.D	S	
pepsis	Rogério Parentoni Martins	pepsis@icb.ufmg.br	Ph.D	S	
redondo	Rodrigo Redondo	redondo@icb.ufmg.br		N	
rodrigo	Rodrigo Antônio de Souza	rasavassi@hotmail.com	Bsc	N	

Figura 3.4: Interface do Sistema de Informação para o Sítio 4

É importante ressaltar que a interface de entrada de dados se adapta facilmente a novos esquemas e novos tipos de coleta de campo, como também pode ser alterada caso ocorram mudanças dos dados coletados. Em sua estrutura interna, essa interface possui um catálogo de dados que permite mapear os campos de um arquivo *csv* tanto para tabelas específicas do banco de dados, como para determinados metadados definidos na linguagem EML. O catálogo essencialmente indica, para cada campo dos arquivos *csv*, qual é a sua respectiva tabela no banco de dados e como é sua representação no documento EML a ser gerado, sendo que campos

que valorados com nulo não são representados no documento. Com essas características, a interface de entrada de dados pode ser utilizada por qualquer um dos sítios do Programa PELD.

3.3 Provedor de Serviços

3.3.1 Protocolo OAI-PMH

O padrão OAI estabelece o protocolo de colheita OAI-PMH *(OAI Protocol for Metadata Harvesting)* para interoperação entre provedores de dados e serviços. O protocolo OAI-PMH também possui regras para colheita de dados XML sob comunicação HTTP. As regras definem como requisitar dados, com o uso de verbos (ou comandos) definidos para funções específicas, e quais as respostas geradas. A Tabela 2.1 apresenta a lista de verbos definidos pelo protocolo OAI-PMH.

Originalmente projetado para transportar dados no formato *Dublin Core*, esse protocolo foi adaptado neste trabalho para transportar dados EML, conteúdo principal do processo de interoperação em nossa arquitetura. Em uma interoperação OAI-PMH, são também incluídos metadados de uso específico do protocolo OAI-PMH. Por exemplo, na Figura 3.6, podemos distinguir o cabeçalho OAI-PMH (linhas 1 a 9), o rodapé OAI-PMH (linhas 15 e 16), o conteúdo da interoperação (linhas 10 a 14) que inclui, mais especificamente, os dados EML propriamente ditos (linhas 11 a 13). Um exemplo de conteúdo EML pode ser visto na Figura 2.3. O documento OAI-PHM da Figura 3.6 foi gerado como resposta para a seguinte requisição OAI-PMH:

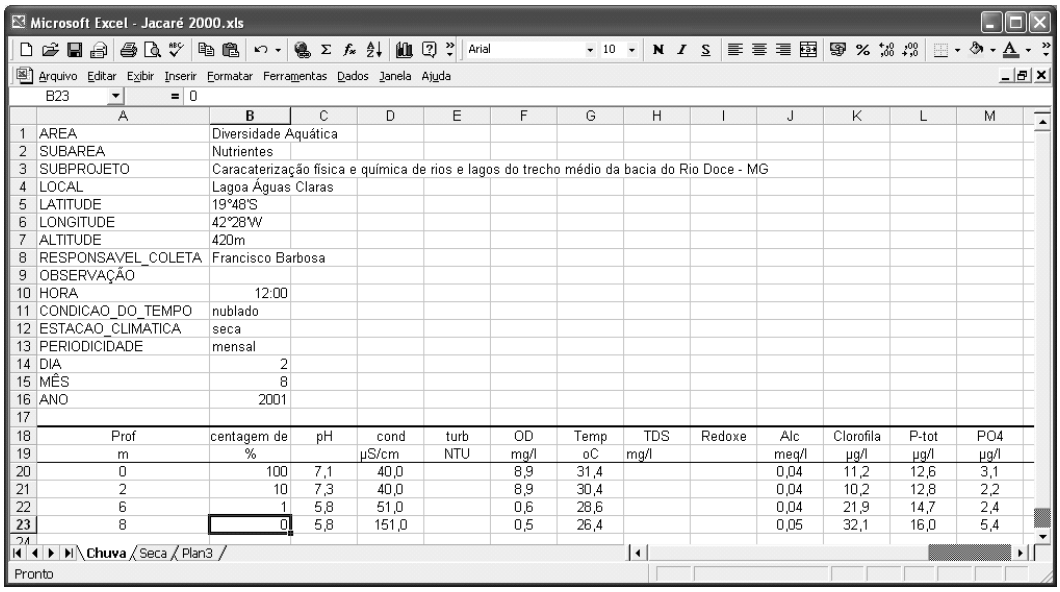
```
http://www.lbd.dcc.ufmg.br:80/cgi-bin/bdigpeld/
ODL-DBUnion-1.2/DBUnion/bdigpeld/union.pl?verb=GetRecord
&metadataPrefix=oai_eml&identifier=oai:test1:54
```

Nessa requisição é solicitado ao repositório `www.lbd.dcc.ufmg.br:80/cgi-bin/bdigpeld`, pelo verbo `GetRecord`, o registro cujo identificador é `oai:test1:54`, descrito no padrão de metadados `oai_eml`.

3.3.2 Infraestrutura ODL

O processo de interoperação de acordo com a arquitetura proposta neste trabalho é implementado através da infraestrutura ODL *(Open Digital Libraries)*, que pode ser considerada uma extensão do padrão OAI [47]. A infraestrutura ODL é composta de componentes abertos para viabilizar a interoperação em bibliotecas digitais e implementar serviços de busca, navegação, recomendação, etc. Embora existam diversos componentes já desenvolvidos e prontos para uso, somente três deles foram utilizados neste trabalho, os quais são apresentados na Tabela 3.1.

Através dos componentes da Tabela 3.1, implementamos os serviços de interoperação e busca. Mais detalhadamente, para colheita de dados nos provedores de dados, nossa arquitetura utiliza o componente ODL-Union Catalog, cujas requisições são atendidas pela camada



Microsoft Excel - Jacaré 2000.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	AREA	Diversidade Aquática											
2	SUBAREA	Nutrientes											
3	SUBPROJETO	Caracaterização física e química de rios e lagos do trecho médio da bacia do Rio Doce - MG											
4	LOCAL	Lagoa Águas Claras											
5	LATITUDE	19°48'S											
6	LONGITUDE	42°28'W											
7	ALTITUDE	420m											
8	RESPONSAVEL_COLETA	Francisco Barbosa											
9	OBSERVAÇÃO												
10	HORA	12:00											
11	CONDICAO_DO_TEMPO	nublado											
12	ESTACAO_CLIMATICA	seca											
13	PERIODICIDADE	mensal											
14	DIA	2											
15	MÊS	8											
16	ANO	2001											
17													
18	Prof	centagem de	pH	cond	turb	OD	Temp	TDS	Redoxe	Alc	Clorofila	P-tot	PO4
19	m	%		µS/cm	NTU	mg/l	oC	mg/l		meq/l	µg/l	µg/l	µg/l
20	0	100	7,1	40,0		8,9	31,4			0,04	11,2	12,6	3,1
21	2	10	7,3	40,0		8,9	30,4			0,04	10,2	12,8	2,2
22	6	1	5,8	51,0		0,6	28,6			0,04	21,9	14,7	2,4
23	8	0	5,8	151,0		0,5	26,4			0,05	32,1	16,0	5,4

Figura 3.5: Arquivo *csv* com dados de nutrientes de uma lagoa

```
1  <?xml version="1.0" encoding="UTF-8" ?>
2   <xoai:GetRecord xmlns="http://www.openarchives.org/OAI/1.1/OAI_GetRecord" ... >
3   <responseDate>2005-04-23T21:17:42-03:00</responseDate>
4   <requestURL>http://www.lbd.dcc.ufmg.br:80/cgi-bin/bdigpeld/ODL-DBUnion-1.2/DBUnion/bdigpeld/
5            union.pl?verb=GetRecord&metadataPrefix=oai_eml&identifier=oai:test1:54</requestURL>
6   <record>
7     <header>
8        <identifier>oai:test1:54</identifier>  <datestamp>2005-04-19T11:31:40+00:00</datestamp>
9     </header>
10    <metadata>
11        <eml>
12        ...
13        </eml>
14    </metadata>
15  </record>
16 </xoai:GetRecord>
```

Figura 3.6: Trecho de um documento OAI-PMH

OAI dos provedores de dados, implementada por meio do componente OAI-XMLFile. Os dados recebidos são armazenados no repositório de dados central. O serviço de busca, apresentado na Subseção 4.2, é implementado pelo componente ODL-IRDB que trabalha como uma máquina de busca que combina consultas booleanas baseadas em verdadeiro ou falso com características do modelo vetorial [16]. O componente ODL-IRDB gera suas entradas de índices a partir de colheitas feitas no componente ODL-Catalog Union, que acessa o repositório central de dados. As requisições entre os componentes ODL-IRDB e ODL-Catalog Union são realizadas por meio do protocolo XOAI [47]. Já o serviço de navegação da BDiG-PELD foi desenvolvido à parte, embora interaja com a infraestrutura ODL através do protocolo XOAI. Os detalhes de implementação do serviço de navegação são apresentados na Subseção 4.3.

A infraestrutura ODL permitiu a rápida prototipação da nossa arquitetura além de facilitar a incorporação de novos serviços, como, por exemplo, os de georreferenciamento. Um caso

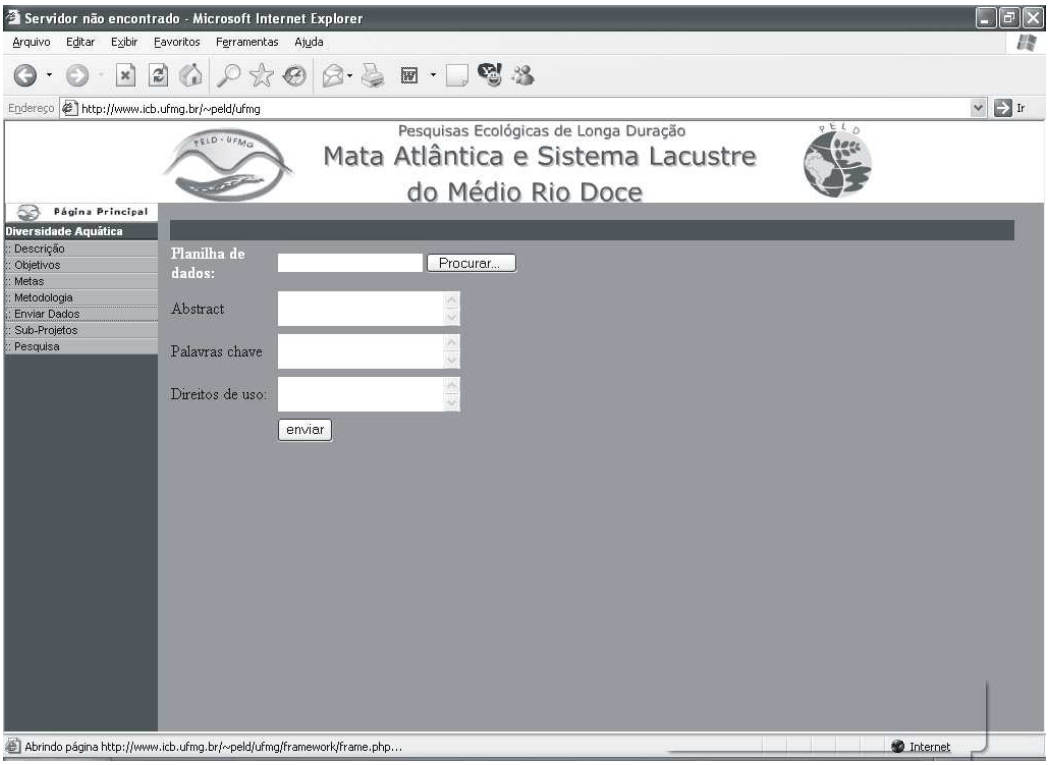


Figura 3.7: Página Web para submissão de arquivos *csv*

Componente ODL	Objetivo	Quem utiliza
ODLUnion Catalog	Colhe os dados nos provedores de dados e os armazena em tabela relacionais.	BDiG-PELD
OAI-XMLFile	Permite ao provedor de dados atender requisições da BDiG-PELD cujas respostas são dados EML.	Provedores de dados
ODL-IRDB	Oferece serviços de busca.	BDiG-PELD

Tabela 3.1: Componentes ODL utilizados no ambiente da arquitetura

concreto de rápida prototipação utilizando a infraestrutura ODL é a ETANA-DL[3], biblioteca digital para arqueologia citada na Seção 1.5, cuja implementação básica durou quatro meses de acordo com [43]. Uma outra alternativa para compartilhar dados ecológicos seria a construção de um sistema de informação de acordo com a arquitetura cliente-servidor, com bancos de dados centralizados ou mesmo distribuídos, mas que envolveria muita complexidade devido à diversidade de dados envolvida, a complexidade da infra-estrutura tecnológica e o longo tempo de implementação. Outra razão para a utilização da infraestrutura ODL é o fato ser de código aberto, muito importante para projetos dessa natureza.

[3]feathers.dlib.vt.edu

3.4 Implementação da BDiG-PELD

Para a implementação da BDiG-PELD, foi utilizado uma série de programas na linguagem PERL[4] que constituem a interface de acesso do ambiente. Essa interface se comunica com os componentes ODL utilizados, para os quais envia requisições dos usuários, recebendo seus resultados. Ao receber os dados de resposta, a interface os processa e os apresenta aos usuários. Dentre as principais requisições possíveis de um usuário citamos: 1) seleção de coletas por busca ou navegação, 2) exibição de dados das coletas selecionadas, 3) recuperação dos dados originais das coletas selecionadas e 4) georreferenciamento das mesmas. Os principais programas da interface de acesso, cujo desenvolvimento levou 6 meses, são apresentados na Tabela 3.2.

Programa	Objetivo	#linhas
search.pl	A partir dos termos de consulta do usuário, solicita à máquina de busca ODL, os registros ou coletas elegíveis. A partir do resultado, o usuário pode solicitar a exibição de dados, que ainda podem ser georreferenciados.	345
exibe_metadados.pl	Permite ao usuário a exibição dos metadados de um registro.	94
exibe_dados.pl	Possibilita a recuperação dos dados originais de coleta de um registro.	134
navega.pl	Possibilita a navegação entre os registros ou coletas armazenados na biblioteca digital. A partir dos resultados da navegação, o usuário pode exibir dados e metadados, assim como as facilidades de georreferenciamento.	1138

Tabela 3.2: Principais programas da interface de acesso da BDiG-PELD

Para avaliarmos qual o tamanho do esforço de codificação na implementação da BDiG-PELD usando a infraestrutura ODL, levantamos o total de linhas de código de cada componente ODL utilizado, o total de linhas de código da interface de acesso desenvolvida e também o total de linhas das transformações XSLT [14] usadas para processar e visualizar o conteúdo dos documentos XML gerados. Os dados desse levantamento são apresentados na Tabela 3.3 que também apresenta o percentual de codificação de cada componente em relação ao total de código da BDiG-PELD.

A análise da Tabela 3.3 permite constatar a real reutilização dos componentes ODL para implantação dos serviços básicos em uma biblioteca digital. Nesse aspecto, os componentes ODL constituem 63% de toda a codificação da BDiG-PELD. O restante envolve, basicamente, a interface de acesso (12%) e os programas XSLT (25%). É importante ressaltar que, já na fase de implementação, percebemos um desempenho ruim na utilização dos componentes ODL, principalmente quando o acesso a eles envolve o protocolo XOAI. Uma maneira de se minimizar esse problema de desempenho é permitir o acesso diretamente aos dados mantidos pelos

[4] www.perl.org

Componente	#Linhas de Código	Percentual
ODL-IRDB	8135	38%
ODL-DBUnion	5381	25%
OAI-XMLFile	4264	20%
Interface de acesso	3382	16%
Programas XSLT	7015	25%

Tabela 3.3: Quantidade de linhas de código

componentes, ou seja, eliminando-se o uso do protocolo XOAI. Isso pode ser feito por meio de consultas SQL submetidas diretamente aos bancos de dados mantidos por cada componente. Essa alteração será fruto de trabalhos futuros, pois não pôde ser realizada durante o desenvolvimento da versão-piloto.

Todos os componentes da Tabela 3.3 foram instalados no servidor de aplicações Web Apache[5] do Laboratório de Bancos de Dados da UFMG.

[5]www.apache.org

Capítulo 4

Serviços da BDiG-PELD

Neste capítulo apresentamos os serviços oferecidos pela BDiG-PELD. Este capítulo está divido em seções da seguinte forma. A Seção 4.1 descreve a especificação de requisitos conforme a abordagem 5S. A Seção 4.2 apresenta o serviço de busca da BDiG-PELD. Por sua vez, a Seção 4.3 descreve o serviço de navegação. Finalmente, a Seção 4.4 faz uma detalhamento das facilidades de georreferenciamento do ambiente.

4.1 Especificação de Requisitos

Conforme já mencionado, bibliotecas digitais necessitam de modelos para especificar adequadamente suas complexas interações e os requisitos de seus usuários. Basicamente, a construção de bibliotecas digitais envolve algumas decisões importantes e que se bem definidas podem ajudar muito nesse processo. Segundo [26], essas decisões são: (1) que tipo de conteúdo multimídia será apoiado, (2) como a informação é organizada e estruturada, (3) quais são as comunidades de usuários, e (4) quais serviços e facilidades poderão ser fornecidos a elas. Para entendermos melhor os requisitos da biblioteca digital proposta neste trabalho, adotamos a abordagem 5S [26, 27] para sua especificação.

A abordagem 5S permite modelar bibliotecas digitais levando em consideração cinco perspectivas diferentes e complementares: os tipos de dados multimídia suportados pela biblioteca digital *(Streams)*, como essas informações são estruturadas e organizadas *(Structures)*, os detalhes do modelo de recuperação de informação adotado, além das características da interface de usuário da biblioteca *(Spaces)*, os aspectos comportamentais da biblioteca digital *(Scenarios)* e as diferentes comunidades envolvidas *(Societies)* [26, 27]. A Tabela 4.1 mostra exemplos dessas perspectivas.

A seguir, com base em [45], definimos mais detalhadamente as perspectivas da abordagem 5S e especificamos melhor a BDiG-PELD, sob cada uma delas, ainda que informalmente.

4.1.1 *Societies*

A perspectiva *Societies* define, para uma biblioteca digital, um conjunto de entidades e os relacionamentos entre elas [27]. Essas entidades incluem pessoas, bem como componentes

Perspectiva	Exemplo
Streams	Texto, vídeo, imagem, áudio.
Structures	Coleções, catálogos, hipertextos, documentos, metadados.
Spaces	Vetorial, probabilístico.
Scenarios	Busca, navegação, auto-arquivamento.
Societies	Gerentes, educadores, pesquisadores.

Tabela 4.1: Exemplos de uso das perspectivas 5S

de hardware e software. Por exemplo, na BDiG-PELD, temos grupos de pesquisadores em biodiversidade que incluem biólogos, bolsistas, estudantes de pós-graduação, todos espalhados pelo Brasil, bem como pesquisadores das agências de fomento e parceiros (CNPq, Governo Estadual, etc). A comunidade em educação (professores de ecologia, educadores, alunos, pesquisadores em educação, etc.) também pode ser incluída sob essa perspectiva. Há também pessoal técnico envolvido, dentre eles: técnicos de campo e laboratório, guarda florestais, etc. Outra comunidade é formada por especialistas de outras áreas (economistas, prefeituras, entidades não governamentais, etc.) que também interagem com a BDiG-PELD, pois utilizam os resultados do Programa PELD em suas ações e políticas.

As relações entre essas comunidades também são analisadas sob essa perspectiva, determinando as regras e papéis a serem seguidos por suas entidades. Por exemplo, o CNPq, principal agência de fomento do programa PELD, estabelece que todos os dados coletados no programa devem ser de domínio público em um prazo máximo de dois anos a partir da data de coleta. Antes disso somente alguns dados podem ser disponibilizados publicamente e para se ter acesso a eles os autores devem ser contatados. O CNPq também requer o envio de relatórios com toda a produção científica sobre o projeto, assim como o andamento das atividades em todos os sítios de pesquisa do Programa.

4.1.2 *Structures*

Essa perspectiva diz respeito à organização das partes perante o todo. Em bibliotecas digitais, essa perspectiva define como organizar as estruturas de dados, como, por exemplo, um livro deve ser organizado em capítulos com seções, as quais contém subseções.

Quando aplicada à BDiG-PELD, essa perspectiva permite a observação de algumas características importantes, entre as quais: grande volume de dados e a presença de informações diversas nas coletas de campo, tais como, taxonomia, local físico, data e hora. Ainda observamos que os objetos digitais, ou melhor, as coletas de campo representadas na BDiG-PELD dizem respeito a sub-áreas bem diferentes da biodiversidade, tais como, genética, aquática, faunística e botânica. Cada uma dessas áreas possui vários projetos de pesquisa, como, por exemplo, o Parque Estadual do Rio Doce, sítio ecológico 4 do programa, que possui 16 projetos ao todo, identificados a partir de um levantamento inicial.

Citamos um exemplo de coleta de um desses projetos para percebemos o caráter histórico e, às vezes, o trabalho envolvido em uma coleta. Por exemplo, para levantamentos dos dados físicos e químicos em lagos é necessário que amostras de água sejam coletadas mensalmente

e analisadas em laboratório. Esse exemplo, mesmo que específico a um dos 12 sítios de pesquisa do Programa, dá uma noção sobre as particularidades das coletas, bem como o caráter histórico do dados gerados pelo Programa.

Para atender à essas características, decidimos que cada objeto do nosso repositório central representa uma coleta individual, sendo constituído pelo dados: projeto de pesquisa, data, local, taxonomia envolvida, dados amostrais propriamente ditos, entre outros. Todo esse conjunto de dados é descrito com metadados da linguagem EML, já apresentada na Subseção 2.3. Os objetos digitais mantidos na BDiG-PELD são separados por sítio ecológico e possuem uma identificação única dentro do sítio ecológico ao qual ele está associado, bem como em relação a todo o repositório.

4.1.3 *Scenarios*

Sob essa perspectiva, são definidas as interações de uma biblioteca digital com suas entidades, principalmente sua comunidade alvo. Como parte do processo para construção de sistemas de informação, essa perspectiva é considerada muito útil, pois define o que deve acontecer às *Streams* em *Spaces* através de *Structures*. Ou ainda, conforme [27], definem os fluxos de dados e trabalho em uma biblioteca digital.

Na BDiG-PELD, alguns *scenarios* contemplam desde o processo de coleta de dados, ou sua geração em experimentos de laboratório, até sua recuperação através de serviços da biblioteca digital. Alguns deles são apresentados a seguir:

1. De uma maneira geral, os registros das coletas, independentemente de origem e tipo, devem manter datas, locais, responsáveis, tipo de coleta, classificação taxonômica, etc. Felizmente, os tipos de coleta que um sítio realiza são previsíveis e com procedimentos bem definidos, o que torna gerenciável, embora complexo, o cenário de armazenamento digital dos objetos sobre coletas.

2. Possibilidade de navegação entre os objetos armazenados no repositório central da BDiG-PELD com agrupamento das coletas por local, data e taxonomia, em qualquer combinação e ordem. Por exemplo, navegar até as coletas de uma determinada espécie em um local específico. É importante oferecer alguns dados estatísticos, como, por exemplo, quantidade de ocorrências por grupo de coletas a cada passo da navegação. O serviço de navegação também deve combinar os dados de vários sítios PELD sob uma mesma interface.

3. Busca de coletas por um ou mais termos, mas com a possibilidade de recuperação de todos os dados associados às ocorrências resultantes da busca. Essa recuperação deve ser possível, também, com o uso do serviço de navegação.

4. Os dados propriamente ditos, quando recuperados a partir do repositório central, devem ser claramente interpretados por um usuário externo ao sítio que o gerou, bem como podem ser utilizados e analisados em outros experimentos.

5. Facilidades de georreferenciamento de uma seleção de ocorrências de coletas, ou todas, obtidas pelos serviços de busca ou de navegação.

4.1.4 *Spaces*

Sob essa perspectiva, podemos descrever os detalhes do modelo de recuperação de informação adotado em uma biblioteca digital, assim como ajuda a definir as características da aparência das interfaces de usuário. No caso da BDig-PELD, além do espaço vetorial utilizado em seu serviço de busca, outro importante aspecto, sob a perspectiva *Spaces*, é a distribuição geográfica das coletas, as quais podem ser visualizadas por meio de facilidades de georreferenciamento, como discutido na Seção 4.4.

4.1.5 *Streams*

Essa perspectiva relaciona-se às seqüências de elementos arbitrários (imagens, bits, caráteres, etc.) que representam conteúdo estático ou dinâmico. Na parte estática, um *stream* corresponde a qualquer conteúdo de informação que é interpretado como seqüências de elementos básicos do mesmo tipo. Um tipo comum de *stream* estático é um texto (seqüência de caracteres). O tipo de um *stream* estático define sua semântica e área de aplicação. Já um *stream* dinâmico representa um fluxo de informação ou seqüências de mensagens codificadas, sendo assim importantes para representar comunicação em qualquer ambiente que contenha uma biblioteca digital [23].

No caso da BDiG-PELD, a perspectiva *Streams* representa a enorme quantidade de dados textuais (em formato XML) em seus objetos, os quais representam coletas de campo diversas e históricas, bem como os dados gráficos ou cartográficos que podem ser visualizados através das facilidades de georreferenciamento.

4.2 Serviço de Busca

A BDiG-PELD oferece aos seus usuários um serviço de busca implementado através do componente ODL-IRDB da infraestrutura ODL e de uma interface própria, semelhante a dos serviços de busca convencionais. Para utilizar o serviço de busca o usuário deve informar, no campo dos formulários de busca, um ou mais termos e solicitar que a busca seja feita. São dois os formulários, o primeiro na página inicial da BDiG-PELD apresentada na Figura 4.1 e outro apresentado na Figura 4.2. A Figura 4.1 apresenta a página inicial da BDiG-PELD[1] que possui o formulário de busca à sua direita, onde o usuário pode especificar sua consulta. Já a Figura 4.2 apresenta o segundo formulário de busca, alcançado por meio da opção busca da barra de *links*, presente em todas as páginas, ou como resultado de uma busca iniciada pelo primeiro formulário.

Os resultados de uma busca são apresentados na página mostrada na Figura 4.3, a partir da qual todos os dados de uma coleta podem ser exibidos, bastando para isso clicar no

[1] www.lbd.dcc.ufmg.br/cgi-bin/bdigpeld/index.pl

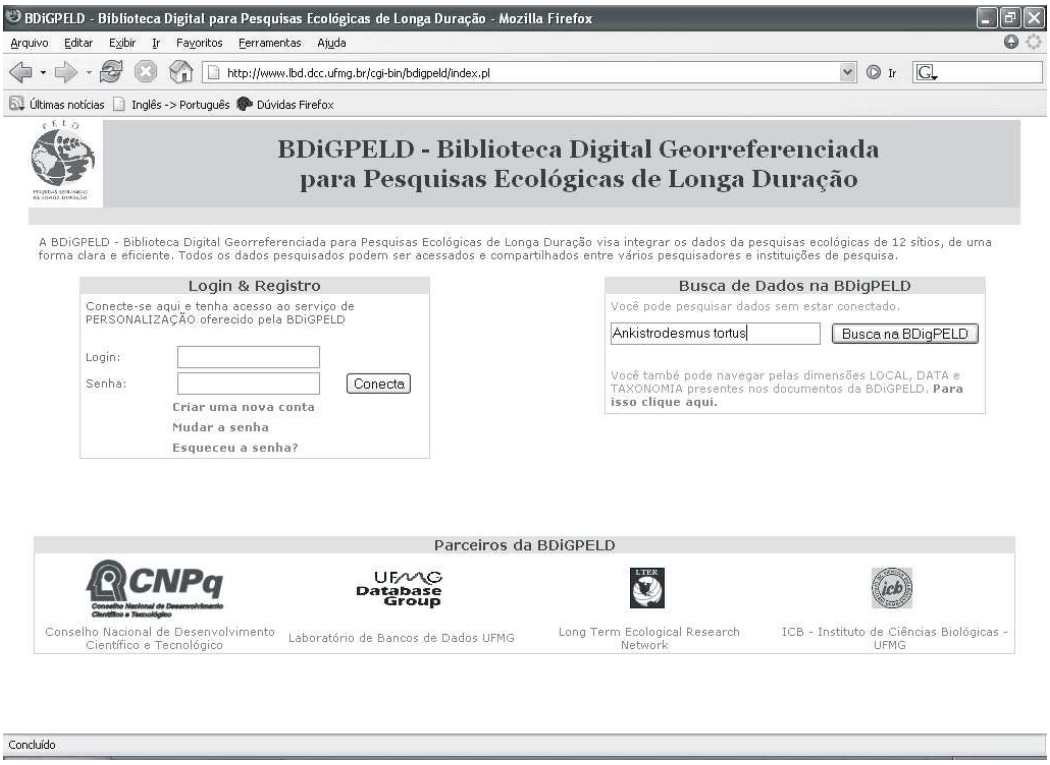


Figura 4.1: Pagina Inicial da interface da BDiG-PELD com formulário de busca

atalho *(hiperlink)* associado à ocorrência. Cada ocorrência dessa página possui o título da coleta, autor, instituição e resumo, se o mesmo existir.

A seleção de uma ocorrência através do atalho disponibilizado exibe os principais dados de uma coleta, como podem ser vistos nas Figuras 4.4, 4.5, 4.6 e 4.7. Mais especificamente, a Figura 4.4 apresenta os dados gerais de uma coleta (título, autor, local, hora da coleta, etc.). Já a Figura 4.5 exibe os dados da classificação taxonômica das espécies envolvidas em uma coleta. A Figura 4.6, por sua vez, apresenta os dados do projeto de pesquisa de uma coleta (nome, local do projeto, pessoal, etc.). Finalmente, a Figura 4.7, exibe os dados de uma coleta (nome, descrição, tipos de dados, etc.). Todas as Figuras de 4.4 a 4.7 são resultado de transformações de documentos EML, realizadas por meio da linguagem XSLT, e refletem os principais conjuntos de dados de uma coleta.

O serviço de busca ainda permite exibir e recuperar os dados originais de uma coleta. A exibição é possível através de um atalho constante da página que exibe os dados sobre a coleta, conforme pode ser visto ao final da página apresentada na Figura 4.7. O resultado do acesso a esse atalho pode ser verificado na página da Figura 4.8, a qual exibe os dados originais da coleta em formato tabular. Para a recuperação desses dados diretamente em um arquivo, basta selecionar o segundo atalho também disponibilizado ao final da página da Figura 4.7. O resultado dessa recuperação, em formato *csv*, é apresentado na planilha da Figura 4.9.

As ocorrências de uma busca também podem ser marcadas para georreferenciamento, conforme descrito na Subseção 4.4.

Figura 4.2: Segundo formulário de busca da interface da BDiG-PELD

4.3 Serviço de Navegação

4.3.1 Módulo de Navegação

Serviços de busca são muito úteis para recuperar informação com base em palavras-chaves. No entanto, serviços de navegação permitem uma maior liberdade e poder ao usuário quando o objetivo é descobrir novas informações sob determinadas categorias. Pensando nisso, a BDiG-PELD oferece um serviço de navegação sofisticado que combina acesso a múltiplas categorias de dados. Através do serviço de navegação, implementado na BDiG-PELD por meio de um *data mart* [49], é possível fazer a integração semântica dos objetos obtidos através da interoperação entre várias fontes de dados, conforme sugere Obrst [39].

O processo de carga do *data mart* envolve a extração periódica, em todos os registros mantidos no repositório do ambiente da BDiG-PELD, dos dados que representam as categorias ou dimensões de navegação desejadas, que são local geofísico, data, projeto e classificação taxonômica. Esses dados são representados por metadados específicos da EML e, após suas extrações, são combinados nas tabelas do *data mart*. A Figura 4.10 ilustra, na notação gráfica da ferramenta DBDesinger [5], o esquema relacional desse *data mart*. As tabelas periféricas apresentadas nessa figura são tabelas dimensão de um *data mart* e a tabela central é a sua tabela fato.

Embora nossa estratégia de integração de objetos digitais tenha sido realizada por meio de um *data mart*, fizemos uso efetivo da semântica implícita entre os objetos, expressa por

Figura 4.3: Página com resultados de uma busca

meio de metadados. Tal estratégia, segundo Obrst em [39], ainda é usual e bem sucedida, mas se destina, essencialmente, à interoperação entres sistemas com bom nível de acoplamento, como, por exemplo, sistemas baseados em bancos de dados relacionais, nos quais a integração é realizada com operações cujas estruturas e sintaxes são muito próximas. Isso diverge um pouco da crescente necessidade de integração com semânticas mais explícitas, encampadas, atualmente, pelo uso de ontologias. Trabalhos relacionados à integração e interoperação entre sistemas por meio de ontologias podem ser encontrados em [38, 39].

Outra alternativa para prover o serviço de navegação seria a utilização do componente ODL-Browse da infraestrutura ODL [47]. No entanto, esse componente só contempla uma dimensão ou categoria de dados por vez. Para trabalharmos com múltiplas dimensões simultaneamente seria necessário um esforço maior de implementação, sem a garantia da desejada integração semântica. Por essa razão, implementamos nosso próprio serviço de navegação.

4.3.2 Interface de Navegação

O serviço de navegação é iniciado pela página mostrada na Figura 4.11, na qual podemos perceber três quadros para as categorias ou dimensões de navegação, que são taxonomia, local e data. Apesar da dimensão projeto ter sido prevista no módulo de navegação ela não foi incluída para encurtar o tempo de implementação da interface. Nos trabalhos futuros, essa dimensão será incluída.

A partir dessa página o usuário pode, simultaneamente, e em tempo real, selecionar

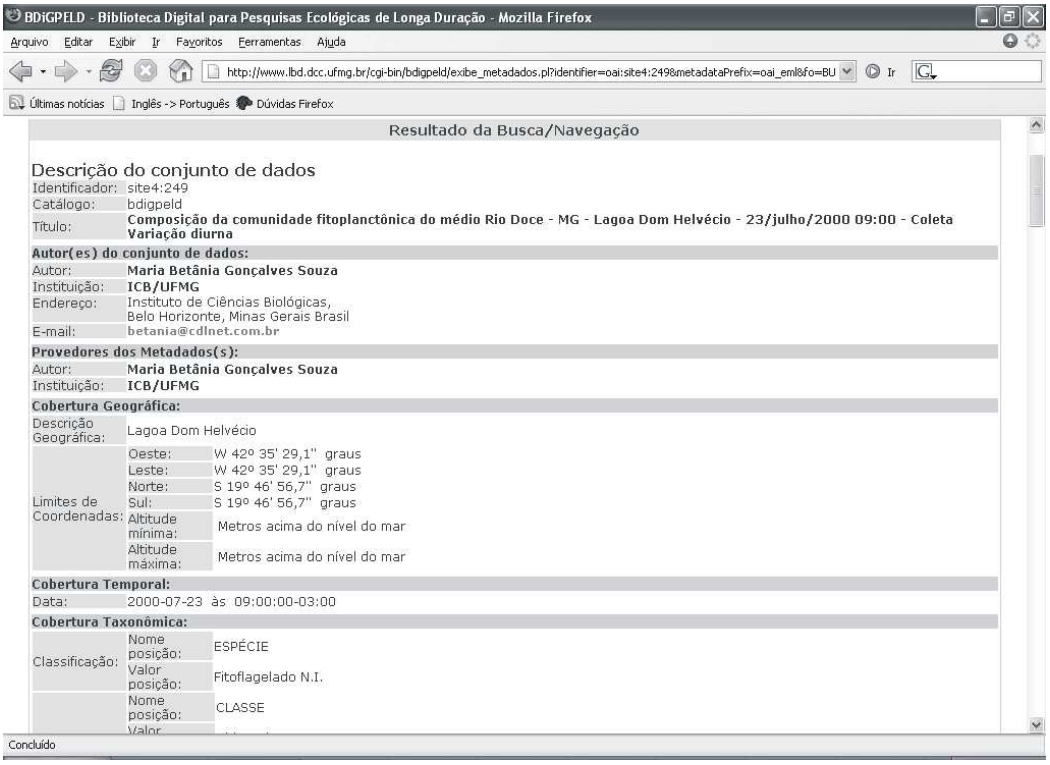


Figura 4.4: Dados gerais de uma coleta

atalhos *(links)*, os quais representam subcategorias possíveis dentro de uma categoria de navegação (taxonomia, local e data). Mais especificamente, essas categorias e suas subcategorias são apresentadas a seguir.

Taxonomia :

Reino —> Divisão —> Sub-Divisão —> Filo —> Sub-Filo —> Super-Classe —> Classe —> Sub-Classe —> Super-Ordem —> Ordem —> Sub-Ordem —> Infra-Ordem —> Super-Família —> Família —> Tribo —> Sub-Tribo —> Gênero —> Espécie —> Sub-Espécie —> Infra-Espécie —> Variedade

Local :

Sítio Ecológico —> Local no Sítio

Data :

Data Inicial —> Ano e/ou Data Inicial —> Mes, e/ou
Data Final —> Ano e/ou Data Final —> Mes

4.3.2.1 Navegação por Taxonomia

Um exemplo de navegação em algumas subcategorias de Taxonomia é apresentado na página da Figura 4.12, na qual aparecem 7 subcategorias selecionadas simultaneamente (divisão, filo,

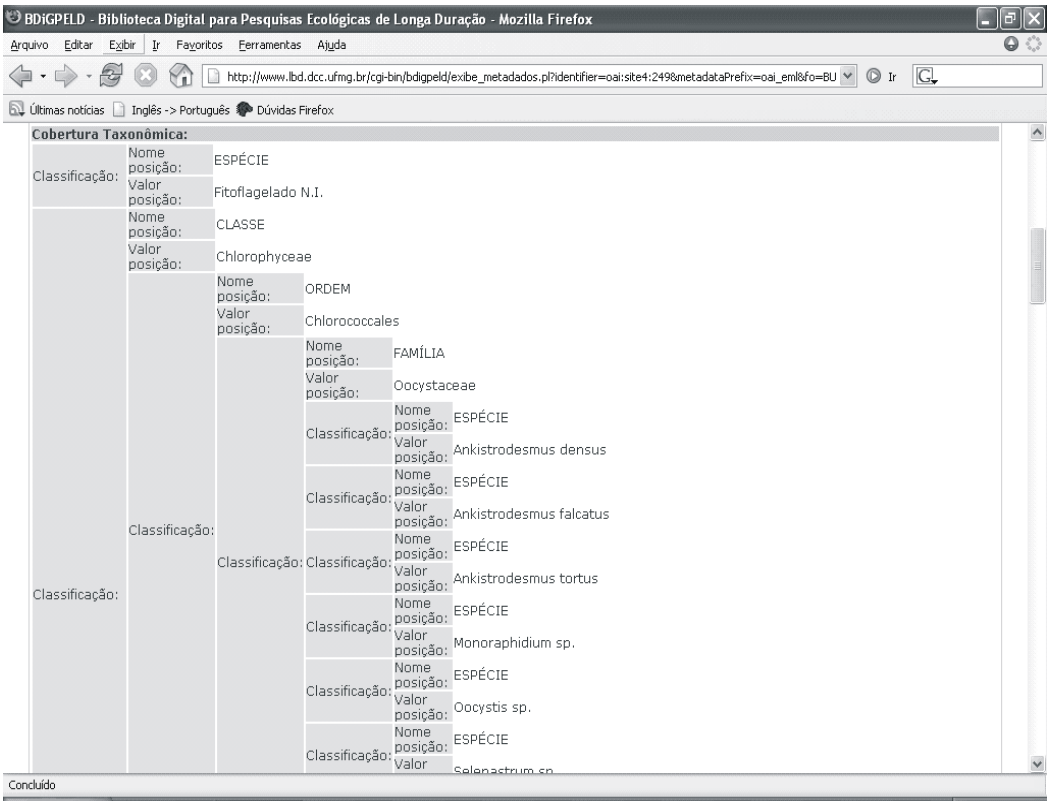


Figura 4.5: Dados da classificação taxonômica de uma coleta

classe, sub-classe, família, gênero e espécie), dispostas em ordem de hierarquia da esquerda para direita.

As ocorrências ou instâncias resultantes das subcategorias selecionadas podem ser acessadas por *links*, estabelecendo um filtro ou restrição para a categoria corrente, como, por exemplo, o filtro a partir de uma espécie, apresentado na página da Figura 4.13. No exemplo dessa figura, a instância *"Ankistrodemus tortus"* da subcategoria Espécie foi estabelecida como filtro para a categoria taxonomia. A partir desse filtro, qualquer outra categoria, incluída na mesma navegação, deve considerar a espécie *"Ankistrodemus tortus"* como critério para os resultados. Esse mesmo comportamento pode ser aplicado às outras categorias.

4.3.2.2 Navegação por Taxonomia e Local

O usuário também pode navegar por duas dimensões simultaneamente, como é o caso do exemplo apresentado na página da Figura 4.14. Nela podemos observar que o usuário selecionou a subcategoria "Sítio" da categoria Local. Ainda nessa figura é possível ver o resultado do filtro estabelecido na página da Figura 4.14, ou seja, são apresentados, como resultado final, todos os sítios ecológicos onde ocorrem a espécie *"Ankistrodemus tortus"*.

Além disso, durante a realização da navegação é apresentada a quantidade de coletas existentes para o combinação entre as subcategorias usadas e filtros estabelecidos. Dessa forma, o usuário pode fazer análises de distribuição de coletas cruzando subcategorias e filtros

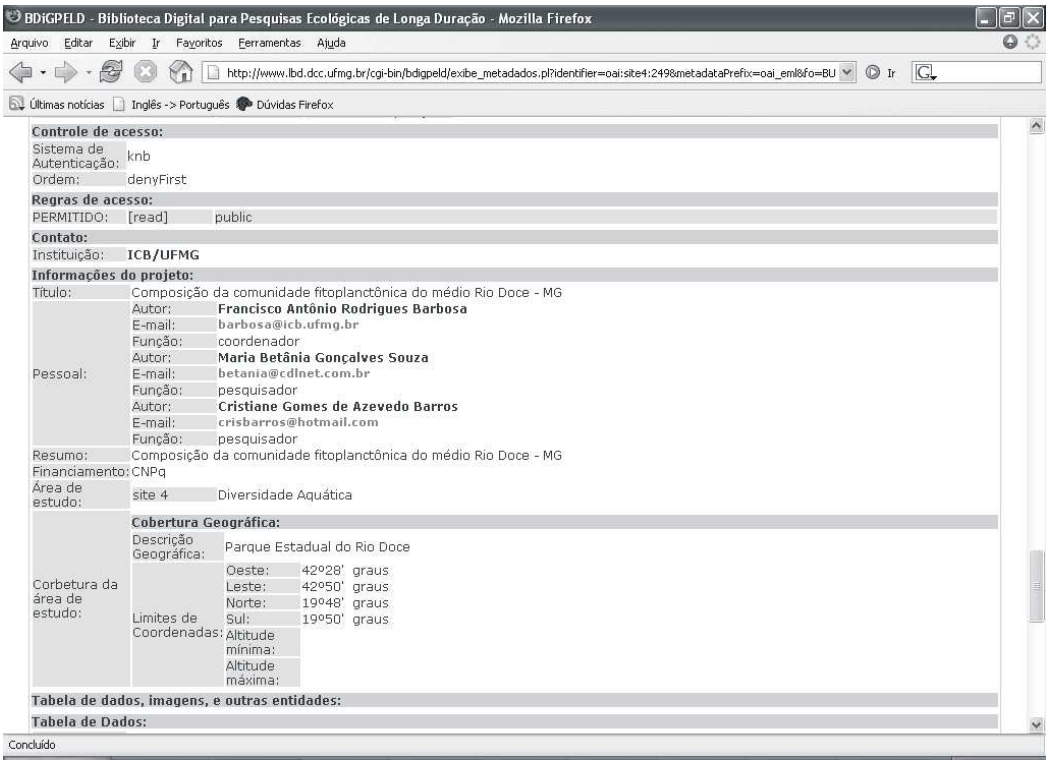


Figura 4.6: Dados do projeto de pesquisa

de forma fácil e intuitiva. Por exemplo, na Figura 4.14 há o quadro "Resultado da Navegação" que mostra que existem quatro coletas para a espécie *"Ankistrodemus tortus"* no "Parque do Rio Doce". Nesse caso, foram escolhidas somente as subcategorias Local e Espécie, ambas de categorias diferentes, além do filtro *"Ankistrodemus tortus"*, para a subcategoria espécie.

4.3.2.3 Navegação por Taxonomia, Local e Data

Como já era previsto, o usuário também pode incluir em uma navegação a categoria Data, combinando três categorias simultaneamente. Um exemplo para essa combinação é apresentado na página da Figura 4.15, na qual o usuário estabelece um filtro, sobre a categoria Data, restringindo os resultados às coletas realizadas em 2002. A partir dessa mesma figura também pode ser observado que o quadro "Resultado da Navegação" reflete, agora, uma nova distribuição de coletas, a qual envolve as três categorias disponíveis para navegação, além do filtro *"Ankistrodemus tortus"*, para espécie.

A partir dos resultados de uma navegação, o usuário pode listar as coletas associadas, bastando, para isso, pressionar o botão "Lista Amostras", depois de selecioná-las no quadro "Resultado da Navegação". O resultado dessa operação é apresentado na página da Figura 4.16, cujos *links* permitem, ainda, a exibição de dados, no mesmo formato apresentado pelo serviço de busca (Seção 4.2), ou o georreferenciamento das coletas selecionadas.

Para remover subcategorias ou filtros em uma navegação, basta clicar nos atalhos associados aos mesmos. Ao fazer isso, o resultado da navegação é imediatamente atualizado.

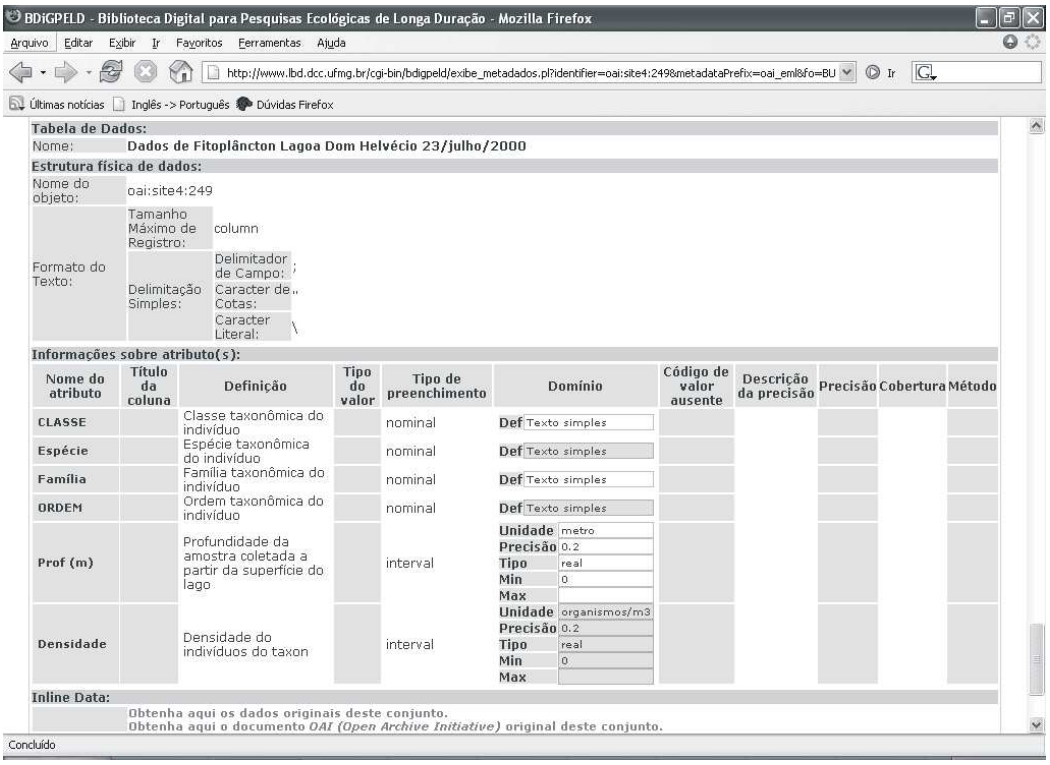


Figura 4.7: Dados sobre atributos de coleta

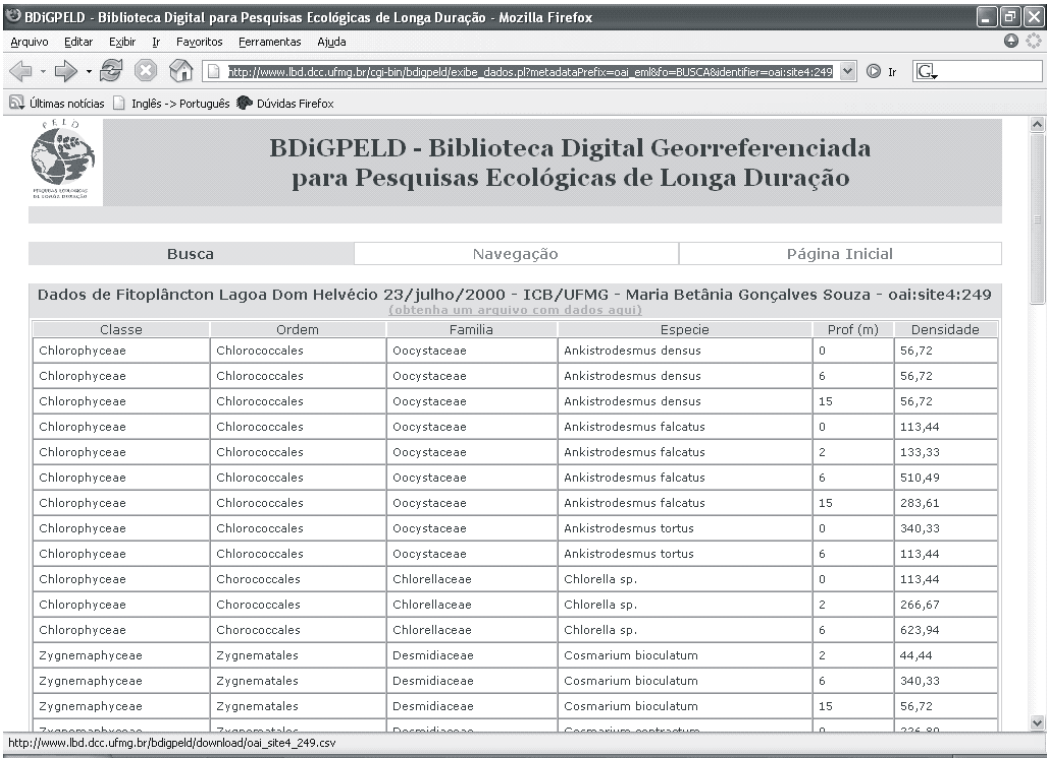
4.4 Facilidades de Georreferenciamento

4.4.1 Locus

As facilidades de georreferenciamento oferecidas pela BDiG-PELD utilizam o Locus [21, 22], localizador geográfico desenvolvido no Laboratório de Bancos de Dados da UFMG, que provê, com base em um *gazetteer*, um serviço de localização de lugares através de referências espaciais diretas e indiretas. Hill [30] define um *gazetteer* como um catálogo de nomes de lugares (um dicionário toponímico) capaz de fornecer um vocabulário de termos geográficos, acompanhados de suas respectivas localizações.

O Locus foi modelado a partir de uma ontologia de lugar. Essa ontologia reflete a forma como as pessoas percebem e usam a informação geográfica. Nela, o conceito de lugar considera aspectos cognitivos na sua representação, e não apenas sua geometria. Entre os conceitos mais importantes representados estão a hierarquia de subdivisões territoriais, o sistema de endereçamento postal e locais de referência conhecidos da população, como, por exemplo, acidentes geográficos, edificações, serviços e outros. A arquitetura do Locus pode ser observada na Figura 4.17. O *gazetteer* é o componente responsável pelo armazenamento dos registros de lugares. Cada registro possui no mínimo três dados básicos: nome, tipo e localização geográfica.

O acesso ao sistema pode ocorrer via serviço WFS ou através de *servlets* Java que implementam interfaces para consulta e apresentação de resultados. A BDiG-PELD utiliza o



Classe	Ordem	Familia	Especie	Prof (m)	Densidade
Chlorophyceae	Chlorococcales	Oocystaceae	Ankistrodesmus densus	0	56,72
Chlorophyceae	Chlorococcales	Oocystaceae	Ankistrodesmus densus	6	56,72
Chlorophyceae	Chlorococcales	Oocystaceae	Ankistrodesmus densus	15	56,72
Chlorophyceae	Chlorococcales	Oocystaceae	Ankistrodesmus falcatus	0	113,44
Chlorophyceae	Chlorococcales	Oocystaceae	Ankistrodesmus falcatus	2	133,33
Chlorophyceae	Chlorococcales	Oocystaceae	Ankistrodesmus falcatus	6	510,49
Chlorophyceae	Chlorococcales	Oocystaceae	Ankistrodesmus falcatus	15	283,61
Chlorophyceae	Chlorococcales	Oocystaceae	Ankistrodesmus tortus	0	340,33
Chlorophyceae	Chlorococcales	Oocystaceae	Ankistrodesmus tortus	6	113,44
Chlorophyceae	Chorococcales	Chlorellaceae	Chlorella sp.	0	113,44
Chlorophyceae	Chorococcales	Chlorellaceae	Chlorella sp.	2	266,67
Chlorophyceae	Chorococcales	Chlorellaceae	Chlorella sp.	6	623,94
Zygnemaphyceae	Zygnematales	Desmidiaceae	Cosmarium bioculatum	2	44,44
Zygnemaphyceae	Zygnematales	Desmidiaceae	Cosmarium bioculatum	6	340,33
Zygnemaphyceae	Zygnematales	Desmidiaceae	Cosmarium bioculatum	15	56,72

Figura 4.8: Dados originais de uma coleta

serviço WFS[2] na fase de inserção dos objetos no Locus e *servlets* durante as consultas.

4.4.2 Interface de Georreferenciamento

Um exemplo da interface de georreferenciamento, integrada ao serviço de busca, pode ser observado na Figura 4.18. A tela à esquerda apresenta o resultado de uma busca textual na BDiG-PELD pelo termo *zooplâncton*, cujas ocorrências foram georreferenciadas na tela da direita.

Outra possibilidade de georreferenciar amostras é através do serviço de navegação que permite selecionar coletas através da combinação de categorias (tempo, local e taxonomia). Após realizar uma navegação, o usuário pode escolher, entre os objetos recuperadas, aquelas que deseja georreferenciar. Por exemplo, se uma espécie específica for acessada na navegação, podem ser georreferenciadas todas as coletas relativas a ela, mesmo sendo de sítios ecológicos diferentes.

[2] http://www.opengeospatial.org/specs/

http://www.lbd.dcc.ufmg.br/bdigpeld/download/oai_site4_249.csv - Microsoft Internet Explorer

	Classe	Ordem	Familia	Especie	Prof (m)	Densidade
2	Chlorophyceae	Chlorococcales	Oocysaceae	Anksroesmus ensus	0	56,72
3	Chlorophyceae	Chlorococcales	Oocysaceae	Anksroesmus ensus	6	56,72
4	Chlorophyceae	Chlorococcales	Oocysaceae	Anksroesmus ensus	15	56,72
5	Chlorophyceae	Chlorococcales	Oocysaceae	Anksroesmus falcaus	0	113,44
6	Chlorophyceae	Chlorococcales	Oocysaceae	Anksroesmus falcaus	2	133,33
7	Chlorophyceae	Chlorococcales	Oocysaceae	Anksroesmus falcaus	6	510,49
8	Chlorophyceae	Chlorococcales	Oocysaceae	Anksroesmus falcaus	15	283,61
9	Chlorophyceae	Chlorococcales	Oocysaceae	Anksroesmus orus	0	340,33
10	Chlorophyceae	Chlorococcales	Oocysaceae	Anksroesmus orus	6	113,44
11	Chlorophyceae	Chorococcales	Chlorellaceae	Chlorella sp.	0	113,44
12	Chlorophyceae	Chorococcales	Chlorellaceae	Chlorella sp.	2	266,67
13	Chlorophyceae	Chorococcales	Chlorellaceae	Chlorella sp.	6	623,94
14	Zynemaphyceae	Zynemaales	Desmaceae	Cosmarum boculaum	2	44,44
15	Zynemaphyceae	Zynemaales	Desmaceae	Cosmarum boculaum	6	340,33
16	Zynemaphyceae	Zynemaales	Desmaceae	Cosmarum boculaum	15	56,72
17	Zynemaphyceae	Zynemaales	Desmaceae	Cosmarum conracum	0	226,89
18	Zynemaphyceae	Zynemaales	Desmaceae	Cosmarum conracum	2	222,22
19	Zynemaphyceae	Zynemaales	Desmaceae	Cosmarum conracum	6	794,1
20	Zynemaphyceae	Zynemaales	Desmaceae	Cosmarum conracum	15	113,44
21	Zynemaphyceae	Zynemaales	Desmaceae	Cosmarum sp.	0	170,16
22	Zynemaphyceae	Zynemaales	Desmaceae	Cosmarum sp.	6	56,72
23	Crypophyceae	Crypomonaales	Crypomaaceae	Crypomonas spp.	0	226,89
24	Crypophyceae	Crypomonaales	Crypomaaceae	Crypomonas spp.	2	133,33
25	Crypophyceae	Crypomonaales	Crypomaaceae	Crypomonas spp.	6	453,77
26	Crypophyceae	Crypomonaales	Crypomaaceae	Crypomonas spp.	15	113,44
27				Foflaelao N.I.	0	2665,91
28				Foflaelao N.I.	2	400
29				Foflaelao N.I.	6	1361,32
30				Foflaelao N.I.	15	567,21
31	Cyanophyceae	Nosocales	Nosocaceae	Lmnohr sp.	2	88,89
32	Cyanophyceae	Nosocales	Nosocaceae	Lmnohr sp.	15	56,72

Figura 4.9: Dados originais de uma coleta recuperados em formato *csv*

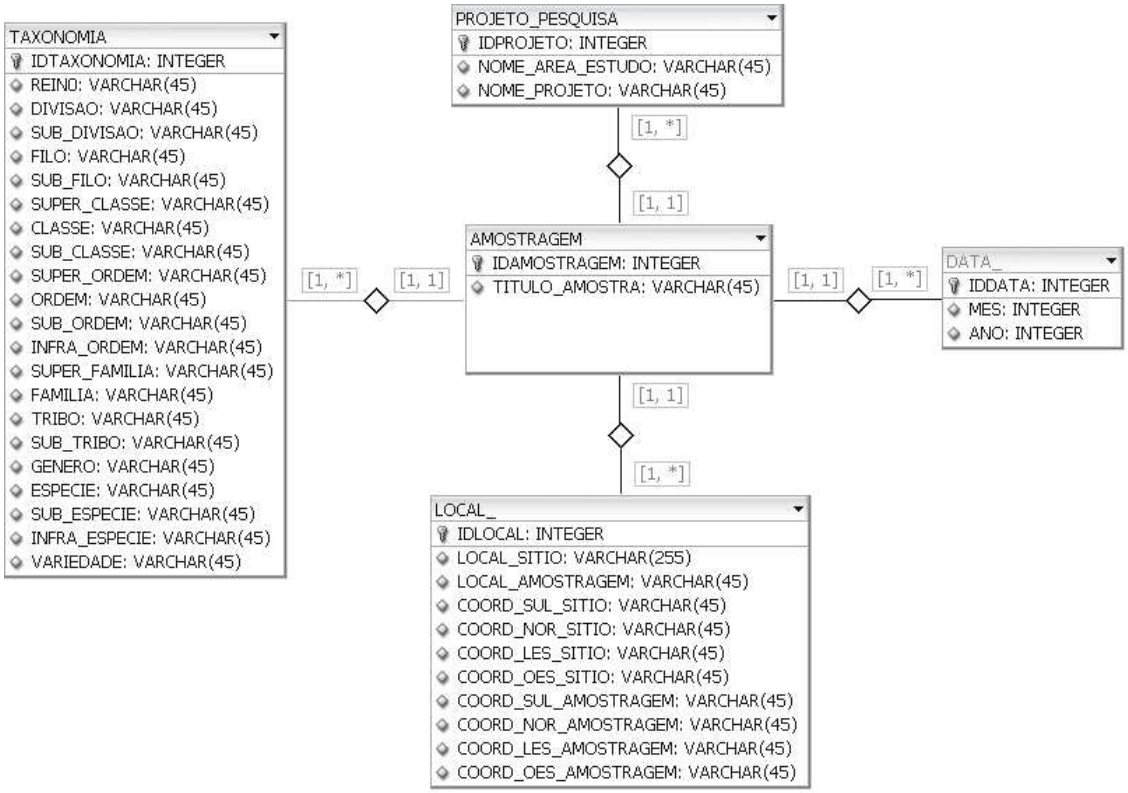


Figura 4.10: Esquema relacional do *data mart* do serviço de navegação

Figura 4.11: Página inicial do serviço de navegação

Figura 4.12: Navegação com várias subcategorias de taxonomia

Figura 4.13: Filtro a partir de uma instância da subcategoria espécie

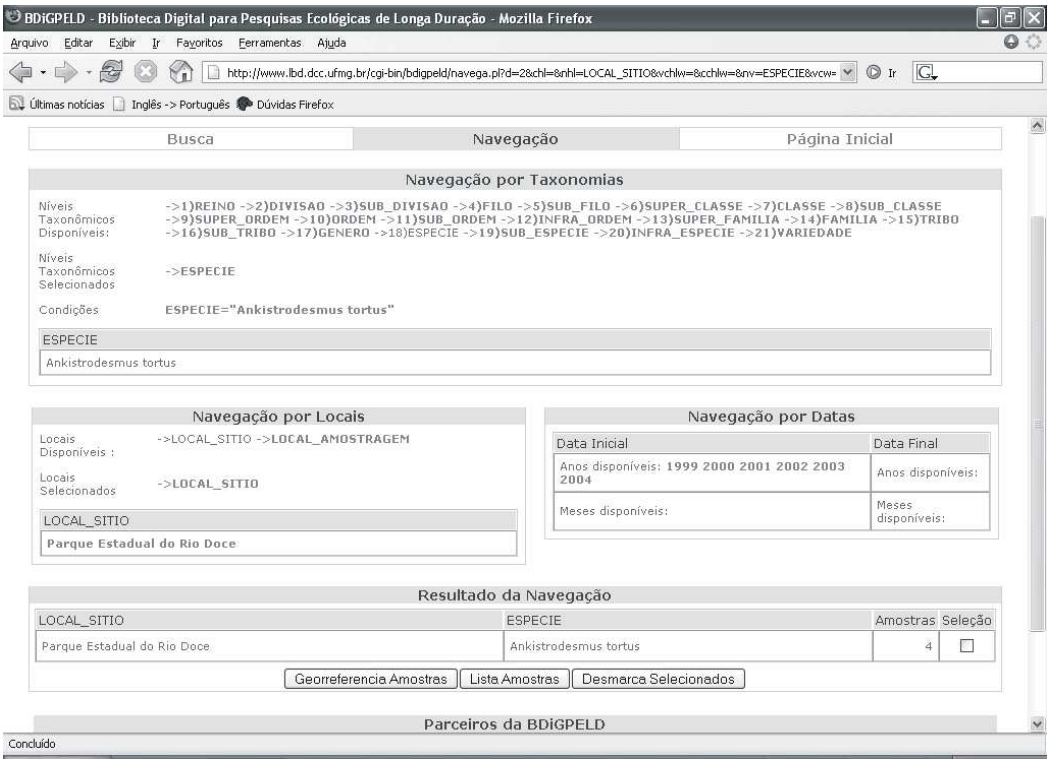


Figura 4.14: Navegação com a combinação das categorias Taxonomia e Local

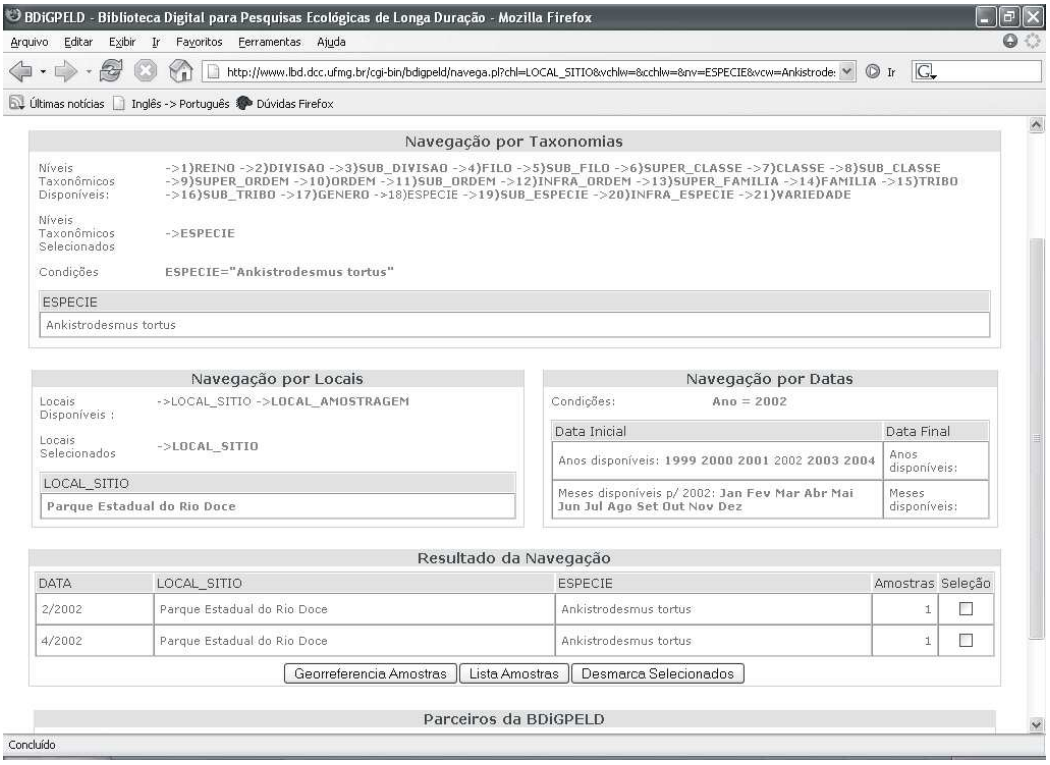


Figura 4.15: Navegação com a combinação das categorias Taxonomia, Local e Data

Figura 4.16: Lista de amostras obtidas através do Serviço de Navegação

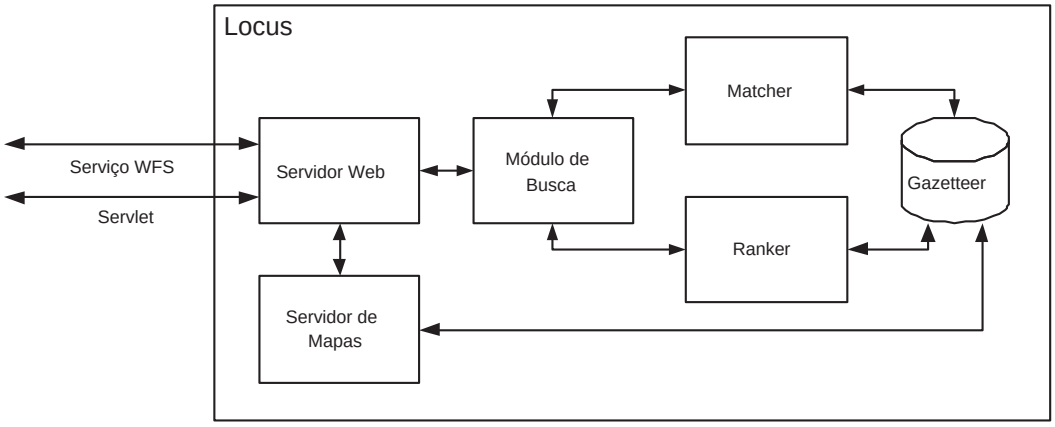


Figura 4.17: Arquitetura do módulo Locus

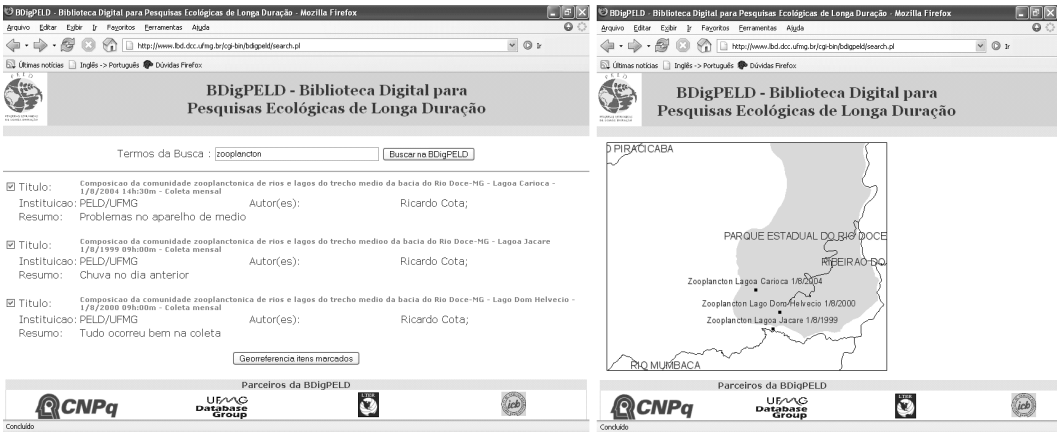


Figura 4.18: Interfaces de busca e georreferenciamento

Capítulo 5

Avaliação Experimental

Neste capítulo apresentamos uma avaliação experimental da BDiG-PELD com o objetivo de validar sua efetividade como ambiente para integração de dados do Programa PELD. O capítulo está dividido da seguinte forma. A Seção 5.1 descreve o experimento realizado. Já a Seção 5.2, por sua vez, apresenta os resultados obtidos no experimento. Finalmente, a Seção 5.3 discute e analisa os resultados obtidos.

5.1 Descrição do Experimento

Uma versão-piloto da BDiG-PELD, segundo a arquitetura apresentada neste trabalho, foi toda implementada para o sítio 4 do Programa PELD, cuja base é o Parque Estadual do Rio Doce, o qual não possuía um banco de dados que pudesse torná-lo um provedor de dados. Assim sendo, tivemos que desenvolver o banco de dados, a interface de entrada de dados e um sistema de informação via Web para manutenção do banco de dados. Esse sistema permitiu manipular e consultar os dados básicos do sítio, tais como, áreas de pesquisa, nome dos pesquisadores, locais de coleta, etc. Para criação do banco de dados e implementação da interface de entrada de dados, fizemos um intenso levantamento de dados no sítio 4, onde identificamos dezesseis[1] tipos de coleta, cada qual representando um subprojeto de pesquisa, sendo que para cada um foi definido um arquivo com modelo próprio para preenchimento dos dados. As coletas de campo foram apontadas em seu respectivos arquivos para posterior submissão ao provedor de dados através da interface de entrada de dados.

Ao receber um arquivo *csv*, a interface de entrada de dados (Subseção 3.2) identificou qual o tipo da coleta e a processou, alimentando o banco de dados local e gerando os dados para coleta posterior pelo provedor de serviços, no caso a BDiG-PELD. Com essa abordagem agilizamos o processo de entrada de dados, tanto por permitir o apontamento autônomo dos dados, como por utilizar algumas coletas já apontadas em planilhas eletrônicas, cujas estruturas procuramos preservar. Essa abordagem também nos poupou tempo por não termos que desenvolver interfaces específicas para cada tipo de coleta do sítio. No entanto, ela poderia

[1] www.icb.ufmg.br/∼peld/ufmg

ser questionada quanto aos critérios de usabilidade e qualidade dos dados gerados, o que nos levou a realizar um experimento para avaliação.

No experimento, solicitamos a 7 usuários, pesquisadores das áreas de diversidade Genética, Botânica e Aquática, que preenchessem, sem qualquer ajuda, os dados de algumas coletas em arquivos *csv* (ou planilhas como esses arquivos são conhecidos pelos usuários) e que respondessem a um questionário antes de submetê-los através da interface de entrada de dados. As perguntas constantes do questionário são listadas a seguir:

1. Como você classificaria a sua experiência em submissão de dados através de planilhas ANTES de começar o experimento?
2. Como você classificaria o seu conhecimento das planilhas que utilizou, ANTES de submetê-las?
3. Como você classificaria o seu conhecimento das planilhas que utilizou, DEPOIS de submetê-las?
4. Qual o grau de facilidade de preenchimento das planilhas?
5. Quão confortável você se sentiu utilizando as planilhas?
6. Você acha que planilhas são úteis para submeter corretamente os dados de coletas do Programa PELD ao banco de dados local?
7. Você acha que a planilha que você usou representa corretamente os dados das coletas da sua linha de pesquisa?
8. Você acha que a ordem dos dados nas planilhas está adequada para o preenchimento?
9. Escreva no espaço abaixo alguns comentários a respeito das planilhas e da interface de submissão ou sobre a sua experiência ao usar o serviço.

Também foi solicitado que o tempo de preenchimento de cada planilha fosse registrado. As perguntas do questionário tinham como objetivo obter as seguintes informações: conhecimento da planilha utilizada (perguntas 1 a 3), grau de facilidade e conforto do processo de entrada de dados (perguntas 4 e 5), e grau de representatividade da planilha em relação ao tipo de coleta (perguntas 6, 7 e 8). Para as perguntas de 1 a 7, as respostas poderiam ser expressas no intervalo de 1 a 7, indicando escalas que variavam, de acordo com a pergunta, de baixo a alto, pouco a muito, pouco útil a muito útil e menos representativa a mais representativa. Para a pergunta 8 a resposta deveria ser sim ou não. O questionário inclui ainda um campo para comentários sobre o processo de entrada de dados.

Os usuários envolvidos no experimento, embora em número reduzido, representam 3 das mais importantes áreas do programa e que contemplam, juntas, 7 linhas de pesquisas com inúmeros participantes. Outros fatores que reforçam a representatividade do experimento são o número de amostras ou coletas apontadas, ao todo 168, e a complexidade das mesmas, pois em algumas delas são encontradas até 143 espécies de seres vivos e em outras o processo de coleta foi bastante demorado.

Pergunta/Questionário	Q1	Q2	Q3	Q4	Q5	Q6	Média
1	7	2	2	6	7	4	4,7
2	7	7	7	6	7	5	6,5
3	7	7	7	6	7	6	6,7
4	7	6	7	6	7	6	6,5
5	4	2	2	6	5	4	3,8
6	7	4	3	6	5	7	5,3
7	4	7	6	7	6	7	6,2
8	Não	Sim	Sim	Sim	Não	Sim	
Quantidade de Planilhas	28	23	29	2	6	80	Total 168
Área de Diversidade	Gen.	Aqu.	Aqu.	Bot.	Gen.	Aqu.	Total
Tempo de preenchimento total (mim)	172,0	36,0	69,7	360,0	73,0	199,9	910,6
Tempo médio (min)	6,1	1,6	2,7	180,0	12,1	2,5	Média 5,4

Tabela 5.1: Resultados do experimento de entrada de dados

5.2 Resultados

5.2.1 Aspectos de Usabilidade

Após o preenchimento dos dados nas planilhas, foram recebidos 6 questionários, pois 2 usuários responderam um único questionário. No entanto, das 168 planilhas preenchidas, apenas foram submetidas 118, gerando 118 documentos EML. Os resultados são apresentados na Tabela 5.1. Os usuários foram também instruídos a fazer a submissão sem ajuda. Para qualquer erro na submissão seriam apresentados seu motivo e o que deveria ser feito. Se o erro persistisse, eles deveriam procurar por ajuda, o que ocorreu principalmente devido ao preenchimento incorreto de taxonomias.

Uma análise mais criteriosa dos resultados da Tabela 5.1 nos permite destacar as seguintes observações:

1. Pelas médias altas nas perguntas relacionadas ao conhecimento das planilhas utilizadas (perguntas 1 a 3), estimamos que os pesquisadores as conhecem bem. Isso é explicado pelo fato deles próprios terem participado da definição dos seus esquemas.

2. Para o grupo de perguntas relacionadas ao grau de facilidade e conforto (perguntas 4 e 5), os pesquisadores consideram, pelas médias, alto o grau de facilidade e médio o conforto em utilizá-las. O alto grau de facilidade se explica pela fluência dos pesquisadores em utilizar planilhas eletrônicas em seus trabalhos e pelo fato de conhecerem as planilhas utilizadas. Já o conforto médio pode ser atribuído ao trabalho em redigitar dados de campo e coleta já apontados em outros meios, como cadernos de campo e até mesmo outras planilhas, cujos esquemas podem ser diferentes daqueles das planilhas utilizadas.

3. As planilhas foram consideradas úteis pelos pesquisadores e bem representativas dos seus dados, conforme os valores médios das perguntas 6 e 7, respectivamente.

4. De uma maneira geral, a opinião quanto à ordem dos campos nas planilhas foi considerada boa (pergunta 8), embora alguns pesquisadores tenham discordados disso, pois tiveram que preencher os dados em uma seqüência diferente da habitual. É importante ressaltar que a ordem proposta facilita a utilização dos dados por pesquisadores alheios a eles, bem como seu entendimento.

5. O tempo médio aproximado de 5,4 minutos para preenchimento de uma planilha não é considerado alto, embora um dos pesquisadores tenha levado 6 horas para preencher duas planilhas, pois o preenchimento foi feito durante o experimento em campo, conforme observação em seu questionário. Esse caso particular corresponde a um experimento normalmente demorado que trata da mensuração de características físicas de árvores em um hectare. Se retirarmos o respectivo questionário da amostra o tempo médio cai para 3,30 minutos, aproximadamente.

Quanto aos comentários deixados nos questionários, percebemos dois fatos interessantes. O primeiro é a percepção que a BDiG-PELD deveria permitir a análise dos dados e o armazenamento estruturado dos mesmos, como em sistemas desenvolvidos especificamente para uma aplicação. O segundo é a visão particular de alguns pesquisadores sobre seus objetos digitais, ou melhor, coletas. Na BDiG-PELD os objetos são orientados por evento de coleta, ou seja, cada evento de coleta torna-se um objeto digital. Para alguns pesquisadores, um objeto deveria representar várias coletas, como é o caso da área de Genética que realiza diversas coletas sobre um mesmo ser vivo. Os dois casos são visões particulares dos objetos da BDiG-PELD que se confrontam com a visão compartilhada e descritiva que se pretende oferecer à comunidade usuária de uma biblioteca digital. No entanto, essas observações são pertinentes e devem ser consideradas na evolução da BDiG-PELD em trabalhos futuros.

5.2.2 Qualidade dos Dados

Após a entrega dos questionários, os usuários deveriam fazer a submissão das planilhas conforme instruídos. Qualquer erro encontrado seria informado pela própria interface[2] com uma recomendação de solução. Se eventualmente o erro persistisse, então o usuário deveria procurar ajuda. Os erros encontrados nas 118 submissões estão listados na Tabela 5.2. Consideramos boa a qualidade dos dados, pois se trata da primeira entrada de dados efetiva realizada pelos usuários da BDiG-PELD. O erro que ocorreu mais vezes foi o de preenchimento incorreto das taxonomias, representando 81,9% do total, já que a interface de entrada de dados só aceita as taxonomias cadastradas previamente no banco de dados local, embora o mais comum tenha sido a grafia incorreta dessas taxonomias. Não há ocorrências de campos obrigatórios não preenchidos, pois os dados específicos da coleta podem ou não ser preenchidos pelo usuário, ao contrário dos dados básicos da coleta, como, por exemplo, responsável, data, hora, taxonomia e local, todos obrigatórios. Os outros erros encontrados são pouco expressivos para outras análises.

[2] www.icb.ufmg.br/∼peld/ufmg

Erro	Quant.	Percentual planilhas	Percentual total
Taxonomia não cadastrada	104	21,1%	81,9%
Não existe o local de coleta	16	13,6%	12,6%
Data inválida	2	1,7%	1,6%
Hora inválida	3	4%	3,1%
Coluna inesperada na planilha	1	0,8%	0,8%

Tabela 5.2: Avaliação da qualidade dos dados

5.3 Discussão

A partir dos resultados da experimentação e da experiência nela adquirida, observamos que a construção de um banco de dados local foi essencial para identificar claramente os tipos de coleta realizados e quais metadados EML deveriam ser utilizados. A interface de entrada de dados também permitiu, que sítios PELD com fontes de dados pouco estruturadas, pudessem participar efetivamente do ambiente de interoperação. No entanto, essa interface pode ainda evoluir bastante. Por exemplo, as planilhas textuais podem ser preenchidas em campo através de um *PDA (Personal Digital Assistant)*, tornando direto o processo de entrada de dados, já que não se exige a redigitação na submissão. Outra melhoria é criar uma ferramenta que permita fazer manutenção do catálogo de dados, descrito na Subseção 3.2.2, pois isso é feito atualmente de forma manual.

O esforço despendido no processo de criação do provedor de dados a partir de fontes pouco estruturadas foi, em grande parte, devido à falta de engajamento dos usuários nas atividades definidas. Percebemos a falta de um líder para coordenar o agendamento de reuniões e acompanhar o cumprimento dessas atividades, o que dificultou o processo como um todo. Também consideramos que o levantamento de dados foi longo, aproximadamente 5 meses, já que envolveu todos os tipos de coleta do sítio, objeto da experimentação, e cerca de 17 pesquisadores. Em suma, todo o trabalho para criar um provedor de dados, cujo ápice foi esta avaliação experimental, tomou aproximadamente 60% do tempo envolvido neste trabalho, cuja duração foi de um ano, de março de 2004 a março de 2005.

Outro fator muito importante no processo de criação de provedores de dados de natureza diversa, é o uso de uma linguagem de metadados com grande poder de descrição, como é o caso da EML, cuja amplitude permite descrever desde objetos bibliográficos, com o uso do formato *Dublin Core*, até mesmo a descrição vetorial de objetos espaciais. Da linguagem EML, só utilizamos algo em torno de 30% dos seus campos de metadados de acordo com nossa avaliação, número suficiente para descrever, com riqueza de detalhes, todos os tipos de coleta utilizados nesta avaliação trimestral. Por outro lado, essa baixa utilização evidencia a complexidade do formato frente às necessidades reais dos sítios do Programa PELD, o que pode causar problemas de utilização e adoção do padrão. Finalmente, é importante ressaltar que, embora a versão-piloto da BDiG-PELD envolva somente um provedor de dados, ela é capaz de demonstrar a funcionalidade da arquitetura proposta e a sua capacidade de agregar os demais sítios de pesquisa do Programa sem grandes esforços adicionais de implementação.

Capítulo 6

Conclusões e Trabalhos Futuros

Esta dissertação apresenta um ambiente, baseado em bibliotecas digitais, para integração de dados do Programa PELD (Pesquisas Ecológicas de Longa Duração), composto por 12 sítios representativos da biodiversidade brasileira. Os dados do programa possuem natureza diversa e são provenientes de fontes pouco estruturadas. Esse ambiente utiliza o protocolo OAI-PMH para coleta de dados dos diversos sítios PELD, e posterior inclusão desses dados no repositório de uma biblioteca digital, denominada BDiG-PELD (Biblioteca Digital Georreferenciada do PELD).

Pelo fato das fontes de dados disponíveis serem pouco estruturadas, foi necessário padronizá-las para que se pudesse viabilizar o processo de interoperação através do protocolo OAI-PMH. Esse protocolo estabelece a existência de dois tipos de componentes: os provedores de dados, que publicam seus repositórios de dados, e os provedores de serviços, que coletam esses dados e os disponibilizam por meio de seus serviços. No ambiente proposto nesta dissertação, cada sítio do Programa PELD representa um provedor de dados e a BDiG-PELD representa o provedor de serviços.

A preparação dos provedores de dados envolveu, em si, a implementação de uma interface de entrada de dados descentralizada que combina arquivos textuais, bancos de dados e repositórios XML no padrão OAI. Essa interface permite que os dados de coleta sejam carregados diretamente no banco de dados local de um sítio e que sejam publicados na forma de documentos EML *(Ecological Metadata Language)*, linguagem própria para descrição de dados ecológicos. Uma vez publicados, os dados podem ser coletados e integrados no repositório central da BDiG-PELD.

A interoperação entre provedores de dados e a BDiG-PELD foi viabilizada com a utilização da infraestrutura ODL, desenvolvida na *Virginia Polytechnic Institute and State University*. Considerada uma extensão do protocolo OAI-PMH, essa infraestrutura é composta por componentes abertos para implementação de serviços de interoperação e busca, entre outros. A infraestrutura ODL permitiu conceber plenamente a arquitetura do nosso ambiente.

A BDiG-PELD também foi especificada com a ajuda da abordagem 5S, atualmente, uma das poucas iniciativas que definem um modelo para desenho e prototipação de bibliotecas digitais. A partir dessa especificação, pudemos entender a abrangência da BDiG-PELD bem

como os requisitos de seus usuários, possibilitando a criação de serviços específicos para a sua comunidade fim.

No tocante a serviços, a BDiG-PELD agrega, a partir do seu repositório central, serviços que combinam busca textual e navegação com facilidades de georreferenciamento a partir de uma única interface de acesso, dispensando o uso alternado de diferentes sistemas, fato comum em muitas aplicações da área de biodiversidade. Esses serviços foram implementados com o uso da infraestrutura ODL, de um módulo próprio de navegação e do localizador geográfico Locus. O componente de navegação utiliza um *data mart* para viabilizar navegações em várias categorias e integrar semanticamente os dados dos provedores de dados. A integração entre a infraestrutura ODL, o módulo de navegação e o Locus também envolveu aspectos de interoperação, resolvidos com os protocolos HTTP e WFS.

Para demonstrar a efetividade do ambiente proposto, uma versão-piloto da BDiG-PELD foi criada com base nos dados do sítio do Programa PELD sediado no Parque Estadual do Rio Doce. A partir de um experimento criterioso, com usuários reais, avaliamos, sob os aspectos de usabilidade e qualidade de dados, as características que permitem a um sítio, com dados pouco estruturados, integrar esse ambiente. Ressaltamos que, embora a versão-piloto da BDiG-PELD envolva somente um provedor de dados, ela é capaz de demonstrar a funcionalidade da arquitetura proposta e a sua capacidade de agregar os demais sítios de pesquisa do Programa sem grandes esforços adicionais de implementação.

A partir da análise dos resultados e da experiência obtida nesta dissertação, podemos concluir que o esforço de criação de provedores de dados toma boa parte da concepção de um ambiente de interoperação, mas que a abordagem de interface de entrada de dados descentralizada e a existência de uma linguagem de metadados específica oferecem um suporte essencial para que o processo de interoperação seja bem sucedido. Embora esse esforço tenha sido grande, ganhamos com a utilização da infraestrutura ODL, cujos componentes abertos permitiram agregar, de forma fácil e transparente diversos serviços, como, por exemplo, o de georreferenciamento. Avaliamos que soluções de interoperação através de bibliotecas digitais abertas são efetivas e permitem agregar serviços de qualidade aos seus usuários.

Esta dissertação abre inúmeras perspectivas para trabalhos futuros. Inicialmente, pretendemos agregar à BDiG-PELD um serviço de personalização, assim como melhorar os serviços de georreferenciamento de modo a permitir aos usuários a criação de visões com mapas temáticos, como, por exemplo, mapas de vegetação, clima e solo, sobrepostos ou não. Esses serviços também deverão ser avaliados sob os aspectos de usabilidade. Pretendemos, ainda, a partir da BDiG-PELD, desenvolver um arcabouço genérico para ser aplicado a outras áreas do conhecimento, que permita oferecer diferentes serviços sob a concepção de bibliotecas digitais. Tal arcabouço pode propor uma extensão da infraestrutura ODL, incluindo novas operações básicas que possibilitem a manipulação de objetos espaciais e a agregação de serviços de georreferenciamento.

Um outro tópico a ser considerado é o estabelecimento de políticas de segurança para acesso aos dados mantidos no repositório da BDiG-PELD, já que a EML prevê esse tipo de facilidade em sua especificação. Embora os pesquisadores do Programa PELD sejam

obrigados a tornar públicos os dados de coleta com o prazo máximo de dois anos, alguns desses dados podem ser considerados muito sensíveis, o que requer um controle no acesso a eles. Outro aspecto a ser abordado diz respeito à melhoria de desempenho das interações entre os componentes ODL. Já na fase de implementação da interface de acesso, pudemos perceber um baixo desempenho nessas interações, causado, essencialmente, pelo protocolo próprio de comunicação entre componentes, o protocolo XOAI. Para resolver esse problema, a interface de acesso deve ser capaz de acessar os dados mantidos diretamente pelos componentes, evitando interações entre eles, de modo a tornar o ambiente escalável. Além disso, deve ser considerada ainda a criação de uma ferramenta para atualizar o catálogo de dados da interface de entrada, o que atualmente é feito de forma manual. Tal melhoria tornaria mais versátil a incorporação de novos tipos de coleta ou mesmo a alteração dos tipos já existentes.

Finalmente, coerente com os desafios das principais bibliotecas digitais no mundo, a BDiG-PELD deve procurar cuidar das questões de sustentabilidade e preservação. Destinada a ser a solução de compartilhamento de dados para o Programa PELD e por ter se demonstrado efetiva e econômica, a BDiG-PELD merece agora evoluir do estágio de protótipo para se tornar um produto. Para isso é importante contar com a participação de todos os sítios do programa e com o financiamento de agências de fomento. Ao ser tratada como produto, a BDiG-PELD poderá abordar aspectos essenciais para preservação de seus dados, tais como, manutenção, suporte e evolução.

Referências Bibliográficas

[1] BioCASE - Biological Collection Access Service for Europe. http://www.biocase.org. Acessado em 18/02/2005.

[2] Computer Science Teaching Center. http://www.cstc.org. Acessado em 16/07/2005.

[3] Darwin Core2 Review. http://darwincore.calacademy.org. Acessado em 18/07/2005.

[4] DiGIR - Distributed Generic Information Retrieval. http://digir.sourceforge.net/. Acessado em 18/07/2005.

[5] Fabulous Force Database Designer 4. http://fabforce.net. Acessado em 18/02/2004.

[6] Open Archives Initiative - Frequently Asked Questions (FAQ). http://www.opearchives.org/documents/FAQ.html. Acessado em 09/02/2004.

[7] Partner Interface Process Technical Architecture, RosettaNet. http://www.rosettanet.org. Acessado em 16/07/2005.

[8] RDF Primer W3C Recommendation 10 February 2004. World Wide Web Consortium, 2000. http://www.w3.org/TR/rdf-primer/. Acessado em 17/07/2005.

[9] Taxonomic Databases Working Group. http://www.tdwg.org. Acessado em 18/07/2005.

[10] Web Services Description Language (WSDL) 1.1. http://www.w3.org/TR/wsdl. Acessado em 18/02/2004.

[11] XMLSpy - Altova . http://www.altova.com. Acessado em 18/06/2005.

[12] XML Schema Definition, 2000. World Wide Web Consortium. http://www.w3.org/TR/xmlschema-0/. Acessado em 20/03/2005.

[13] Ecological Metadata Language (EML) Specification, 2004. Knowledge Network for Biocomplexity. http://knb.ecoinformatics.org/software/eml/eml-2.0.0/. Acessado em 10/03/2004.

[14] S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web - From Relations to Semistructured Data and XML*. Morgan Kaufman Publishers, San Francisco, California, 2000.

[15] D. L. Atkins, T. Ball, G. Bruns, and K. C. Mawl. A domain-specific language for form-based services. *IEEE Transactions on Software Engineering*, 25(3):334–346, 1999.

[16] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, New York, NY, 1999.

[17] E. G. Barros, L. A. de Souza, R. G. Cota, K. A. V. Borges, A. H. F. Laender, M. A. Gonçalves, and F. A. R. Barbosa. Uma Biblioteca Digital Georreferenciada para Dados Ecológicos. In *XX Simpósio Brasileiro de Bancos de Dados*, pages 175–189, Universidade Federal de Uberlândia, MG, Brazil, 2005.

[18] E. G. Barros and A.H.F. Laender. Uma Biblioteca Digital para o PELD Brasil. In *Livro de Resumos do Simpósio Internacional sobre Projetos Ecológicos de Longa Duração*, pages 57–59, Julho 2004.

[19] T. Bray, J. Paoli, C. M. Sperberg-McQueen, and E. Maler. Extensible Markup Language (XML) 1.0. W3C Recommendation, World Wide Web Consortium, 2000. http://www.w3.org/TR/REC-xml.

[20] R. da S. Torres, C. B. Medeiros, M. A. Gonçalves, and E. A. Fox. A Digital Library Framework for Biodiversity Information. In *International Journal on Digital Libraries*, volume 6, page 3, February 2006.

[21] L. A. de Souza, T. M. Delboni, K. A. V. Borges, C. A. Davis Jr., and A. H. F. Laender. Locus: Um Localizador Espacial Urbano. In *VI Simpósio Brasileiro de GeoInformática*, Campos do Jordão, SP, 2004.

[22] L. A. de Souza, C. A. Davis Jr., K. A. V. Borges, T. M. Delboni, and A. H. F. Laender. The Role of Gazetteers in Geographic Knowledge Discovery on the Web. In *LA-WEB'05 - Third Latin American Web Congress*, pages 157–165, Buenos Aires, Argentina, 2005.

[23] D. P. V. del Pozo, L. V. e Silva, A. H. F. Laender, and M. A. Gonçalves. Modelagem de Bibliotecas Digitais usando a Abordagem 5S: Um Estudo de Caso. In *SBBD*, pages 350–362, Brasília, DF, Brazil, 2004.

[24] M. Döring, R. Giovanni, D. Hobern, D. Vieglais, A. Güntsch, S. Blum, J. Wieczorek, and J. Torre. The integration of DiGIR and BioCASe. 2004. http://www.tdwg.org/2004meet/paperabstracts/TDWG_2004_Papers_Doering_1.htm.

[25] E. A. Fox, J. L. Eaton, G. McMillan, N. A. Kipp, P. Mather, T. McGonigle, and W. Schweiker. Networked digital library of theses and dissertations: An international effort unlocking university resources. *D-Lib Magazine*, 3(9), 1997. http://www.dlib.org/dlib/september97/theses/09fox.html.

[26] M. A. Gonçalves and E. A. Fox. 5SL: A Language for declarative specification and generation of digital libraries. In *JCDL '02: Proceedings of the 2th ACM/IEEE-CS joint conference on Digital libraries*, pages 263–272, 2002.

[27] M. A. Gonçalves, E. A. Fox, L. T. Watson, and N. A. Kipp. Streams, structures, spaces, scenarios, societies (5S): A formal model for digital libraries. *ACM Trans. Inf. Syst.*, 22(2):270–312, 2004.

[28] M. F. Goodchild. The Alexandria Digital Library Project - Review, Assessment and Prospects. *D-Lib Magazine*, 10(5), 2004. http://www.dlib.org/dlib/may04/goodchild/05goodchild.html. Acessado em 22/02/2005.

[29] S. Hanard. The self-archiving initiative. *Nature - On Line*. http://www.nature.com/nature/debates/e-access/Articles/harnad.html. Acessado em 20/07/2005.

[30] L. L. Hill. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, ECDL*, pages 280–290, London, UK, 2000. Springer-Verlag.

[31] L. L. Hill. Georeferencing in Digital Libraries. *D-Lib Magazine*, 10(5), 2004. http://www.dlib.org/dlib/may04/hill/05hill.html. Acessado em 22/02/2005.

[32] L. L. Hill, J. Frew, and G. Janée. Issues in Georeferenced Digital Libraries. *D-Lib Magazine*, 10(5), 2004. http://www.dlib.org/dlib/may04/janee/05janee.html. Acessado em 22/02/2005.

[33] C. A. Joly. Strengthening of the BIOTA/FAPESP information system and study of the development of a GIS for the program. Technical report, BIOTA/FAPESP, 1998. http://sinbiota.cria.org.br/info/projeto_sinbiota.pdf.

[34] M. B. Jones, C. Berkley, J. Bojilova, and M. Schildhauer. Managing Scientific Metadata. volume 5, pages 59–68, Piscataway, NJ, USA, 2001. IEEE Educational Activities Department.

[35] A. H. F. Laender, M. A. Gonçalves, and P. A. Roberto. BDBComp: building a digital library for the Brazilian computer science community. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 23–24, New York, NY, USA, 2004. ACM Press.

[36] M. Gudgin and M. Hadley and N. Mendelsohn and J.J. Moreau and H. F. Nielsen. SOAP Version 1.2 Part 1: Messaging Framework. World Wide Web Consortium, W3C Recommendation, 2005. http://www.w3.org/TR/soap12-part1. Acessado em 16/07/2005.

[37] C. H. Marcondes and L. H. Sayão. Integração e interoperabilidade no acesso a recursos informacionais em Ciência e Tecnologia: a proposta da biblioteca digital brasileira. *Ciência da Informação*, 3(30):23–33, Setembro/Dezembro 2001.

[38] R. S. Mello and C. A. Heuser. Aplicação de ontologias a dados semi-estruturados. In *Centro Latinoamericano de Estudios en Informatica*, Cidade do México, México, 2000. Actas (published in CD).

[39] L. Obrst. Ontologies for semantically interoperable systems. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 366–369, New York, NY, USA, 2003. ACM Press.

[40] S. Payette and C. Lagoze. Flexible and Extensible Digital Object and Repository Architecture (FEDORA). In C. Nikolaou and C. Stephanidis, editors, *ECDL*, volume 1513 of *Lecture Notes in Computer Science*, pages 41–59. Springer, 1998.

[41] E. Pazinato, C. Baptista, and R. Miranda. GeoLocalizador: Um Sistemas de Referência Espaço-Temporal Indireta utilizando um SGBD Objeto-Relacional. In *IV Simpósio Brasileiro de GeoInformática*, pages 49–56, Belo Horizonte, MG, 2002.

[42] J. H. Porter, D. Henshaw, and S. Stafford. Research Metadata in Long-Term Ecological Research (LTER). In W. S. Michener, J. H. Porter, and S. Stafford, editors, *IEEE Metadata Meeting - Data and Information Management in the Ecological Sciences*, Albuquerque, New Mexico, 1997.

[43] U. Ravindranathan, R. Shen, M. A. Gonçalves, W. Fan, E. Fox, and J. W. Flanagan. Prototyping Digital Libraries Handling Heterogeneous Data Sources. The ETANA DL Case Study. In *Proceedings of the 8th European Conference on Research and Advanced Technology for Digital Libraries, ECDL*, pages 186–197, Bath, Uk, 2004. Springer.

[44] U. Seeliger, C. Cordazzo, and F. Barbosa. *Os Sites e o Programa Brasileiro de Pesquisas Ecológicas de Longa Duração.* U. Seeliger and C. Cordazzo and F. Barbosa, Belo Horizonte-MG, Brazil, 2002.

[45] R. Shen, M. A. Gonçalves, W. Fan, and E. Fox. Requirements gathering and modeling of domain-specific digital libraries with the 5s framework: An archaeological case study with etana. In *9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, 2005. (a ser publicado).

[46] H. V. Sompel and C. Lagoze. The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, 6(2). http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html. Acessado em 12/07/2005.

[47] H. Suleman. *Open Digital Libraries.* PhD thesis, Virginia Polytechnic Institute and State University, 2002. Blacksburg, Virginia.

[48] H. Suleman and E. A. Fox. A Framework for Building Open Digital Libraries. *D-Lib Magazine*, 7(12), 2001. http://www.dlib.org/dlib/december01/suleman/12suleman.html. Acessado em 20/03/2004.

[49] I. H. Witten and E. Frank. *Data mining: practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.