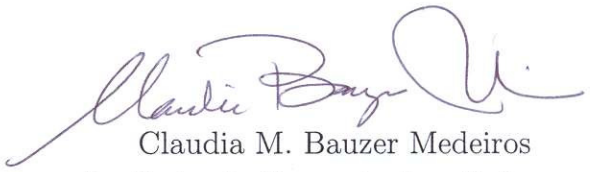
Um Serviço de Gerenciamento de Coletas para Sistemas de Biodiversidade

Este exemplar corresponde à redação final da Dissertação devidamente corrigida e defendida por Joana E. Gonzales Malaverri e aprovada pela Banca Examinadora.

Campinas, 5 de maio de 2009.



Claudia M. Bauzer Medeiros
Instituto de Computação - Unicamp
(Orientadora)

Dissertação apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP
Bibliotecária: Maria Fabiana Bezerra Müller – CRB8 / 6162

Gonzales Malaverri, Joana Esther
G589s Um serviço de gerenciamento de coletas para sistemas de biodiversidade/Joana Esther Gonzales Malaverri-- Campinas, [S.P. : s.n.], 2009.

Orientador : Claudia Maria Bauzer Medeiros
Dissertação (Mestrado) - Universidade Estadual de Campinas, Instituto de Computação.

1.Sistemas de informação gerencial. 2. Diversidade biologica. 3. Banco de dados. 4. Serviços na Web. 5.Metadados. I. Medeiros, Claudia Maria Bauzer. II. Universidade Estadual de Campinas. Instituto de Computação.
III. Título.

(mfb/imecc)

Título em inglês: A service for data collection management for biodiversity information systems

Palavras-chave em inglês (Keywords): Management information systems, Biodiversity, Data bases, Web services, Data about data.

Área de concentração: Banco de Dados

Titulação: Mestre em Ciência da Computação

Banca examinadora: Profa. Dra. Claudia Maria Bauzer Medeiros (IC-Unicamp)
Profa. Dra Maria Camila Nardini Barioni (Univ. Federal do ABC)
Profa. Dra. Eliane Martins (IC-Unicamp)

Data da defesa: 24/04/2009

Programa de Pós-Graduação: Mestrado em Ciência da Computação

TERMO DE APROVAÇÃO

Dissertação Defendida e Aprovada em 24 de abril de 2009, pela Banca examinadora composta pelos Professores Doutores:

Maria Camila Nardini Barioni

Profª. Drª. Maria Camila Nardini Barioni
Centro de Matemática e de Computação / Universidade Federal do ABC

Eliane Martins

Profª. Drª. Eliane Martins
IC / UNICAMP

Profª. Drª. Claudia Maria Bauzer Medeiros
IC / UNICAMP

Instituto de Computação
Universidade Estadual de Campinas

Um Serviço de Gerenciamento de Coletas para Sistemas de Biodiversidade

Joana E. Gonzales Malaverri[1]

Abril de 2009

Banca Examinadora:

- Claudia M. Bauzer Medeiros
  Instituto de Computação - Unicamp (Orientadora)
- Maria Camila Nardini Barioni
  Universidade Federal do ABC
- Eliane Martins
  Instituto de Computação - Unicamp
- Rodolfo Jardim de Azevedo (Suplente)
  Instituto de Computação - Unicamp
- Antonia Cecília Zacagnini Amaral (Suplente)
  Instituto de Biologia - Unicamp

[1]Este trabalho teve o apoio financeiro da CAPES, 08/2007 à 03/2009, CNPq e Microsoft Research

Resumo

A pesquisa em biodiversidade requer correlacionar dados sobre seres vivos e seus habitats. Essas correlações podem ser de vários tipos, integrando desde relacionamentos espaciais a especificações ambientais (descrições do habitat ou ecossistema). Sistemas de Biodiversidade são sistemas de software complexos que permitem aos pesquisadores realizar tais análises. A complexidade desses sistemas varia com os dados manipulados, os usuários visados e o ambiente em que são executados. Um problema que deve ser enfrentado, especialmente no ambiente *Web*, é a heterogeneidade dos dados, agravado pela diversidade dos vocabulários dos usuários.

Esta dissertação contribui para resolver tal problema, apresentando um modelo de banco de dados que permite organizar informações de biodiversidade a partir de padrões mundiais de dados. O modelo proposto combina informações coletadas em campo com dados de catálogos de museus. Para a especificação do modelo, contamos com a participação de ecólogos e biólogos. Para permitir o acesso e recuperação de informação de forma transparente, o banco de dados foi encapsulado em um serviço *Web*, que é invocado por aplicações cliente. A validação do banco de dados e do serviço utiliza dados reais fornecidos pelos parceiros do projeto BIO-CORE, um projeto de pesquisa que envolve pesquisadores em Computação e Biologia da UNICAMP e USP.

Abstract

Biodiversity research requires correlations of data on living beings and their habitats. Such correlations can be of different types, considering factors such as spatial relationships or environmental descriptions (e.g., description of habitat and ecossystems). Biodiversity information systems are complex pieces of software that allow researchers to perform these kinds of analysis. The complexity of these systems varies with the data used, the target users, and the environment where the systems are executed. One of the problems to be faced, especially on the Web, is the heterogeneity of the data aggravated by the diversity of user vocabularies.

This research contributes to solve this problem by presenting a database model that organizes the biodiversity information using consensual data standards. The proposed model combines information collected in the field with that from museum data catalogues. The model was specified with the assistance of biologists and ecologists. The database was encapsulated in a *Web* service to ensure transparency in using, accessing and recovering the information. The service is invoked by client applications. The database and service were validated using real data, provided by the BIO-CORE project partners. BIO-CORE is a research project that involves computer and biology researchers from UNICAMP and USP.

Agradecimentos

A Deus pelo amor infinito, por ter me acompanhado e iluminado sempre a cada passo e por tudo que eu tenho recebido dele.

A professora Claudia Bauzer Medeiros, pelo privilégio de tê-la como orientadora. Obrigada pela confiança, atenção e paciência ao longo de todo o meu mestrado. Agradeço pelos ensinamentos e as críticas que permitiram explorar melhor meu potêncial.

A toda minha família, por sempre ter me oferecido todo o seu apoio e carinho. Agradeço em especial a minha mãe e irmãs por todo o amor, força, e por sempre ter acreditado em mim.

Aos meus tios Lucho, Gladys e família que estiveram presentes desde os inícios da minha carreira. Obrigada por todo o carinho e compreensão.

A todos os meus amigos, deste maravilhoso pais, Brasil, que me acolheram com carinho e compartilharam comigo momentos inesquecíveis. Por toda sua bondade, amizade e paciência, muito obrigada.

Aos meus amigos e colegas do Laboratório de Sistemas de Informação (LIS), pelas contribuições, sugestões e inumeras correções, que ajudaram no desenvolvimento desta dissertação.

Aos meus amigos que estão distantes, que mesmo à distância souberam me motivar e ofereceram sempre o melhor dos seus conselhos.

Aos biólogos parceiros do projeto BioCORE, que contribuíram com informações e sugestões no decorrer deste trabalho.

As agências de fomento CAPES, CNPq e Microsoft Research, pelo suporte financeiro no desenvolvimento deste trabalho.

Lista de Abreviações

ABCD: *Access Biological Colections Data.*

BIO-CORE: *Biodiversity and Computing Research.*

BioCase: *Biological Collection Access Services.*

DiGIR: *Distributed Generic Information Retrieval.*

EML: *Ecological Metadata Language.*

FGDC: *Federal Geographic Data Committee.*

GBIF: *Global Biodiversity Information Facility.*

ITIS: *Integrated Taxonomic Information System.*

NBII: *National Biological Information Infrastructure.*

OBIS-SEAMAP: *Ocean Biogeographic Information System Spatial Ecological Analysis of Megavertebrate Populations.*

OGC: *Open Geospatial Consortium.*

TAPIR: *TDWG Access Protocol for Information Retrieval.*

TDWG: *Taxonomic Database Working Group.*

WeBIOS: *Web Service Multimodal Tools for Biodiversity Research, Assessment and Monitoring.*

WOODS: *WOrkflOw-based Spatial Decision Support System.*

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Estudos de biodiversidade manipulam uma grande variedade de dados, incluindo registros de ocorrências de espécies, dados geográficos, ecológicos, socio-econômicos e outros. Dentre os maiores desafios enfrentados pelos pesquisadores de biodiversidade estão: 1) a identificação e avaliação de descontinuidades críticas no conhecimento da biodiversidade, tanto taxonômicas quanto geográficas; 2) o planejamento de meios efetivos de levantamento e descrição de organismos em grupos criticamente importantes; 3) a mineração de dados em coleções existentes e 4) a concepção de novas abordagens para uso das informações existentes [1]. Esses mesmos desafios, enfrentados por qualquer grupo de pesquisadores de biodiversidade que trabalha com coletas na natureza, se tornam ainda mais complexos quando dados providos por grupos de pesquisa distintos precisam ser integrados.

Esse cenário tem motivado diversos esforços na coleta e organização de dados. O resultado é uma grande quantidade de informações, que requer novas soluções de gerenciamento e análise das características das espécies e suas interações. Os *Sistemas de Informação de Biodiversidade* [50, 26] surgiram com este objetivo. O escopo de tais sistemas vai desde a recuperação de informações textuais, como descrições literais, à combinação de informações taxonômicas com a distribuição espacial de uma ou mais espécies. Em geral, há pouca flexibilidade para as consultas nesses sistemas. Por exemplo, não fornecem a possibilidade de consultas exploratórias para a mineração de informação dos relacionamentos entre espécies.

Sistemas de informação de biodiversidade, em geral, permitem consultas sobre dados armazenados em sistemas de banco de dados tradicionais ou estruturados em *XML*. Um registro de ocorrência de espécie armazena dados sobre alguma observação ou coleta de seres vivos, incluindo informações sobre a classificação taxonômica, responsáveis pela coleta, local e demais características da coleta [30]. Consultas básicas de interesse dos biólogos estão relacionadas à classificação taxonômica das espécies, à localização geográfica

e o habitat, por quem e como foram realizadas. Na maioria dos casos esses sistemas são projetados para cobrir cenários específicos: quer para catalogação de espécies nos museus, quer para projetos que abrangem coletas de material em campo para seu posterior estudo em laboratórios. Para cada um destes cenários são implementados bancos de dados diferentes que funcionam com módulos independentes.

Em mais detalhe, há dois tipos básicos de informação manipulada em sistemas de biodiversidade: a) registros constantes de catálogos e acervos de museus e b) registros que documentam coletas e observações feitas em campo. Ambos tipos de registros contêm informações sobre espécies: sua identificação, quando foram coletadas, onde, como e por quem. Enquanto o primeiro tipo de registros trata de coleções catalogadas, o segundo é mais comum em coleções de (um ou vários) grupos de pesquisa, em que os seres vivos observados não estão disponíveis em um acervo comum. Assim, por exemplo, registros de catálogos freqüentemente contêm informações sobre a forma de conservação de um exemplar. Já registros de coletas priorizam informações sobre o processo de coleta e, muitas vezes, aspectos ecológicos. Com isto, muitas informações que poderiam ser compartilhadas ficam repetidas em sistemas distintos, gerando possíveis problemas de duplicação e integridade.

O objetivo da dissertação é fornecer aos pesquisadores em biodiversidade um Repositório de Coletas onde possam armazenar seus dados integrando as duas formas de trabalho. Esse repositório deve poder abrigar todos os principais tipos de dados manipulados por projetos de biodiversidade e acervos. Para flexibilizar a recuperação de informação, sua interface de acesso é realizada via um serviço *Web*, combinando registros de ocorrência de espécies e registros de catálogos.

Esta dissertação de mestrado foi desenvolvida como parte do projeto BIO-CORE [2]. Este é um projeto na área de biodiversidade baseado em serviços *Web*, conduzido conjuntamente por pesquisadores de Computação e de Biologia da UNICAMP e da USP. O objetivo do BIO-CORE é prover aos pesquisadores de biodiversidade um sistema que suporte consultas exploratórias multimodais sobre fontes de dados heterogêneas de biodiversidade (dados textuais de espécies, imagens, dados geográficos, ontologias e anotações), acessados via serviços *Web*.

As principais contribuições deste trabalho são:

- Levantamento e organização da informação fornecida pelos usuários-alvo, ecólogos do Laboratório Inseto-Plantas e biólogos do Museu de Zoologia, parceiros do projeto BIO-CORE;

- Especificação e implementação do Repositório de Coletas para armazenamento de informação de biodiversidade, considerando a manipulação integrada dos varios tipos de registros manipulados por sistemas de biodiversidade;

- Especificação de um Serviço de Coletas usando tecnologia *Web* para realizar consultas ao repositório;

- Implementação parcial do serviço proposto, usando dados reais fornecidos pelos usuários-alvo.

Uma parte desta pesquisa está no artigo "*A Tool Based on Web Services to Query Biodiversity Information*", aceito como *short paper* no "*5th International Conference on Web Information Systems and Technologies*".

O restante deste documento está organizado como segue. O Capítulo 2 apresenta os principais conceitos e trabalhos relacionados à pesquisa. O Capítulo 3 descreve o Repositório de Coletas, os cenários que o modelo de banco de dados suporta, e uma descrição sucinta das entidades que formam parte do modelo. O Capítulo 4 apresenta a especificação do Serviço de Coletas proposto. O Capítulo 5 discute aspectos de implementação. Finalmente, o Capítulo 6 apresenta as conclusões e trabalhos futuros. O apêndice contém detalhes sobre o banco de dados implementado.

Capítulo 2

Trabalhos Correlatos

Sistemas de Informação de Biodiversidade (SIB) são sistemas que gerenciam grandes conjuntos de dados geográficos, assim como grandes bancos de dados relacionados a coleções de espécies [17]. A maioria dessas aplicações aproveita características e facilidades fornecidas pelo desenvolvimento de ferramentas, serviços, técnicas, frameworks, entre outros, disponíveis na *Web*. Alguns dos desafios relacionados à área de pesquisa em biodiversidade incluem: a heterogeneidade e grande volume de dados com os quais se deve lidar; limitações espaço-temporais na distribuição das coleções de espécies; e incorporação de georeferenciamento correto às coletas [50].

Um outro desafio está relacionado ao compartilhamento e transmissão dos dados de biodiversidade entre comunidades de pesquisa [26]. Assim, uma quantidade considerável de projetos está adotando padrões de metadados, como é o caso do Darwin Core proposto em [53]. A transmissão de dados baseados nestes padrões originou o desenvolvimento de protocolos como DiGIR (*Distributed Generic Information Retrieval*) [51] e BioCASE (*Biological Collection Access Services*) [21], para permitir o acesso às fontes de dados de biodiversidade distribuídas na *Web*. Tais esforços são combinados ao desenvolvimento de ferramentas que auxiliem a combinação e visualização dos dados em mapas. Um exemplo desse tipo de aplicações são os servidores de mapas.

O principal objetivo deste capítulo é dar uma visão geral dos sistemas de informação de biodiversidade, abrangendo as tecnologias que estão associadas à disseminação e compartilhamento de dados na Internet. Os trabalhos e fundamentos estudados neste capítulo serviram como base para especificar o repositório de dados e o serviço de coletas alvos desta dissertação.

Este capítulo está organizado como segue. A Seção 2.1 descreve de forma sucinta algumas aplicações de biodiversidade e os projetos que estão direcionando esforços para o estabelecimento de mecanismos que ajudem à formação de comunidades de pesquisa na área de biodiversidade. A Seção 2.2 apresenta os padrões de metadados e os protocolos

propostos para a organização e transferência de informação. A Seção 2.3 descreve uma visão geral do projeto BIO-CORE, ao qual a dissertação está associada. Finalmente a Seção 2.4 apresenta o resumo deste capítulo.

2.1 Aplicações para o gerenciamento de informação em Biodiversidade

Há um grande número de projetos que visam desenvolver meios para publicar e gerenciar dados disponíveis na *Web* relacionados à pesquisa em biodiversidade. Como mencionado no Capítulo 1, os sistemas existentes se dedicam ao gerenciamento de registros de acervos de museus ou de coletas de dados em campo. Muitos desses projetos foram propostos com a finalidade de gerenciar as coleções de Museus de História Natural e Herbários. Porém, a heterogeneidade dos dados com os quais lidam é um dos fatores mais relevantes que devem ser considerados nos projetos de sistemas de informação de biodiversidade. Um exemplo do primeiro tipo de projeto é o speciesLink [11]. Este sistema *Web* tem por objetivo integrar a informação primária sobre biodiversidade, ou seja, informação catalogada sobre coleções biológicas e observações documentadas de organismos vivos na natureza, disponíveis em museus, herbários e coleções microbiológicas, publicando-a de forma livre e aberta na Internet. Outro exemplo é Specify [7], um projeto que visa fornecer uma plataforma computacional que utiliza serviços *Web* como suporte para o gerenciamento das coleções de dados, incluindo descrição geográfica da coleta, dados dos coletores e algumas operações que devem ser realizadas sobre o acervo como empréstimos, intercâmbios, adesões e doações.

Um outro tipo de aplicações de biodiversidade são os programas desenvolvidos para gerenciar dados de coletas de campo. Um exemplo é o projeto Biota [12] que foi um dos primeiros em se interessar pelos registros de ocorrências realizadas pelos biólogos no campo e propor um banco de dados para gerenciar inventários de biodiversidade para o projeto ALAS (*Artropodos de La Selva*). Um outro exemplo deste tipo de sistema é o SinBiota [49] que gerencia registros de observações de campo realizadas por grupos de pesquisa financiados pela FAPESP, no estado de São Paulo.

Em paralelo, projetos como o GBIF (*Global Biodiversity Information Facility*) [23], ITIS (*Integrated Taxonomic Information System*) [28], *Species 2000* [44], TDWG (*Taxonomic Database Working Group*)[53], NBII (*National Biological Information Infrastructure*) [40], entre outros, estão direcionando esforços para estabelecer aplicações e padrões para a integração e a interoperabilidade de dados das coleções biológicas para torná-las disponíveis na *Web*. GBIF, por exemplo, é uma organização mundial cujo objetivo é disponibilizar informação sobre biodiversidade por meio de uma rede global distribuída

de bancos de dados interoperáveis respeitando a propriedade intelectual dos fornecedores de dados.

Uma característica comum das aplicações de biodiversidade é a sua concentração no nível taxonômico de espécies. Isso ocorre porque as espécies são a base de um sistema de agrupamento hierárquico conhecido como árvore taxonômica, usado pelos cientistas para classificar formas de vida [39]. Assim, um outro conjunto considerável de aplicações de biodiversidade lida com o gerenciamento de informações taxonômicas e a distribuição geográfica das espécies. Esse é o caso de *The Tree of Life* [33], *Catalogue of Life* [8], OBIS-SEAMAP [27], e TaiBIF [48]. O projeto *The Tree of Life* é um esforço internacional para prover informação sobre a diversidade de organismos na terra, suas características e evolução histórica. Já o projeto *Catalogue of Life* visa fornecer um catálogo mundial de taxonomia das espécies vivas unificando essa informação em um sistema de banco de dados que seja mundialmente acessível. Já o projeto OBIS-SEAMAP é um banco de dados com referência espacial para coleções de espécies marinhas, que podem ser visualizadas usando aplicações que envolvem mapas. O objetivo do projeto TaiBIF é integrar a informação de biodiversidade do Taiwan, abrangendo lista de espécies, imagens, características geográficas, informação ambiental, informação encontrada na literatura, informação fornecida por expertos do domínio e uma lista de instituições e organizações relevantes. Todos esses projetos utilizam tecnologia *Web* para a publicação da informação.

Uma outra abordagem encontrada na literatura são ferramentas que permitem a identificação de espécies baseadas no conceito de guias de campo. Um guia de campo é um livro desenhado para ajudar na identificação de espécies [39]. Por exemplo, *Electronic Field Guide* (EFG) [39], é uma ferramenta que permite aos cientistas redigir e gerar suas próprias guias de campos e sofisticadas chaves de identificação taxonômica, que podem ser publicadas e compartilhadas na Internet.

A Figura 2.1 resume as características mais relevantes das aplicações de biodiversidade descritas nesta seção. A coluna *Objetivo* descreve em linhas gerais o objetivo principal de cada aplicação. A coluna *Protocolos* apresenta os protocolos de comunicação que são usados por essas aplicações. A seguir, a coluna *Ferramentas de desenvolvimento* mostra as tecnologias utilizadas para o desenvolvimento dessas aplicações e a coluna *Padrões de metadados* mostra os padrões de metadados usados por algumas aplicações. Já as colunas *Espacial* e *Temporal* permitem conhecer quais são as aplicações que incorporam estas características. Na coluna *Framework para gerenciamento de conteúdo* mostra as aplicações que utilizam ferramentas para gerenciar seu conteúdo na *Web*. Finalmente, a coluna *Estratégia de Banco de Dados* descreve os sistemas de banco de dados que são usados por estas aplicações.

Como podemos observar a maioria usa o protocolo DiGIR [1] como mecanismo para

[1] http://digir.sourceforge.net

compartilhar e recuperar registros de dados das organizações participantes destes projetos. DiGIR é compatível com o Darwin Core [54], que é um padrão que permite a representação de dados de coleções de observação de espécies. Geralmente a linguagem usada como meio de desenvolvimento é Java e algumas das aplicações incorporam conceitos de sistemas para gerenciamento de conteúdo. Propriedades espaciais e temporais estão se tornando cada vez mais importantes. Os trabalhos [11, 39] utilizam serviços *Web* usando SOAP. Entretanto, se consideramos uma definição mais abrangente, no qual um serviço *Web* pode ser considerado como uma aplicação acessível por meio da *Web* [4], então todos os trabalhos, utilizam esse tipo de serviços.

Tabela 2.1: Tabela das principais características das aplicações de biodiversidade

Aplicações de Biodiversidade	**Objetivo**	**Tecnologías envolvidas**		**Padrão de metadados**	**Extensão**		**Framework para gerenciamento de conteúdo**	**Estratégia de Banco de dados**
		Protocolos	Ferramentas de desenvolvimento		Espacial	Temporal		
speciesLink	Integrar informação primária de acervos de coleções biológicas	DiGIR, SOAP	Perl, PHP e Java	Darwin Core	Mapas de distribuição de espécies	Data da observação e coleta de espécies	Nenhum	Distribuído, PostgreSQL/ PostGIS
The Tree of Life	Coleção de informação sobre evolução e diversidade da vida na terra	HTTP	Java, Jakarta Tapestry, Hibernate	Nenhum	Nenhum	Nenhum	SIM (Desenvolvimento próprio)	Orientado a Objetos, MySQL
Catalogue of Life	Catalogar a taxonomia das diferentes espécies conhecidas	SOAP	PHP	-	Nenhum	Nenhum	Nenhum	Distribuído, MySQL
OBIS-SEAMAP	Repositório global de dados georeferenciados para espécies marinhas	DiGIR, OpenDAP	Java	Darwin Core, FGDC	Mapas de distribuição de espécies	Data de observação da espécie	Plone	Relacional, PostgreSQL/ PostGIS
TAIBIF	Integrar informação da biodiversidade da Tailándia em um sitio *Web*	DiGIR	Java, AJAX	Darwin Core	Mapas de distribuição de espécies	Nenhum	Nenhum	Disribuído
Biota	Gerenciar inventarios de biodiversidade	TCP/IP	-	Nenhum	Apenas registro da localidade da coleta	Data da observação e coleta de espécies	Nenhum	Relacional, 4DServer
SinBiota	Gerenciar registros de observação de campo do estado de São Paulo	HTTP	-	Próprio (Fagundes, 99)	Mapas de distribuição de espécies	Data da observação e coleta de espécies	Nenhum	Relacional, Oracle e ArcInfo
Specify	Gerenciar coleções de dados de biodiversidade	DiGIR	Delphi	Darwin Core	Apenas registro da localidade da coleta	Data da observação e coleta de espécies	Plone	Relacional, Microsoft SQL Server e Microsoft Access
EFG	Ferramenta de identificação baseado no conceito de guia de campo	SOAP	Java, HTML	Nenhum	Nenhum	Nenhum	Nenhum	Orientado a Objetos, eXcelon, MySQL

2.2 Mecanismos para disseminação e compartilhamento de informação na Internet

2.2.1 Padrões de Metadados para biodiversidade

A pesquisa em biodiversidade exige a manipulação de uma grande variedade de dados, como os registros de ocorrências de espécies, dados geográficos, ecológicos, socioeconômicos, entre outros. Os cientistas envolvidos neste domínio cada vez mais estão utilizando uma ampla variedade de padrões para coletar dados sobre tópicos diversos por exemplo, funções da comunidade de bactérias marinhas. Este esforço mundial está resultando no armazenamento de dados heterogêneos em sistemas de bancos de dados independentes e dispersos por toda a comunidade de pesquisa [29]. É fundamental o compartilhamento de informação para a realização de estudos mais abrangentes, possibilitando a análise de diversos tipos de espécies e incorporando elementos geográficos [30].

Há várias abordagens clássicas para resolver o problema de compartilhamento da informação. Algumas oferecem uma visão global unificada, mantendo os dados como estão [5, 10, 45]. Criar um esquema único, convertendo esquemas e dados fisicamente para esta nova organização é outra alternativa [6]. Há também o desenvolvimento de camadas de software que realizem traduções entre pedidos externos e os vários sistemas internos (mediadores) [9, 31].

Um meio comum para facilitar o acesso e disseminação da informação na Internet são os metadados. Segundo [24], os metadados são dados estruturados sobre um objeto que suportam funções associadas a esse objeto específico. Eles facilitam o compartilhamento, a recuperação e a transferência de dados [43].

Alguns dos padrões de metadados responsáveis pela descrição dos dados de ocorrência de biodiversidade são o Darwin Core (e suas diferentes versões) [54] e o ABCD (*Access Biological Colections Data*) [52]. O objetivo do Darwin Core é facilitar o intercâmbio de informação sobre a ocorrência geográfica de espécies e a existência de espécimes em coleções. Campos básicos do Darwin Core incluem a especificação do nome do organismo, onde, quando e quem fez a coleta. A Tabela 2.2.1 apresenta alguns elementos contidos na especificação do padrão Darwin Core.

Já o padrão ABCD é um esquema comum de dados que permite estruturar e especificar unidades de coleções biológicas, isto é, informação de espécies vivas e preservadas e das observações feitas em campo. O ABCD está destinado a apoiar o intercâmbio e a integração de dados de coleções biológicas. O Darwin Core é um padrão de metadados não-hierárquico, ideal para os registros de ocorrência de espécies, enquanto que o ABCD traz elementos adicionais aos fornecidos pelo Darwin Core. Em [15] apresenta-se um novo padrão de metadados que integra atributos pertencentes a diversos padrões mundiais de

metadados já consolidados e utilizados por muitos sistemas de informação. Na atualidade, a maioria das aplicações que lida com informação primária de biodiversidade vem utilizando cada vez mais os padrões de metadados, como é o caso do Darwin Core e o ABCD, fomentados e desenvolvidos por organizações internacionais.

Tabela 2.2: Exemplos de campos do padrão Darwin Core, retirado de [30]

Campo	**Descrição**	**Exemplo**
ScientificName	Táxon de mais baixo nível no qual o organismo foi identificado	Ctenomys sociabilis (Genus + SpecificEpithet)
CollectingMethod	O nome ou breve descrição do método ou protocolo usado na coleta	armadilha de raios UV, rede de arrastão
Collector	Nome(s) do(s) coletor(es)	Erica P. Anseloni
DecimalLatitude	Latitude do local no qual o organismo foi coletado, em graus decimais	23, 41

O Darwin Core é baseado no padrão Dublin Core[2] proposto inicialmente para metadados de obras impressas e objetos digitais em geral (por exemplo, videos, sons, imagens, textos e documentos na *Web*). A comunidade de pesquisa ecológica desenvolveu o padrão EML (*Ecological Metadata Language*) [36], o qual permite a representação de dados ecológicos. O padrão FGDC (*Federal Geographic Data Committee*) [13] e o ISO19115/ISO19119 [41] visam fornecer um conjunto de definições para organizar e descrever dados geoespaciais [15]. Embora o padrão FGDC lide com dados geoespaciais, ainda não oferece suporte para os dados provenientes da pesquisa biológica. Assim, alguns elementos do EML foram incorporados dentro de um perfil biológico ao FGDC para tornar este padrão mais abrangente para os pesquisadores da área de ecologia [22].

2.2.2 Protocolos para compartilhamento de dados de biodiversidade

Para possibilitar a disseminação e o compartilhamento de informação, são necessários mecanismos que forneçam conectividade entre as aplicações [30]. A primeira iniciativa direcionada ao compartilhamento de informação na Internet foi o HTTP (*Hypertext Transfer*

[2]http://dublincore.org/

Protocol) que é um protocolo de comunicação na camada de aplicação segundo o Modelo OSI (*(Open Systems Interconnection)*) [57]. Outro protocolo comum é o SOAP (*Simple Object Access Protocol*) baseado em documentos XML [58]. SOAP é independente da plataforma de desenvolvimento e visa fornecer um meio de comunicação para tráfego das mensagens entre diversos protocolos.

As comunidades de cientistas de biodiversidade também precisam compartilhar seus dados para dar à sua pesquisa um alcance global. Assim, torna-se necessário o uso de protocolos de comunicação neste domínio. DiGir [51] é um protocolo que disponibiliza um ponto de acesso às fontes de dados distribuídas. Está baseado em HTTP, XML e UDDI, três elementos que possibilitam o transporte, representação, publicação e gerência dos dados e serviços *Web*. Além disso, DiGir utiliza o padrão Darwin Core [54] como uma especificação funcional.

BioCase [21] é um outro protocolo desenvolvido para fornecer conectividade entre bases de coleções biológicas. É baseado em HTTP e XML e utiliza o padrão ABCD [52] para transmitir dados na rede BioCase. A principal diferença entre os dois padrões é que o DiGir lida com o esquema Darwin Core, enquanto que BioCase permite que o provedor de dados selecione um esquema conceitual de metadados, comumente o esquema ABCD [26]. Ambos protocolos são uma proposta do grupo *Biodiversity Information Standards*, anteriormente conhecido como *Taxonomic Database Working Group* (TDWG) [53].

Uma nova abordagem, conhecida como TAPIR (TDWG *Access Protocol for Information Retrieval*), está sendo fomentada pelo GBIF [23] para unificar os protocolos DiGir e BioCASE, para aumentar a interoperabilidade entre esses protocolos. TAPIR [53] especifica um protocolo padrão que está baseado em esquemas XML e serviços *Web*, sendo que a informação é transmitida via HTTP. Uma característica deste protocolo é a independência que possui da estrutura lógica do banco de dados que se utilize. Protocolos como DiGIR, por exemplo, precisam utilizar um esquema único (o Darwin Core) para poder acessar e transmitir a informação.

Em linhas gerais, os projetos que desejem utilizar TAPIR deverão mapear o esquema da sua estrutura de dados para uma camada de abstração de dados. DarwinCore e ABCD são exemplos de camadas de abstração de dados para biodiversidade. Assim, os diferentes bancos de dados participantes podem ser mostrados dentro de uma visão unificada dos dados. A Figura 2.1 mostra um exemplo do mapeamento entre uma fonte de dados e a camada de abstração que neste caso é representada pelo Darwin Core. Na parte esquerda temos a representação de um repositório que contém campos para descrever a localização e classificação taxonômica de espécie. No lado direito temos alguns dos elementos que o Darwin Core possui. Cada um dos campos do repositório é casado com os elementos do Darwin Core. Por exemplo, os campos *GUID*, *Latitude* e *Longitude* que pertencem à tabela *Specimen* correspondem aos campos *GUID*, *DecimalLatitude* e *DecimalLongitude*

respectivamente.

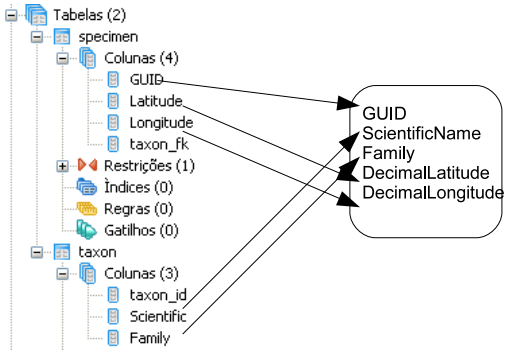


Figura 2.1: Mapeamento da Fonte de Dados para a Camada de Abstração, inspirado em [53]

As organizações como GBIF esperam que os fornecedores de dados que utilizam os protocolos DiGIR, BioCASE ou outros migrem gradativamente para o TAPIR, os novos fornecedores usem esse protocolo, e novos desenvolvimentos sejam feitos com base no TAPIR. Por exemplo, o projeto speciesLink [11] ainda usa o protocolo DiGIR mas planeja migrar para o TAPIR. Na atualidade ainda é difícil avaliar estatisticamente a utilização desses tipos de protocolos porque não existe um meio formal para a análise destas atividades. Novas implementações desenvolvidas pelo TDWG estão considerando as ontologias como um meio para melhorar a interoperabilidade em sistemas de biodiversidade.

Como dados de biodiversidade estão fortemente ligados ao espaço, o uso de serviços e padrões espaciais vem sendo adotado por sistemas de biodiversidade. Assim, outro esforço, dirigido pela OGC (*Open Geospatial Consortium*) [41], visa o desenvolvimento e a implementação de padrões de informação e serviços *Web* geoespaciais. Os serviços disponibilizados pela OGC incluem *Web Feature Service* (WFS), que permite que os clientes obtenham e atualizem dados geoespaciais codificados em GML (*Geography Markup Language*). GML é uma linguagem baseada em XML que permite a representação de características geográficas [14]. O *Web Map Service* (WMS), também disponibilizado pela OGC, define um serviço de mapas em duas dimensões a partir de dados geoespaciais, descritos em um formato portável e multi-plataforma para armazenamento em um repositório ou para transferência entre aplicações. O *Web Coverage Service* (WCS), da OGC, possibilita o acesso aos dados que representam fenômenos com variação contínua no espaço, permitindo a manipulação de dados modelados como campos geográficos.

O projeto OBIS-SEAMAP [27], por exemplo, adotou a especificação WMS da OGC para recuperar imagens oceanográficas do servidor disponibilizado pela NASA [3] e o DiGir

[3]http://seablade.jpl.nasa.gov/de.shtml

como protocolo para disseminação, busca e recuperação de informação entre as redes participantes do projeto.

2.3 O projeto BIO-CORE

O principal objetivo do BIO-CORE [2] é oferecer um sistema para cientistas que trabalham com questões ambientais e de biodiversidade. A disponibilização na *Web* permitirá seu uso por equipes distintas de pesquisadores. O BIO-CORE teve sua origem no projeto de sistema de biodiversidade *WeBios* [1], sendo ambos propostos por pesquisadores em computação e em biodiversidade. Alguns dos resultados obtidos no *WeBios* foram o desenvolvimento de um Serviço de Ontologias [18] e um Serviço Ecologicamente Ciente [30] baseado também em ontologias e repositórios distribuídos. Basicamente o BIO-CORE possui a mesma arquitetura do *WeBios*, mas além de permitir gerenciar dados de coletas na natureza, também incorpora informação do Museu de Zoologia da UNICAMP. Uma outra característica de BIO-CORE é que visa fornecer um ambiente para ensino e compartilhamento de informação entre pesquisadores da biologia e computação.

As fontes de dados do BIO-CORE incluem imagens (fotos de seres vivos ou seus habitats), dados geográficos (mapas de regiões com ocorrência de espécies), ontologias e metadados específicos do domínio (descrições do habitat e ecossistema). A idéia é permitir que cientistas realizem um trabalho exploratório que considere seus conhecimentos sobre espécies, as interações entre elas e seus habitats e outras correlações. As diversas fontes de dados serão acessadas por meio de serviços *Web*, enquanto a formulação das consultas, pré-processamento e visualização dos resultados executarão como uma aplicação cliente remota.

O principal diferencial em relação aos demais Sistemas de Biodiversidade é permitir em uma mesma consulta, a combinação de predicados baseados em conteúdo, espaciais e textuais tradicionais (de ontologias e metadados) [16]. Os sistemas disponíveis publicamente não atacam estas questões simultaneamente, concentrando-se apenas em dados de imagem ou em dados espaciais. A Figura 2.2 ilustra em linhas gerais a arquitetura proposta para o *WeBios* e estendida para o BIO-CORE. Ela é composta de três camadas principais: o *Serviço de Armazenamento*, os *Serviços de Suporte* e os *Serviços Avançados*. O módulo *Aplicação Cliente*, como o nome indica, deve ser executado como tal e utiliza Serviços *Web* para acessar os dados provenientes dos demais módulos. Esta aplicação é responsável por reunir, processar e exibir os dados ao usuário.

Os Serviços de Armazenamento encapsulam os repositórios de dados que são acessados pelos serviços das camadas superiores. Os Serviços de Suporte mostrados na Figura 2.2 incluem um serviço para consulta de imagens baseado em conteúdo; um Serviço de Metadados que ajuda a estabelecer relacionamentos entre fontes de dados distintas; um

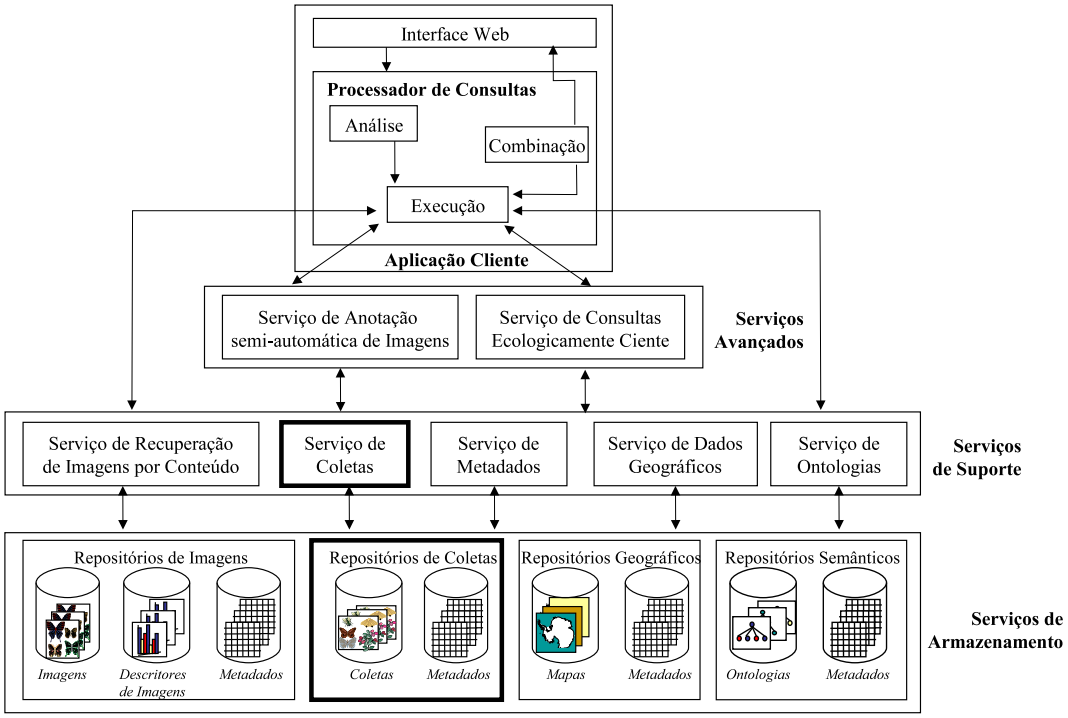


Figura 2.2: Arquitetura do Sistema BIO-CORE, retirado de [1]

Serviço de Coletas que visa possibilitar o gerenciamento e consultas dos dados de coletas de espécies e finalmente um Serviço de Ontologias que permite o acesso e gerenciamento a ontologias que descrevam conceitos de algum domínio específico, resolvendo casos de ambigüidade de termos e refinando as consultas [18]. Os Serviços Avançados visam combinar as funcionalidades de dois ou mais serviços de suporte, provendo funcionalidades mais complexas.

Como parte dos serviços avançados estão o Serviço de Consultas Ecologicamente Ciente [30], que combina relações ecológicas entre espécies, descritas em ontologias, com a recuperação de dados geográficos relacionados à ocorrência de espécies; e um Serviço de Anotação Semi-automática de Imagens baseado em ontologias, que auxilia a busca e anotação de imagens, integrando abordagens baseadas no conteúdo da imagem com ontologias e palavras-chave.

Esta proposta de mestrado se refere à especificação e implementação do Repositório e do Serviço de Coletas destacados na figura 2.2. O desenvolvimento destes serviços será baseado nas características que fornecem as tecnologias de código de fonte aberto e Serviços *Web* que visam garantir a interoperabilidade na Internet.

2.4 Resumo

Este capítulo apresentou uma descrição dos principais conceitos e as aplicações disponíveis para o gerenciamento de biodiversidade, junto com as abordagens e tecnologias que utilizam. Além disso, o capítulo descreveu brevemente os padrões e mecanismos para compartilhamento de dados. Todo este arcabouço de conhecimentos será utilizado para a especificação do serviço e o repositório de coletas que é parte do projeto BIO-CORE.

Capítulo 3

Modelo Proposto para Armazenamento de Dados de Biodiversidade

Este capítulo descreve o modelo de banco de dados proposto para gerenciar dados de biodiversidade. A seção 3.1 apresenta uma visão geral do modelo e os cenários que foram estudados para seu projeto. A seção 3.2 apresenta o modelo proposto e descreve as suas características mais relevantes. A seção 3.3 apresenta exemplos de consultas baseadas em nosso modelo. Em seguida, a seção 3.4 mostra alguns exemplos da informação que é armazenada. Finalmente, a seção 3.5 apresenta o resumo do capítulo.

3.1 Visão Geral

Uma parte importante de nosso trabalho foi o projeto do Repositório de Coletas que contém informação sobre dados de biodiversidade, unificando em um único modelo registros de acervos e de observações de espécies. Uma característica no desenvolvimento do modelo foi que este deveria permitir o intercâmbio de informações entre diferentes grupos de pesquisa. Por meio de consultas à informação armazenada, de acordo com o modelo, pode-se responder sobre características da classificação de uma espécie coletada e também determinar outros elementos envolvidos na sua catalogação, como procedimentos de amostragens da coleta de espécies, dados relacionados à sua localização geográfica, entre outros.

O trabalho foi conduzido em cooperação com nossos usuários alvo, biólogos de duas áreas de pesquisa: ecologia e biologia marinha. O primeiro grupo realiza trabalho de campo para coletar dados sobre interações entre insetos e plantas e emprega métodos específicos para a organização dessas coletas (Laboratório Inseto-Planta). Já os biólogos

marinhos realizam coletas de pequenos animais marinhos. Estes últimos comandam um grande projeto que visa reorganizar o Museu de Zoologia da UNICAMP e, portanto, estão familiarizados com as necessidades e métodos para gerenciar a catalogação de espécies.

Assim, o modelo de banco de dados reflete uma visão dual para o gerenciamento de dados sobre biodiversidade. De um lado, suporta o armazenamento e a manipulação de dados de observação e de coletas de espécies feitas em campo. Por outro lado, também atende às necessidades de catalogação do Museu, cuja função está mais próxima às dos gestores de bibliotecas (físicas e digitais).

3.1.1 Cenário Genérico para o Modelo

O modelo de banco de dados proposto armazena dois tipos de informação: registros de catálogos e de acervos de museus e registros de coletas e de observações feitas em campo. O primeiro tipo de registros trata de coleções catalogadas, enquanto que o segundo envolve coleções de (um ou vários) grupos de pesquisa, em que os seres vivos coletados ou observados não estão disponíveis em um acervo comum. Toda a informação obtida é registrada em estruturas de armazenamento próprias de cada grupo de pesquisa, como planilhas em *Excel*, bases de dados, catálogos ou fichas ou algum outro tipo de meio digital.

Em linhas gerais, o procedimento que os biólogos realizam para a pesquisa de campo e catalogação de espécies abrange um conjunto de passos que começam com a coleta de amostras em campo e finalizam com o armazenamento da informação em repositórios de dados e em coleções físicas de museus. Cada coleta em campo é feita utilizando uma metodologia, em uma localidade e período específicos, por determinados coletores. Em campo realizam-se anotações sobre a informação referente às coletas, que são levadas ao laboratório para estudo. Essas anotações referem-se ao lugar onde está sendo realizada a coleta, quem é o responsável, a data, o que e o como está sendo coletado.

As amostras são o material recolhido em campo, por exemplo um conjunto de plantas, um pote de areia, um grupo de insetos, entre outros. Também podem ser uma observação que os biólogos realizam, como uma marca de pisada, um vídeo, uma gravação (por exemplo, o canto de um pássaro) ou uma foto. A informação base que é registrada pelos biólogos refere-se a: o que (identificação de espécies), onde (caracterização da localização), como (metodologia de coleta), quem (coletou ou identificou) e quando (datas em que são realizadas as coletas). Outras informações de destaque que podem ser armazenadas envolvem as condições ambientais, o habitat onde se coletam as amostras, as características das espécies, por exemplo propriedades anatômicas, e os relacionamentos ecológicos entre espécies, por exemplo, competição, predação, mutualismo, entre outros [38].

Em geral, as amostras são separadas em diferentes *substratos*, que podem conter ma-

terial biológico e não biológico. O substrato representa parte da amostra onde estão os seres a serem estudados. Por exemplo, na coleta de pequenos animais em baldes de areia, o substrato é a areia. Em alguns casos, um substrato também pode ser um animal alvo, como um rato ou uma borboleta.

Diferenças metodológicas de coleta são observadas em cada ambiente de pesquisa. Por exemplo, para um certo tipo de coleta de pequenos animais realizada em uma praia, retira-se uma quantidade de areia, com o auxílio de pás ou cilindros (metodologia). Essa areia (substrato) é acondicionada em sacos e levada ao laboratório para triagem, onde os animais são separados do substrato com o auxílio de peneiras de diferentes tamanhos de malhas. A medida destas malhas determina o tamanho do animal a ser retido e separado para estudos. Coletas realizadas em outros ambientes também consistem em amostragem de determinadas quantidades de substratos para posterior separação dos animais. Por exemplo, as coletas em infralitoral (mar adentro) dependem de equipamentos maiores e mais pesados para que possam atingir o fundo marinho e amostrar substratos em diferentes profundidades. Neste caso, podem ser utilizados dragas de arrasto, redes de pesca e/ou pegadores de fundo.

Cada amostra coletada pode conter mais de um indivíduo de uma ou mais espécies. Os indivíduos devem ser separados de acordo com a espécie a que pertencem. Cada indivíduo é chamado de *espécime*. Um *lote* representa um conjunto de um ou mais indivíduos da mesma espécie. Pode haver ou não necessidade de registro individual de cada espécime; não havendo essa necessidade, registra-se somente o lote. Caso contrário, o lote e cada espécime são registrados.

Finalmente, os lotes são enviados à catalogação, em algum acervo, onde de novo há registro das propriedades de interesse (o que, quem, como, quando, onde). A Figura 3.1 ilustra em uma seqüência de passos o cenário geral dos procedimentos de coletas e catalogação.

Após as amostras serem coletadas (1), e segundo o ambiente de pesquisa, os biólogos separam as amostras em substratos biológicos e não biológicos (2). Os substratos biológicos são divididos em lotes (3). Esses conjuntos podem ser registrados em um banco de dados de coletas de um grupo de pesquisa (6) ou podem ser enviados ao Museu (4) para serem registrados no banco de dados do Museu (5.a) e preservados em coleções físicas, cujo registro está no banco de dados (5.b). Há também a possibilidade de se registrar diretamente as espécies amostradas em um banco de dados de coletas (6).

3.1.2 Cenários de Trabalho dos usuários de BioCore

A seção 3.1.1 apresentou uma descrição geral dos procedimentos que os biólogos realizam para coleta e/ou catalogação de espécies. Partindo desse cenário geral, esta seção descreve

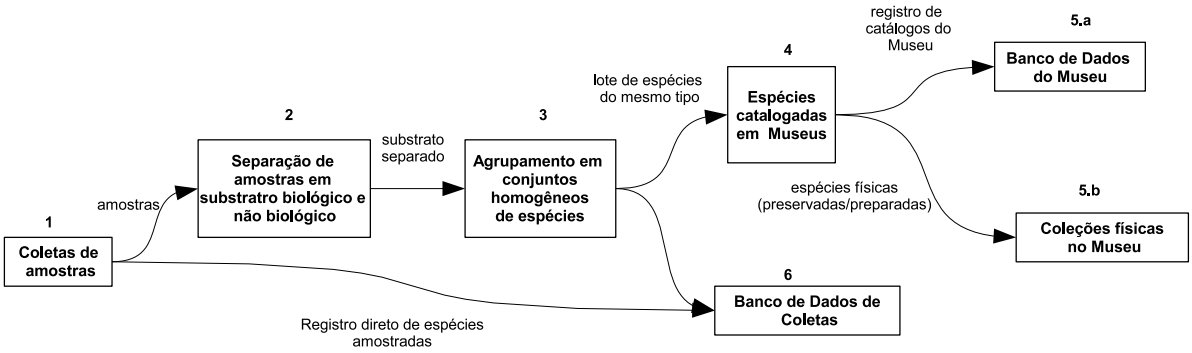


Figura 3.1: Representação do cenário geral dos procedimentos de coleta e catalogação - cópia de tela do WOODSS [47]

dois cenários específicos que foram estudados para a especificação do modelo, tendo sido necessário realizar um levantamento de informação que permitisse entender qual é a forma de trabalho adotada dos usuários-alvo.

No caso do Laboratório Inseto-Planta, o objetivo é identificar espécies de insetos que interagem com espécies de plantas. O procedimento de coleta utilizado atualmente é: (1) recolher flores em campo, que em grande maioria contêm larvas de insetos, e (2) aguardar que os insetos cheguem à fase adulta para sua identificação [42].

Os *capítulos* são um conjunto de flores de uma planta. Deve-se registrar informação sobre o capítulo (ou capítulos) e os insetos nele encontrados. Quando os biólogos encontram uma planta rara, sem capítulos suficientes para compor uma amostra para identificação de insetos, coletam a planta para ter o registro de localização da sua espécie. A Figura 3.2 apresenta o processo correspondente, representado em um diagrama de atividades. É possível verificar que, após a coleta em campo, o pesquisador deve retornar ao laboratório para acomodar as flores das plantas amostradas em potes fechados. Junto com as flores estão as larvas dos insetos que alimentam-se dos tecidos florais. Quando os insetos chegam à fase adulta, os biólogos iniciam o processo de identificação desses insetos e toda a informação obtida é registrada em uma base de dados.

No caso do Museu de Zoologia, o procedimento é diferente. O Museu recebe depósitos (lotes ou coletas) de animais para catalogar. Neste processo, conhecido como *tombamento*, os animais são separados em lotes que agrupam números diferentes de indivíduos de um mesmo táxon, provenientes de uma mesma localidade e coletados em uma mesma data. Um táxon descreve o nível taxonômico de uma espécie [30]. A cada um destes lotes (com um ou mais exemplares) é atribuído um número de registro de catálogo (ou número de tombo). Após essa separação e identificação, os dados e características referentes ao lote são registrados em um banco de dados e em um livro "Tombo" impresso. A Figura 3.3 ilustra este processo, indicando que após receber as coletas dos animais, o pesquisador

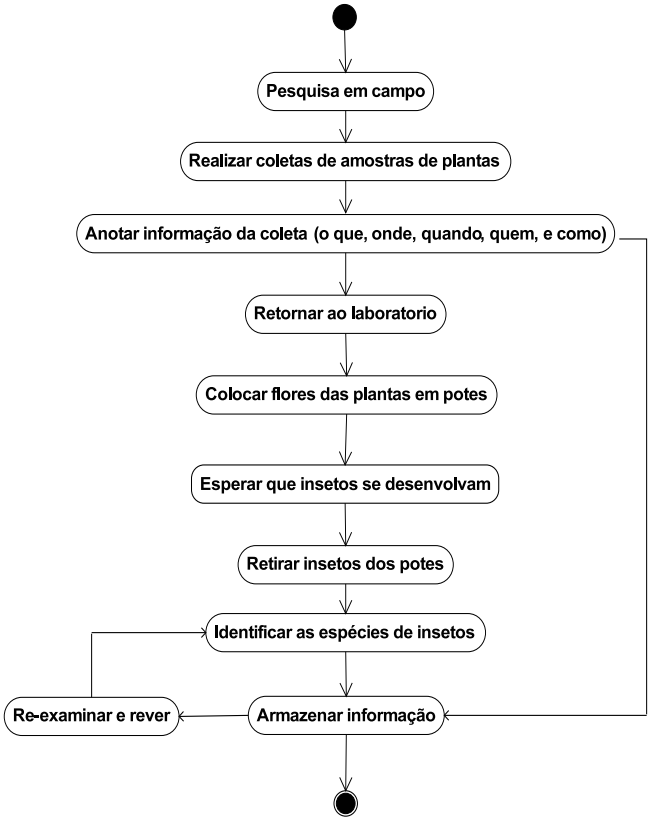


Figura 3.2: Diagrama de atividades para o processo de coletas do laboratório Inseto-Plantas

deve separá-los em lotes de espécies para registro em catálogo. No entanto, existem casos em que os animais já chegam organizados em lotes e, assim, o responsável somente atribui ao lote um número de tombo e faz o registro.

Um detalhe importante é que o Museu de Zoologia não precisa promover coletas em campo. Seu objetivo é criar um acervo para pesquisa. Portanto, diferentes pesquisadores podem depositar os animais provenientes de seus trabalhos no Museu. O processo que sempre acontece é a catalogação. Ressalte-se que tais processos não terminam. Uma coleta (ou um elemento tombado) pode ser re-estudado, subdividido e ter a identificação modificada. Isso significa que as informações podem variar temporalmente, inclusive quanto à identificação das espécies.

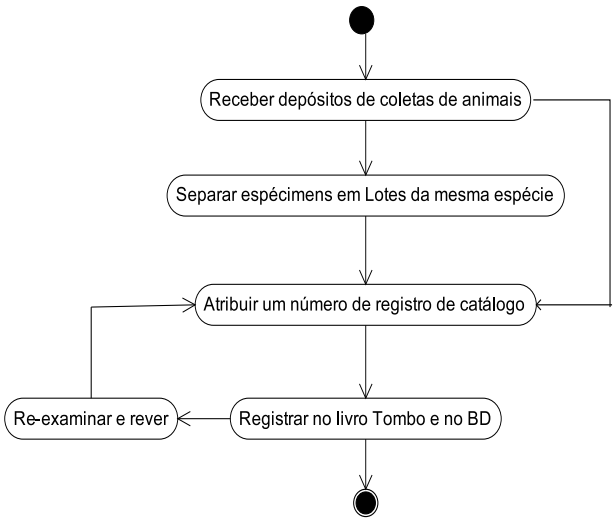


Figura 3.3: Diagrama de atividades para o processo de catalogação do Museu de Zoologia

3.2 O Modelo de Banco de Dados proposto para o BioCore

O modelo proposto na dissertação é uma extensão de um modelo anterior projetado para o laboratório Inseto-Plantas, havendo sido alterado por duas razões: (1) para que comportasse a informação do Museu de Zoologia e (2) para responder a questionamentos que surgiram em entrevistas com os biólogos. Este novo modelo incorpora elementos que são parte do padrão Darwin Core [54], visando a interoperabilidade futura de nosso trabalho com outros projetos semelhantes, como o OBIS-SEAMAP, o speciesLink, entre outros mencionados no Capítulo 2. O projeto proposto partiu da definição de um subconjunto de metadados de interesse no Darwin Core, acrescentando outros campos relevantes, especificados por nossos usuários alvo. O Darwin Core é um padrão que possui uma boa documentação e foi fácil de adaptá-lo ao modelo.

Tabela 3.1: Elementos de interesse do padrão Darwin Core

Elementos	Descrição
InstitutionCode	Código da instituição. Exemplo: ZUEC
CollectionCode	Código da coleção. Exemplo: BIV, POL
CatalogNumber	Número do registro no livro tombo
BasisOfRecord	É uma descrição que indica se o registro é uma observação ou objeto. Por exemplo: Foto, Son, Pegada, Fezes, entre outros
Kingdom	O nome do reino (Kingdom) no qual o organismo é classificado
Phylum	O nome do filo (Phylum) no qual o organismo é classificado
Class	O nome da classe (Class) no qual o organismo é classificado
Order	O nome da ordem (Order) no qual o organismo é classificado
Family	O nome da familia (Family) no qual o organismo é classificado
Genus	O nome do gênero (Genus) no qual o organismo é classificado
SpecificEpithet	Um adjetivo ou substantivo em latim escrito em minúsculas que segue o gênero, permitindo distinguir uma espécie de outra
AuthorYearOfScientificName	Autor e data de identificação de uma espécie
DateIdentified	Data em que um identificador identificou uma espécie
EarliestDateCollected	Data de início da coleta
LatestDateCollected	Data de fim da coleta
ContinentOcean	Nome do continente ou oceano onde se realizou a coleta
Country	Nome do país onde se realizou a coleta
StateProvince	Nome do Estado onde se realizou a coleta
County	Nome do Município onde se realizou a coleta
Longitude	Longitude do local da coleta
Latitude	Latitude do local da coleta
CoordinatePrecision	O limite superior da distância (em metros) a partir da Latitude e Longitude do local da coleta, descrevendo um círculo dentro do qual a localidade está descrita
Sex	Indica o sexo de um especimem
PreparationType	O tipo de preparação que os biólogos realizam nas espécies. Exemplo: taxidermia
IndividualCount	Quantidade de indivíduos de uma mesma espécie
PreviousCatalogNumber	Um número de catálogo anterior

A Tabela 3.1 apresenta os dados do Darwin Core adotados no modelo. A coluna *Elementos* lista os elementos de interesse e a coluna *Descrição* descreve brevemente o significado de cada um desses elementos. Embora a maioria dos elementos listados na tabela pertençam à versão 1.2 do padrão Darwin Core (também conhecido como Darwin-CoreV2), foi necessário adicionar alguns elementos da versão 1.4, por se adaptarem melhor às necessidades dos nossos usuários. A versão 1.2 foi escolhida por ser a mais estável e porque foi usada nos projetos [18] e [30] que estão relacionados a esta dissertação. Os elementos que pertencem à versão 1.4 são destacados em negrito.

A Figura 3.4 mostra o diagrama E-R simplificado do modelo proposto. O modelo abrange um amplo espectro de consultas, tanto no acervo do museu quanto do conjunto

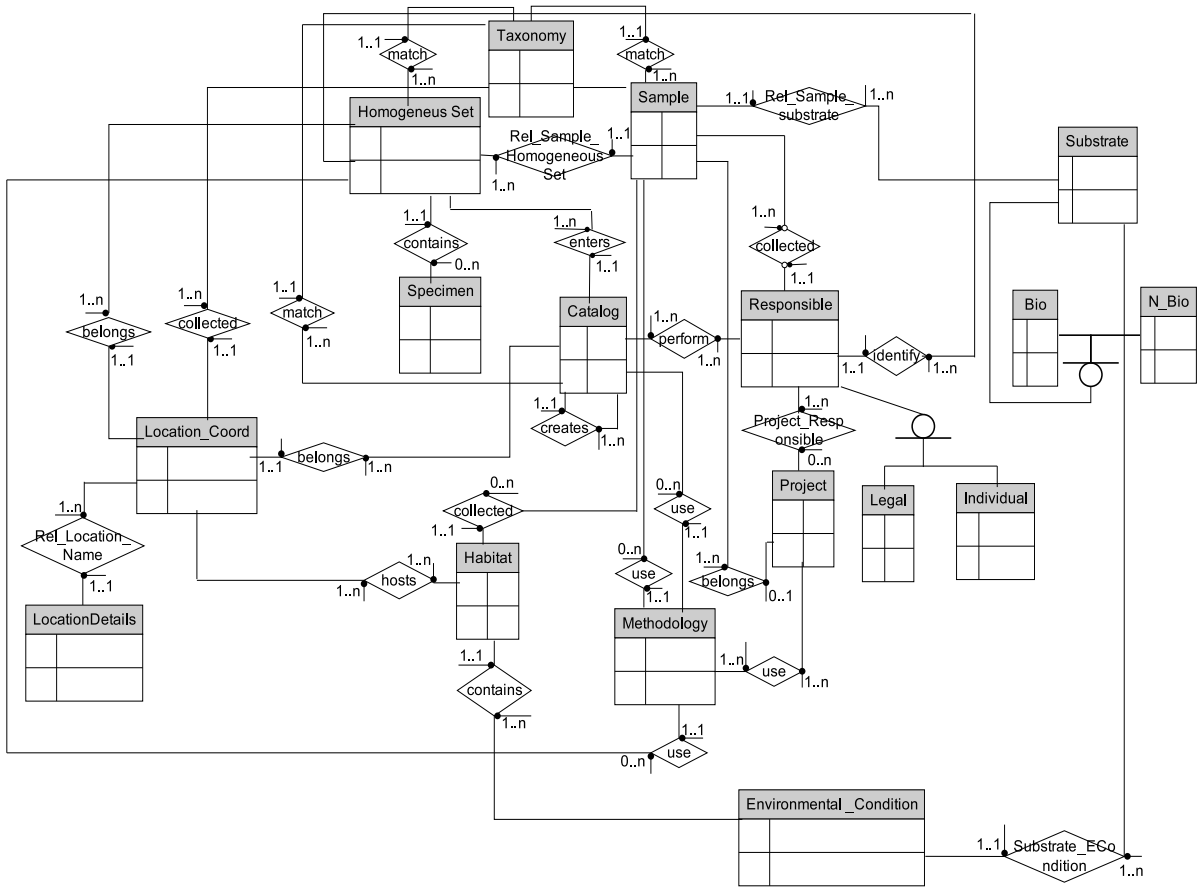


Figura 3.4: Modelo Diagrama Entidade-Relacionamento simplificado do Banco de Dados

de observações em campo. As entidades centrais do modelo de banco de dados são: *Sample* (Amostra), que corresponde às observações e aos registros de coletas de campo; *Homogeneous_Set* (Lote), que armazena os registros sobre um conjunto de indivíduos do mesmo tipo extraídos das coletas feitas em campo; e *Catalog* (Tombo), que engloba os registros do museu.

Exemplificando, os pesquisadores recolhem baldes com areia do mar (uma *Sample*) onde encontram pequenos animais marinhos, que são classificados em lotes, que são conjuntos homogêneos de espécies (cada conjunto é um *Homogeneous_Set*), e atribuem um número a este lote. De um lote, são selecionados espécimes que são extraídos para serem catalogados e preservados no museu. Cada uma dessas coletas possui uma metodologia associada, descrevendo o procedimento de coleta, e está vinculada a projetos desenvolvidos pelo grupo de pesquisa.

Em outro exemplo, podemos ter um grupo de capítulos de plantas que inicialmente podem ser amostrados (*Sample*), depois tornam-se um lote (*Homogeneous_Set*) e finalmente

são catalogados no Herbário. Uma alteração nesse exemplo acontece quando os biólogos utilizam esses capítulos (que são tratados como substratos) e os armazenam em um pote. Nesses capítulos estão as larvas (juvenis) dos insetos que consomem os tecidos florais. Quando eles se transformam em adultos, os biólogos identificam as espécies dos insetos encontrados. Finalmente esses insetos se transformam em um lote (*Homogeneous_Set*) (ver Figura 3.2).

Os registros de *Sample*, *Homogeneous_Set* e *Catalog* precisam responder ao mesmo tipo de consulta: o *que* (espécie identificada), o *como* (foi coletado, preservado ou catalogado), por *quem* (coletor ou identificador), *quando* e *onde*. As respostas a estas perguntas precisam de um contexto. Por exemplo, se as consultas se referem a observações feitas em campo, registros no catálogo, ou suas interligações. Além disso, o *que* (referente à informação taxonômica) é freqüentemente incompleto e pode mudar como resultado da evolução do conhecimento científico sobre as espécies. O mesmo pode acontecer com *onde*, sempre que se usa nome de lugares ao invés de coordenadas. A entidade *Specimen* representa cada um dos indivíduos do lote registrados em *Homogeneous_Set*. O registro de um indivíduo componente do lote é opcional, dependendo da necessidade do biólogo de estudar individualmente cada ser coletado.

A entidade *Taxonomy* corresponde aos registros dos táxons. As entidades *Homogeneous_Set*, *Catalog* e *Sample* são ligadas às suas respectivas classificações taxonômicas pelo relacionamento *match*. Há necessidade, em alguns estudos, de se descrever as interações que existem entre as amostras (por exemplo, um grupo de plantas) e os lotes, por exemplo, um grupo de insetos. O relacionamento *Rel_Sample_HomogeneousSet* permite essa descrição.

Uma amostra pode envolver (*Rel_Sample_Substrate*) um substrato (*Substrate*) que pode ser biológico ou não. A especialização das entidades *Bio* e *N_Bio* representa os substratos biológicos e não-biológicos, respectivamente. A entidade *Environmental_Condition* representa a informação referente às variáveis ambientais do local onde se realizou uma coleta e está associada (*contains*) a um *Habitat*. Um substrato também pode envolver condições ambientais, sendo *Substrate_ECondition* que faz essa ligação.

A entidade *Habitat* armazena as informações sobre o habitat ao qual pertencem as coletas e a entidade *Location_Coord* descreve as coordenadas geográficas onde são realizadas as coletas. Uma localidade abriga (*hosts*) um habitat. As amostras são coletadas (*collected*) em uma localidade. As entidades *Homogeneous_Set* e *Catalog* são ligadas às suas respectivas localidades pelos relacionamentos *belongs*.

A entidade *LocationDetails* complementa (*Rel_Location_Name*) os dados de *Location_Coord*. *Responsible* corresponde a cada entidade (Física ou Jurídica) que realiza coletas (*collected*), identifica (*identify*) ou cataloga (*perform*) às espécies.

Project é a entidade que engloba os dados dos projetos realizados. Um projeto deve

ter (*Project_Responsible*) um ou mais responsáveis. A entidade *Methodology* contém os métodos utilizados nas coletas. Cada projeto utiliza (*use*) uma metodologia. As amostras são coletadas (*belongs*) para um projeto. Uma atividade para coletar uma amostra também poderia usar (*use*) uma metodologia independente de projeto. A Figura 3.5 apresenta o modelo com todos os elementos que o compõem. Para um melhor entendimento desses campos, o apêndice A apresenta o dicionário de dados, com a descrição dos atributos e o objetivo de cada entidade.

No modelo foi necessário repetir campos nas entidades *Catalog*, *Homogeneous_Set* e *Sample*. Isso devido aos seguintes fatos: i) pesquisadores de diferentes grupos podem depositar diretamente no museu espécies para sua catalogação; ii) a quantidade de indivíduos catalogados em um lote pelo museu pode corresponder apenas a um subconjunto da quantidade de indivíduos efetivamente identificada neste lote; e iii) pode haver necessidade de se registrar todo o procedimento da coleta.

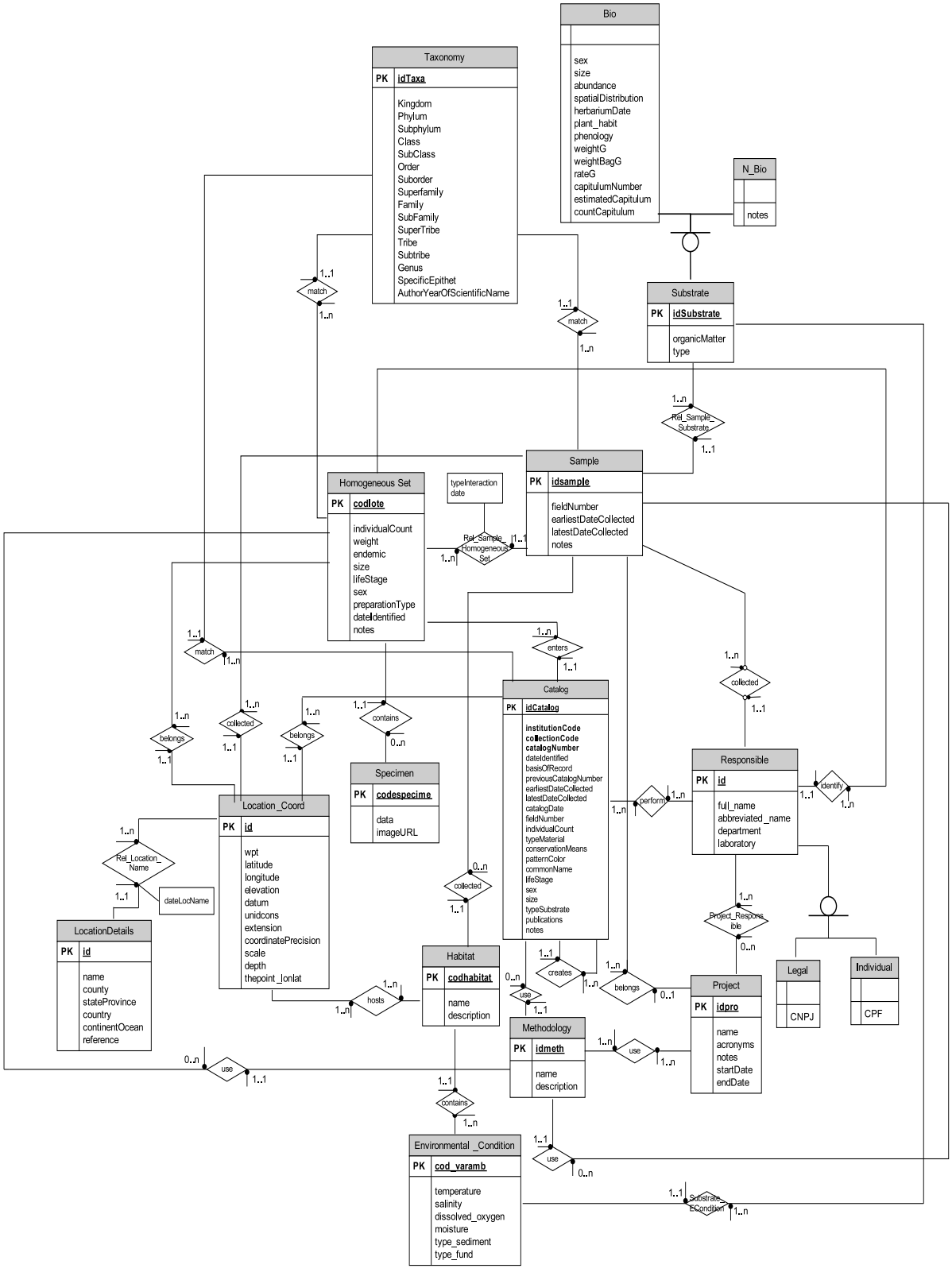


Figura 3.5: Diagrama Entidade-Relacionamento do banco de dados baseado no padrão Darwin Core versão 1.2

3.3 Exemplos de Consultas

O desenvolvimento do modelo exigiu a análise das consultas dos usuários, aqui classificadas como simples e complexas. Em analogia ao trabalho desenvolvido em [32] consideramos consultas simples aquelas que acessam uma única tabela e cujos predicados sejam igualmente simples, ou seja, não exigem o desenvolvimento de funções específicas. Consultas simples correspondem a consultas padrão da álgebra relacional, para uma única tabela.

Um exemplo de consulta simples, utilizando o modelo seria "Mostrar as características catalogadas de uma coleção C". Para obter o resultado desejado, esta consulta precisa pesquisar somente a tabela *Catalog* e retornar as tuplas correspondentes à coleção C, ou seja, apenas uma operação de seleção. Um outro exemplo é "Apresentar as espécies que pertencem a um táxon específico cujo valor é X". Esta consulta é realizada na tabela *Taxonomy*, retornando os registros correspondentes do nível taxonômico especificado.

Já consultas complexas envolvem a junção de várias tabelas e/ou os predicados exigem a implementação de funções específicas, podendo envolver relacionamentos espaciais ou temporais. Em [20] detalha-se um conjunto de consultas que utilizam este tipo de predicados. Além disso, neste tipo de consultas o número de tuplas pode ser bastante volumoso e como há muitas lacunas nos dados armazenados, o processamento de junções precisa considerar diversos tipos de valor nulo. Por estes motivos, consultas usando junções são tratadas como consultas complexas. Embora junções sejam operações básicas em álgebra, muitas das junções exigidas nas consultas dos sistemas de biodiversidade podem envolver predicados que incluem o desenvolvemento de funções específicas.

Por exemplo, o modelo descrito na seção 3.2 permite rastrear um registro de coleta catalogado a partir de junções de múltiplas relações. Rastrear um registro de coleta do catálogo significa obter informação relacionada sobre o *que* (espécie e taxonomia), *quem* (identificou e/ou coletou), *quando* (data de coleta, identificação e catalogação), *como* (a metodologia) e *onde* (localização da coleta). Um caso mais complexo que poderia se apresentar neste rastreamento é quando, por algum motivo, esse registro tornou-se um novo item. Neste caso é preciso obter o número de catálogo anterior (*previousCatalogNumber*) para realizar o rastreamento correto desse registro.

Outros exemplos incluem predicados espaciais, temporais ou ambos. Alguns exemplos de consultas que envolvem predicados espaciais são: a) "Selecione as amostras feitas ao norte do município de Ubatuba", b)"Recupere as espécies localizadas em um raio de 10 quilômetros das coordenadas(X,Y)", e c) "Obtenha os habitats que estão contidos em uma região M onde realizaram-se coletas".

Exemplos de consultas que envolvem predicados temporais complexos são: a) "Quais amostras foram coletadas no mesmo período que a amostra S", b) "Quais espécies foram agrupadas em lotes antes do segundo semestre do ano 2000". Um exemplo combinando

ambos predicados é: "Quais são as espécies amostradas nas regiões adjacentes a uma região R durante um certo período".

3.4 Exemplo da informação armazenada pelos usuários

A seguir apresentamos um exemplo de uma coleta do laboratório Inseto-Planta e um exemplo da catalogação de espécies fornecidas pelos biólogos. Esses exemplos correspondem a dados a serem armazenados no Repositório de Coletas segundo o modelo. Nas tabelas os campos de chaves estrangeiras e relacionamentos foram omitidos. Dados incompletos são representados por "...". No mapeamento do modelo ER para o banco de dados relacional, decidiu-se tratar as especializações da entidade *SUBSTRATE* em uma única tabela.

Como explicado na seção 3.1.2, o laboratório Inseto-Planta realiza coletas de plantas de onde extraem-se lotes de insetos para sua identificação. As figuras 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9 e 3.10 ilustram um exemplo de informação armazenada para uma coleta. Neste exemplo, representamos a informação de uma amostra de planta da espécie *Proteopsis argentea*, coletada em 16/02/1995, sendo que os biólogos envolvidos são Paulo I.K.L. Prado, Thomas M. Lewinsohn e Bruno D. Buys. As colunas *tabela*, *campo* e *valor* correspondem ao nome da tabela, ao nome do campo e ao seu conteúdo, respectivamente.

A tabela 3.2 corresponde à entidade *Sample* e contém informação sobre o período em que foi realizada essa amostra, no caso "16/02/1995". A entidade *Substrate* é mapeada na tabela 3.3, que complementa os dados contidos na entidade *Sample* descrevendo características relevantes da amostra, como a fenologia da planta amostrada, sua distribuição espacial, o número de capítulos que essa amostra possui e o tipo de substrato, entre outros. No exemplo, a amostra tem fenologia *flor/botão*, distribuição *Mancha esparsa*, possui *26 capítulos* e o substrato é do tipo *Planta*. Essa planta pertence à familia *Asteraceae*, à tribu *Vernonieae*, ao gênero *Proteopsis* e ao epíteto específico *argentea*. A classificação taxonômica completa dessa espécie de planta, assim como o autor dessa classificação são ilustradas na tabela 3.4.

Tabela 3.2: Informação das Amostras coletadas

Tabela	Campo	Valor
SAMPLE	fieldNumber	...
	earliestDateCollected	16/02/1995
	latestDateCollected	16/02/1995
	notes	Alíquota migrada de PIC95350

Tabela 3.3: Informação do Substrato da Amostra - no banco de dados as especializações da entidade Substrato são tratadas em uma só tabela

Tabela	Campo	Valor
SUBSTRATE	organicMatter	...
	type	Planta
	sex	...
	plant_habit	Erva em roseta, escapo ca. 0,8 m
	phenology	flor/botão
	abundance	Comum
	spatialDistribution	Mancha esparsa
	herbariumDate	...
	weightG	...
	weightBagG	...
	rateG	9,69
	capitulumNumber	26
	estimatedCapitulum	40,95
	countCapitulum	...
	size	...
	notes	...

Tabela 3.4: Taxonomia da Planta, substrato da amostra (taxonomia retirada do ITIS [28])

Tabela	Campo	Valor
TAXONOMY	Kingdom	Plantae
	Phylum	Magnoliophyta
	Subphylum	...
	Class	Magnoliopsida
	SubClass	Asteridae
	Order	Asterales
	Suborder	...
	Superfamily	...
	Famiily	Asteraceae
	SubFamily	Cichorioideae
	SuperTribe	...
	Tribe	Vernonieae
	SubTribe	Lychnophrinae
	Genus	Proteopsis
	SpecificEpithet	argentea
	AuthorYearOfScientificName	Mart. & Zucc. Ex Sch. Bip.

Da amostra foram retirados 15 indivíduos, todos adultos. Essas e outras características do lote são descritas na tabela 3.5. Esse lote de insetos, detalhado na tabela 3.6, foi identificado como sendo da classe *Insecta*, da ordem *Lepidoptera*, de gênero *Adaina* e epíteto específico *bipunctata*, o autor é *Möschler* e o ano dessa classificação é *1890*.

A coleta dessa amostra foi realizada na latitude *17S41,64'* e longitude *44S11,51'*, dentro do município de *Joaquim Felício* na *Serra do Cabral*. Essas características são descritas nas tabelas 3.8 e 3.9 respectivamente. O habitat da amostra é *Encosta predregosa úmida*, como ilustrado na tabela 3.7. Finalmente, a tabela 3.10 apresenta informação de um dos pesquisadores envolvidos nessa coleta. No exemplo, refere-se ao pesquisador *Paulo I.K.L. Prado*, indicando o departamento e laboratório do qual faz parte: *Zoologia* e *Inseto-Plantas* respectivamente.

Tabela 3.5: Informação do lote de espécies retirado da amostra

Tabela	Campo	Valor
HOMOGENEOUS_SET	weight	...
	individualCount	15
	endemic	...
	size	...
	lifeStage	Adulto
	preparationType	...
	sex	...
	dateIdentified	...
	notes	...

Tabela 3.6: Taxonomia do lote de insetos (taxonomia retirada do ITIS [28])

Tabela	Campo	Valor
TAXONOMY	Kingdom	Animalia
	Phylum	Arthropoda
	Subphylum	...
	Class	Insecta
	SubClass	Dicondylia
	Order	Lepidoptera
	Suborder	...
	Superfamily	Pterophoroidea
	Famiily	Pterophoridae
	SubFamily	Pterophorinae
	SuperTribe	...
	Tribe	...
	SubTribe	...
	Genus	Adaina
	SpecificEpithet	bipunctata
	AuthorYearOfScientificName	Möschler 1890

Tabela 3.7: Informação do Habitat do local onde foi coletada a amostra

Tabela	Campo	Valor
HABITAT	name	Encosta predregosa úmida
	description	Microhabitat

Tabela 3.8: Informação da localização da amostra coletada

Tabela	Campo	Valor
LOCATION_COORD	wpt	esp1-20
	latitude	17S41,64'
	longitude	44S11,51'
	elevation	1001
	datum	...
	unidcons	...
	extension	...
	coordinatePrecision	...
	scale	...
	depth	...
	thepoint_lonlat	...
	dateLocName	...

Tabela 3.9: Informação geral relacionada à localização da amostra

Tabela	Campo	Valor
LOCATIONDETAILS	name	Serra do Cabral
	county	Joaquim Felício
	stateProvince	MG
	country	Brasil
	continentOcean	...
	reference	Prox. Matinha

Tabela 3.10: Informação dos pesquisadores que realizaram a coleta

Tabela	Campo	Valor
RESPONSIBLE	full_name	Paulo I.K.L. Prado
	abbreviated_name	...
	department	Zoologia
	laboratory	Inseto-Plantas
	CPF	...

No caso do Museu de Zoologia temos informação da espécie catalogada *Hemipholis elongata*. A entidade *Catalog* é representada na tabela 3.11. Ela armazena informação sobre o código da instituição, o código da coleção a que pertence essa espécie, o número do catálogo que foi atribuído ao lote de espécies, a via de conservação, número de campo que os biólogos atribuem a um espaço específico onde realizam a coleta e o tipo de substrato, entre outros dados. No exemplo, o registro catalogado pertence à coleção *Ophiuroidea*,

com número de catálogo *1*, a via de conservação é *Úmida*, o número de campo é *264* e o tipo de substrato é *Lama com areia fina*.

Tabela 3.11: Informação do catálogo

Tabela	Campo	Valor
CATALOG	institutionCode	ZUEC
	collectionCode	Ophiuroidea
	catalogNumber	1
	catalogDate	...
	dateIdentified	1997
	earliestDateCollected	07/1993
	latestDateCollected	07/1993
	fieldNumber	264
	conservationMeans	Úmida
	basisOfRecord	...
	previousCatalogNumber	...
	typeMaterial	...
	individualCount	1
	patternColor	...
	commonName	...
	lifeStage	Adulto
	sex	...
	size	...
	typeSubstrate	Lama com areia fina
	publication	...
	notes	Tese Mônica Angélica Varella Petti

Tabela 3.12: Informação da taxonomia da espécie (taxonomia retirada do ITIS [28])

Tabela	Campo	Valor
TAXONOMY	Kingdom	Animalia
	Phylum	Echinodermata
	Subphylum	Eleutherozoa
	Class	Ophiuroidea
	SubClass	...
	Order	Ophiurida
	Suborder	Gnathophiurina
	Superfamily	...
	Famiily	Ophiactidae
	SubFamily	...
	SuperTribe	...
	Tribe	...
	SubTribe	...
	Genus	Hemipholis
	SpecificEpithet	elongata
	AuthorYearOfScientificName	(Say, 1825)

A espécie catalogada, ilustrada na tabela 3.12, é da familia *Ophiactidae*, o gênero é *Hemipholis* e o epíteto específico é *elongata*. O local da coleta foi realizada no município de *Ubatuba*, nas coordenadas *23S22'* (latitude) e *44W53'* (longitude). As tabelas 3.13 e 3.14 apresentam os dados geográficos respectivamente. O habitat da coleta é *Bentônico Marinho*, descrita na tabela 3.15. As espécies catalogadas pertencem ao projeto *Estrutura e dinâmica da Macrofauna Bentônica*, ilustrada na tabela 3.16. A metodologia, tabela 3.17, usada para coletar as espécies é *van Veen* e o coletor responsável é *Mônica Angélica Varella Petti*, como ilustrado na tabela 3.18.

Tabela 3.13: Informação da localização da espécie catalogada

Tabela	Campo	Valor
LOCATION_COORD	wpt	...
	latitude	23S22'
	longitude	44W53'
	elevation	...
	datum	...
	unidcons	...
	extension	...
	coordinatePrecision	...
	scale	...
	depth	15
	thepoint_lonlat	...
	dateLocName	...

Tabela 3.14: Informação geral relacionada à localização

Tabela	Campo	Valor
LOCATIONDETAILS	name	...
	county	Ubatuba
	stateProvince	São Paulo
	country	Brasil
	continentOcean	...
	reference	...

Tabela 3.15: Informação do habitat da espécie catalogada

Tabela	Campo	Valor
HABITAT	name	Bentônico Marinho
	description	...

Tabela 3.16: Informação do projeto ao qual pertence a espécie catalogada

Tabela	Campo	Valor
PROJECT	name	Estrutura e dinâmica da Macrofauna Bentônica
	acronyms	..
	notes	...
	startDate	...
	endDate	...

Tabela 3.17: Informação da metodologia adotada para coletar espécie

Tabela	Campo	Valor
METHODOLOGY	name	van Veen
	description	...

Tabela 3.18: Informação do pesquisador que realizou a coleta

Tabela	Campo	Valor
RESPONSIBLE	full_name	Mônica Angélica Varella Petti
	abbreviated_name	...
	department	Museu de Zoologia
	laboratory	...
	CPF	...

3.5 Resumo

Este capítulo descreveu o modelo de banco de dados para o Repositório de Coletas, focando na descrição dos elementos que o compõem. O modelo descrito reflete uma visão dual para gerenciar dados de biodiversidade, sendo produto da análise realizada com os usuários. Por um lado, permite o registro de coletas realizadas em campo e por outro permite a catalogação de espécies.

Um dos desafios enfrentados na modelagem do banco de dados foi unificar os cenários de trabalho destes dois grupos de biólogos. Outro desafio foi o estudo e entendimento dos dados dos arquivos mantidos pelos usuários, resultado das coletas e observações em campo e catalogação de espécies. Assim, foi necessário entender o nome dos campos

e a informação que representam. Para oferecer ao leitor uma visão mais específica da abrangência do modelo, descreveram-se os cenários analisados para sua especificação.

Com o intuito de que o modelo fosse geral e permitisse no futuro o intercâmbio de informação entre diferentes grupos de pesquisa, o modelo foi baseado no padrão Darwin Core. Vários problemas adicionais de modelagem não foram abordados, como a evolução das hierarquias taxonômicas. Estes problemas serão deixados para trabalhos futuros.

Finalmente, o capítulo exemplifica algumas consultas que podem ser respondidas pelo modelo de banco de dados e apresenta exemplos que correspondem a dados a serem armazenados.

Capítulo 4

O Serviço de Coletas

Este capítulo descreve os elementos que integram o Serviço de Coletas. A Seção 4.1 apresenta uma breve descrição do serviço proposto. A Seção 4.2 apresenta a especificação do Serviço e a Seção 4.3 descreve os tipos de usuários que interagem com o serviço. A Seção 4.4 apresenta a descrição do seu funcionamento. Finalmente a Seção 4.5 apresenta o resumo do capítulo.

4.1 Visão Geral

O Serviço de Coletas é uma ferramenta baseada em serviços *Web* para consultar registros de biodiversidade. Similar às aplicações descritas na seção 2.1, mas com objetivos próprios das necessidades do projeto BIO-CORE, o serviço visa permitir um meio para realizar consultas ao repositório de coletas utilizando tecnologia *Web*. A especificação das funcionalidades deste serviço é resultado do estudo realizado sobre algumas aplicações pertencentes ao domínio de biodiversidade e de reuniões com nossos parceiros do projeto BIO-CORE.

A Figura 4.1 ilustra uma visão em alto nível da arquitetura proposta para o Serviço de Coletas. A arquitetura tenta suprir alguns requisitos próprios dos sistemas de biodiversidade e de outros sistemas em geral como: integração dos dados e adoção de padrões para o compartilhamento da informação. No nível inferior, o Repositório de Coletas, cujo modelo foi discutido no Capítulo 3, armazena os dados dos registros de coletas e acervos. O Serviço *Web* de Coletas encapsula operações que permitem a execução das solicitações feitas pelo usuário por meio de uma Aplicação Cliente.

As operações de atualização, remoção e inserção não foram especificadas como parte do Serviço de Coletas. Na dissertação, pressupõe-se que essas funcionalidades são fornecidas por um módulo independente aos usuários que possuam privilégios avançados de acesso aos dados. Uma possível extensão do trabalho é a especificação e implementação de um

módulo responsável pelo gerenciamento do Repositório de Coletas integrado ao serviço.

A implementação do Serviço de Coletas adotou a linguagem de programação Java e o ambiente de desenvolvimento *Eclipse 3.2*. Como a arquitetura proposta é baseada em serviços, houve a necessidade de utilizar um servidor *Web* e um *framework* para a implementação desses serviços.

O servidor *Web* utilizado foi o Apache Tomcat 6.0; o Axis 2 foi escolhido como ferramenta para desenvolver o serviço de Coletas. Dentre as principais características deste *framework* estão a implementação do protocolo SOAP (*Simple Object Acess Protocol*) e implementação de classes para agilizar a publicação e comunicação de serviços *Web*.

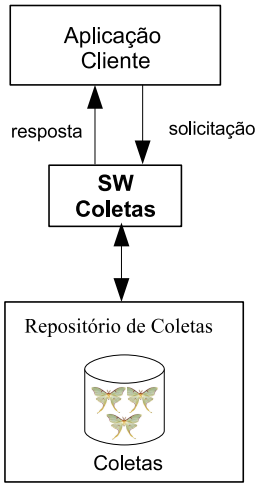


Figura 4.1: Arquitetura do Serviço de Coletas (inspirado em [18])

4.2 Especificação do Serviço

O Serviço de Coletas é composto por um conjunto de operações que permitem obter informações do Repositório de Coletas. As operações foram definidas baseadas no estudo das necessidades observadas dos usuários e no levantamento de informação realizados a alguns sistemas de biodiversidade ([11],[7],[12]). As funções especificadas tentam abranger o máximo de consultas que os biólogos possam solicitar. Solicitações mais complexas podem ser atentidas por composição das operações oferecidas.

A interface do Serviço provê operações que são divididas em i) operações de propósito geral e ii) operações específicas. No primeiro grupo as operações permitem atender solicitações sobre o banco de dados e o serviço. No segundo, as operações atendem solicitações diretas sobre os dados armazenados no banco de dados. Há também a possibilidade que um usuário expanda uma consulta com o objetivo de abranger mais resultados que mante-

nham a intenção do usuário. Neste caso, o serviço de Coletas invoca as operações providas pelo Serviço de Expansão de Consultas.

Esse serviço visa oferecer um meio para reescrever consultas feitas pelos usuários usando mecanismos de processamento e ontologias. Este trabalho é parte de um mestrado em andamento [55]. O Serviço de Expansão de Consulta define a operação **enviaSExp**, por meio da qual o serviço de Coletas envia a *string* da consulta para sua reescrita.

Na interface do serviço destacam-se 4 parâmetros: o parâmetro *flagE* das operações, indica se a operação deve requisitar expansão de consulta ou não; o parâmetro *nomeTaxon* indica o nome científico de uma espécie (em latim); o parâmetro *tipoContexto*, que estabelece a que contexto pertence a consulta admite apenas 1 de 3 valores válidos: *catalog, sample* e *homogeneous_set*. Por exemplo, se uma consulta se refere a informações de coletas feitas em campo, o contexto é *sample*; se a consulta envolve registros no catálogo, o contexto é *catalog*. O parâmetro *predicadoConj* fornece cláusulas de um predicado conjuntivo do tipo <atributo,operador,valor>, onde operador pode ser: =, <, >, <=, >= e <>. A seguir serão descritas as operações oferecidas pelo serviço.

4.2.1 Operações de propósito geral

- *getListaOperacoesServico()*: Lista todas as operações que fazem parte do serviço.
- *getVersaoServico()*: Obtém a versão do serviço.
- *getListaTabelas()*: Retorna todos os nomes das tabelas que pertencem ao banco de dados.
- *getListaAtributosTabela(nomeTabela)*: Retorna a lista dos nomes dos atributos e o seu significado (o tipo, se é chave primária ou chave estrangeira e o tamanho do campo) da tabela (*nomeTabela*).
- *getListaValoresAtributos(nomeTabela,listaNomesAtributos)*: Retorna os valores dos atributos da tabela (*nomeTabela*) que constam da lista *listaNomesAtributos*. O parâmetro *listaNomesAtributos* pode ser definido como "*", neste caso, retornando a tabela inteira.
- *getListaRegistrosFiltrados(nomeTabela,predicadoConj,flagE)*: Retorna os valores de todos os atributos da tabela (*nomeTabela*) que satisfazem ao predicado (*predicadoConj*).
- *executeQuery(strCon)*: Executa uma consulta escrita em SQL. O parâmetro corresponde ao *string* da consulta.

4.2.2 Operações específicas

- *getInformacoesEspecies(tipoContexto,nomeTaxon,flagE)*: Obtém toda a informacão armazenada sobre uma espécie no banco de dados de acordo com o contexto informado. Isso corresponde a obter dados sobre o que, onde, como, por quem e quando. A execução desta operação realiza uma busca nas tabelas que armazenam os dados sobre: informação taxonômica da espécie (*Taxonomy*)– "o que"; a localização geográfica de coleta associada à espécie (*Location_Coord* e *LocationDetails*)– "onde"; a metodologia da coleta (*Methodology*)– "como"; o responsável pela identificação e/ou coleta dessa espécie (*Responsible*)– "quem"; e as datas de coleta e identificação que são armazenadas nas tabelas centrais (*Sample, Catalog* e *Homogeneous_Set*).

  Por exemplo, deseja-se recuperar toda a informação relacionada à espécie *Adaina bipunctata* existente no catálogo. Para recuperar essas informações, a operação é invocada da seguinte maneira: getInformacoesEspecies('catalog','Adaina bipunctata',false). Neste caso, como o contexto informado é *catalog*, será realizada uma junção da tabela *Catalog* com as tabelas citadas anteriormente.

- *getListaNomesEspeciesRaio(tipoContexto,x,y,raio,flagE)*: Obtém os nomes das espécies localizadas dentro de um raio a partir de uma coordenada x,y, de acordo com um contexto específico. Os parâmetros correspondem à longitude *(x)*, à latitude *(y)* e ao raio *(raio)* que está definido em quilômetros.

- *getListaLocalizacaoEspecie(tipoContexto,nomeTaxon,flagE)*: Retorna os nomes das regiões onde há registros de coletas de uma determinada espécie de acordo com o contexto especificado.

- *getListaNomeCientificoEspecies(tipoContexto,predicadoConj,flagE)*: Retorna os nomes científicos das espécies a partir de um contexto específico (*tipoContexto*) e de um predicado (*predicadoConj*) que permita filtrar os resultados.

- *getListaNivelTaxonomicoInferior(nivelTaxon,valorTaxon,flagE)*: Retorna todos os descendentes que pertencem a um determinado taxon segundo a hierarquia taxonômica registrada. O parâmetro (*nivelTaxon*) indica o nome de um nível taxonômico e o parâmetro (*valorTaxon*) indica o valor atribuído a esse nível taxonômico. Por exemplo, deseja-se obter os níveis taxonômicos inferiores das espécies que pertencem à ordem *Lepidoptera*. A operação pode ser invocada da seguinte forma: getListaNivelTaxonomicoInferior('order','Lepidoptera',false). Estas informações estão armazenadas na tabela *Taxonomy*.

- *getListaTaxonomiaEspecie(nomeTaxon,flagE)*: Retorna a hierarquia taxonômica de uma espécie.

- *getListaNomeEspeciesRegiao(tipoContexto,nomeRegiao,predicadoEsp,flagE)*: Retorna os nomes das espécies e suas respectivas coordenadas geográficas de coletas em uma região (*nomeRegiao*), a partir da especificação de um predicado espacial (*predicadoEsp*) e que pertençam a um determinado contexto. No caso, o predicado espacial pode ser: *in*, isto é espécies coletadas em uma região; *adj*, que significa espécies coletadas em regiões adjacentes a uma região específica; *inexc*, que são as espécies coletadas somente em uma região, mas não em outras; e *notin*, aquelas espécies que não foram coletadas nessa região. O parâmetro (*nomeRegiao*) corresponde ao nome da região e o parâmetro (*predicadoEsp*) indica um predicado do tipo espacial.

- *getListaInteracoesEspecies(tipoContexto,nomeTaxon1,tipoInteracao,nomeTaxon2-,flagE)*: Retorna os nomes das espécies e as interações registradas entre elas de acordo com o contexto especificado (*tipoContexto*), a partir da combinação dos parâmetros passados. Caso o valor de um dos parâmetros seja "*" a operação irá considerar todos os valores do campo correspondente, porém, irá manter as restrições dos outros campos. Por exemplo, considerando que o contexto é *sample* temos: i) getListaInteracoesEspecies('sample',*, 'predação', 'B'): devolve todos os nomes de espécies que possuem a relação de predação com a espécie B coletada. ii) getListaInteracoesEspecies('sample','A', *, 'B'): devolve todos os tipos de interação entre as espécies A e B das coletas. iii) getListaInteracoesEspecies('sample',*, 'predação', *): devolve todos os pares de espécies que possuem predação como tipo de interação nas coletas.

- *getListaIncidenciaEspeciesTempo(tipoContexto,nomeRegiao,nomeTaxon,dataInicio,dataFim,flagE)*: Retorna a incidência (número) de espécies de nome *nomeTaxon* de uma determinada região em um intervalo de tempo específico de acordo com o contexto (*tipoContexto*). Os valores dos campos (*nomeRegiao*) e (*nomeTaxon*) podem ser definidos como "*", desde que não ocorram em ambos campos ao mesmo tempo. Por exemplo: i) getListaIncidenciaEspeciesTempo('homogeneous_set',*, 'A', 01/01/2001, 01/01/2002): devolve a incidência da espécie A dentro do intervalo de 01/01/2001 a 01/01/2002 ocorridas em todas as regiões registradas no banco de dados que pertencem ao contexto *homogeneous_set*. ii) getListaIncidenciaEspeciesTempo('catalog','R', *, 01/01/2001, 01/01/2002): devolve a incidência de todas as espécies catalogadas ocorridas em uma região R dentro do intervalo de tempo especificado registradas no banco de dados. Os parâmetros (*dataInicio*) e (*dataFim*) representam respectivamente a data inicial e a data final do período considerado.

4.3 Descrição dos tipos de usuários que interagem com o Serviço

O serviço fornece todas as operações necessárias para que um usuário recupere a informação do banco de dados ou para a criação de uma consulta quando não existir uma função específica que atenda à sua solicitação. São dois tipos de usuários que podem interagir com o serviço de Coletas: i) usuários com perfil técnico e ii) usuários leigos. O primeiro tipo inclui desenvolvedores de outros projetos de biodiversidade que utilizam o serviço para extrair informação do banco de dados. O segundo são usuários gerais que não estão familiarizados com linguagens de consulta, como SQL, e com estruturas de banco de dados. Por exemplo, biólogos, ecólogos e estudantes, que precisam consultar a informação para sua pesquisa.

Suponha um primeiro cenário em que um usuário com perfil técnico deseja realizar uma consulta para obter a classificação taxonômica das amostras coletadas em março de 2000. Neste caso o usuário deve possuir informação sobre as operações que o serviço fornece, as tabelas a serem consultadas e os campos a serem retornados. O procedimento que deve ser realizado é o descrito a seguir:

- Listar as operações que o serviço fornece: *getListaOperacoesServico()*;
- Listar o nome das tabelas do banco de dados: *getListaTabelas()*;
- Obtidos os nomes das tabelas o usuário deverá escolher o conjunto de tabelas a consultar. No caso, as tabelas necessárias são *SAMPLE* e *TAXONOMY*;
- A seguir, precisa-se obter informações sobre os campos dessas tabelas: *getListaAtributosTabela(SAMPLE)* e *getListaAtributosTabela(TAXONOMY)*;
- Para montar a consulta é preciso primeiro descobrir quais amostras foram feitas em março de 2000. Uma possível consulta em SQL seria:

```
SELECT * FROM ‘‘SAMPLE’’ s
WHERE (s).‘‘earliestDateCollected’’.‘‘month’’=03
AND (s).‘‘earliestDateCollected’’.‘‘year’’=2000
OR (s).‘‘latestDateCollected’’.‘‘month’’=03
AND (s).‘‘latestDateCollected’’.‘‘year’’=2000;
```

- Recuperando as amostras que foram realizadas nessa data o usuário precisa recuperar a informação taxonômica desse grupo de amostras. Para isso, usa-se a junção das tabelas *SAMPLE* e *TAXONOMY*. Dessa forma, a consulta ficaria:

```
SELECT * FROM ‘‘SAMPLE’’ s, ‘‘TAXONOMY’’ t
WHERE (s).‘‘earliestDateCollected’’.‘‘month’’=03
AND (s).‘‘earliestDateCollected’’.‘‘year’’=2000
OR (s).‘‘latestDateCollected’’.‘‘month’’=03
AND (s).‘‘latestDateCollected’’.‘‘year’’=2000
AND s.‘‘fktaxa’’ = t.‘‘idTaxa’’;
```

- Construído o SQL, o usuário pode invocar a função *executeQuery* e passar o *string* da consulta gerada por ele.

Um segundo cenário apresenta-se para o usuário padrão, aquele que não tem conhecimento para criar uma sentença SQL. Assim, precisa-se prover uma interface gráfica que possibilite a criação do SQL, abstraindo os detalhes técnicos para sua construção. O escopo deste trabalho não abrange o desenvolvimento desse tipo de recurso. Não entanto, há na literatura trabalhos que abordam a geração de consultas utilizando interfaces gráficas, como apresentado em [19, 34, 46].

Os únicos mecanismos providos pelo serviço são as operações que servirão de meios para a construção da consulta. Essas operações podem ser invocadas por ferramentas implementadas com esse propósito. O conjunto de operações descritas como sendo de propósito geral permitem que ferramentas utilizadas para a criação de consultas em SQL obtenham as informações necessárias para outras operações. Cabe ao desenvolvedor utilizar as operações de acordo com suas necessidades.

4.4 Descrição do Funcionamento do Serviço

O serviço de Coletas recebe diretamente as solicitações feitas pelos usuários e encarrega-se de acessar o repositório, extraindo somente a informação necessária para responder à consulta. Essa informação é obtida pela invocação de operações providas pelo serviço. A Figura 4.2 apresenta uma visão do serviço e os passos para processar uma solicitação de consulta.

Inicialmente uma aplicação Cliente envia a solicitação de consulta de um usuário ao Serviço de Coletas (1). Esta consulta pode ser ou não expandida, segundo indicação do usuário. Caso não seja expandida, o serviço disponibiliza uma conexão com o Repositório de Coletas, executa a consulta (2) e recebe um resultado (3), os dados são retornados à aplicação cliente (4). Caso contrário, o serviço de Coletas encaminha uma solicitação para o Serviço de Expansão de Consultas [55] que utiliza ontologias para reescrever a consulta (5). Essas ontologias são gerenciadas pelo Serviço de Ontologias “Aondê” [18] permitindo abranger a descrição de uma maior quantidade de termos de biodiversidade. A consulta

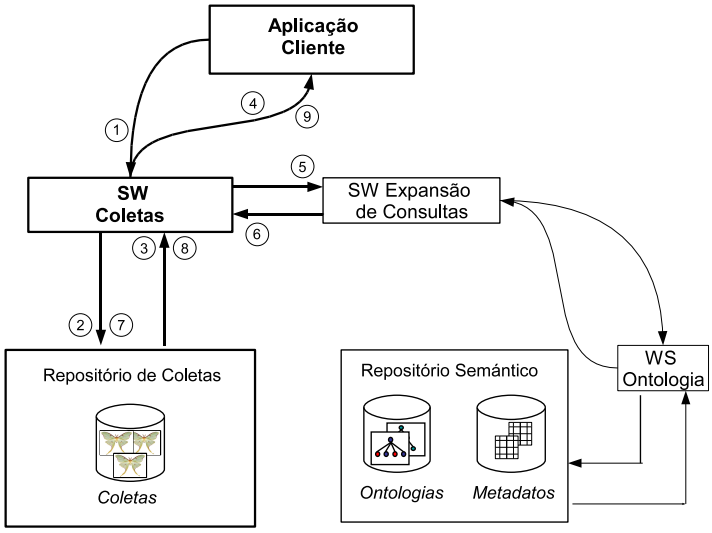


Figura 4.2: Processo de Execução do serviço de Coletas

reescrita é enviada de volta ao serviço de Coletas (6), que finalmente executa a consulta (7) e retorna o resultado ao cliente (8,9).

A Figura 4.3 apresenta em um diagrama de seqüência geral os passos para executar uma consulta sem considerar sua expansão. Uma consulta sem expansão é uma consulta básica escrita em SQL que é feita ao Repositório de Coletas. A aplicação cliente deve conhecer o esquema do banco de dados. Os únicos predicados aceitáveis são aqueles que envolvem campos do banco de dados.

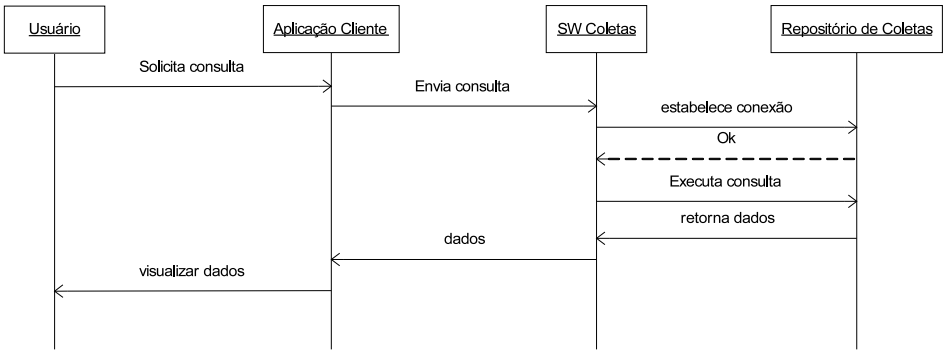


Figura 4.3: Seqüência de atividades para a execução de uma consulta sem expansão

Por exemplo, considere a seguinte consulta sem pedido de expansão: a) "Obter a informação completa da espécie *Hemipholis elongata* registrada no catálogo". Esta consulta pode ser resolvida por meio da operação *getInformacoesEspecies*. No caso, se deseja toda a informação da espécie *Hemipholis elongata* catalogadas. Assim, os valores dos parâmetros

da operação são: *tipoContexto* = 'catalog', *nomeTaxon = 'Hemipholis elongata'* e *flagE* = 'false'.

A seqüência de passos que o serviço realiza para executar essa consulta é ilustrada na Figura 4.4.

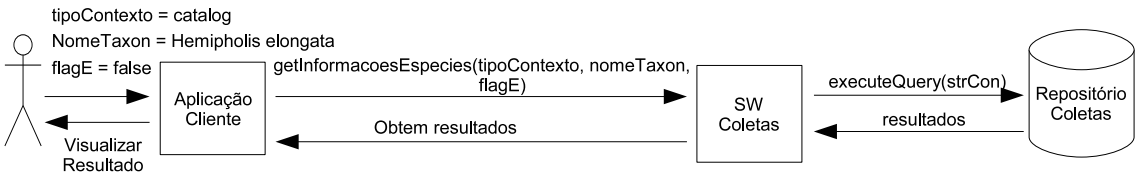


Figura 4.4: Exemplo de execução de uma consulta

O usuário solicita uma consulta e estabelece os parâmetros de seleção, de acordo com as opções providas na aplicação cliente. Além disso, o usuário precisa especificar se deseja expandir ou não a consulta. A aplicação cliente invoca a operação *getInformacoesEspecies* passando os valores atribuídos aos parâmetros. Essa operação cria a consulta em SQL. A *string* da consulta é passada à operação *executeQuery* para execução no banco de dados. Finalmente, os resultados são retornados ao usuário.

As Tabelas 4.1 e 4.2 mostram os resultados desta operação. As tabelas apresentam informação referente a o "que" (informação taxonômica), "quem" (identificador e coletor), "como" (metodologia), "onde" (localidade da coleta) e "quando" (data de identificação e da coleta). A coluna *Locality* está vazia porque essas coletas foram feitas mar adentro no município de Ubatuba. Note que a operação exigiu combinar dados das tabelas *Taxonomy*, *Location_Coord*, *LocationDetails*, *Responsible*, *Methodology* e *Catalog* do banco de dados subjacente.

Tabela 4.1: Resultado parcial da consulta (a)

Class	**Order**	**Family**	**Genus**	**Specific Epithet**	**Author-Year**	**Collector**	**Identifier**	**Methodology**
Ophiuroidea	Ophiurida	Ophiactidae	Hemipholis	elongata	(Say, 1825)	Mônica Angélica Varella Petti	Michela Borges	van Veen
Ophiuroidea	Ophiurida	Ophiactidae	Hemipholis	elongata	(Say, 1825)	Adilson Fransozo	Michela Borges	van Veen

Tabela 4.2: Continuação: Resultado parcial da consulta (a)

Latitude	**Longitu_de**	**Depth (m)**	**County**	**Locality**	**Habitat**	**Data Identificação**	**Data Coleta**
23.366667	44.883333	15	Ubatuba		Bentônico Marinho	1997	01/09/93
23.757778	45.231389	20	Caragua_tatuba		Bentônico Marinho - Infralitoral	2003	30/07/01

Considere um outro exemplo onde se deseja: b) "Recuperar os nomes de espécies para as quais existem amostras coletadas em um raio de 10 quilômetros da coordenada (X,Y)". A operação do serviço invocada é *getListaNomesEspeciesRaio*.

A Figura 4.5 ilustra o procedimento que o serviço de Coletas realiza para processar essa consulta. O usuário precisa especificar os valores para os parâmetros. No caso, *tipoContexto* refere-se a *sample*, as coordenadas latitude (y) é *23S23'19"* e longitude (x) é *44W49'58"*, o raio é 10 quilômetros, e também um argumento que define que a consulta não será expandida (*flagE*=false). A aplicação cliente envia esses parâmetros à operação *getListaNomesEspeciesRaio* que cria o SQL da consulta. A *string* da consulta é passada à operação *executeQuery* para execução no banco de dados. Após seu processamento os dados são retornados ao usuário. Esta consulta utiliza operadores espaciais, implementados como funções no PostGIS.

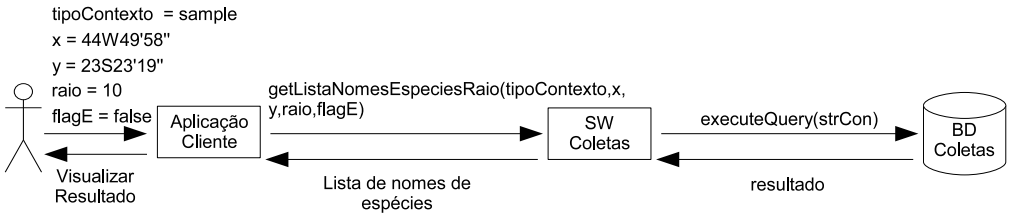


Figura 4.5: Exemplo de execução de uma consulta

O resultado desta consulta é mostrado na Tabela 4.3.

Tabela 4.3: Resultado parcial da consulta (b)

Espécie
Amphiodia atra
Amphiodia planispina
Amphiodia pulchella
Amphiodia riisei
Amphipholis januarii

A Figura 4.6 ilustra o procedimento geral para uma consulta com expansão, com invocação do Serviço de Expansão de Consultas. Após o serviço de Coletas receber a consulta com pedido de expansão, ele faz uma solicitação ao serviço de Expansão de Consultas. Este último reescreve a consulta e retorna a consulta expandida ao serviço de Coletas, que finalmente se encarrega de sua execução.

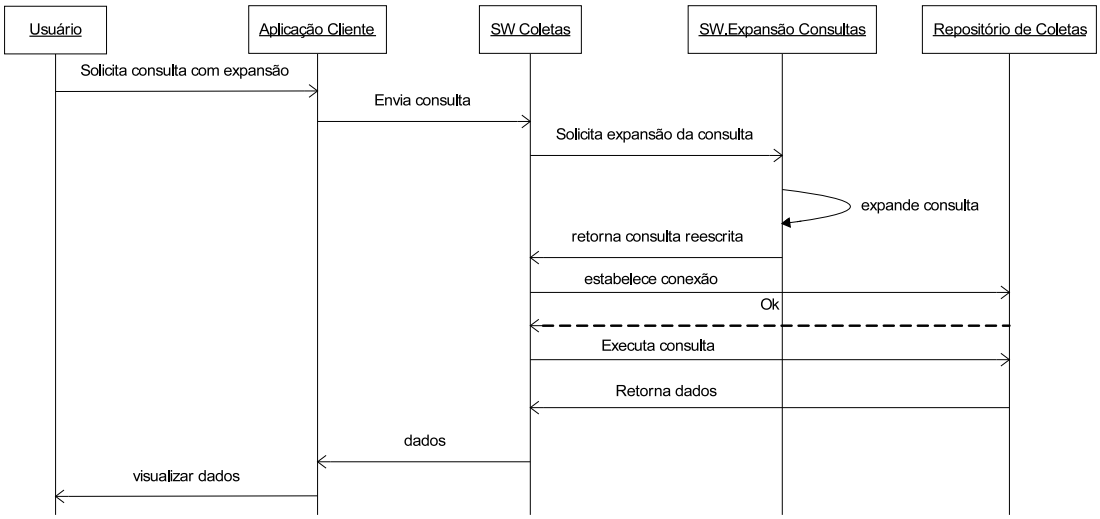


Figura 4.6: Diagrama de seqüência geral para a execução de uma consulta com expansão

Por exemplo, considere a seguinte consulta: "Retornar informação sobre a taxonomia dos insetos que pertencem à ordem *lepidoptera*". Para atender esta solicitação é preciso invocar as operações: *getListaTabelas* e *getListaRegistrosFiltrados*. A primeira operação é chamada para identificar o nome da tabela que pode ser consultada para obter as taxonomias dos insetos. A segunda permite criar o SQL e definir que se deseja expandir a consulta. Os valores dos parâmetros enviados para a operação *getListaRegistrosFiltrados* são: *nomeTabela*='Taxonomy', *predicadoConj*={Order='lepidoptera'} e *flagE*=true. A consulta em SQL criada é a seguinte:

```
SELECT * FROM ‘‘TAXONOMY’’ t
```

```
WHERE t.‘‘Order’’ = ’lepidoptera’
```

Como um usuário solicitou que a consulta seja expandida, o serviço invoca a operação *enviaSExp* permitindo passar ao serviço de Expansão de Consultas o *string* da consulta para sua expansão. Uma possível consulta reescrita e retornada para ser executada pelo Serviço de Coletas é a seguinte:

```
SELECT * FROM ‘‘TAXONOMY’’ t
WHERE t.‘‘Superfamily’’ in (’Gracillarioidea’, ’Hesperioidea’,
                            ’Micropterigoidea’, ’Papilionoidea’)
```

Neste caso *Gracillarioidea*, *Hesperioidea*, *Micropterigoidea*, e *Papilionoidea* são subclasses (ontológicas) da ordem *lepidoptera*. Estes exemplos foram baseados em [56].

A Figura 4.7 ilustra este procedimento. O usuário estabelece os valores dos parâmetros e especifica que deseja uma consulta com expansão. Por meio de uma aplicação cliente são invocadas as operações *getListaTabelas* e *getListaRegistrosFiltrados*. Como a consulta precisa ser expandida, o serviço de Coletas invoca a função *enviaSExp* do serviço de Expansão, passando como argumento o *string* da consulta a ser expandida. O serviço de Expansão de Consultas processa e retorna uma consulta reescrita ao serviço de Coletas, que utiliza a operação *executeQuery* para sua execução no banco de dados.

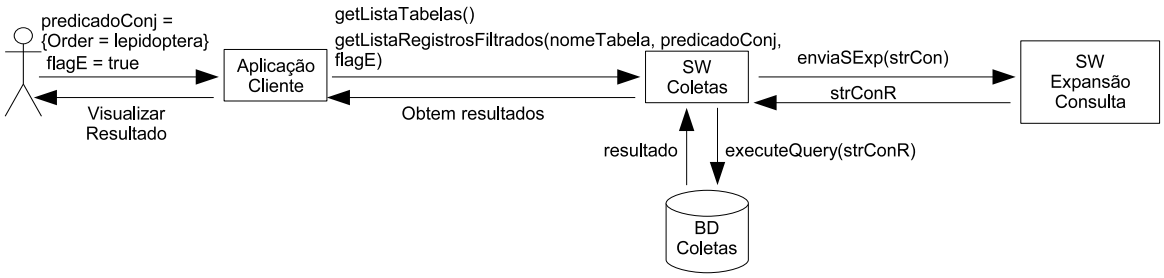


Figura 4.7: Exemplo de execução de uma consulta com pedido de expansão

4.5 Resumo

Este capítulo apresentou uma visão geral e a descrição das funcionalidades do serviço de coletas. O serviço visa fornecer um meio para realizar consultas ao repositório de coletas. O desenvolvimento do serviço usa tecnologias baseadas em serviços *Web* e ferramentas de código aberto. O capítulo ilustrou também alguns exemplos dos procedimentos que devem ser seguidos para execução de uma consulta. O próximo capítulo descreve os aspectos de implementação do repositório e do serviço de coletas.

Capítulo 5

Aspectos de Implementação

5.1 Implementação do Banco de Dados Proposto

Existem diversos SGBDs disponíveis como software livre, que oferecem bom desempenho, compatibilidade, portabilidade, entre outras características. Exemplos deles são MySQL [3] e o PostgreSQL [25]. Optamos pelo PostgreSQL. A escolha foi baseada principalmente nos projetos desenvolvidos no LIS [18, 30], que usam PostgreSQL para implementação do seus repositórios. Esses projetos estão relacionados a esta dissertação. Além disso, o PostgreSQL possui características espaciais implementadas em uma extensão conhecida como PostGIS, permitindo que o banco de dados suporte tipos de dados geométricos e operações sobre estes dados.

Para possibilitar a conexão ao banco de dados foi utilizado o conector postgresql-8.3-603.jdbc4 especificado para implementações em Java. Esse conector ou *driver* é uma classe que implementa as funcionalidades para manipulação das características do SGBD.

O primeiro passo para a criação do banco de dados foi o mapeamento do modelo E-R descrito na seção 3.2 para o modelo relacional. Houve necessidade de se criar campos geográficos para a tabela *Location_Coord*. Nesta tabela foi criado o campo *thepoint_lonlat* que é do tipo geométrico *POINT* para representar as coordenadas (latitude e longitude) da coleta.

A criação do tipo de dado espacial da tabela *Location_Coord* foi feita em duas etapas, por requisito do PostGIS [25]. Na primeira etapa construiu-se a tabela com os campos básicos. Na segunda, utilizou-se a função *Addgeometrycolumn* para adicionar o atributo espacial *thepoint_lonlat* nessa tabela. A sentença SQL para adicionar o tipo geométrio é a seguinte:

```
SELECT AddGeometryColumn( 'public', 'LOCATION_COORD',
       'thepoint_lonlat', 4291, 'POINT', 2 );
```

Houve necessidade de se implementar uma função (*converter*) para conversão de latitude e longitude, fornecidos pelos cientistas em formato *graus*, *minutos* e *segundos*, para o formato *graus decimais*. Após essa conversão os graus decimais são transformados no tipo geométrico *POINT*. Neste caso, utilizamos a função *PointFromText* que permite converter um dado numérico em um dado espacial. A sentença SQL gerada é a seguinte:

```
UPDATE LOCATION_COORD
SET thepoint_lonlat = PointFromText('POINT(' || converter(longitude)
                          || ' ' || converter(latitude) || ')',4291)
```

Para abranger um maior conjunto de consultas que possibilitem a realização de operações espaciais, adicionou-se uma tabela *GeomCounty* contendo a geometria dos municípios do estado de São Paulo. Com isso, foi possível realizar diversas consultas tomando como base as coordenadas das coletas (um ponto) e as localidades (um polígono) às quais pertencem essas coletas. A tabela *GeomCounty* foi relacionada à tabela *LOCATION_COORD*. Os arquivos dos municípios estão disponíveis em formato *shape* [1], sendo necessário transformá-los para um formato que o PostGIS aceite. Por isso, o arquivo *shape* foi transformado em um arquivo com estrutura relacional escrita utilizando SQL.

O ambiente de trabalho considerado nesta dissertação é centralizado, não requerendo a implementação de mecanismos para manter a integridade referencial em tabelas distribuídas, como descrito em [32]. Os dados dos biólogos, inicialmente em tabelas do *Microsoft Access*, foram migrados para o banco de dados proposto, utilizando um programa implementado em C++. Este processo de migração é parte de um trabalho de iniciação científica.

5.2 Implementação das Consultas

As consultas que foram implementadas surgiram das demandas dos parceiros do projeto BIO-CORE. Por meio das consultas pode-se determinar, por exemplo, quais são as localidades onde há maior incidência de um determinado grupo de espécies, obter as espécies catalogadas no museu, consultar as espécies que estão contidas dentro de uma região mas não em outras, entre outras consultas.

Dois pontos importantes devem ser considerados no caso da implementação das consultas utilizando predicados temporais. O primeiro é que seguindo a definição apresentada por [20] e [37] e utilizada em [35], os dados temporais são tratados como atributos alfanuméricos tradicionais. O segundo é que foi preciso tratar os dados temporais como um

[1] ftp://geoftp.ibge.gov.br/mapas/malhas_digitais/municipio_2007/

tipo composto [25], pois os registros de coletas eventualmente chegam com datas incompletas, seja para catalogação ou mesmo no caso dos projetos de coletas realizados em campo. A seguir apresentam-se algumas consultas típicas, divididas em simples e complexas, que foram implementadas.

5.2.1 Consultas simples

Considere a seguinte consulta: "Recuperar toda a informação da classificação taxonômica das espécies de um táxon X". Suponha que o táxon refere-se ao campo Família cujo valor é *Ophiactidae*. As operações do serviço que permitem obter o resultado desejado para esta consulta são: *getListaTabelas()*, que retorna os nomes das tabelas do banco de dados, e para retornar as tuplas desejadas se invoca a operação *getListaRegistrosFiltrados(nomeTabela,predicadoConj,flagE)*: onde *nomeTabela* é *Taxonomy*, *predicadoConj* é {Family=Ophiactidae} e *flagE* é false. A consulta SQL resultante da operação *getListaRegistrosFiltrados* é a seguinte:

```
SELECT t."Kingdom", t."Phylum", t."Subphylum",
       t."Class", t."SubClass", t."Order",
       t."Suborder", t."Superfamily", t."Family",
       t."SubFamily", t."SuperTribe", t."Tribe",
       t."Subtribe", t."Genus", t."SpecificEpithet",
       t."AuthorYearOFScientificName"
FROM "TAXONOMY" t
WHERE t."Family" = 'Ophiactidae'
```

As Tabelas 5.1 e 5.2 apresentam os resultados obtidos após a execução da consulta "Recuperar toda a informação da classificação taxonômica das espécies de um táxon X". Alguns campos estão vazios porque a informação dos usuários está em processo de migração.

Tabela 5.1: Resultado da consulta

Kingdom	Phylum	Subphy_ lum	Class	Sub_ Class	Order	Sub_ order	Superfa_ mily	Family
	Echino_ dermata		Ophiuroi_ dea		Ophiurida			Ophiac_ tidae
	Echino_ dermata		Ophiuroi_ dea		Ophiurida			Ophiac_ tidae
	Echino_ dermata		Ophiuroi_ dea		Ophiurida			Ophiac_ tidae
	Echino_ dermata		Ophiuroi_ dea		Ophiurida			Ophiac_ tidae
	Echino_ dermata		Ophiuroi_ dea		Ophiurida			Ophiac_ tidae

Tabela 5.2: Continuação: Resultado da consulta

SubFamily	SuperTribe	Tribe	Subtribe	Genus	Specific_ Epithet	AuthorYearOF_ ScientificName
				Hemipholis	elongata	(Say, 1825)
				Hempiholis	elongata	(Say, 1825)
				Ophiactis	lymani	Ljungman, 1872
				Ophiactis	savignyi	(Müller & Troschel, 1842)
				Ophiactis	brasiliensis	Manso, 1988

5.2.2 Consultas complexas

Considere a consulta envolvendo um predicado espacial: "Quais são as coordenadas das coletas de espécies catalogadas contidas na região de São Sebastião". No caso, o contexto da solicitação são as coletas (*sample*). A operação solicitada ao serviço que atende essa solicitação é *getListaNomeEspeciesRegiao(tipoContexto,nomeRegiao,predicadoEsp,flagE)*, onde *tipoContexto* é *sample*, *nomeRegiao* é São Sebastião, *predicadoEsp* é **in** e *flagE* é false. A consulta em SQL resultante da operação *getListaNomeEspeciesRegiao* é a seguinte:

```
SELECT distinct(t."Genus") || ''|| '  ' || ''||
       t."SpecificEpithet" as Especies,
       x(lc.thepoint_lonlat) as Longitude,
       y(lc.thepoint_lonlat) as Latitude
FROM "CATALOG" c, "TAXONOMY" t,
     "LOCATION_COORD" lc, "GeomCounty" m
```

```
WHERE within(lc."thepoint_lonlat", m."the_geom")='t'
      AND upper(m.county) = upper('São Sebastião')
      AND c."fk_location" = lc."id"
      AND c."fk_taxa" = t."idTaxa"
```

A Tabela 5.3 lista os resultados dessa consulta após sua execução no banco de dados.

Tabela 5.3: Resultado da consulta

Espécie	Longitude	Latitude
Amphipholis januarii	-45.6752777777778	-23.7741666666667
Amphipholis squamata	-45.6752777777778	-23.7741666666667
Amphipholis squamata	-45.6644444444444	-23.7808333333333
Ophiactis lymani	-45.6752777777778	-23.7741666666667
Ophiactis savignyi	-45.6752777777778	-23.7741666666667
Ophiactis savignyi	-45.6647222222222	-23.7811111111111
Ophiactis savignyi	-45.6644444444444	-23.7808333333333

A consulta avalia se existem pontos de coletas que estão dentro do polígono que pertence ao município de São Sebastião. Caso isso aconteça, retorna as coordenadas (Latitude/Longitude) específicas das coletas e os nomes de espécies que foram coletadas nesses pontos. A avaliação dos pontos que estão dentro do município de São Sebastião é feito por meio da função *within*. Essa função retorna verdadeiro se há uma geometria que está contida em outra.

Considere uma consulta envolvendo predicados temporais: "Quais são os nomes das espécies que foram coletadas na mesma data inicial que a amostra S". Considere que S faz referência à amostra de identificador 10. No caso, o serviço não oferece uma operação específica para resolver esta consulta. Porém, é possível que o usuário crie o *string* em SQL para a sua execução, como descrito na Seção 4.3. Supondo que o usuário tenha um perfil técnico, o conjunto de operações que permitem atender esta solicitação são:

- getListaTabelas(), retorna os nomes das tabelas;
- getListaAtributosTabela('SAMPLE'), retorna as características dos atributos da tabela *SAMPLE*;
- getListaAtributosTabela('TAXONOMY'), retorna as características dos atributos da tabela *TAXONOMY*;
- executeQuery(strCon), executa a consulta. Um possível valor para *strCon* é:

```
SELECT distinct(t."Genus") || ''|| '  ' || ''||
       t."SpecificEpithet" as "Especie"
FROM "SAMPLE" s, "TAXONOMY" t
WHERE (s)."earliestDateCollected"."day" in
            (SELECT (a)."earliestDateCollected"."day"
             FROM "SAMPLE" a
             WHERE a."idsample" = 10) AND
      (s)."earliestDateCollected"."month" in
            (SELECT (b)."earliestDateCollected"."month"
             FROM "SAMPLE" b
             WHERE b."idsample" = 10) AND
      (s)."earliestDateCollected"."year" in
            (SELECT (c)."earliestDateCollected"."year"
             FROM "SAMPLE" c
             WHERE c."idsample" = 10) AND
      s."idsample" <> 10 AND
      s."fk_taxa" = t."idTaxa"
```

A Tabela 5.4 apresenta os resultados obtidos após a execução dessa consulta.

Tabela 5.4: Resultado da consulta

Espécie
Amphiodia atra
Amphiodia pulchella
Amphiodia riisei
Amphipholis subtilis
Hemipholis elongata
Ophiophragmus lutkeni

Esta consulta retorna os nomes de espécies que foram coletadas em uma mesma data que a amostra cujo identificador é 10. A consulta implementa subconsultas para recuperar o dia, O mês e o ano da coleta dessa amostra. Já a consulta principal avalia se existem coletas feitas nessa data.

Um exemplo envolvendo predicados espaciais e temporais é: "Quais são os nomes das espécies coletadas nas regiões adjacentes ao município de Caraguatatuba durante o mês de Agosto de 2002". O serviço não fornece uma operação específica para atender esta consulta. Porém, um usuário mais técnico poderia construir o *string* da consulta combinando um conjunto de operações. A sequência de operações que seriam necessárias para resolver esta solicitação são:

- getListaTabelas(), retorna os nomes das tabelas;
- getListaAtributosTabela('TAXONOMY'), retorna as características dos campos da tabela *TAXONOMY*;
- getListaAtributosTabela('SAMPLE'), retorna as características dos campos da tabela *SAMPLE*;
- getListaAtributosTabela('LOCATION_COORD'), retorna as características dos campos da tabela *LOCATION_COORD*;
- getListaAtributosTabela('GeomCounty'), retorna as características dos campos da tabela *GeomCounty*;
- executeQuery(strCon), executa a consulta. Uma possível consulta em SQL é:

```
SELECT distinct(t."Genus") || ''|| '  ' || ''||
       t."SpecificEpithet" as "Especie",
       lc.latitude, lc.longitude
FROM "LOCATION_COORD" lc, "SAMPLE" s, "TAXONOMY" t
WHERE lc.fk_gid in (SELECT t2.gid
              FROM "GeomCounty" t1, "GeomCounty" t2
              WHERE TOUCHES(t1.the_geom, t2.the_geom)='t'
              AND upper(t1.county) = upper('Caraguatatuba'))
     AND lc."id" = s."fk_location"
     AND s."fk_taxa" = t."idTaxa"
     AND ("s")."earliestDateCollected"."month" = 8
     AND ("s")."earliestDateCollected"."year" = 2002
```

A Tabela 5.5 apresenta os resultados após a consulta ser executada.

Tabela 5.5: Resultado da consulta

Espécie
Amphiodia atra
Amphiodia pulchella
Amphipholis januarii
Hemipholis elongata
Ophiactis lymani
Ophiophragmus lutkeni

No caso, esta consulta é executada em dois passos. O primeiro seleciona os identificadores que pertencem aos municípios adjacentes ao município de Caraguatatuba. A relação de vizinhança entre duas geometrias é feita por meio da função *touch*. O segundo seleciona as espécies que foram coletadas nesses municípios a partir dos identificadores obtidos na subconsulta, em Agosto de 2002.

Em todo este grupo de consultas supõe-se que a invocação das operações do serviço é feita por um usuário que possui um perfil técnico. Usuários leigos, por exemplo biólogos, precisam de uma interface que permita gerar as consultas em SQL. As consultas foram feitas tendo como base os registros da coleção das *Ophiuroidea* fornecidas pelos biólogos do Museu de Zoologia.

5.3 Implementação do Serviço

Um protótipo do serviço *Web* de coletas foi implementado utilizando a linguagem Java. Foi utilizado o *framework* Apache Axis para a implementação do serviço de Coletas. Para ativação e disponibilização do serviço utilizamos o servidor Apache Tomcat.

As operações do serviço foram agrupadas em operações de propósito geral (*WSGeneral*) e operações de propósito específico (*WSCollectionRepository*), cuja especificação foi apresentada na Seção 4.2. Isso permitiu deixar em evidência quais serviços são voltados à obtenção da informação sobre o próprio serviço e o banco de dados e aqueles que são voltados especificamente para o contexto do Repositório de Coletas. A Figura 5.1 ilustra a estrutura utilizada para a implementação do serviço.

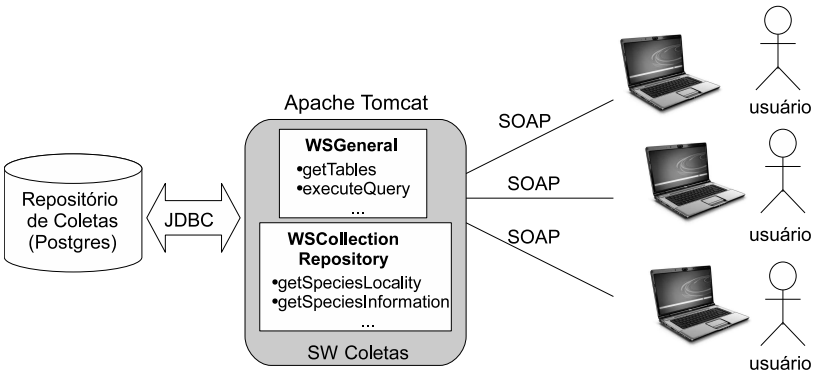


Figura 5.1: Protótipo do serviço de Coletas

A Tabela 5.6 apresenta os métodos disponíveis para realizar ações de propósito geral.

Tabela 5.6: Métodos de propósito geral do serviço de Coletas

Métodos	**Parâmetros**	**Retorno**
getServiceOperations		String[] operationNames
getServiceVersion		String serviceVersion
getTables		String[] tablesNames
getTableAttributeFeatures	String tableName	String[][] attributeFeatures
getAttributeValues	String tableName, String[] attributeNames	String[][] attributeValues
executeQuery	String sqlQuery	String[][] resultSet
getFilteredRecords	String tableName,	String[][] resultSet
	String[][] attributeOpeValues,	
	Boolean flagExpansion	

A Tabela 5.7 apresenta os métodos disponíveis para atender ações específicas sobre os dados do repositório.

Tabela 5.7: Métodos de propósito específico do serviço de Coletas

Métodos	**Parâmetros**	**Retorno**
getSpeciesInformation	String typeContext,String taxonName,	String[][] resultSet
	Boolean flagExpansion	
getSpeciesNameRadius	String typeContext, String x, String y,	String[] speciesName
	Double r, Boolean flagExpansion	
getSpeciesLocality	String typeContext,	String[] localityNames
	String taxonName,	
	Boolean flagExpansion	
getScientificName	String typeContext,	String[] scientificNames
	String[][] attributeOpeValues,	
	Boolean flagExpansion	
getDownLevelTaxon	String levelTaxon, String valueTaxon,	String[][] taxon
	Boolean flagExpansion	
getTaxonomicHierarchy	String taxonName,	String[][] taxonomicHierarchy
	Boolean flagExpansion	
getSpeciesNameLocality	String typeContext,	String[][] resultSet
	String localityName,	
	String spatialOp,	
	Boolean flagExpansion	
getRelationshipSpecies	String typeContext,	String[][] resultSet
	String taxonName1,	
	String RelationshipType,	
	String taxonName2,	
	Boolean flagExpansion	
getSpeciesIncidence	String typeContext,	String[][] resultSet
	String localityName,	
	String taxonName, Date startDate,	
	Date endDate,Boolean flagExpansion	

5.4 Resumo

No decorrer deste capítulo foram apresentados os aspectos de implementação referentes ao Repositório de Coletas e ao protótipo desenvolvido para o Serviço *Web* de Coletas proposto na dissertação. Alguns exemplos de consultas com invocação às operações do serviço também foram descritas. O próximo capítulo apresenta as conclusões e as possíveis extensões do trabalho.

Capítulo 6

Conclusões e Extensões

Esta dissertação apresentou a especificação e a implementação de um banco de dados (Repositório de Coletas) e de um Serviço *Web* para permitir a execução de consultas no repositório. Tanto o repositório quanto o serviço são elementos que compõem o projeto BIO-CORE, um Sistema de Biodiversidade que está sendo desenvolvido em parceria entre pesquisadores do Instituto de Biologia e do Instituto de Computação da UNICAMP.

O Repositório de Coletas proposto suporta informação de registros de coletas e de observações feitas em campo e registros de catálogos e acervos de museus. Isso possibilita integrar a informação proveniente das pesquisas realizadas em projetos de biodiversidade e de acervos. A pesquisa feita até agora nos leva a acreditar que não existe um modelo de banco de dados semelhante ao nosso, abrangendo registros de coletas em campo e registros para catalogar espécies no museu. De fato, quando provêm ambas funcionalidades, os sistemas de biodiversidade invocam módulos independentes que funcionam com bancos de dados separados.

O Repositório de Coletas tem suporte para dados espaciais e características temporais são tratados nesta dissertação como atributos alfanuméricos tradicionais. O Serviço de Coletas visa apoiar a recuperação de informação por meio de consultas. A implementação desse serviço foi baseada em tecnologia *Web* e em aplicativos de código aberto disponíveis na Internet.

O único módulo disponibilizado pelo serviço é o de consultas, sendo que as funcionalidades para o gerenciamento do Repositório de Coletas foram deixadas para uma implementação futura. A abordagem utilizada no trabalho foi considerar um repositório centralizado de dados.

Para a especificação do Serviço e do Repositório de Coletas foi preciso estudar o ambiente de trabalho dos usuários alvo do projeto BIO-CORE, definindo o tipo de informação que eles precisam armazenar e conseqüentemente as consultas que eles precisam realizar. Além disso, foi preciso realizar um levantamento de aspectos e de características que os

sistemas de informação de biodiversidade possuem. As tabelas do banco de dados foram preenchidas com dados fornecidos pelos biólogos.

Uma das dificuldades mais importantes foi a integração de cenários de dois grupos de pesquisa diferentes. Um grupo é formado por ecólogos que trabalham com projetos de biodiversidade e estão interessados nas interações entre espécies, no caso entre insetos e plantas. O outro grupo é de biólogos marinhos que formam parte da equipe que coordena o Museu de Zoologia e estão encarregados da catalogação e preservação de espécies dentro das coleções no museu. Isso significa que houve necessidade de identificar claramente quais são as características comuns a ambos grupos e quais os diferenciam. A partir dessa identificação, projetou-se um esquema de banco de dados que fosse o mais abrangente possível.

Um outro desafio foi projetar um banco de dados que no futuro permitisse a interação com outros sistemas pertencentes a este domínio. Isso levou à adoção de um padrão de metadados. O padrão escolhido foi o Darwin Core (versão 1.2), muito utilizado por projetos de sistemas de biodiversidade e pelos projetos relacionados a esta dissertação. Embora exista uma definição central de elementos do Darwin Core, projetos como OBIS-SEAMAP o adaptaram às suas necessidades. A conseqüência é que há diferentes versões desse padrão.

Para a implementação das consultas espaciais houve necessidade de adicionar campos do tipo geométrico ao repositório, os quais inicialmente não haviam sido considerados. Assim, importou-se um arquivo em formato *shape* que contém as geometrias dos municípios do estado de São Paulo ao banco de dados. Utilizando ferramentas do PostGIS, esse *shape* foi transformado em um arquivo escrito em SQL permitindo realizar operações espaciais sobre esses dados. As coordenadas geográficas das coletas foram transformadas ao tipo geométrico *POINT*.

As principais contribuições da dissertação são:

- Levantamento das características de diversos sistemas de informação de biodiversidade disponíveis, unificando e reaproveitando algumas das principais características, de forma a integrar sistemas e dados de diferentes propósitos;

- Estudo e organização da informação fornecida pelos usuários-alvo, parceiros do projeto *BIO-CORE*;

- Especificação e implementação do Repositório de Coletas para armazenamento de informação de biodiversidade, considerando a manipulação integrada dos vários tipos de registros manipulados por sistemas de biodiversidade. Esta integração não ocorre em sistemas similares e permite a reutilização de informação provinda de várias fontes. Não só isto permite integrar pesquisas, mas facilita a identificação e correção de erros de digitação, nomenclatura e outros;

- Especificação e implementação parcial do Serviço *Web* de Coletas para realizar consultas ao repositório, considerando exemplos reais fornecidos pelos biólogos parceiros do projeto.

As extensões desta dissertação podem ser tanto de pesquisa quanto práticas. Algumas possíveis direções futuras são listadas a seguir:

- Tratamento e integração de dados de sensores coletados no meio ambiente, por exemplo temperatura, no Repositório de Coletas;
- Estudo de técnicas que permitam a representação da evolução das hierarquias taxonômicas das espécies;
- Implementação de um módulo para permitir interação com o Serviço de Expansão de Consultas;
- Estudo das vantagens de armazenamento de dados de biodiversidade em estruturas *XML* versus bancos de dados relacionais;
- Implementação de uma interface de consulta na qual os usuários possam definir diretamente as consultas que desejam realizar;
- Especificação e implementação de uma interface amigável ao usuário que permita por exemplo, a representação gráfica das geometrias junto com informação textual;
- Melhoria das consultas implementadas utilizando técnicas de otimização de consultas;
- Implementação de um módulo para utilizar o protocolo TAPIR, muito usado em projetos de biodiversidade, como mecanismo para transferência de informação;
- Estabelecimento de regras básicas para realizar processos de recuperação, melhorias da performance e refinamento do banco de dados;
- Incorporação do Repositório e Serviço de Coletas ao sistema BioCore.

Referências Bibliográficas

[1] Webios. Web Service Multimodal Tools for Strategic Biodiversity Research, Assessment and Monitoring. http://www.lis.ic.unicamp.br/projects/webios, 2005. (accessed May 20, 2008).

[2] Bio-core. tools, models and techniques to support research in biodiversity. http://www.lis.ic.unicamp.br/projects/bio-core/, 2008. (accessed November 20, 2008).

[3] MySQL AB. Mysql. http://www.mysql.com/, 1995. (accessed November 11, 2008).

[4] G. Alonso, F. Casati, H. Kuno, and V. Machiraju. *Web Services: Concepts, Architectures and Applications*. Springer, 2004.

[5] G. Aslan and D. McLeod. Semantic heterogeneity resolution in federated databases by metadata implantation and stepwise evolution. *The VLDB Journal*, 8:120–132, 1999.

[6] C. Batini, M. Lenzerini, and S. B. Navathe. A comparative analysis of methodologies for database schema integration. *ACM Comput. Surv.*, 18(4):323–364, 1986.

[7] J. Beach. Specify Biodiversity Collections Software. http://www.specifysoftware.org/Specify/, 2007. (accessed July 30, 2008).

[8] F.A. Bisby, Y.R. Roskov, M.A. Ruggiero, T.M. Orrell, L.E. Paglinawan, P.W. Brewer, N. Bailly, and J. van Hertum. Catalogue of Life. www.catalogueoflife.org/annual-checklist/2007/, 2007. (accessed September 22, 2008).

[9] G. Blair, L. Blair, V. Issarny, P. Tuma, and A. Zarras. The role of software architecture in constraining adaptation incomponent-based middleware platforms. In *Middleware '00: IFIP/ACM International Conference on Distributed systems platforms*, pages 164–184, Secaucus, NJ, USA, 2000. Springer-Verlag New York, Inc.

[10] S. Castano, V. De Antonellis, and S. di Vimercati. Global Viewing of Heterogeneous Data Sources. *IEEE Trans. on Knowl. and Data Eng.*, 13(2):277–297, 2001.

[11] Centro de Referência em Informação Ambiental (Cria). specieslink. http://splink.cria.org.br, 2001. (accessed July 5, 2008).

[12] R.K. Colwell. *Biota. The Biodiversity Database Manager*. Sinauer Associates, 1996.

[13] The Federal Geographic Data Committee. Federal Geographic Data Committee (FGDC). http://www.fgdc.gov, 2002. (accessed July 30, 2008).

[14] Open Geospatial Consortium. Geography Markup Language. http://www.opengeospatial.org/standards/gml.

[15] A. da S. Fagundes. Projeto e Implementação de um Banco de Metadados para o Sistema de Informação de Biodiversidade do Estado de São Paulo. Master's thesis, Instituto de Computação - Unicamp, 1999.

[16] R. da Silva Torres. *Ambiente de Gerenciamento de Imagens e Dados Espaciais para Desenvolvimento de Aplicações em Biodiversidade*. PhD thesis, Instituto de Computação - Unicamp, 2004.

[17] R. da Silva Torres, C.B. Medeiros, M.A. Gonçalves, and E.A. Fox. A Digital Library Framework for Biodiversity Information Systems. *International Journal on Digital Libraries*, 6(1):3 – 17, February 2006.

[18] J. Daltio. Aondê: Um Serviço Web de Ontologias para Interoperabilidade em Sistemas de Biodiversidade. Master's thesis, Instituto de Computação - Unicamp, 2007.

[19] S. A. de Souza, E. L. de Campos, and A. D. Dos Santos. Uma Ferramenta para a Definição de Consultas Baseada em Entidades e Papéis. In *IEEE Latin America Transactions*, volume 4, June 2006.

[20] G. Faria. Um Banco de Dados Espaço-Temporal para Desenvolvimento de Aplicações em Sistemas de Informação Geográfica. Master's thesis, Instituto de Computação - Unicamp, Março 1998.

[21] The Biological Collection Access Service for Europe. Biocase. http://www.biocase.org/index.shtml. (accessed July 30, 2008).

[22] A. Frondorf, M. Jones, and S. Stitt. Linking the FGDC Geospatial Metadata Content Standard to the Biological/Ecological Sciences. *Proceedings of the Third IEEE Computer Society Metadata Conference. IEEE. Betheeda, MD*, 1999.

[23] GBIF. Global Biodiversity Information Facility. URL: http://www.gbif.org, 2004. (accessed July 30, 2008).

[24] J. Greenherg. Metadata and the World Wide Web. *Encyclopedia of Library and Information Science*, 2003.

[25] PostgreSQL Global Development Group. Postgresql. http://www.postgresql.org/, 1996. (accessed July 30, 2008).

[26] R. Guralnick and D. Neufeld. Challenges Building Online GIS Services to Support Global Biodiversity Mapping and Analysis: Lessons from the Mountain and Plains Database and Informatics project. *Biodiversity Informatics*, 2:56–69, 2005.

[27] P.N. Halpin, A.J. Read, B.D. Best, K.D. Hyrenbach, E. Fujioka, M.S. Coyne, L.B. Crowder, S.A. Freeman, and C. Spoerri. OBIS-SEAMAP: developing a biogeographic research data commons for the ecological studies of marine mammals, seabirds, and sea turtles. *Marine Ecology Progress Series*, 316:239–246, 2006.

[28] ITIS. Integrated Taxonomic Information System. http://www.itis.gov/, 2007. (accessed July 30, 2008).

[29] M. Jones, C. Berkley, J. Bojilova, and M. Schildhauer. Managing Scientific Metadata. *IEEE Internet Computing*, 5(5):59–68, 2001.

[30] L.C. Gomes Jr. Uma Arquitetura para Consultas a Repositórios de Biodiversidade na Web. Master's thesis, Instituto de Computação - Unicamp, May 2007.

[31] Kristian Ellebaek Kjaer. A survey of context-aware middleware. In *SE'07: Proceedings of the 25th conference on IASTED International Multi-Conference*, pages 148–155, Anaheim, CA, USA, 2007. ACTA Press.

[32] A.A. Kondo. Gerenciamento de Rastreabilidade em Cadeias Produtivas Agropecuárias. Master's thesis, Instituto de Computação - Unicamp, Abril 2007.

[33] D.R. Maddison and K.S. Schulz. The Tree of Life Web Project. *Zootaxa*, 1668, 2007.

[34] D. Mark, K. John, and R. S. F. An interactive visual query environment for exploring data. In *UIST '97: Proceedings of the 10th annual ACM symposium on User interface software and technology*, pages 189–198, New York, NY, USA, 1997. ACM.

[35] S. Matias. Processamento de Consultas ao Banco de Dados de Biodiversidade do BIOTA. Master's thesis, Instituto de Computação - Unicamp, 2000.

[36] P. McCartney and M. Jones. Using XML-encoded Metadata as a Basis for Advanced Information Systems for Ecological Research. *Proc. 6th World Multiconference Systemics, Cybernetics and Informatics*, 7:379–384, 2002.

[37] C.B. Medeiros and M. Botelho. Tratamento do Tempo em SIG. In *In Proceedings of GIS Brazil '96*, Curitiba, PR, Brazil, Maio 1996. (in portuguese).

[38] P.J. Morin. *Community ecology*. Wiley-Blackwell, 1999.

[39] R.A. Morris, R.D. Stevenson, and W. Haber. An architecture for electronic field guides. *J. Intell. Inf. Syst.*, 29(1):97–110, 2007.

[40] NBII. National biological information infrastructure (nbii). http://www.nbii.gov/portal/server.pt, 1993. (accessed May 20, 2008).

[41] OGC. Open Geospatial Consortium (OGC). http://www.opengeospatial.org/, 2008. (accessed July 10, 2008).

[42] J. Martinez Perdigueiro and A. Donalisio. Relatório de Iniciação Científica para modelagem de um banco de dados para o laboratório Inseto-Plantas. Technical report, Instituto de Computação - UNICAMP, 2006.

[43] N. Press. Understanding metadata. Technical report, National Information Standards, 2004.

[44] Species 2000 project. Species 2000. http://www.sp2000.org/, 2008. (accessed September 22, 2008).

[45] M.P Reddy. A methodology for integration of heterogeneous databases. *IEEE Trans. on Knowl. and Data Eng.*, 6(6):920–933, 1994.

[46] L. B. Schmitz. Construção de um gerador gráfico de consultas SQL via Web utilizando a plataforma .NET. Technical report, Centro Universitário Luterano de Palmas, 2004.

[47] L.A. Seffino. WOODSS - Spatial Decision Support System based on Workflows. Master's thesis, Instituto de Computação - Unicamp, Julio 1998.

[48] K.T. Shao, C.I. Peng, E. Yen, K.C. Lai, M.C. Wang, J. Lin, H. Lee, Y. Alan, and S.Y. Chen. Integration of biodiversity databases in taiwan and linkage to global databases. *Data Science Journal*, pages 2–10, 2007.

[49] SinBiota. São Paulo Biodiversity System. http://sinbiota.cria.org.br/, 2001. (access Oct, 2008).

[50] J. Soberón and T. Peterson. Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions: Biological Sciences*, 359(1444):689–698, 2004.

[51] SourceForge.NET. Distributed Generic Information Retrieval (DiGIR). http://sourceforge.net/projects/digir, 1999. (accessed February 3, 2008).

[52] Taxonomic Databases Working Group (TDWG). Access Biological Colections Data (ABCD). http://bgbm3.bgbm.fu-berlin.de/TDWG/acc/, 2007. (accessed July 30, 2008).

[53] Taxonomic Databases Working Group (TDWG). Biodiversity Information Standards. http://www.tdwg.org/, 2007. (accessed June 20, 2008).

[54] Taxonomic Databases Working Group (TDWG). DarwinCore. http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome, 2007. (accessed July 30, 2008).

[55] B. Vilar. Processamento Semântico de Consultas de Biodiversidade usando Ontologias. Master's thesis, Instituto de Computação - Unicamp.

[56] B. Vilar, J.G. Malaverri, and C.B. Medeiros. A Tool Based on Web Services to Query Biodiversity Information. In *5th International Conference on Web Information Systems and Technologies (Webist 2009)*, March 2009.

[57] W3C. HTTP - Hypertext Transfer Protocol. http://www.w3.org/Protocols/, 2003. (accessed February 3, 2008).

[58] W3C. SOAP specifications. http://www.w3.org/TR/soap/, 2003. (accessed February 3, 2008).

Apêndice A

Dicionário de Dados

Tabela	Descrição	Campo	Tipo	Descrição
SAMPLE	Representa as amostras que foram coletadas em uma determinada localidade de uma viagem especifica que realizam os coletores. No caso do Laboratório Inseto-Planta as amostras representam plantas. No caso do museu de zoologia representam um conjunto de espécies.	idsample	integer	Chave primária
		fieldNumber	character varying(20)	Número de Registro de Campo assinado pelos biólogos
		earliestDateCollected	Composite Types	Data de início da coleta
		latestDateCollected	Composite Types	Data final da coleta. Se a coleta for realizada no mesmo dia colocar neste campo o valor de earliestDateCollected.
		notes	character varying(100)	Algumas anotações que os biólogos desejam realizar sobre essa amostra
		fk_idcollector	integer	Chave estrangeira de Responsible
		fk_location	integer	Chave estrangeira de Location_Coord
		fk_codenmcon	integer	Chave estrangeira de Environmental_Condition
		fk_project	character varying(10)	Chave estrangeira de Project
		fk_habitat	integer	Chave estrangeira de Habitat
		fk_taxa	integer	Chave estrangeira de Taxonomy
		fk_meth	integer	Chave estrangeira de Methodology

Figura A.1: Descrição da tabela Sample

Tabela	Descrição	Campo	Tipo	Descrição
HOMOGENEOUS SET	Representa um conjunto de indivíduos de uma mesma espécie.	codlote	character varying(10)	Chave primária do Lote(um lote representa uma agrupação de indivíduos da mesma espécie.)
		weight	double precision	Representa o peso do lote de espécies
		notes	character varying(100)	Alguma observação que os biólogos desejam realizar sobre esse lote de indivíduos
		individualCount	integer	Quantidade de indivíduos de uma mesma espécie
		endemic	boolean	Se a espécie é endêmica ou não
		size	character varying(20)	Medidas realizadas aos animais
		lifeStage	character varying(20)	Fase de desenvolvimento
		preparationType	character varying(20)	O tipo de preparação que os biólogos realizam nas espécies. Exemplo: taxidermia
		sex	char(1)	Sexo do lote
		dateIdentified	Composite Types	Data de identificação de um lote de espécies
		fk_idsample	integer	Chave estrangeira de Sample
		fk_responsible	integer	Chave estrangeira de Responsible
		fk_taxa	integer	Chave estrangeira de Taxonomy
		fk_catalog	integer	Chave estrangeira de Catalog
		fk_meth	integer	Chave estrangeira de Methodology
		fk_location	integer	Chave estrangeira de Location_Coord

Figura A.2: Descrição da tabela Homogeneous_Set

Tabela	Descrição	Campo	Tipo	Descrição
CATALOG	Representa o caderno onde é realizado o tombamento das espécies no museu de zoologia	idCatalog	integer	A chave primária está formado pelos campos: institutionCode, collectionCode e catalogNumber
		institutionCode	character varying(10)	Código da instituição. Exemplo: ZUEC
		collectionCode	character varying(30)	Código da coleção. Exemplo: BIV, POL
		catalogNumber	character varying(10)	Número do registro no tombo
		catalogDate	date	Data em que se realizou um tombamento
		dateIdentified	Composite Types	Data no qual um identificador identificou uma espécie
		conservationMeans	character varying(50)	Forma de preservação do animal. Por exemplo: úmida (álcool, formol), seco ou lâmina
		basisOfRecord	character varying(20)	É uma descrição que representa se o registro é uma observação ou objeto. Por exemplo: Foto, Son, Pegada, Fezes, entre outros
		previousCatalogNumber	character varying(10)	Um número de catálogo anterior
		typeMaterial	character varying(20)	Espécimes tipos, representativos de determinada espécie. Por exemplo: holótipo, parátipo, neótipo, lectótipo
		patternColor	character varying(50)	O padrão de coloração de uma espécie registrada. Por exemplo, Pardo acinzentado
		lifeStage	character varying(20)	Fase de desenvolvimento
		sex	char(1)	Sexo do lote
		size	character varying(20)	Medidas realizadas às espécies
		typeSubstrate	character varying(40)	Descrição do tipo de substrato
		commomName	character varying(50)	Nome comum da espécie registrada
		fieldNumber	character varying(20)	Número de Registro de Campo assinado pelos biólogos
		publication	character varying(50)	Título de uma publicação que usa informação do catálogo
		notes	character varying(100)	Observações que realizam os curadores do museu
		earliestDateCollected	Composite Types	Data de início da coleta
		latestDateCollected	Composite Types	Data final da coleta. Se a coleta for realizada no mesmo dia colocar neste campo o valor de earliestDateCollected.
		individualCount	integer	Quantidade de indivíduos de uma mesma espécie a serem catalogados
		fk_cataloguedBy	integer	Chave estrangeira de Responsible (identifica quem catalogou)
		fk_location	integer	Chave estrangeira de Location_Coord
		fk_taxa	integer	Chave estrangeira de Taxonomy
		fk_collector	integer	Chave estrangeira de Responsible (identifica o coletor)
		fk_identifier	integer	Chave estrangeira de Responsible (identifica quem identificou a espécie)
		fk_meth	integer	Chave estrangeira de Methodology

Figura A.3: Descrição da tabela Catalog

Tabela	Descrição	Campo	Tipo	Descrição
TAXONOMY	Registra a classificação taxonômica das espécies	idTaxa	integer	Chave primária
		Kingdom	character varying(20)	Nível Taxonômico
		Phylum	character varying(20)	Nível Taxonômico
		Subphylum	character varying(20)	Nível Taxonômico
		Class	character varying(20)	Nível Taxonômico
		SubClass	character varying(20)	Nível Taxonômico
		Order	character varying(20)	Nível Taxonômico
		Suborder	character varying(20)	Nível Taxonômico
		Superfamily	character varying(20)	Nível Taxonômico
		Famiily	character varying(20)	Nível Taxonômico
		SubFamily	character varying(20)	Nível Taxonômico
		SuperTribe	character varying(20)	Nível Taxonômico
		Tribe	character varying(20)	Nível Taxonômico
		SubTribe	character varying(20)	Nível Taxonômico
		Genus	character varying(20)	Nível Taxonômico
		SpecificEpithet	character varying(20)	Nível Taxonômico
		AuthorYearOfScientificName	character varying(50)	Autor e data do nome científico da

Figura A.4: Descrição da tabela Taxonomy

Tabela	Descrição	Campo	Tipo	Descrição
LOCATION_COORD	Representa a localização onde foi realizada a coleta.	id	Integer	Chave Primária.
		wpt	character varying(20)	É o código atribuído pelo GPS para uma localização geográfica, ao fazer um registro no campo
		latitude	character varying(20)	Latitude
		longitude	character varying(20)	Longitude
		elevation	character varying(60)	Altitude
		datum	character varying(10)	Referencial usado pelo GPS no campo
		unidcons	character varying(30)	Unidade de conservação para assinalar as localidades que são reservas
		extension	character varying(30)	Atributo do processo de localização no campo que indica o espaço ocupado pela amostra ou amostras
		coordinatePrecision	character varying(60)	O limite superior da distância (em metros) a partir da Latitude e Longitude descrevendo um círculo dentro do qual a localidade está descrita
		scale	character varying(20)	Escala cartográfica (no GPS) utilizada no dimensionamento de mapas
		depth	character varying(60)	Profundidade da qual se tirou uma amostra (em metro)
		fk_index	Integer	Chave estrangeira de Locationdetails
		thepoint_lonlat	geometry	Ponto da coleta
		dateLocName	date	Representa o histórico das ligações entre um determinado local geográfico e os nomes atribuídos a ele durante o tempo.
		fk_gid	integer	Chave estrangeira da tabela GeomCounty

Figura A.5: Descrição da tabela Location_Coord

Tabela	Descrição	Campo	Tipo	Descrição
LOCATIONDETAILS	Representa os designadores dos locais de amostragem.	id	Integer	Chave Primária.
		name	character varying(10)	Nome do local onde foi feita a coleta
		county	character varying(20)	Nome do Município onde foi feita a coleta de amostras. Por exemplo: Itirapina, Rifaina, etc.
		stateProvince	character varying(20)	Nome do Estado onde é feita a coleta
		country	character varying(20)	Nome do país onde se realizou a coleta
		continentOcean	character varying(20)	Nome do continente ou oceano onde se realizou a coleta
		reference	character varying(30)	Descrição de algum ponto que pode servir para referenciar o local

Figura A.6: Descrição da tabela LocationDetails

Tabela	Descrição	Campo	Tipo	Descrição
HABITAT	Descreve o habitat onde foi realizada a coleta de amostras.	codhabitat	integer	Chave Primária.
		name	character varying(60)	Nome do habitat. Por exemplo: cerrado, pasto, cerradão, entre outros
		description	character varying(20)	Descrição do habitat

Figura A.7: Descrição da tabela Habitat

Tabela	Descrição	Campo	Tipo	Descrição
HOSTS	Tabela de relacionamento entre Habitat e Location_Coord(N-N)	id	Integer	Chave Primária
		fk_codhabitat	integer	Chave estrangeira de Habitat
		fk_idlocation	integer	Chave estrangeira de Location_Coord

Figura A.8: Descrição da tabela Hosts

Tabela	Descrição	Campo	Tipo	Descrição
ENVIRONMENTAL_CONDITION	Representa as condições ambientais de um habitat e/ou amostras	cod_varamb	integer	Chave primária
		temperature	double precision	Valor que representa a temperatura do lugar onde foi realizada a coleta de amostras
		salinity	double precision	Valor que representa a salinidade do lugar onde foi realizada a coleta de amostras
		dissolved_oxygen	double precision	Valor que representa a quantidade de oxigênio dissolvido no lugar onde foi realizada a coleta de amostras
		type_fund	character varying(20)	Descreve o tipo de fundo onde foi realizada a coleta de amostras. Por exemplo: fundo rochoso, inconsolidado, consolidado.
		moisture	double precision	Umidade do local onde se realizou a coleta
		type_sediment	character varying(10)	Uma classificação. Exemplo: areia, argila
		fk_codHabitat	integer	Chave estrangeira de Habitat

Figura A.9: Descrição da tabela Environmental_Condition

Tabela	Descrição	Campo	Tipo	Descrição
SPECIMEN	Registra os indivíduos de cada espécie.	codespecime	integer	Código que identifica o indivíduo da espécie
		data	date	Uma descrição geral
		imageURL	character varying(30)	Endereço da imagem tirada do espécime
		fk_codLote	character varying(10)	Chave estrangeira de Homogeneous_Set

Figura A.10: Descrição da tabela Specimen

Tabela	Descrição	Campo	Tipo	Descrição
PRO_USE_MET	Tabela de relacionamento entre Project e Methodology(N-N)	idProMet	integer	Chave Primária
		fk_idmeth	integer	Chave estrangeira de Methodology
		fk_idproject	character varying(10)	Chave estrangeira de Project

Figura A.11: Descrição da tabela Pro_use_Met

Tabela	Descrição	Campo	Tipo	Descrição
SUBSTRATE	Representa parte da amostra onde localiza-se as espécies a serem estudadas	idSubstrate	integer	Chave primária
		organicMatter	double precision	A quantidade de matéria orgânica determinada em um substrato
		type	character varying(40)	Descrição do tipo de substrato
		notes	character varying(100)	Algumas anotações que os biólogos desejam realizar sobre esse substrato não biológico
		sex	character(1)	Identifica o sexo. No caso do Laboratório Inseto-Plantas, identifica o sexo da planta.
		plant_habit	character varying(20)	Forma ou modo de crescimento de uma planta. Por exemplo, se é um arbusto, erva, etc.
		phenology	character varying(20)	Estado da amostra ao longo da vida (ex: estado vegetativo, reprodutivo: flor, fruto, etc)
		abundance	character varying(20)	Descrição da quantidade das amostras coletadas (ex: comum, rara, média)
		spatialDistribution	character varying(20)	Descrição da distribuição espacial da amostra coletada (ex:mancha, isolada, esparsa, mancha densa, etc.)
		herbariumDate	date	Data em que uma amostra foi depositada no Herbário
		weightG	double precision	Peso em gramas da amostra
		weightBagG	double precision	Peso em gramas da sacola de papel
		rateG	double precision	O valor porcentual do peso da amostra (Alíquota)
		capitulumNumber	integer	Número de capítulos que existe na alíquota (rateG)
		estimatedCapitulum	double precision	Número de capítulos estimados que a amostra possui
		countCapitulum	double precision	Contagem manual que os biólogos realizam dos capítulos
		size	character varying(20)	Medidas realizadas a uma amostra
		fk_sample	integer	Chave estrangeira de Sample
		Bio	boolean	Representa se o substrato é biológico ou não biológico

Figura A.12: Descrição da tabela Substrate

Tabela	Descrição	Campo	Tipo	Descrição
RESPONSIBLE	Corresponde a cada pessoa do laboratório que realiza coletas ou o tombamento	id	integer	Chave Primária
		full_name	character varying(80)	Nome completo do pesquisador
		abbreviated_name	character varying(40)	Nome abreviado
		department	character varying(30)	Nome do departamento que pertence o pesquisador
		laboratory	character varying(20)	Nome do laboratório do pesquisador
		CPF	character varying(20)	Número que identifica uma pessoa física
		CNPJ	character varying(20)	Número que identifica uma pessoa juridica

Figura A.13: Descrição da tabela Responsible

Tabela	Descrição	Campo	Tipo	Descrição
METHODOLOGY	Armazena o registro das metodologias utilizadas nas coletas, nos projetos, etc.	idmeth	integer	Chave Primária
		name	character varying(50)	Nome da metodologia que foi utilizada
		description	character varying(100)	Descrição da metodologia

Figura A.14: Descrição da tabela Methodology

Tabela	Descrição	Campo	Tipo	Descrição
PROJECT	Armazena dados dos projetos realizados no laboratório	idpro	character varying(10)	Chave Primária
		name	character varying(50)	Nome assinado ao projeto
		acronyms	character varying(30)	Siglas para assinar ao Projeto
		notes	character varying(100)	Alguma anotação que o biólogo realiza sobre o projeto
		startDate	date	Data de início do projeto
		endDate	date	Data do fim do projeto

Figura A.15: Descrição da tabela Project

Tabela	Descrição	Campo	Tipo	Descrição
PROJECT_RESPONSIBLE	Tabela de relacionamento entre Project e Responsible(N-N)	id	integer	Chave Primária
		fk_idresponsible	integer	Chave estrangeira de Responsible
		fk_idproject	character varying(10)	Chave estrangeira de Project

Figura A.16: Descrição da tabela Project_Responsible