UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Busca multimodal para apoio à pesquisa em biodiversidade

Este exemplar corresponde à redação final da Dissertação devidamente corrigida e defendida por Gabriel de Souza Fedel e aprovada pela Banca Examinadora.

Campinas, 13 de Abril de 2011.

Este exemplar corresponde à redação final da Tese/Dissertação devidamente corrigida e defendida por: Gabriel de Souza Fedel
e aprovada pela Banca Examinadora.
Campinas, 25 de julho de 2011
COORDENADOR DE PÓS-GRADUAÇÃO
CPG-IC

Prof. Dr. Julio César Lopez Hernández
Coord. Subst. de Pós-Graduação
Instituto de Computação/Unicamp
Matr. 28.620-1

Claudia Maria 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

Fedel, Gabriel de Souza
F316b Busca multimodal para apoio à pesquisa em biodiversidade/Gabriel de Souza Fedel-- Campinas, [S.P. : s.n.], 2011.

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

1.Banco de dados. 2.Sistemas de recuperação da informação. 3.Biodiversidade. 4.Imagens - Recuperação. 5.Ontologia. I. Medeiros, Claudia Maria Bauzer. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Título em inglês: Multimodal search to support research on biodiversity

Palavras-chave em inglês (Keywords): 1.Databases. 2.Information storage and retrieval systems. 3.Biodiversity. 4.Image database. 5.Ontology.

Área de concentração: Banco de Dados

Titulação: Mestre em Ciência da Computação

Banca examinadora: Profa. Dra. Claudia Maria Bauzer Medeiros (IC – UNICAMP)
Prof. Dr. André Santanchè (IC – UNICAMP)
Prof. Dr. Roberto Marcondes Cesar Junior (IME - USP)

Data da defesa: 13/04/2011

Programa de Pós-Graduação: Mestrado em Ciência da Computação

TERMO DE APROVAÇÃO

Dissertação Defendida e Aprovada em 13 de abril de 2011, pela Banca examinadora composta pelos Professores Doutores:

Roberto M C Junior

Prof. Dr. Roberto Marcondes Cesar Junior
IME / USP

André Santanchè

Prof. Dr. André Santanchè
IC / UNICAMP

Profª. Drª. Claudia Maria Bauzer Medeiros
IC / UNICAMP

Instituto de Computação
Universidade Estadual de Campinas

Busca multimodal para apoio à pesquisa em biodiversidade

Gabriel de Souza Fedel[1]

Abril de 2011

Banca Examinadora:

- Claudia Maria Bauzer Medeiros
  Instituto de Computação - Unicamp (Orientadora)
- Roberto Marcondes Cesar Junior
  Instituto de Matemática e Estatística - USP
- André Santanchè
  Instituo de Computação - Unicamp
- Ariadne Maria Brito Rizzoni Carvalho
  Instituto de Computação - Unicamp (Suplente)
- Luís Felipe de Toledo Ramos Pereira
  Instituto de Biologia - Unicamp (Suplente)

[1]Suporte financeiro de: Bolsa da CAPES Mar/2009 – Jul/2010, Bolsa Fapesp Ago/2010 – Fev/2011 Processo 2009/11233-0

Resumo

A pesquisa em computação aplicada à biodiversidade apresenta muitos desafios, que vão desde o grande volume de dados altamente heterogêneos até a variedade de tipos de usuários. Isto gera a necessidade de ferramentas versáteis de recuperação. As ferramentas disponíveis ainda são limitadas e normalmente só consideram dados textuais, deixando de explorar a potencialidade da busca por dados de outra natureza, como imagens ou sons.

Esta dissertação analisa os problemas de realizar consultas multimodais a partir de predicados que envolvem texto e imagem para o domínio de biodiversidade, especificando e implementando um conjunto de ferramentas para processar tais consultas. As contribuições do trabalho, validado com dados reais, incluem a construção de uma ontologia taxonômica associada a nomes vulgares e a possibilidade de apoiar dois perfis de usuários (especialistas e leigos). Estas características estendem o escopo da consultas atualmente disponíveis em sistemas de biodiversidade. Este trabalho está inserido no projeto BioCORE, uma parceria entre pesquisadores de computação e biologia para criar ferramentas computacionais para dar apoio à pesquisa em biodiversidade.

Abstract

Research on Computing applied to biodiversity present several challenges, ranging from the massive volumes of highly heterogeneous data to the variety in user profiles. This kind of scenario requires versatile data retrieval and management tools. Available tools are still limited. Most often, they only consider textual data and do not take advantage of the multiple data types available, such as images or sounds.

This dissertation discusses issues concerning multimodal queries that involve both text and images as search parameters, for the domanin of biodiversity. It presents the specification and implementation of a set of tools to process such queries, which were validate with real data from Unicamp's Zoology Museum. The ain contribuitions also include the construction of a taxonomic ontology thath includes species common names, and support to both reserachers and non-experts in queries. Such features extend the scop of queries available in biodiversity information systems. This research is associated with the Biocore project, jointly conducted by reserachers in computing and biology, to design and develog computational tools to support research in biodiversity.

Agradecimentos

Agradeço inicialmente aos meus pais, Silvio e Márcia, que sempre me deram amor e se empenharam ao máximo para me formar um cidadão crítico e transformador. Em segundo aos meus amigos, de todos os lugares, que sempre foram portos seguros para os momentos de tempestade.

À Profa. Dra. Claudia Bauzer Medeiros, pela orientação, oportunidades, desafios e pelo apoio proporcionados desde que entrei na Unicamp.

Aos amigos e colegas do LIS por esses dois anos de amizade, companheirismo, convivência e pela ajuda sempre prestativa nas horas necessárias.

Agradeço à banca examinadora pelas sugestões eaos professores Felipe de Toledo e Michela Borges do Instituto de Biologia da Unicamp que colaboraram uito para a execução do meu mestrado.

Aos amigos do laboratório RECOD, em especial à Jefersson Alex, que me ajudaram muito com o módulo de recuperação de imagens por conteúdo e com o GP Framework.

As agências CNPq (Projeto Biocore) e CAPES pelo apoio.

Agradeço também o apoio da FAPESP dentro do processo 2009/11233-0.

Por fim agradeço à Unicamp, e todos que a constituem: estudantes, funcionários e professores, por nesses anos que passei aqui terem colaborado para minha formação intelectual e pessoal.

Sumário

Lista de Figuras

Capítulo 1

Introdução

O termo Biodiversidade se refere à descrição da riqueza e a variedade do mundo natural [45]. O conhecimento da biodiversidade nos ecossistemas é importante para desenvolvimento de políticas públicas efetivas para a preservação, monitoramento de alterações ambientais em diferentes escalas e aproveitamento sustentável da diversidade biológica. Os dados disponíveis são coletados em vários lugares do mundo por muitos grupos de pesquisadores, sendo publicados em formatos distintos e especificados em inúmeros padrões. Este cenário é caracterizado por sua heterogeneidade intrínseca – não apenas de dados e modelos conceituais utilizados, como também de necessidades e perfis dos especialistas que coletam e analisam os dados.

As dificuldades citadas têm motivado esforços na coleta de dados visando a soluções de gerenciamento computacional que permitam aos pesquisadores recuperar e analisar os dados armazenados. Sistemas de Informação de Biodiversidade [16, 42] vêm sendo desenvolvidos com tais objetivos, apresentando vários desafios na sua especificação e implementação, inclusive pela heterogeneidade de dados e usuários.

Os dados nestes sistemas são de diversas naturezas, podendo ser textuais (como nome da espécie ou tipo de habitat), imagens (como fotos da espécie), sons (gravações de ruídos de animais), croquis, entre outros. A abordagem mais comum para recuperar esses dados, a busca textual sobre texto e metadados, não considera particularidades como o conteúdo de uma imagem ou de um arquivo de áudio, o que poderia refinar o resultado. A busca textual também não motiva o uso desses sistemas por leigos, pois os sistemas em geral exigem conhecimento especializado, como terminologia científica. Além disso, como normalmente só é utilizada uma modalidade para recuperação desses dados, o potencial de fusão dessas modalidades é descartado.

Somado ao uso limitado dos dados disponíveis, há pouca preocupação em disponibilizar esses dados de maneira diferenciada para os diferentes tipos de usuários interessados. A princípio, podemos identificar pelo menos dois grandes tipos de usuários interessados

nestes dados: pesquisadores em biodiversidade e leigos. Embora existam propostas de sistemas para buscas multimodais, estas costumam ser restritas a um tipo específico de usuário.

O objetivo desta dissertação é estudar os problemas associados ao uso de multimodalidade para recuperação de dados de biodiversidade e facilidades para diversos tipos de usuário no acesso a esses dados. Para isso apresentamos o Sinimbu - um sistema que permite consultas multimodais combinando texto e conteudo de imagem para recuperar dados de coleta (observações na natureza), de coleções (acervos de museus) e imagens. O nome Sinimbu significa camaleão em tupi-guarani, denotando que há vários modos de interação que se adaptam a perfis e necessidades do usuário. As principais contribuições desta dissertação, publicadas em um resumo estendido [17], são:

- análise da situação atual dos sistemas de biodiversidade e busca multimodal;
- exploração da multimodalidade dos dados para processar consultas, criando ferramentas para combinar resultados - permitindo combinar a tradicional busca baseada em metadados com busca por imagens e nome vulgar;
- tratamento diferenciado para os dois tipos principais de usuários: leigos e pesquisadores;
- criação e utilização de uma ontologia taxonômica com nomes vulgares para estender os tipos de consulta;
- validação da proposta com implementação de um protótipo baseado em dados reais do Museu de Zoologia da Unicamp, com 40000 registros e 1200 imagens.

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 Sinimbu em detalhes apresentando as estruturas de dados, arquitetura e protótipos de tela. O Capítulo 4 apresenta aspectos de implementação do desenvolvimento do Sinimbu. Por fim, o Capítulo 5 apresenta as conclusões e trabalhos futuros.

Capítulo 2

Trabalhos Relacionados e Conceitos Básicos

Este capítulo aborda Sistemas de Informação de Biodiversidade, o projeto BioCORE, Ontologias e Recuperação de Dados. Também é apresentado uma série de trabalhos de busca multimodal para mostrar o estado atual desta área. A Figura 2.3 resume os sistemadas estudados.

2.1 Sistemas de Informação de Biodiversidade

Um Sistema de Informações de Biodiversidade (SIB) é um sistema de informação que gerencia grandes conjuntos de dados geográficos assim como grandes bases de dados relativos a espécies (como coleções de história natural, registros de observação de campo e dados experimentais) [13]. Tais sistemas têm como objetivo auxiliar a pesquisa em biodiversidade, ajudando a busca de novas espécies, busca por interações entre as espécies, o planejamento para preservação de espécies, dentre outras funcionalidades.

O estudo, a conservação e o uso sustentável da biodiversidade requerem um tratamento interdisciplinar, além de um ambiente de colaboração global [9]. A utilização de um SIB para tal finalidade exige gerenciar e correlacionar dados de ocorrência de espécies com diferentes outros tipos de informação, como dados geográficos, dicionários de termos geográficos e topônimos, catálogos de nomes científicos, registros históricos e vários outros [15].

Os sistemas que atuam na área de biodiversidade possuem variações na finalidade e no escopo. O OBIS (*Ocean Biogeographic Information System*) [10], por exemplo, trata de informações referentes à biodiversidade marinha, enquanto que o KBIS (*The Korean Bird Information System*) [33] tem como objetivo construir um sistema cooperativo para aquisição, gerenciamento e compartilhamento de informações sobre pássaros coreanos, por

meio de grupos especialistas e usuários não-especialistas. No Brasil existe um trabalho que também visa coletar informações de pássaros de maneira colaborativa, o WikiAves [2], que permite armazenamento de informações textuais, visuais, sonoras e geográficas sobre pássaros brasileiros.

A principal iniciativa brasileira, o SinBIOTA [1], desenvolveu um conjunto de serviços e ferramentas para acesso, por pesquisadores, à coletas realizadas por pesquisadores do estado de São Paulo. Contando com algumas centenas de milhares de registros, o sistema oferece a possibilidade de busca textual a partir de parametros de informação taxonômica, localização e dados de coleta em geral. Oferece saídas textuais, listas de registros ou distribuição geográfica de espécies.

Também existem trabalhos que visam auxiliar a identificação de novas espécies. Em Thi [41], por exemplo, é apresentado um trabalho que busca associar nomes vulgares de bambu, da Indochina, aos nomes científicos por meio de características morfológicas.

2.2 Projeto BioCORE

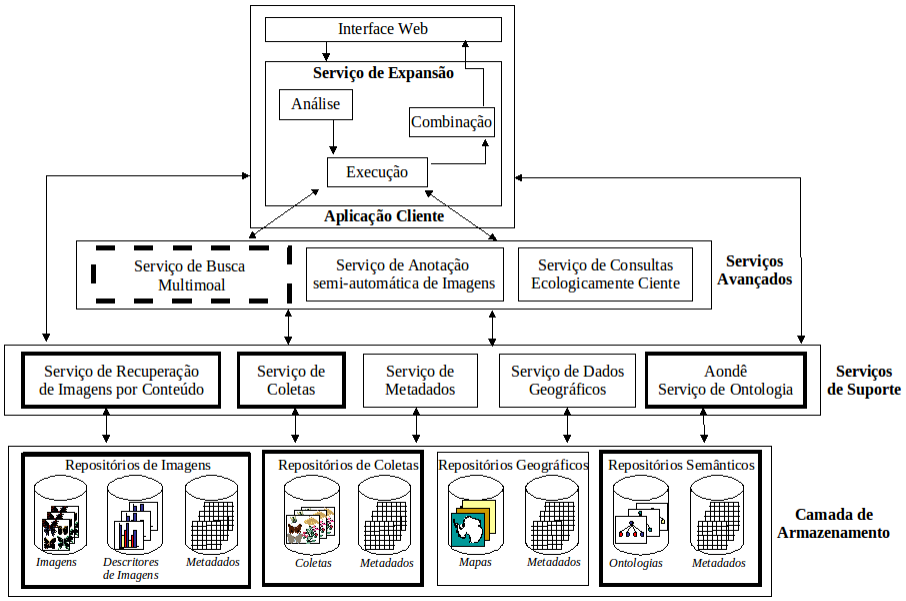

Figura 2.1: Arquitetura do Projeto BioCORE

O projeto BioCORE [4] é uma continuação do projeto WeBios [3]. Ambos são um esforço de cientistas da computação e pesquisadores em biodiversidade. Entre seus objetivos podemos destacar a especificação e desenvolvimento de ferramentas computacionais para permitir que pesquisadores em biodiversidade manejem e compartilhem dados, ajudando-

os na construção de modelos complexos e na modelagem de ecossistemas, incluindo a descoberta de relacionamentos e interações entre as espécies.

O BioCORE tem ainda o objetivo de resolver questões teóricas e de implementação relativas ao gerenciamento de dados científicos, como por exemplo, imagens, vídeos, dados geográficos e dados de coleções de espécies. As ferramentas e os dados gerados e manipulados com o projeto estão sendo utilizados para dar origem ao museu virtual de zoologia da UNICAMP. Este museu será derivado da coleção existente não-digital.

O museu virtual espera atingir inicialmente dois grupos de usuários: pesquisadores em biodiversidade e visitantes. O primeiro estará buscando informações no acervo com intuito de encontrar dados para sua pesquisa. O segundo utilizará o museu para conhecer o acervo, e buscar informações sobre espécies, porém sem interesse em informações especializadas.

A Figura 2.1 apresenta a arquitetura do projeto BioCORE, composta por 4 camadas principais: armazenamento, serviços de suporte, serviços avançados e aplicação cliente. A *aplicação cliente* fornece uma interface entre o usuário e os serviços disponíveis. Os serviços são categorizados como de *suporte* e *avançados*. O primeiro grupo, de *serviços de suporte*, fornece acesso aos dados dos repositórios localizados na *camada de armazenamento* e compõe os serviços de: recuperação de imagens por conteúdo, coletas, metadados, dados geográficos e ontologias. O segundo conjunto, de *serviços avançados*, é composto pelos serviços de anotação semi-automática de imagens e de consultas ecologicamente cientes. Serviços avançados são resultados de chamadas aos serviços de suporte.

A *camada de armazenamento* possui um conjunto de repositórios responsável pelo gerenciamento de informações sobre imagens, mapas, coletas e ontologias. Estes recursos permitem associar informações coletadas por pesquisadores, gerenciar imagens para posteriores análise e processamento. As ontologias fornecem a possibilidade de guardar informações conceituais e empregá-las em consultas. Além disso, cada repositório mantém um conjunto de metadados para facilitar o gerenciamento e a recuperação das informações.

Grande parte das camadas de suporte presentes no BioCORE já está implementada. O *Repositório Semântico*, da Camada de Armazenamento, e *Aondê - Serviços de Ontologia*, dos Serviços de Suporte são desenvolvidos no trabalho de mestrado apresentado em Daltio [14]. O *Serviço de Consultas Ecologicamente Ciente* dos Serviços Avançados é resultado do mestrado de Gomes [24]. Já o *Repositório de Coletas*, da Camada e Armazenamento, e o *Serviço de Coletas*, dos Serviços de Suporte, foram implementados em Malverri [28]. O *Serviço de Expansão* da Aplicação Cliente foi desenvolvido no mestrado de Villar [44].

A dissertação, como detalhado no capítulo 3, desenvolveu novas ferramentas que deverão ser acopladas ao BioCORE para permitir buscas multimodais envolvendo, inclusive, busca por nomes vulgares.

2.3 Ontologias

Uma ontologia é uma especificação de uma conceitualização [20]. Por conceitualização, entende-se a visão e o entendimento que se tem da realidade. Dessa forma, ontologias procuram capturar a semântica de um domínio pelo desenvolvimento de primitivas de representação do conhecimento, habilitando que máquinas possam (parcialmente) interpretar o significado dos relacionamentos entre os conceitos em um domínio [40].

A característica de processamento em máquinas também é evidenciada na definição de Quiou [35], para o qual uma ontologia "... descreve entidades em um mundo e seus relacionamentos, combinando o entendimento humano de símbolos com a capacidade de processamento por máquina". Para Noy [32], é a descrição formal de um domínio, projetada para o compartilhamento entre diferentes aplicações e expressa em uma linguagem que pode ser usada para o raciocínio.

A capacidade de descrever formalmente um conhecimento e de raciocinar sobre ele torna possível adotar uma ontologia como forma de representar um domínio, padrão ou consenso de um grupo e, posteriormente, utilizar a mesma representação por meio de mecanismos e ferramentas que a processem. Assim, nota-se uma disposição para o uso de ontologias como uma forma de representar concepções, regras e diferentes recursos em tarefas que podem comportar serviços automatizados.

Ontologias restringem o conjunto de possíveis mapeamentos entre símbolos e seus significados [38]. Isso ocorre por meio da utilização de conceitos, propriedades, relacionamentos, restrições, axiomas e instâncias:

- Conceitos são abstrações de objetos ou partes enumeráveis do universo representado;
- Propriedades permitem traçar as características dos conceitos, sob a forma de atributos que os caracterizem;
- Restrições tornam possível aproximar cada conceito da realidade representada, impondo condições a serem satisfeitas;
- Relacionamentos representam as diferentes relações existentes entre os conceitos, como as relações de herança e de parte/todo, bem como quaisquer alternativas que possam especificar a interação entre dois objetos;
- Axiomas que são classes de inferências e permitem estabelecer as relações do domínio;
- Instâncias representam casos, ou exemplos de conceitos, que utilizam os componentes supracitados com a atribuição de valores que as tornem únicas.

Existem vários motivos para se utilizar ontologias, entre eles podemos citar [32]:

- Compartilhar um entendimento comum da estrutura da informação entre pessoas ou agentes de *software*: se um conjunto de aplicativos está contextualizado em um domínio comum ou em domínios relacionados, é possível utilizar agentes para consultar as informações abrangidas por diferentes ontologias para responder a consultas de forma eficiente, ainda que o conhecimento não esteja diretamente armazenado na estrutura de sua aplicação;

- Habilitar o reuso do domínio do conhecimento: se um grupo de pesquisa trabalha com uma área do conhecimento, é possível que outros grupos, envolvidos em áreas relacionadas, aproveitem o conhecimento produzido pelo primeiro grupo e reutilizem as informações anteriormente produzidas;

- Tornar explícitas as suposições do domínio: permite manter uma independência entre os recursos realizados por programação e a evolução da ontologia e do conhecimento sobre a área descrita;

- Separar o domínio do conhecimento do conhecimento operacional: é possível descrever uma tarefa, como a configuração de produtos a partir de seus componentes, e permitir a realização automática dessa tarefa, como a configuração independente de produtos e componentes;

- Analisar o domínio do conhecimento: ter uma especificação de um domínio é útil à medida em que é possível realizar uma análise formal dos termos e assim poder reutilizar ou estender tal domínio.

Particularmente em Sistemas de Informação de Biodiversidade, é possível encontrar diferentes finalidades conferidas às ontologias. Trabalhos como OBSERVER [30], SEEK [31] e Animal Diversity Web [34] incluem o uso de ontologias para permitir a consulta e a análise de dados em fontes de informação múltiplas e heterogêneas. Outros exemplos de uso de Ontologias em SIB's são apresentados em Hyam [22] e Liao [27], que apresentam o uso de ontologias para integrar vocabulários de diferentes fontes referentes à biodiversidade, tais fontes muitas vezes possuem termos diferentes para o mesmo conceito. Além disso, é possível encontrar aplicações específicas, como descrever recursos e serviços para automatizar processos, contextualizar e inferir informações para o processamento de consultas, dentre outros.

A dissertação, como descrito a partir do capítulo 3, construiu uma ontologia taxonômica associada a nomes vulgares para ampliar o leque de consultas do BioCORE.

2.4 Recuperação de Dados

Além das questões tradicionais de processamento de consultas, um dos grandes desafios da recuperação de dados é conseguir entender o que o usuário busca. Quando se trabalha com grandes quantidades de dados as dificuldades de se realizar buscas aumenta e traz alguns problemas, como a grande quantidade de resultados, indexação, entrada e saída. A dissertação visa combinar dois tipos de recuperação: busca textual e em imagens, analisadas a seguir.

2.4.1 Recuperação por busca textual

Um dos meios mais utilizados para recuperação de dados é a busca textual, normalmente pela definição de predicado sobre campos armazenados. A forma mais comum de realizar busca textual é por palavras-chave, que podem estar associadas a uma tupla e/ou a uma imagem. A área de recuperação de informações também trata de busca textual, porém utilizando critérios estatísticos e mineração em textos. De acordo com Manning [29] recuperação de informações é a busca por materiais (normalmente documentos) de uma natureza desestruturada (normalmente texto) que satisfaz uma necessidade de informação e que está armazenado em grandes coleções. A necessidade de informação do usuário, citada por Manning [29], não é trivial de ser descoberta e está entre os maiores problemas da área.

2.4.2 Recuperação de imagens por conteúdo

Quando os dados envolvidos em predicados de busca são imagens, uma das maneiras de se recuperá-los é por meio do seu conteúdo. Para trabalhar com o conteúdo das imagens, são extraídas suas características, que normalmente são armazenadas em vetores. As características extraídas podem ser referentes a textura, cor ou forma. Um descritor de conteúdo de uma imagem é um par $< f, v >$ onde v é o vetor de características da imagem e f a função de distância utilizada para comparar dois vetores [43].

Uma consulta baseada em conteúdo normalmente recebe uma imagem como consulta e retorna as imagens mais similares à imagem consultada, ou seja, retorna as imagens com vetores de características mais próximos do vetor de característica da imagem fornecida. Um dos grandes problemas da recuperação de imagens por conteúdo é o *gap* semântico que existe entre o que o usuário deseja buscar e o que o sistema retorna. A maioria das aplicações com recuperação de imagens por conteúdo que possuem bom desempenho utilizam um domínio específico e trabalham com imagens já pré-processadas.

2.4.3 Programação Genética e o GP Framework

Programação Genética (GP) [26] é uma técnica de aprendizado de máquina que tenta resolver problemas baseado em princípios de biologia evolutiva. A estrutura básica no GP é o indivíduo, que representa uma possível solução para um problema dado. Os indivíduos são programas que durante o processo evolutivo passam por sucessivas recombinações, sendo progressivamente refinados.

No fim do processo espera-se encontrar as soluções mais apropriadas para o problema em questão, dentro do universo dos parâmetros escolhidos. Como os indivíduos são calculados por meio de uma função de adaptação, a programação genética pode ser entendida como uma busca por um indivíduo que melhor resolva o problema em um espaço de todas as soluções possíveis.

Esta função tem o objetivo de combinar valores de similaridade por diferentes descritores. Desta maneira cada nó interno da árvore que representa o indivíduo é um operador aritmético, considerando que as folhas são valores de similaridade de um descritor d entre uma imagem I_j e I_k. Um exemplo de um indivíduo GP é apresentado na figura 2.2. Esse indivíduo corresponde à função $f(d_{1I_jI_k}, d_{2I_jI_k}, d_{3I_jI_k}) = (\frac{d_{1I_jI_k} - d_{3I_jI_k}}{d_{2I_jI_k}}) + \sqrt{d_{2I_jI_k} * d_{3I_jI_k}}$.

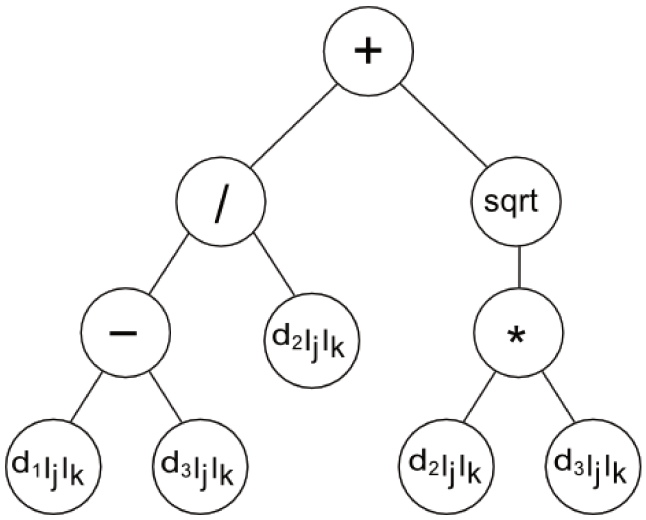

Figura 2.2: Um exemplo de indivíduo em programação genética

O GP framework [12] utiliza programação genética para combinar descritores e encontrar as necessidades do usuário. Essa ferramenta combina diferentes descritores de imagens, usando programação genética, para encontrar uma função que permita retorne da melhor maneira as imagens buscadas pelo usuário. A dissertação usou o GP framework para construir descritores de imagens e assim permitir busca por conteúdo.

2.5 Consultas Multimodais

Esta seção apresenta alguns trabalhos sobre consultas multimodais. Tais consultas envolvem mais de um tipo de dado, como metadados, ontologias, imagem ou som e permite vários modos de intereção. Tais consultas buscam aumentar a qualidade da recuperação dos dados por meio do uso da combinação de várias informações. Os próximos parágrafos descrevem exemplos de pesquisa em consultas multimodais, quer associadas a sistemas genéricos ([8, 5, 6, 46, 19, 25, 18, 36, 37, 39]) , quer sistemas que utilizem dados biológicos ([21, 7]).

O desenvolvimento de interfaces multimodais é uma área em crescimento, existindo inclusive uma conferência internacional dedicada a tal assunto - a International Conference on Multimodal Interfaces, já na décima primeira edição [23]. Esta conferência cobre uma vasta gama de assuntos, estando associada ao grupo especial de interfaces da ACM (SIGCHI). Exemplos de tópicos cobertos são Processamento multimodal e multimídia, Interfaces Multimodais de Entrada e Saída, Aplicações Multimodais, entre outros.

Em Atnafu [8] é apresentado um novo modelo de dados para imagens que suporta de maneira eficiente descrição de imagens baseadas em metadados e o conteúdo dessas imagens, e utiliza tal modelo para recuperação de imagens em uma base de dados de imagem. Por meio do modelo apresentado é possível realizar consultas multimodais baseadas em contexto, semântica e conteúdo de imagens.

Outro trabalho que também aborda o uso de ontologias, metadados e imagens para recuperação de informações é Addis [5], porém o enfoque é mais específico para o domínio de museus e galerias de artes. A arquitetura da ferramenta permite realizar busca combinando imagens, metadados e conceitos, sendo que existe um módulo específico para navegar entre os conceitos. O processo de busca pode ser realizado de maneira interativa.

Consultas multimodais também podem ser utilizadas sem envolver imagens. Em Ammir [6], por exemplo, é apresentado um sistema que permite busca multimodal em dados de vídeos. O trabalho utiliza realimentação de relevância para melhorar o processo e os resultados da busca.

Outro contexto em que consultas multimodais são utilizadas é para se recuperar músicas. Em Zhang [46], por exemplo, é apresentada uma nova medida de similaridade para música (multimodal e adaptativa). Neste trabalho as informações da música são divididas em informações de conteúdo (referente ao conteúdo do arquivo digital, que tem informações de tempo e timbre) e informações sociais (como título e comentários). Um dos diferenciais deste trabalho é a possibilidade do usuário personalizar a consulta indicando qual conteúdo deve ser utilizado.

O trabalho proposto em Freitas [19], além de utilizar imagens, palavras-chave e ontologias para recuperação, também apresenta como realizar um processo de anotação com-

binando essas informações. Além disso, propõe uma arquitetura para uso de múltiplas ontologias e múltiplos descritores de imagens. A solução apresentada, o ambiente Onto-SAIA, tem a característica importante de sistematizar a anotação de imagens (para que o resultado da abordagem baseado em palavras-chave funcione). A anotação funciona com base em sugestões ligadas a anotações feitas em outras imagens, e anotações relacionadas ao conteúdo da imagem. O protótipo permite busca de imagens por conteúdo, palavras-chaves ou ambos.

No trabalho apresentado em Kesorn [25] utiliza-se um modelo baseado em ontologias para reestruturar os conceitos semânticos presentes em legendas de imagens, buscando atingir um resultado mais eficaz de recuperação destas imagens. O uso de processamento semântico visa superar os problemas de busca textual e adicionar conceitos indiretamente relevantes. Fotos de esportes retiradas da Web são utilizadas nos experimentos, que mostram que técnicas baseadas em ontologias podem melhorar significativamente a efetividade dos sistemas de recuperação de imagens.

Outro trabalho que permite busca por palavras-chave (conceitos) e imagens é Ferecatu [18]. O trabalho utiliza ontologias para gerar conceitos a partir de palavras-chave associadas à imagem. Estes conceitos visam permitir comparar palavras-chave que são diferentes lexicograficamente, porém conceitualmente similares. O sistema também utiliza realimentação de relevância para melhorar os resultados, apresentando melhorias na apresentação dos resultados ao usuário.

Na pesquisa por trabalhos correlatos realizada, verificou-se a grande presença de trabalhos que utilizam a multimodalidade para melhorar os sistemas de recuperação de imagens [36, 37, 39]. Nestes trabalhos os autores partem do princípio que os métodos tradicionais de recuperação de imagens (baseado em texto e baseado em conteúdo), possuem limitações que degradam a qualidade dos resultados. Como solução combinam, de diferentes maneiras, o uso do conteúdo textual e do conteúdo das imagens, para uma melhora na recuperação das imagens.

Os trabalhos anteriores não são voltados a sistemas envolvendo conceitos biológicos. A ferramenta C-DEM [21] manipula dados biológicos e tem como finalidade minerar dados de *Drosophila melanogaster*, conhecidas como moscas de frutas. A mineração/busca ocorre a partir de uma imagem de consulta e/ou genes e/ou palavras-chaves. A Busca por genes ocorre a partir de uma cadeia de caracteres que os identifica. A base de dados utilizada (BDGP) nessa ferramenta possui mais de 70.000 fotografias digitais documentando a expressão de padrões de mais de 3000 genes, anotados com um conjunto padrão de termos. Para se realizar a busca, o C-DEM gera uma representação da base em forma de grafo e executa um algoritmo desenvolvido especialmente para busca no grafo, para encontrar os dados mais próximos da consulta. O processo de busca no C-DEM pode ser realizado de maneira iterativa, permitindo que o usuário refine sua busca com o resultado

de sua consulta inicial. Os autores afirmam que este é o único, até sua publicação, que combina dados multimídia, consulta flexível e ordenação dos resultados para bases de dado biológicas com imagens.

Em Arpah [7] também é apresentado um sistema multimodal no contexto de biodiversidade. O domínio de dados desse trabalho são imagens de Monogenea (grupo de animais). Para armazenar informações de identificação da espécie e das imagens é apresentado uma ontologia taxonômica. No sistema apresentado é possível buscar os dados utilizando termos, por meio de consultas SPARQL.

Na figura 2.3 apresentamos uma tabela que elenca e compara os sistemas multimodais estudados. A tabela resume alguns dos artigos representativos dos sistemas analisados.

2.6 Conclusões

Este capítulo apresentou conceitos que serão utilizados ao longo do trabalho e alguns das pesquisas mais recentes na área de busca multimodal. Podemos verificar que a área de busca multimodal é uma área com grande enfoque, e que na maioria dos casos envolve o uso de conteúdo de imagens somado ao uso de texto. No próximo capítulo apresentamos o sistema proposto que foi desenvolvido ao longo do mestrado.

Artigo	Domínio de Dados	Dados	Consultas possíveis	Uso de Ontologia	Outras Funcionalidades
[8]	Dados Médicos	Imagens (conteúdo) e metadados	Contexto, semântica e conteúdo	-	Permitir uso de operadores de similaridade
[5]	Museus e galerias de artes	Ontologias, metadados e imagens (conteúdo)	Conteúdo, metadados e conceitos	Para armazenar conceitos e relações que são usadas na busca	Navegar entre conceitos, buscar de forma interativa
[6]	NIST TRECVID 2004	Vídeo, imagem (conteúdo das cenas), fonemas falados, conceitos, texto (extraído via OCR), fala (extraída do som), texto do closed caption	Texto	-	Realimentação de relevância que sugere novos termos para incluir na consulta
[46]	Coleções de música extraídas do Youtube	Informações de conteúdo (relacionadas ao som) e informações sociais (título, comentários, etc.)	Texto (informações da música) e som (música em formato digital)	-	Personalização da consulta pelo usuário
[19]	Variado	Imagens, anotações e ontologias	Imagem e texto	Para sugerir anotações e para expandir as palavras-chave fornecidas na consulta	permite utilizar multiplas ontologias e Multiplos descritores de imagens. permite sistematizar a anotação de imagens
[21]	Drosophila melanogaster	Imagens, genes, anotações	Imagem, genes, texto	-	Permite busca iterativa, de maneira que o usuário refine o resultado
[25]	Fotos de esporte	ontologia, imagens, legendas	Texto (conceitos)	Para determinar conceitos presentes nas legendas	Desenvolveu um método para evitar ambiguidade nas ontologias
[7]	Dados de monogeneam	Imagens, ontologias	Termos (via SPARQL)	Para armazenar informações das imagens	-
[18]	Variado (extraído de www.alinari.com)	Imagens, palavras-chave e ontologias	Imagens e texto (conceitos)	Para gerar conceitos chave	-
[36]	Dados Médicos	Palavras chaves visuais e textuais (Imagem e texto)	Texto e Imagem	-	Expansor de consultas
[37]	Variado	Texto (anotações) e imagem (blobs e regiões)	Texto e Imagem	-	Anotação automática esquema de refinamento gradual
[39]	Variado	Texto (legenda das imagens) e imagem	Texto e imagem (resultante da consulta por texto)	-	Fase de pré-processamento que realiza casamento de conceitos (semântico)

Figura 2.3: Comparação de sistemas de consultas multimodais

Capítulo 3

O Sistema Sinimbu

Este capítulo apresenta em detalhes o sistema desenvolvido e implementado ao longo do mestrado. Na seção 3.1 apresentamos uma visão inicial do sistema proposto, o Sinimbu, na seção 3.2 apresentamos as estruturas de dados utilizadas e como elas estão interligadas, na seção 3.3 apresentamos a arquitetura e os módulos que compõem o sistema, na seção 3.4 apresentamos os protótipos de tela, e por fim na seção 3.5 apresentamos a conclusão do capítulo.

3.1 Visão Geral

Este trabalho implementa uma arquitetura de software que possibilita buscar coletas de espécies utilizando combinação de imagens e termos textuais. Os termos textuais podem ser: (a) nomes vulgares e (b) metadados (como termo taxonômico, local de coleta e responsável pela coleta). Por meio dessas modalidades, é possível realizar buscas que não estão normalmente disponíveis nos sistemas de informação de biodiversidade.

O Sinimbu considera um tratamento diferenciado para dois tipos básicos de usuários: leigos e pesquisadores em biodiversidade. Essa diferenciação busca permitir que o sistema seja utilizado por um número maior de usuários, considerando as suas diferenças de interesse e perfis.

A combinação na consulta de imagens e termos textuais visa resolver os problemas apresentados por buscas baseadas somente em texto ou somente em conteúdo de imagens. A busca combinada também representa uma possibilidade nova para que os pesquisadores em biodiversidade analisem os dados armazenados, além de ser um atrativo para que leigos utilizem o sistema, garantindo maior divulgação das informações.

Mais especificamente, o Sinimbu permite que se realizem consultas por: termos que identificam uma espécie ou sua coleta; uma imagem de uma espécie; e também pela combinação de termo e imagem de consulta. A imagem é processada via busca por

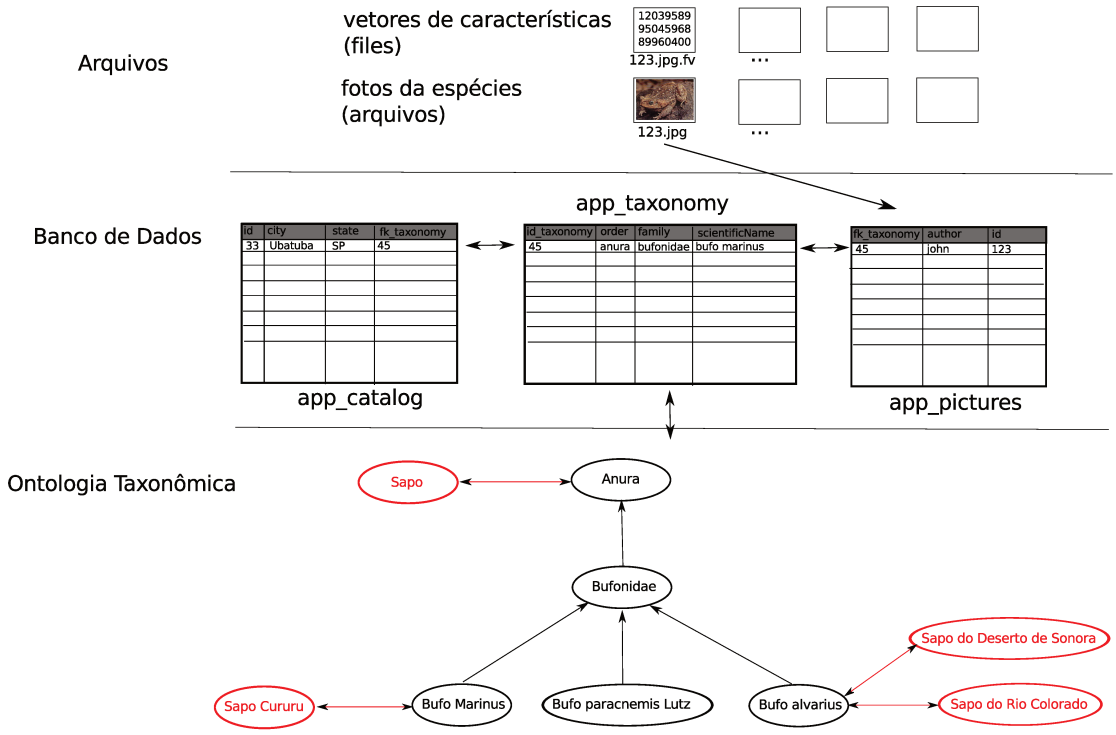

Figura 3.1: Estruturas de Dados

conteúdo. No processamento da consulta textual são utilizados dados de coleta e uma ontologia taxonômica estendida com nomes vulgares para buscar os termos. Os nomes vulgares foram associados aos termos taxonômicos das espécies, podendo estar ligados em diferentes níveis (como Classe, Ordem ou Filo).

3.2 Estruturas de Dados

Os dados utilizados pelo Sinimbu estão armazenados em várias estruturas diferentes, apresentadas na Figura 3.1. Basicamente as estruturas de armazenamento são de 3 tipos: tabelas de um banco de dados PostGreSQL, arquivos (de imagens e seus respectivos descritores) e uma ontologia taxonômica.

O banco de dados é pré-existente à dissertação, havendo sido ampliado com referências às imagens. Tem cerca de 18 tabelas, das quais apenas 5 são utilizadas pelo Sinimbu.

Neste texto, o termo "id de espécie" denota o identificador de um registro da tabela *app_taxonomy* (*id_taxa*), ou seja, o identificador de um registro que contém a descrição taxonômica de uma espécie (Reino, Filo, Classe, etc.). Este é um conhecimento funda-

mental para a ligação e identificação dos registros. O *id_taxa* é único, sendo um número sequencial criado automáticamente na construção da tabela *app_taxonomy* para identificação unívoca de uma espécie.

As tabelas do banco de dados usadas pelo Sinimbu, cujo esquema estão no apêndice A, são as seguintes:

- **app_taxonomy:** tabela que contém descrições taxonômicas das espécies armazenadas no banco de dados. Cada registro desta tabela contém todos os graus taxonômicos da espécie além de armazenar o ano e o autor responsáveis por identificar a espécie. Um registro desta tabela contém os seguintes campos :
  - idTaxa
  - Kingdom
  - Phylum
  - Subphylum
  - Class
  - SubClass
  - Order
  - Suborder
  - Superfamily
  - Family
  - SybFamily
  - SuperTribe
  - Tribe
  - SubTribe
  - Genus
  - SpecificEpithet
  - AuthorYearOfScientificName
- **app_catalog:** armazena os registros de coleta propriamente ditos, com informações como: local, responsável pela coleta e data. Também contém chave estrangeira que liga a coleta à espécie correspondente da tabela *app_taxonomy*. Um registro desta tabela contém os seguintes campos :
  - idCatalog

  - institutionCode
  - collectionCode
  - catalogNumber
  - catalogDate
  - dateIdentified
  - conservationMeans
  - basisOfRecord
  - previousCatalogNumber
  - typeMaterial
  - patternColor
  - lifeStage
  - sex
  - size
  - typeSubstrate
  - fieldNumber
  - publication
  - notes
  - earlistDateCollected
  - latestDateCollected
  - indivudalCount
  - fk_cataloguedBy
  - fk_location
  - fk_taxa
  - fk_collector
  - fk_identifier
  - fk_meth

- **app_pictures:** responsável por armazenar informações sobre as imagens (fotos de espécies) e a ligação entre as imagens e a tabela de taxonomia. Contém informações sobre a fotografia como autor e se ela é pública ou não, e chave estrangeira para a tabela de taxonomia. Um registro desta tabela contém os seguintes campos :

  - id
  - fileName
  - label
  - idTaxa
  - visible
  - classification
  - author
  - idCatalog

- **app_responsible:** armazena informações sobre um pesquisador, no contexto da dissertação apenas pesquisadores responsáveis por coletas. Um registro desta tabela contém os seguintes campos :

  - id
  - full_name
  - abbreviated_name
  - department
  - laboratory
  - CPF
  - CNPJ

- **app_location:** tabela que armazena informações sobre uma localidade geográfica, como país e estado. Um registro desta tabela contém os seguintes campos :

  - id
  - name
  - county
  - stateProvince
  - country
  - continentOcean
  - reference

Os arquivos de imagens estão divididos em dois grupos: as fotos das espécies e os arquivos com os vetores de características extraídos das fotos. Cada foto é armazenada em um arquivo próprio, que recebe como nome o número de identificação da imagem; este

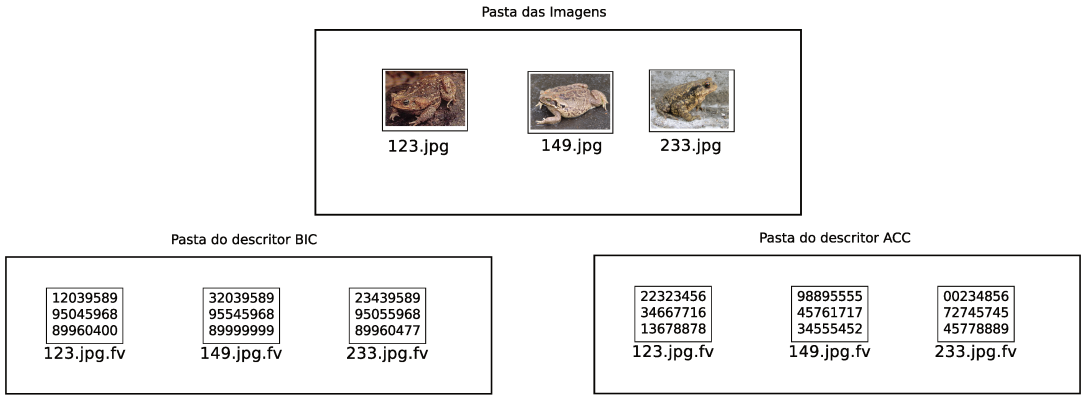

Figura 3.2: Organização das imagens e seus descritores

nome é armazenado na tabela de imagens (*app_pictures*), sendo seu identificador único. Cada imagem está associada a N arquivos com os vetores de características extraídos previamente, onde N é o número de descritores disponíveis no sistema. Os arquivos de vetores de características possuem o mesmo nome do arquivo de imagem, porém são separados em pastas (uma para cada descritor) com extensão ".fv".

Assim, por exemplo, como mostra a figura 3.2, ao arquivo de imagem 123.jpg correspondem os arquivos de descritor 123.jpg.fv, presentes nas pastas correspondentes aos descritores BIC e ACC. Cada nova imagem incluída é processada para extração dos arquivos de vetores de características, inseridos nas pastas correspondentes. A adoção de um novo descritor exige apenas processamento de todos os arquivos da pasta de imagens e a inclusão de uma nova pasta para aquele descritor. Esta estruturação permite estender arbitrariamente o número de descritores utilizados.

O nome do arquivo de imagem é o identificador da imagem. Com isto, é possível navegar de uma imagem aos seus descritores e vice-versa. As chaves estrangeiras de *app_pictures* e *app_taxonomy* permitem navegar de uma espécie às imagens associadas e destas aos registro taxonômicos correspondentes.

3.3 Construção da Ontologia Taxonômica

De acordo com Cullot [11], uma ontologia taxonômica possui uma estrutura hierárquica entre seus termos, de maneira que cada termo leva à um conceito mais especializado que seu pai. A ontologia para este trabalho é central para o processamento de consultas por nome vulgar.

Como ontologias capturam o conhecimento do domónio, são fundamentais para o tratamento de consultas em sistemas, principalmente quando se trata de novos domínios do conhecimento. A vantagem de armazenamento em OWL foi possibilitar associações

diversas com nomes vulgares, de forma mais natural, tendo em vista que um nome vulgar pode associado à vários níveis taxonômicos. Ressalta-se que não foi feito um mero mapeamento entre a tabela e uma árvore, mas sim a expansão da estrutura tabular com conhecimento do domínio.

A ontologia taxonômica foi gerada a partir dos registros da tabela *app_taxonomy*, extraindo-se os termos taxonômicos desta e inserindo-os na ontologia, mantendo a hierarquia dos termos. No apêndice, a tabela A.1, mostra o esquema completo da tabela *app_taxonomy* que serviu de base para a construção da ontologia. A construção da ontologia usou informações do banco de dados e também conhecimento do domónio (como por exemplo a hierarquia taxonômica).

A seguir, associou-se manualmente nomes vulgares aos termos taxonômicos, em diferentes níveis. Esta inserção manual foi necessária pela falta de tabelas de correspondência adequadas em português. A parte inferior da figura 3.1 exemplifica tal associação: a Ordem *Anura* corresponde ao nome vulgar *sapo*, e o binômio *Bufo alvarius* está associado aos nomes vulgares *Sapo do Deserto de Sonora* e *Sapo do Rio Colorado*. Cada termo da taxonomia, assim, pode ter vários nomes vulgares correspondentes por meio da relação de equivalência da ontologia (relação `EquivalentClasses` em OWL). A taxonomia é utilizada para descobrir a correspondência entre nomes vulgares e científicos, por meio de uma consulta SPARQL. Por meio do termo taxonômico resultante pode-se recuperar as informações da espécie na tabela de taxonomias (*app_taxonomy*), realizando-se uma busca simples.

3.4 Arquitetura

A Figura 3.3 mostra a arquitetura do sistema de consulta, que estende a proposta de Freitas [19], aproveitando os serviços de suporte já desenvolvidos, Serviço de Recuperação de Imagens Por Conteúdo e Serviço de Coletas. Nela estão presentes 4 camadas: Serviços de Armazenamento, Serviços de Suporte, Serviço de Busca Multimodal e Interface. O Sinimbu, destacado na figura, abrange o Módulo de Busca Multimodal e o protótipo de interface, com implementação dos módulos indicados, a saber: Extrator de Taxonomias e todos os módulos dentro da caixa “Busca Multimodal”. Também faz parte do trabalho O módulo Programação Genética (GP Framework), foi implementado por outros alunos.

O fluxo de execução é o seguinte: a interface recebe uma solicitação de busca (1). Tal solicitação é passada ao serviço de busca multimodal (2), que encaminha a solicitação para o Gerenciador de Busca Combinada (3). Este invoca os módulos de recuperação (4), que processam a busca a partir dos serviços de suporte (5). Recebidos os resultados parciais o Gerenciador de Busca Combinada consulta o Gerenciador de Tipos de Usuário (7), que determina padrões de visualização de resultados, e integra os resultados que serão

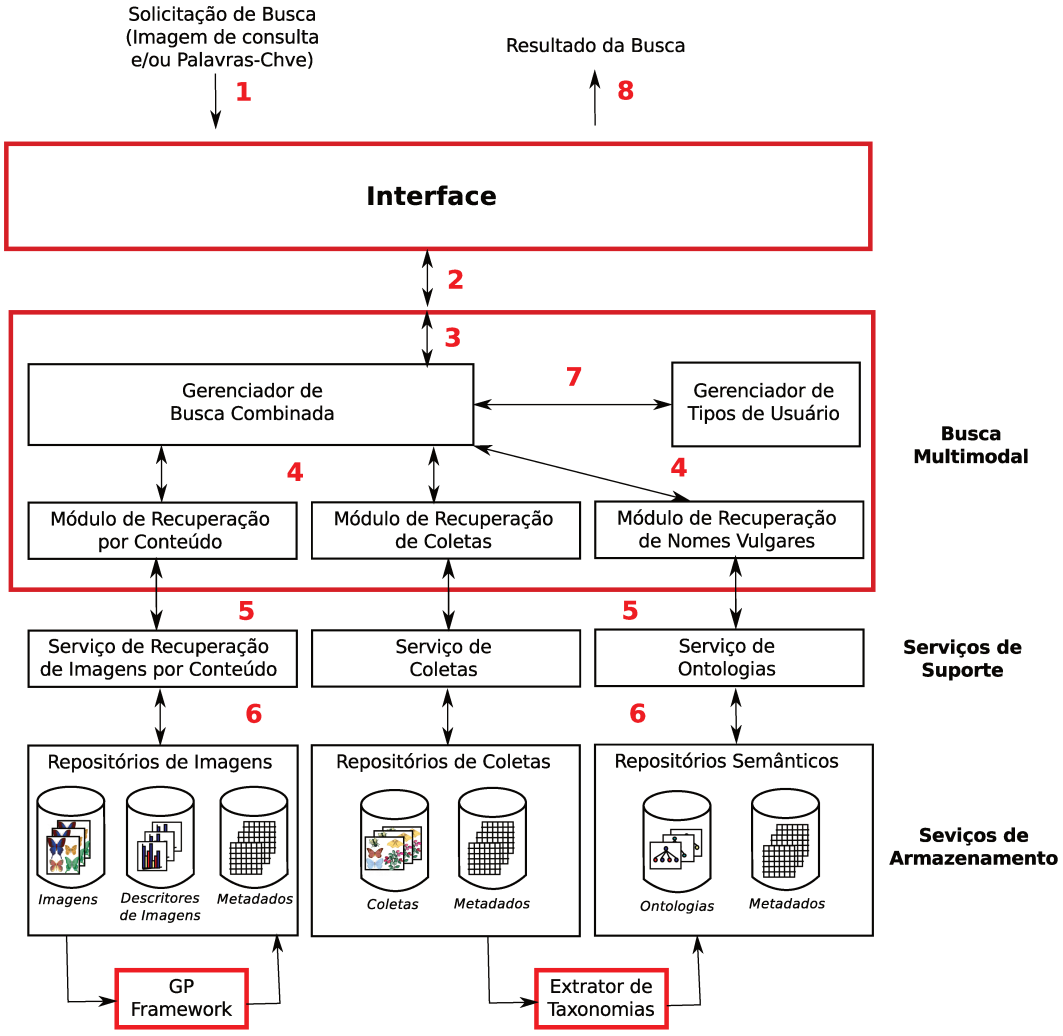

Figura 3.3: Arquitetura do Sinimbu

encaminhados para a Interface, que os apresenta (8). A figura também mostra os 3 tipos de repositórios considerados (6): de imagens, de coletas e semântico (ontologias).

Os módulos de recuperação – Por Conteúdo, De Coletas e De Nomes Vulgares – têm a função de encaminhar as requisições de busca aos módulos de suporte e tratar o retorno destas. O Módulo de Busca Combinada processa os resultados e os integra para repassá-los à Interface.

A seguir apresentamos detalhadamente cada um dos módulos que foram desenvolvidos durante a pesquisa.

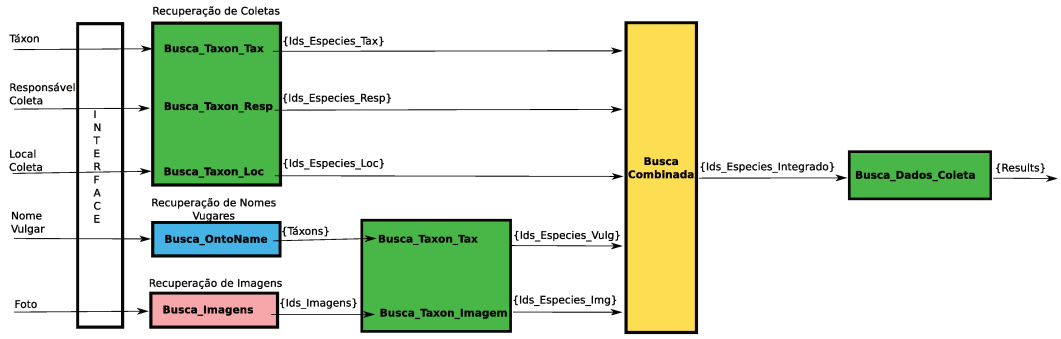

Figura 3.4: Fluxo de Execução dos Módulos

3.4.1 Fluxo de invocação dos módulos

A figura 3.4 detalha o fluxo dos dados dentro dos módulos do Sinimbu. Como podemos ver na figura, o usuário pode entrar com 5 tipos de parâmetros, 4 textuais (Taxon, Responsável por Coleta, Local da Coleta e Nome Vulgar) e 1 de Imagem (Foto da espécie). Esses dados de entrada são repassados para os respectivos módulos.

Taxon, Responsável por Coleta e Local de Coleta vão para o Módulo de Recuperação de Coletas, onde passam respectivamente pelas funções Busca_Taxon_Tax, Busca_Taxon_Resp e Busca_Taxon_Loc. Cada uma dessas funções retorna uma lista de identificadores de espécie das espécies que possuem dados similares aos de entrada: {Ids_Especies_Tax}, {Ids_Especies_Resp} e {Ids_Especies_Loc}.

O Nome Vulgar é encaminhado para o Módulo de Recuperação de Nomes Vulgares, que executa a função Busca_OntoName e retorna uma lista de Taxons que são sinônimos ao Nome Vulgar de entrada. A Foto da espécie é recebida pelo Módulo de Recuperação de Imagens, que por meio da funçao Busca_Imagens retorna uma lista de identificadores das imagens similares à imagem de consulta, {Ids_Imagens}. Os Taxons, resultantes da busca por Nome Vulgar, e os identificadores de imagem, resultantes da busca usando a foto, vão também para o Módulo de Recuperação de Coletas, passando pelas funçoes Busca_Taxon_Tax e Busca_Taxon_Imagem respectivamente e gerando como resultado uma lista de identificadores de espécie para cada: {Id_Especie_Vulg} e {Id_Especie_Img}.

Neste ponto da execução, o sistema possui uma lista de identificadores de espécie para cada dado de entrada. Essas listas vão para o Módulo de Busca Combinada, que as combina ordenando as espécies pela similaridade com os dados de entrada, e gerando uma lista única de identificadores de espécie. Esta última passa então pela função Busca_Dados_Coleta do Módulo de Recuperação de Coletas para retornar a lista de informações de cada espécie.

As próximas subseções descrevem detalhadamente cada um dos módulos e suas funções.

3.4.2 Módulo de Recuperação de Coletas

O Módulo de Recuperação de Coletas é responsável por recuperar os dados de registros de coletas armazenados em tabelas do PostGreSQL. São disponibilizadas 5 funções neste módulo, que permitem acessar os dados das tabelas, a saber:

- **Busca_Taxon_Tax:** Esta função recebe um taxón, que pode ser referente a qualquer nível taxonômico, e retorna registros de coleta que possuam nome taxonômico em que o termo buscado seja uma sub cadeia. O algoritmo `Busca_Taxon_Tax` descreve a função, e basicamente se resume a uma consulta SQL que verifica se o termo consultado é similar a algum nível taxonômico armazenado. A função de similaridade utiliza o predicado `ILIKE` do PostGreSQL;

- **Busca_Taxon_Loc:** Esta função recebe uma localidade (no formato textual, como o nome de uma cidade) e procura dados de coleta em que o termo informado seja uma sub cadeia da sua localidade e retorna então as espécies referentes a essas coletas. O algoritmo `Busca_Taxon_Loc` descreve a função e é basicamente uma consulta SQL;

- **Busca_Taxon_Resp:** Esta função recebe um nome de responsável por coleta, e verifica por coletas realizadas por ele. A partir destas coletas, retorna as espécies referentes a elas. O algoritmo `Busca_Taxon_Resp` descreve a função;

- **Busca_Taxon_Imagem:** Esta função recebe um conjunto de identificadores de imagens e retorna um conjunto de identificadores de espécies referentes às imagens. Seu algoritmo é descrito em `Busca_Taxon_Imagem`;

- **Busca_Dados_Coleta:** Esta função recupera registros de coleta e imagens da espécie a partir de um vetor de identificadores da tabela *app_taxonomy* (ou seja, vetor contendo o id de espécie da tabela), que identificam espécies. Este algoritmo, apresentado em `Busca_Dados_Coleta`, verifica se uma imagem é pública ou não, e se pode ser exibida.

BUSCA_TAXON_TAX

ENTRADA:
T: termo taxonômico buscado
SAÍDA:
Id_E: Vetor com Ids dos Registros de informação taxonômica da tabela app_taxonomy em que ILIKE(app_taxonomy.Phylum, Tax) ou ILIKE(app_taxonomy.Class, Tax) ou ILIKE(app_taxonomy.Order, Tax) ou ILIKE(app_taxonomy.Family,Tax) ou ILIKE(app_taxonomy.Genus + app_taxonomy.SpecificEpithet, Tax)
Dados Utilizados:
app_taxonomy: tabela contendo informação taxonômica

BUSCA_TAXON_LOC

ENTRADA:
L: Localidade a ser buscada
SAÍDA:
Id_E: Vetor com Ids dos Registros de informação taxonômica da tabela app_taxonomy que possuam localidade de coleta similar a L
Dados Utilizados:
app_taxonomy: tabela contendo informação taxonômica
app_collect: tabela contendo registros de coleta
app_location: tabela contendo informação de localidade

1 L_similar = ILIKE(app_location.county, L) ou ILIKE(app_location.stateProvince, L) ou ILIKE(app_location.country, L) ou ILIKE(continentOcean, L) ou ILIKE(name, L) ou ILIKE(reference, L)
2 $Id\_E = \Pi_{id\_taxa}(\Pi_{id\_location}(L\_similar) \bowtie app\_collect) \bowtie app\_taxonomy$

BUSCA_TAXON_RESP

ENTRADA:
R: Responsável a ser buscado
SAÍDA:
Id_E: Vetor com Ids dos Registros de informação taxonômica da tabela app_taxonomy que possuam responsável de coleta similar a R
Dados Utilizados:
app_taxonomy: tabela contendo informação taxonômica
app_collect: tabela contendo registros de coleta
app_responsible: tabela contendo informação de responsáveis de coleta

1 R_similar = ILIKE(app_responsible.full_name, R)
2 $Id\_E = \Pi_{id\_taxa}(\Pi_{id\_responsible}(R\_similar) \bowtie app\_collect) \bowtie app\_taxonomy$

BUSCA_TAXON_IMAGEM

ENTRADA:
Id_I: vetor k dimensional contendo id's de imagens
SAÍDA:
Id_T: vetor k dimensional contendo id's da tabela app_taxon
Dados Utilizados:
app_pictures: tabela com informações das imagens

1 Id_T = $\Pi_{idTaxa}(Id\_I \bowtie app\_pictures)$

BUSCA_DADOS_COLETA

ENTRADA:
Id_T: vetor contendo id's de espécies
SAÍDA:
Result: estrutura de dados contendo imagens e registros de coleta a serem apresentados ao usuário, com a seguinte estrutura:
Result[x], onde x é uma espécies
Foto: img - todas as imagens da espécie X
Col: coleta - todos os registros de coleta da espécie X
Dados Utilizados:
Tabelas app_collect e app_pictures

1 **for** cada Id em Id_T
// Só repassa as imagens que são públicas
2 **if** $reg\_foto.publico == true$, onde reg_foto é o registro de app_pictures com $Id\_taxa == Id$
3 $Result[Id].Foto = reg\_foto$
4 **else**
5 $Result[Id].Foto = \emptyset$
6 $Result[Id].Col = regs\_coleta$, onde regs_coleta são registros de app_collect que tenham $Id\_Taxa == Id$

3.4.3 Módulo de Recuperação de Imagens por Conteúdo

Este módulo é responsável por recuperar imagens armazenadas nos repositórios a partir de uma imagem de consulta. Implementa a função Busca_Imagens, cujo algoritmo é `Busca_Imagens`.

BUSCA_IMAGENS

ENTRADA:
Img : imagem de consulta
N: quantas imagens serão retornadas
SAÍDA:
Tax_Dist: Matriz k x 2, onde k é o número de imagens em app_pictures, indicando a distância de Img à imagem correspondente, sendo:
Tax_Dist[j,1] - identificador taxonômico
Tax_Dist[j,2] - distância da Img à imagem do animal de identificador taxonômico Tax_Dist[j,1]
Dados Utilizados:
Func: função de distância utilizada para cálculo de similaridade
Tabela app_pictures

1 Inicializar Vet_Dist
2 $Contador = 0$
// Extraem-se as características de Img
// segundo todos os descritores disponíveis.
3 **for** cada descritor Desc disponível
4 Extrai_Carac(Img, Desc)
5 **for** cada Img_Atual em app_pictures
6 $Vet\_Dist[Contador].Id\_Taxa = Img\_Atual.Id\_Taxa$
7 $Vet\_Dist[Contador].Dist = Func(Img, Img\_Atual)$
8 $Contador = Contador + 1$
9 Ordenar Vet_Dist pelo atributo Dist
10 **return** as N primeiras posições de Vet_Dist

Quando uma imagem (Img) é repassada para este módulo ocorre a extração de suas características segundo todos os descritores disponíveis no sistema. As características da imagem extraídas são então comparadas com as características das imagens armazenadas no sistema, utilizando uma função que combina as diversas funções de distância disponíveis. Tal função é gerada pelo módulo GP Framework, explicado detalhadamente na seção 3.4.7. A seguir o módulo de Recuperação ordena os resultados segundo sua similaridade e retorna a lista de identificadores das N imagens mais próximas a imagem.

3.4.4 Módulo de Recuperação de Taxonomia

Este módulo é responsável por recuperar taxons que estão associados a um nome vulgar. É implementado pela função Busca_OntoName que a partir de um termo contendo o nome

vulgar, retorna uma lista de termos taxonômicos equivalentes, implementada como uma chamada a um comando SPARQL:

```
SELECT ?equivalentClass WHERE {?equivalentClass owl:equivalentClass db:  name
}
```

onde *name* é o nome vulgar buscado.

Esta consulta SPARQL é processada na ontologia taxonômica. O resultado da consulta é transformado em um vetor com termos taxonômicos equivalentes a *name*.

3.4.5 Módulo de Busca Combinada

O Módulo de Busca Combinada desempenha um papel central no sistema. Seu algoritmo (`Busca_Combinada`) recebe como entrada todos os dados buscados: imagem (Img), termo taxonômico (Tax), responsável pela coleta (Resp), local de coleta (Loc) e nome vulgar (Name). Também recebe qual peso cada um destes terá na busca (se não informado pelo usuário, pesos são iguais). O resultado é obtido usando os serviços de suporte, a partir das funções: Busca_Taxon_Tax, Busca_Taxon_Resp, Busca_Taxon_Loc, Busca_Taxon_Imagem, Busca_Imagem, Busca_OntoName e Busca_Dados_Coleta.

Os resultados das buscas são vetores de id's de espécies, um para cada dado de entrada: Id_T (Taxon),Id_R (Responsável pela coleta),Id_L (Local da Coleta),Id_I (Imagem),Id_N (Nome vulgar) . Terminada a execução das buscas, é necessário que essas informações sejam tratadas de maneira integrada (já que estão em acervos de dados diferentes). A estrutura Id_I, que é uma lista de identificadores de imagens, é ligada aos dados taxonômicos através da função Busca_Taxon_Imagem (linha 24). Para os nomes vulgares, Id_N que é uma lista de termos taxonômicos a ser ligado aos dados taxonômicos usando Busca_Taxon_Tax (linha 26).

Encontrados todos os identificadores de registros que satisfaçam às condições de entrada, é computada a pontuação para cada registro (linhas 27 a 41). A seguir, ordena-se o vetor de identificadores de registros (Registro) pela pontuação. Recupera-se os dados de coleta via a chamada do *Busca_Dados_Coleta*, e retorna-se essa lista como resultado.

BUSCA_COMBINADA

ENTRADA:
Tax: termo taxonômico referente à espécie buscada
Resp: responsável por uma coleta referente à espécie buscada
Loc: Localidade de uma coleta referente à espécie buscada
Img: imagem de consulta da espécie buscada
Name: Nome Vulgar referente à espécie buscada
Peso_Tax , Peso_Resp, Peso_Loc, Peso_Img, Peso_Name:
valor inteiro de peso para cada dado de entrada
SAÍDA:
{Registros} {< Id_Especie_cadastrada, Pontuacao >} onde Pontuacao é a
ponderação obtida na busca. A dimensão deste vetor é o número de espécies distintas
no banco de dados, ou seja o número de linhas da tabela app_taxonomy

1 **if** $Peso\_Tax \neq \emptyset$
2 $Peso\_Tax = 1$
3 **if** $Peso\_Resp \neq \emptyset$
4 $Peso\_Resp = 1$
5 **if** $Peso\_Loc \neq \emptyset$
6 $Peso\_Loc = 1$
7 **if** $Peso\_Img \neq \emptyset$
8 $Peso\_Img = 1$
9 **if** $Peso\_Name \neq \emptyset$
10 $Peso\_Name = 1$
11 **for** cada registro taxonômico RT em app_taxonomy
12 $Registros[RT.id][Pontuacao] = 0$
13 **if** $Peso\_Tax \neq 0$
14 $Id\_T = Busca\_Taxon\_Tax(Tax)$
15 **if** $Peso\_Resp \neq 0$
16 $Id\_R = Busca\_Taxon\_Resp(Resp)$
17 **if** $Peso\_Loc \neq 0$
18 $Id\_L = Busca\_Taxon\_Loc(Loc)$
19 **if** $Peso\_Img \neq 0$
20 $Id\_I = Busca\_Imagem(Img)$
21 **if** $Peso\_Name \neq 0$
22 $Id\_N = Busca\_OntoName(Name)$
23 **if** $Id\_I \neq \emptyset$
24 $Id\_I\_Tax = Busca\_Taxon\_Imagem(Id\_I)$
25 **if** $Id\_N \neq \emptyset$
26 $Id\_N\_Tax = Busca\_Taxon\_Tax(Id\_N)$
27 **if** $Id\_I\_Tax \neq \emptyset$
28 **for** cada Id de Id_I_Tax
29 $Registros[Id].Pontuacao = Registros[Id].Pontuacao + Peso\_Img$
30

30 **if** $Id\_N\_Tax \neq \emptyset$
31 **for** cada Id de Id_N_Tax
32 $Registros[Id].Pontuacao = Registros[Id].Pontuacao + Peso\_Name$
33 **if** $Id\_T \neq \emptyset$
34 **for** cada Id de Id_T
35 $Registros[Id].Pontuacao = Registros[Id].Pontuacao + Peso\_Tax$
36 **if** $Id\_R \neq \emptyset$
37 **for** cada Id de Id_T
38 $Registros[Id].Pontuacao = Registros[Id].Pontuacao + Peso\_Resp$
39 **if** $Id\_L \neq \emptyset$
40 **for** cada Id de Id_T
41 $Registros[Id].Pontuacao = Registros[Id].Pontuacao + Peso\_L$
42 Ordenar Registros por Pontuação
43 $Resultado = Busca\_Dados\_Coleta(Registros)$
44 **return** Resultado

3.4.6 Gerenciador de Tipos de Usuário

A partir dos dados informados na consulta, o Gerenciador de Tipos de Usuário determina quem está utilizando o sistema. A hipótese é que leigos utilizem nome vulgar para busca, e pesquisadores utilizem nomes científicos.

O Módulo de Tipos de Usuários trata os diferentes perfis de usuário do sistema, definindo parâmetros para a consulta (como pesos diferenciados para cada modalidade), e também configurações para a exibição dos resultados. Na versão atual do sistema, tem a funcionalidade única de determinar como os dados serão exibidos dependendo do tipo de usuário que está usando o sistema.

Para leigos, as informações exibidas serão mais focadas nas características da espécie, e darão preferência para fotos da espécie viva, como ilustrado na Figura 3.6. Caso seja um pesquisador, são retornadas informações técnicas e as imagens são preferencialmente de holótipos e parátipos (que são os espécimes que definiram a espécie), como ilustrado na Figura 3.7.

3.4.7 GP Framework

O Módulo de Programação Genética é ativado quando ocorre nova inserção de imagens, para atualizar a função de combinação de descritores utilizada. Diferente dos módulo anteriores, é executado de forma assíncrona, ou seja, separado da consulta.

Relembrando da seção 2.4.2 do capítulo 2, um descritor de imagem é um par $< d, f >$ onde d é um vetor de características (cor, textura ou forma), e f uma função de distância.. É possível fazer a recuperação das imagens utilizando apenas um descritor. Entretanto, combinar alguns descritores que encapsulam informações diferentes pode produzir resultados melhores (noção de descritor composto). A ideia mais simples para essa combinação é fazer um cálculo básico sobre as distâncias calculadas individualmente, como por exemplo uma média aritmética das distâncias dos descritores. Nesse caso, um descritor composto por 3 descritores seria, por exemplo: f(d_1ab, d_2ab, d_3ab) = $(d_1ab + d_2ab + d_3ab)/3$, sendo d_iab, a distância entre as imagens "a" e "b" utilizando o descritor d_i.

Quando ocorre a inserção de uma nova imagem no sistema, o GP Framework calcula uma função de similaridade mais complexa, utilizando indivíduos de programação genética. O GP Framework foi implementado por alunos de doutorado do laboratório RECOD. Um desses alunos (Jefferson Santos) pré-processou todas as imagens do Sinimbu, permitindo assim as consultas por conteúdo da dissertação.

3.4.8 Módulo de Extração de Taxonomia

A ontologia taxonômica é construída a partir dos dados armazenados nas tabelas de espécies (*app_taxonomy*). Assim como o GP Framework, o Módulo de Extração de Taxonomia também é executado de forma assíncrona. Este módulo deve ser ativado pelo usuário quando se inserem novos registros de coletas. Se novas espécies forem registradas quando ocorre a inserção desses dados de coleta, o módulo de extração de taxonomia atualiza as novas espécies na ontologia taxonômica. O algoritmo correspondente está descrito em `Extrai_Taxonomia`.

EXTRAI_TAXONOMIA

SAÍDA:

Arq: Arquivo owl resultante com a ontologia taxonômica

DADOS UTILIZADOS:

app_taxonomy: Tabela com dados taxonomicos das espécies cadastradas

```
1  Gravar cabeçalho owl em Arq
2  for cada registro R em app_taxonomy
3        classeAnt = 'Taxonomia'
4        for cada atributo i de R
5              if (R[i] ≠ ∅)
6                    classeAtual = R[i]
7                    Grava em Arq a classe classeAtual
8                    Grava em Arq a relação subClassOf de classeAtual e classeAnt
9                    classeAnt = classeAtual
10 Grava fim de arquivo owl em Arq
```

3.5 Protótipo de Tela

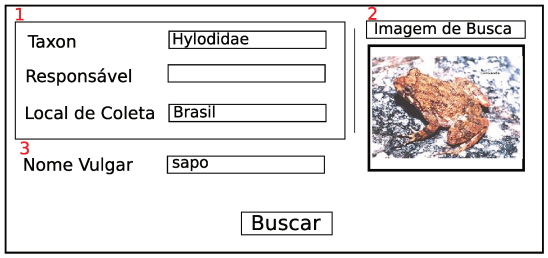

Figura 3.5: Tela de consulta

A Figura 3.5 mostra o protótipo de tela de consulta com um exemplo de consulta que combina imagem, local de coleta e nome comum. A parte 1 identifica onde o usuário pode fornecer metadados relativos ao registro de coleta da espécie: Taxon (uma classificação taxonômica, por exemplo a Família *Hylodidae*), Nome do responsável pela coleta e/ou local de coleta (no caso Brasil). No item 2 o usuário fornece a imagem de consulta. No item 3 o usuário pode fornecer um nome vulgar de consulta (por exemplo, *sapo*).

O resultado mostra ao usuário os dados encontrados, ordenados pela semelhança com os dados de entrada. As Figuras 3.6 e 3.7 ilustram os dois possíveis tipos de visualização dos resultados para, respectivamente, leigos e pesquisadores.

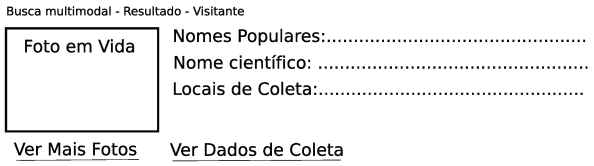

Figura 3.6: Protótipo da tela de resultado de um registro quando o perfil for de leigo

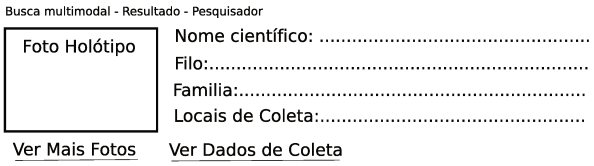

Figura 3.7: Protótipo da tela de resultado de um registro quando o perfil for de pesquisador

3.6 Conclusões

Este capítulo descreveu a arquitetura do sistema de consultas multimodais, detalhando estruturas de dados, algoritmos e ilustrando telas de interface. A proposta permite a extensão dos dados e requisitos de consultas, com estruturas separadas para imagens, dados taxonômicos e dados de coleta. O próximo capítulo analisa aspectos de implementação da arquitetura.

Capítulo 4

Aspectos de Implementação

Este capítulo apresenta os detalhes de implementação do Sinimbu. Na seção 4.1 apresentamos os dados disponíveis e utilizados neste projeto, na seção 4.2 descrevemos as tecnologias utilizadas, na seção 4.3 apresentamos uma série de possibilidades de consultas que o Sinimbu disponibiliza, na seção 4.4 apresentamos um exemplo de sessão de usuário utilizando o Sinimbu e por fim concluímos o capítulo na seção 4.5.

4.1 Dados Utilizados

Os dados disponíveis para este trabalho são basicamente dados de coleta e imagens de espécies. Ao final de 2010, o banco de dados do Museu de Zoologia continha aproximadamente 40000 registros de coletas, referentes a aproximadamente 5000 espécies, com metadados sobre espécimes armazenados em um banco de dados PostGreSQL. Há aproximadamente 1200 imagens já catalogadas.

A ontologia taxonômica constituída tem 5 mil espécies, e foi gerada a partir dos dados de coleta utilizando o módulo de extração de taxonomias. Foram inseridos manualmente aproximadamente 200 nomes vulgares para as espécies. As fontes de dados consultadas para isto foram sites de biodiversidade e documentos de espécies do Brasil.

4.2 Tecnologias Utilizadas

A implementação do sistema Sinimbu envolveu várias tecnologias, escolhidas para suprir da melhor forma as demandas de cada módulo.

Para o armazenamento dos dados de coleta foi utilizado o SGBD PostGreSQL com a extensão PostGIS que garante suporte a dados geográficos. Este é o SGBD definido para implementação do museu virtual de zoologia da UNICAMP. Os módulos GP Framework

e de Extração de Taxonomia são executados à parte, manualmente, não havendo sido prevista interface com o usuário ou sua invocação automática quando o banco de coletas ou de imagens é atualizado.

A ontologia taxonômica foi armazenada usando o formato OWL, gerado pelo extrator de características. Os nomes vulgares foram inseridos na ontologia utilizando a ferramenta Protégé.

Os arquivos de imagem foram armazenados em JPG, um formato que garante grande facilidade de exibição. Os arquivos com os descritores foram armazenados com arquivos binários contendo os vetores de características.

Os Módulos que compõem o serviço de busca multimodal foram desenvolvidos utilizando a linguagem JAVA e bibliotecas auxiliares. O Módulo de Recuperação de conteúdo utilizou tecnologia JNA (Java Native Access) para acessar os descritores e funções de distâncias, que estão implementadas em C. O Módulo de Recuperação de Coleções utilizou o JDBC (Java Database Connectivity) para garantir acesso às tabelas PostGreSQL. O Módulo de Recuperação de Taxonomias utilizou o framework JENA, que garantiu acesso à ontologia taxonômica e o uso de SPARQL.

Os dois módulos externos ao serviço de busca multimodal, Extração de taxonomias e GP Framework, foram desenvolvidos utilizando outras tecnologias. O primeiro foi desenvolvido utilizando a linguagem Python, pois o mestrando já tinha bom conhecimento prévio da linguagem. O segundo, desenvolvido no laboratório RECOD, utilizou C para garantir alto desempenho.

A figura 4.1 mostra as técnologias associadas aos módulos implementados.

4.3 Consultas disponibilizadas

Esta seção apresenta exemplos das consultas que podem ser realizadas no Sinimbu.

4.3.1 Consulta usando metadados

A consulta usando metadados consiste no preenchimento de ao menos um dos atributos: Taxon (identificador científico da espécie), local de coleta ou responsável pela coleta. Suponha por exemplo que o usuário forneça apenas o atributo taxon, com valor "Proceratophrys". O sistema constata que só se solicita busca usando metadados e invoca o módulo de recuperação de coletas, que executa o seguinte comando SQL:

```
SELECT DISTINCT idTaxa FROM app_taxonomy WHERE (app_taxonomy.Phylum ILIKE
'%Proceratophrys%' or app_taxonomy.Class ILIKE '%Proceratophrys%'
or app_taxonomy.Order ILIKE '%Proceratophrys%' or app_taxonomy.Family ILIKE
```

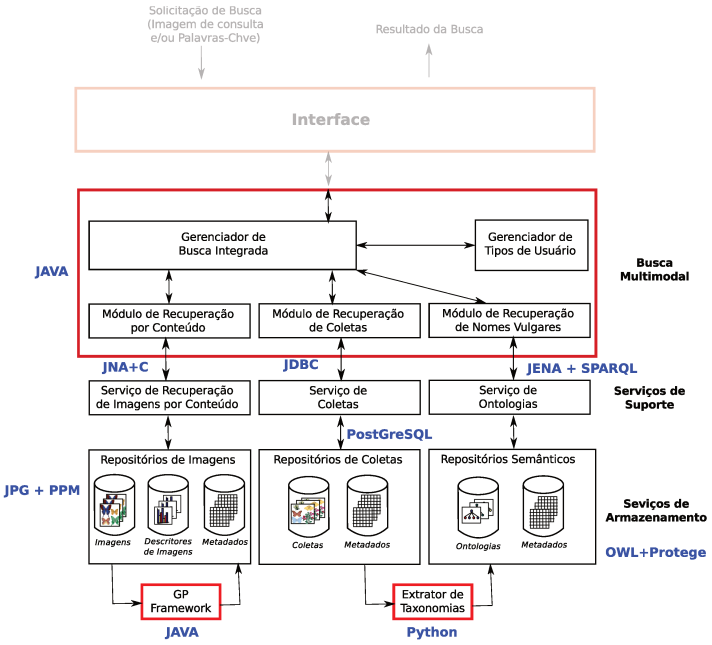

Figura 4.1: Imagens exemplo

```
'%Proceratophrys%' or app_taxonomy.Genus ILIKE '%Proceratophrys%'
or app_taxonomy.SpecificEpithet ILIKE '%Proceratophrys%')
```

O resultado do comando SQL é uma lista de identificadores de taxonomia, que a seguir são ligados aos dados de coleta e às imagens. A lista dos registros de coleta e de imagens, agrupados por espécie, é então encaminhada para o Módulo Gerenciador de Busca Combinada, que os exibe na interface.

4.3.2 Consulta por nome vulgar

A consulta por nome vulgar é realizada de maneira similar à consulta por metadados. O usuário preenche um campo com um nome vulgar, sendo exibidos os dados de espécie que possuem nome vulgar similar ao fornecido pelo usuário.

Em uma situação exemplo onde o usuário fornece o nome vulgar "sapo", o termo é encaminhado para o Gerenciador de Busca Combinada, que o encaminha para o Módulo de Recuperação de Taxonomia. Este último gera a seguinte consulta SPARQL:

```
SELECT ?equivalentClass WHERE {?equivalentClass owl:equivalentClass db:  sapo
}
```

A consulta é executada na ontologia taxonômica, e os resultados são termos taxonômicos equivalentes ao nome vulgar "sapo", no caso a ordem "anura" (ver Figura 3.1). A seguir o termo "anura" é encaminhado novamente ao gerenciador de busca combinada, que busca

o termo “anura” nos dados de coleta, por meio do seguinte comando SQL:

```
SELECT DISTINCT idTaxa FROM app_taxonomy WHERE (app_taxonomy.Phylum ILIKE
'%anura%' or app_taxonomy.Class ILIKE '%anura%' or app_taxonomy.Order ILIKE
'%anura%' or app_taxonomy.Family ILIKE '%anura%' or app_taxonomy.Genus ILIKE
'%anura%' or app_taxonomy.SpecificEpithet ILIKE '%anura%')
```

Os resultados da consulta SQL, uma lista de identificadores de espécie, idTaxa, são a seguir associados aos registros de coleta e às imagens. O resultado é encaminhado ao Gerenciador de Busca Combinada e deste à Interface.

4.3.3 Consulta por imagem

A consulta por imagem ocorre por meio de uma imagem que o usuário fornece, sendo retornadas imagens de espécies que sejam similares à imagem de consulta.

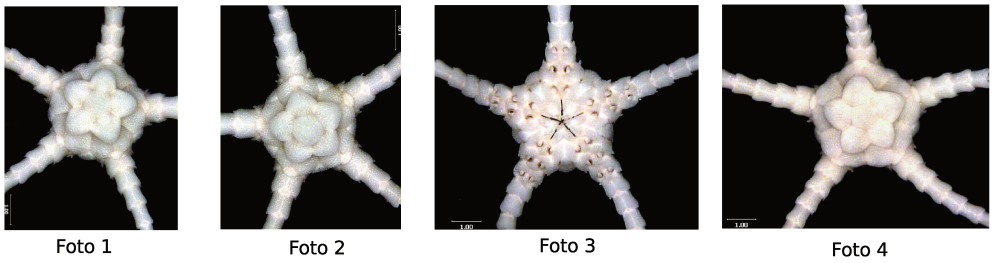

Figura 4.2: Imagens exemplo

Considere por exemplo as imagens apresentadas na figura 4.2 e suponha que a Foto 1 fosse apresentada como imagem de consulta. A imagem é encaminhada ao Gerenciador de busca combinada, que inicia o Módulo de Recuperação por Conteúdo, que extrai os vetores de características da Foto 1 segundo todos os descritores disponíveis no sistema. Os arquivos com os novos vetores de características serão a seguir comparados com os vetores de todas as imagem armazenadas, segundo a função de similaridade adotada, buscando as imagens mais similares à Foto 1.

O Módulo de Recuperação por conteúdo retornará uma lista de identificadores das imagens mais similares à de consulta, que no nosso exemplo contém : “Foto 2” e “Foto 4”. A partir desses identificadores o Gerenciador de Busca Combinada busca os dados das espécies no banco de coletas, devolvendo os resultados para a Interface. No nosso exemplo, são exibidos os dados das espécies referentes às Fotos 2 e 4.

Um ponto importante desta modalidade é que as imagens armazenadas podem ser privadas, ou seja, não disponibilizadas publicamente. Mesmo assim, o conteúdo destas ima-

gens pode ser utilizado para realizar a busca, sendo exibidas apenas as imagens públicas. O Gerenciador de Busca Combinada verifica quais imagens podem ser exibidas.

4.3.4 Consulta Combinada

Além de permitir as consultas separadas (somente metadados, nome vulgar ou imagem), o Sinimbu também permite combinar tais consultas. Apresentamos a seguir um exemplo de consulta combinada.

Considerando a consulta apresentada na Figura 3.5, em que o usuário entrou com o nome vulgar "sapo" e o local de coleta "Brasil", temos na Figura 4.3 um dos possíveis resultados. Ela mostra uma das espécies que satisfaz à consulta do usuário. O resultado completo é uma lista de resultados como os apresentados na figura, ordenados pela proximidade com a consulta.

A Figura 4.3 destaca os itens de resultado: a foto da espécie, os dados da espécie (nome vulgar, nome científico, e local de coleta), a possibilidade de ver outras fotos da espécie e a possibilidade de ver os dados de coleta relativos à espécie em questão.

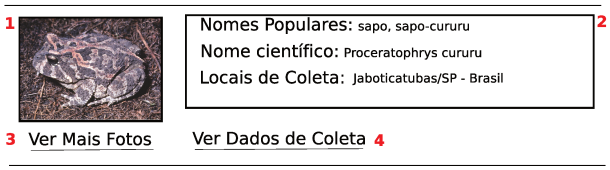

Figura 4.3: Um dos possíveis resultados para a consulta apresentada na Figura 3.5

Para se executar esta consulta, os dados de entrada (metadado, nome vulgar e imagem) são encaminhados ao Gerenciador de Busca Combinada, que encaminha para cada Módulo respectivo o dado de entrada. O metadado local de coleta = "Brasil" é encaminhado para o Módulo de coleta, que retorna uma lista de identificadores de espécies em que registros de coleta indicam o Brasil. O nome vulgar "sapo" é encaminhado para o Módulo de Recuperação Taxonômica, que retorna uma lista de taxons equivalente a "sapo", que é encaminhada ao Módulo de Coleta, resultando em uma lista de identificadores de espécie equivalentes a "sapo". A imagem de consulta é encaminhada ao Módulo de Recuperação por Conteúdo, que retorna uma lista de identificadores de imagens similares à imagem de consulta; tal lista é encaminhada ao Módulo de Coleta para se obter a lista de identificadores de espécies das imagens similares a imagem de consulta.

Depois de recuperados todos os identificadores de espécies referentes a cada parâmetro informado para a consulta, é realizada uma pontuação de cada espécie similar encontrada, para que sejam retornados os resultados mais significativos de maneira ordenada.

4.4 Exemplo de sessão de usuário

Apresentamos nesta seção uma série de exemplos de uso do Sinimbu. Inicialmente considere que um usuário deseja recuperar espécies que tenham registros de coleta na localidade de "Jaboticatubas", como mostra a Figura 4.4. Há 1379 registros que satisfazem a consulta. A Figura 4.5 mostra os 3 primeiros resultados, sendo que apenas o segundo registro tem foto visível. No caso, o primeiro registro não tem foto e o terceiro tem foto protegida.

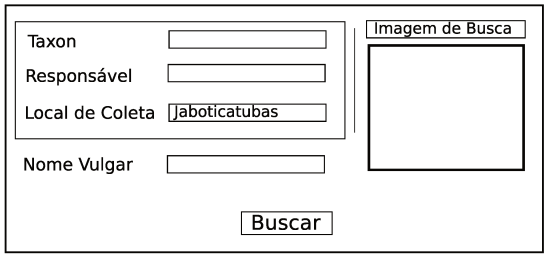

Figura 4.4: Tela da primeira consulta

Figura 4.5: Resultado parcial da primeira consulta

Refinando a busca, agora por espécies com registros de coleta na localidade de "Jaboticatubas" e nome vulgar "Sapo", como mostra a Figura 4.6. Há 1086 registros que satisfazem a esta consulta. O 3 primeiros resultados são apresentados na Figura 4.7, agora restritos a sapo.

Suponha agora que o usuário queira refinar ainda mais a consulta, buscando além de registros com localidade de "Jaboticatubas" e nome vulgar "sapo", incluindo também

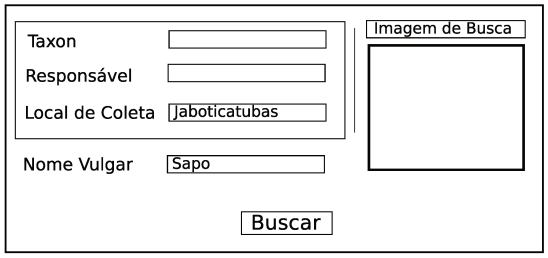

Figura 4.6: Tela da segunda consulta

Figura 4.7: Resultado parcial da segunda consulta

uma imagem de consulta. A Figura 4.8 mostra a tela da consulta. Na Figura 4.9 podemos ver os 3 primeiros resultados desta consulta. Pode-se notar que o resultado possui a ordem invertida dos primeiros registros, em relação ao resultado anterior. Isso ocorre devido à maior similaridade da imagem de consulta com o registro do "Proceratus Cururu".

Ainda a partir da tela 4.9, se o usuário clicar em "Ver dados de coleta", aparecem os dados completos do registro de coleta correspondente. Se clicar em "Ver mais fotos", são mostradas outras imagens do mesmo animal, quando houver.

4.5 Conclusões

Este capítulo descreveu aspectos de implementação da arquitetura do sistema Sinimbu. Foram apresentadas detalhadamente todas as possibilidades de consultas. Também apresentamos um exemplo de sessão de consulta completo, abrangendo as diversas modalidades. O próximo capítulo apresenta as conclusões e trabalhos futuros desta pesquisa.

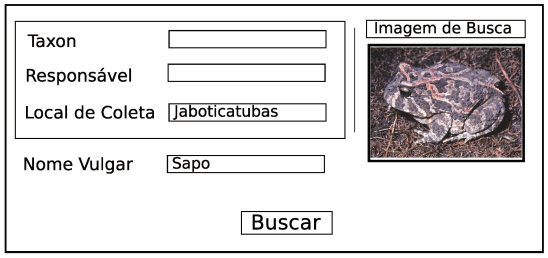

Figura 4.8: Tela da terceira consulta

Figura 4.9: Resultado parcial da terceira consulta

Capítulo 5

Conclusões e Trabalhos Futuros

A heterogeneidade dos dados de biodiversidade e a variedade de tipos de usuário interessados nesses dados apresentam desafios para a recuperação desses dados. Abordagens multimodais podem garantir uma melhor qualidade na recuperação desses dados, bastante heterogêneos. Esta dissertação ataca estes problemas apresentando o Sinimbu que utiliza abordagem multimodal de recuperação, para dados como conteúdo das imagens e nomes vulgares. Isto permite uma recuperação de dados mais abrangente, que auxilia os pesquisadores em biodiversidade em suas tarefas, além de incentivar usuários leigos a acessar esses dados. O trabalho estende assim o escopo de sistemas de informação de biodiversidade tradicionais, na sua maioria limitados a consultas textuais.

5.1 Contribuições

Esta dissertação apresentou o Sinimbu, um conjunto de ferramentas que visa tratar consultas multimodais que envolvem palavras-chave e imagens no domínio de biodiversidade. Também foram abordadas questões como a busca utilizando nomes vulgares e o tratamento diferenciado para leigos e pesquisadores.

Desta forma as contribuições desta dissertação são:

- Estudo e análise da situação atual dos sistemas de busca multimodal e dos sistemas de informação de biodiversidade;
- Especificação e implementação de uma arquitetura (Sinimbu) que processa consultas multimodais em dados de biodiversidade. A arquitetura permite consultas utilizando nome científico da espécie, nome vulgar da espécie, dados de coleta (nome do responsável e local) e imagem da espécie, além de permitir o uso combinado desses atributos;

- O tratamento diferenciado para usuários diferentes;
- Criação de uma ontologia taxonômica com nomes vulgares associados, com isto permitindo busca por nomes vulgares. A possibilidade de buscar dados de biodiversidade utilizando nomes vulgares não está presente na maioria dos sistemas disponíveis na Web. Além disso, também é uma contribuição o uso de uma ontologia para armazenar tal informação, que permite associar termos vulgares a diferentes níveis taxonômicos de maneira eficiente;
- Validação da proposta com implementação de um protótipo, baseado em dados reais do Museu de Zoologia da Unicamp, com 40000 registros e 1200 imagens.

O protótipo possui limitações, porém aponta para possibilidades de uso dos dados de biodiversidade pouco exploradas. As pesquisas realizadas até o momento nos mostram que a arquitetura do Sinimbu é a única que permite consultas multimodais no domínio de biodiversidade.

5.2 Trabalhos Futuros

Durante todo o desenvolvimento do projeto foram identificadas várias possibilidades de avanço e expansão do trabalho que são apresentadas nesta seção.

- **Utilização de sons no *Módulo de Recuperação por conteúdo***: Como a arquitetura proposta é flexível, uma possibilidade simples de implementação é a adicionar um serviço de recuperação de sons. Isto possibilitaria o uso de gravações sonoras como fonte de informação para a busca, uma extensão que tem muitas possibilidades de uso no contexto de biodiversidade e museus virtuais de biologia.
- **Realimentação de relevância**: Uma das maneiras de melhorar um sistema de recuperação é utilizando de realimentação de relevância, que permite ao usuário identificar iterativamente os resultados que estão mais próximo do que ele procura. A partir das indicações do usuário, o sistema identifica propriedades visuais que melhor definem os resultados relevantes. Tal modificação tende a garantir melhoras no resultado, pois considera características de alto nível identificadas pelo usuário para determinar a relevância das imagens.
- **Aperfeiçoamento do Gerenciador de Tipos de Usuário**: Considerando a diferença no perfil dos usuários que utilizam sistemas de biodiversidade, o Gerenciador de Tipos de Usuário poderia ser modificado para um Gerenciador de Perfil de

Usuários. Desta maneira, poderia utilizar um sistema de login para armazenar preferências e configurações de busca específica do usuário. A partir desse gerenciador seria possível configurar parâmetros mais específicos da busca, como qual descritor seria utilizado para as imagens, qual peso seria atribuído aos parâmetros de entrada, qual algoritmo de similaridade e ranqueamento, entre outras possibilidades. É também possível incluir um módulo de realimentação de relevância atrelado ao Gerenciador de Perfil, de maneira que a realimentação ocorra especificamente para cada usuário.

- **Palavras-Chave**: O sistema proposto faz a busca de palavras-chave utilizando primitivas do Banco de Dados e do JENA (no caso das ontologias). Tal funcionamento é limitado e, por exemplo, descarta palavras-chave com algum erro de digitação. Uma solução para tal problema seria a possibilidade de incluir-se um módulo intermediário que extraia previamente todas as palavras-chave do BD e das ontologias, e durante o processo de recuperação use medidas de similaridade textual para recuperar essas palavras-chave. Com isso, a recuperação das palavras-chave poderia ocorrer a partir de termos não exatos.

- **Uso de Dados Geográficos**: Além das modalidades propostas no trabalho, o uso de dados geográficos seria muito útil. Com esta expansão, a localização de onde ocorreu a coleta, informada pelo usuário, poderia ser utilizada na busca, usando funções geográficas de proximidade. A partir dessa nova modalidade também seria possível incluir na interface uma janela para indicar cartograficamente um ponto ou uma região de abrangência para a busca.

- **Atrelar descritores a níveis taxonômicos**: Durante o desenvolvimento da pesquisa, verificou-se que imagens de diferentes espécies têm diferentes características visuais. Uma extensão possível para a obtenção de melhores resultados seria usar diferentes descritores de imagem que fossem adaptados a grupos de taxons, como por exemplo aplicando um descritor de forma para fotos de espécies da ordem *Perciformes* (ordem referente a tipos de peixes), e de cor e forma para espécies da ordem *Lepidoptera* (referente à borboletas).

- **Associação a outas fontes de dados**: Para que haja apoio aos especialistas, será necessário criar ferramentas para associação e busca em outras fontes de dados. As estruturas de dados propostas seguem padrões consensuais, que podem facilitar o intercâmbio de dados com outras coleções. Isto vai envolver também análise do papel do cmapo "id de espécie", usado localmente.

- **Questões de desempenho**: A consulta na ontologia, com 5000 espécies, pode vir a criar gargalos futuros com o aumento dos dados. A parte relativa a busca por

conteúdo já apresenta gargalos no tempod e processamento da consulta. Assim, um desafio é tentar estudar outas estruturas de dados e algoritmos visando desempenho.

Apêndice A

Dicionário de Dados

app_taxonomy		
Campo	**Tipo**	**Descrição**
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
Family	character varying(20)	Nível Taxonômico
SybFamily	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 cinetífico

Figura A.1: Esquema da Tabela app_taxonomy

app_catalog		
Campo	**Tipo**	**Descrição**
idCatalog	integer	Chave primária – formada pelos campos: institution Code, collectionCode e catalogNumber
institutionCode	character varying(10)	Códgio da instituição.
collectionCode	character varying(30)	Código da coleção
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
basisOfRecord	character varying(20)	É uma descrição que representa se o Registro é uma observação ou objeto
previousCatalogNumber	character varying(10)	Um número de catálogo anterior
typeMaterial	character varying(20)	Espécimes tipos, representativos de Determinada espécie
patternColor	character varying(50)	O padrão de coleção de uma espécie Registrada
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
fieldNumber	character varying(20)	Número de Registro de Campo
publication	character varying(50)	Título de uma publicação que usao Informação do catálogo
notes	Composite Types	Observações que realizam os curadores Do museu
earlistDateCollected	Composite Types	Data de início da coleta
latestDateCollected	integer	Data final da coleta.
indivudalCount	integer	Quantidade e indivíduos de uma mesma Espécie
fk_cataloguedBy	integer	Identifica quem catalogou
fk_location	integer	Local de coleta
fk_taxa	integer	Chave estrangeira de app_taxonomy
fk_collector	integer	Identifica o coletor
fk_identifier	integer	Identifica quem identificou a espécie
fk_meth	integer	Chave estrangeira de app_methodology

Figura A.2: Esquema da Tabela app_catalog

app_pictures		
Campo	**Tipo**	**Descrição**
id	integer	Chave primária
fileName	character varying(100)	Nome do arquivo com a imagem
label	character varying(100)	Legenda da Imagem
idTaxa	integer	Nível Taxonômico
visible	boolean	Define se a imagem pode ser visualizada
classification	character varying(100)	Define se a imagem é de holótipo ou Parátipo
author	character varying(100)	Autor da imagem
idCatalog	integer	Registro de coleta a que se refere a Imagem

Figura A.3: Esquema da Tabela app_pictures

app_responsible		
Campo	**Tipo**	**Descrição**
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 Jurídica

Figura A.4: Esquema da Tabela app_responsible

app_location		
Campo	**Tipo**	**Descrição**
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
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.5: Esquema da Tabela app_location

Referências Bibliográficas

[1] SinBiota - Sistema de Informação Ambiental do Biota. http://sinbiota.cria.org.br/, 2011.

[2] Website do WikiAves. http://www.wikiaves.com.br/, 2011.

[3] WeBIOS: Web Service Multimodal Tools for Strategic Biodiversity Research, Assessment and Monitoring. http://www.lis.ic.unicamp.br/projects/webios, started 2005.

[4] BioCORE: Biodiversity and Computing Research. http://www.lis.ic.unicamp.br/projects/bio-core, started 2007.

[5] M. Addis, M. Boniface, S. Goodall, P. Grimwood, S. Kim, P. Lewis, K. Martinez, and A. Stevenson. Sculpteur: Towards a new paradigm for multimedia museum information handling. In *Proceedings of Semantic Web ISWC 2870*, pages 582–596, 2003.

[6] A. Amir, M. Berg, and H. Permuter. Mutual relevance feedback for multimodal query formulation in video retrieval. In *MIR '05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 17–24, New York, NY, USA, 2005. ACM.

[7] A. Arpah, S. Alfred, L. Lim, and K. Sarinder. Monogenean image data mining using taxonomy ontology. In *Networking and Information Technology (ICNIT), 2010 International Conference on*, pages 478 –481, 2010.

[8] S. Atnafu, R. Chbeir, and L. Brunie. Efficient content-based and metadata retrieval in image database. *Journal of Universal Computer Science*, 8(6):613–622, 2002.

[9] D. A. L. Canhos, S. de Souza, and V. P. Canhos. Coleções biológicas e sistemas de informação. Technical report, Centro de Referência em Informação Ambiental -(Cria), 2005.

[10] M. J. Costello and E. V. Berghe. 'ocean biodiversity informatics': a new era in marine biology research and management. In *Marine Ecology Progress Series*, volume 316, pages 203–214, May 2006.

[11] N. Cullot, C. Parent, S. Spaccapietra, and C. Vangenot. Ontologies: A contribution to the DL/DB debate. In *Proc. of the 1st International Workshop on the Semantic Web and Databases, 29th International Conf. on Very Large Data Bases*, pages 109–129, 2003.

[12] R. da S. Torres, A. X. Falcão, M. A. Gonçalves, J. P. Papa, B. Zhang, W. Fan, and E. A. Fox. A genetic programming framework for content-based image retrieval. *Pattern Recognition*, 42(2):283–292, 2009.

[13] R. da Silva Torres, C. B. Medeiros, M. A. Gonçalves, and E. A. Fox. A digital library framework for biodiversity information systems. *Int. J. on Digital Libraries*, 6(1):3–17, 2006.

[14] J. Daltio. Aondê: Um serviço web de ontologias para interoperabilidade em sistemas de biodiversidade (aondê: An ontology web service for interoperability across biodiversity information systems). Master's thesis, Instituto de Computação - Unicamp, August 2007.

[15] J. Daltio, C. B. Medeiros, L. C. G. Jr, and T. M. Lewinsohn. A framework to process complex biodiversity queries. In *Proc. ACM Symposium on Applied Computing (ACM SAC)*, March 2008.

[16] A. 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, December 1999.

[17] G. S. Fedel and C. B. Medeiros. Busca multimodal para apoio à pesquisa em biodiversidade. *Workshop de Teses e Dissertações - Simpósio Brasileiro de Banco de Dados*, 2010.

[18] M. Ferecatu, N. Boujemaa, and M. Crucianu. Semantic interactive image retrieval combining visual and conceptual content description. *ACM Multimedia Systems*, 13(5-6):309–322, 2008.

[19] R. B. Freitas and R. da S. Torres. OntoSAIA: Um ambiente Baseado em Ontologias para Recuperação e Anotação Semi-Automática de Imagens. pages 60–79, October 2005.

[20] T. Gruber. Toward principles for the design for ontologies used for knowledge sharing. In *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Kluwer Academic Publishers, 1993.

[21] F. Guo, L. Li, C. Faloutsos, and E. P. Xing. C-dem: a multi-modal query system for drosophila embryo databases. *Proc. VLDB Endow.*, 1(2):1508–1511, 2008.

[22] R. Hyam. The use of web ontology language (owl) to combine extant controlled vocabularies in biodiversity informatics appears redundant. In *Nature Proceedings*, 2010.

[23] ICMI. International Conference on Multimodal Interfaces. http://icmi2009.acm.org/, 2009.

[24] L. C. G. Jr. An architecture for querying biodiversity repositories on the web. Master's thesis, Instituto de Computação - Unicamp, May 2007.

[25] K. Kesorn and S. Poslad. Semantic restructuring of natural language image captions to enhance image retrieval. *Journal of Multimedia*, 4(5):284–297, Oct 2009.

[26] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

[27] S.-H. Liao, H.-C. Huang, and Y.-N. Chen. A semantic web approach to heterogeneous metadata integration. In J.-S. Pan, S.-M. Chen, and N. Nguyen, editors, *Computational Collective Intelligence. Technologies and Applications*, volume 6421 of *Lecture Notes in Computer Science*, pages 205–214. Springer Berlin / Heidelberg. 10.1007/978-3-642-16693-8_23.

[28] J. E. G. Malaverri. Um serviço de gerenciamento de coletas para sistemas de informação de biodiversidade. Master's thesis, Instituto de Computação - Unicamp, April 2009.

[29] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, July 2008.

[30] E. Mena, A. Illarramendi, V. Kashyap, and A. P. Sheth. Observer: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. *Distributed and Parallel Databases*, 8(2):223–271, 2000.

[31] W. Michener, J. Beach, M. Jones, B. Ludäscher, D. Pennington, R. S. Pereira, A. Rajasekar, and M. Schildhauer. A knowledge environment for the biodiversity and ecological sciences. *J. Intell. Inf. Syst.*, 29(1):111–126, 2007.

[32] N. F. Noy and D. L. McGuinness. Ontology development 101: A guide to creating your first ontology. Technical Report SMI-2001-0880, Stanford University School of Medicine, 2001.

[33] I.-H. Paik, J. Lim, B.-S. Chun, S.-D. Jin, J.-P. Yu, J.-W. Lee, J. Bhak, and W.-K. Paek. The korean bird information system (kbis) through open and public participation. *BMC Bioinformatics*, 10(Suppl 15):S11, 2009.

[34] C. Parr, R. Espinosa, T. Dewey, G. Hammond, and P. Myers. Building a biodiversity content management system for science, education and outreach. *Data Science Journal*, 4:1–11, 2005.

[35] P.-D. Qiou. Design configuration of composite services on semantic web. Master's thesis, Tatung University, 2005.

[36] M. Rahman, S. Antani, R. Long, D. Demner-Fushman, and G. Thoma. Multi-modal query expansion based on local analysis for medical image retrieval. In B. Caputo, H. Müller, T. Syeda-Mahmood, J. Duncan, F. Wang, and J. Kalpathy-Cramer, editors, *Medical Content-Based Retrieval for Clinical Decision Support*, volume 5853 of *Lecture Notes in Computer Science*, pages 110–119. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-11769-5_11.

[37] H. Song, X. Li, and P. Wang. Multimodal image retrieval based on annotation keywords and visual content. In *Control, Automation and Systems Engineering, 2009. CASE 2009. IITA International Conference on*, pages 295 –298, 2009.

[38] L. Stojanovic, S. Staab, and R. Studer. elearning based on the semantic web. In *World Conference on the WWW and the Internet (WebNet 01), Orlando, Florida*, 2001.

[39] J.-H. Su, B.-W. Wang, T.-Y. Hsu, C.-L. Chou, and V. S. Tseng. Multi-modal image retrieval by integrating web image annotation, concept matching and fuzzy ranking techniques. *International Journal of Fuzzy Systems*, 12(2):136–149, 2010.

[40] H. Tangmunarunkit, S. Decker, and C. Kesselman. Ontology-based resource matching in the grid - the grid meets the semantic web. In D. Fensel, K. P. Sycara, and J. Mylopoulos, editors, *International Semantic Web Conference*, volume 2870 of *Lecture Notes in Computer Science*, pages 706–721. Springer, 2003.

[41] M. H. D. Thi, R. V. Lebbe, H. P. Nguyen, and B. L. N. Thi. Indochinese bamboos: biodiversity informatics to assist the identification of "vernacular taxa". In *Proceedings of the International Congress, Paris*, pages 207–211, 2010.

[42] R. S. Torres. *Ambiente de Gerenciamento de Imagens e Dados Espaciais para Desenvolvimento de Aplicacoes em Biodiversidade*. PhD thesis, Instituto de Computação, Unicamp, Campinas, SP, Outubro 2004.

[43] R. S. Torres and A. X. Falcão. Content-based image retrieval: Theory and applications. *Revista de Informática Teórica e Aplicada*, 13:161–185, 2006.

[44] B. S. C. M. Vilar. Processamento de consultas baseado em ontologias para sistemas de biodiversidade. Master's thesis, Instituto de Computação - Unicamp, September 2009.

[45] WWF. Site do WWF Brasil. http://www.wwf.org.br/informacoes/questoes_ambientais/biodiversidade/ acessada em 18/01/2011, 2010.

[46] B. Zhang, Q. Xiang, Y. Wang, and J. Shen. Compositemap: a novel music similarity measure for personalized multimodal music search. In *Proceedings of the seventeen ACM international conference on Multimedia*, MM '09, pages 973–974, New York, NY, USA, 2009. ACM.