Universidade Federal do Tocantins – UFT

Programa de Pós-Graduação em Modelagem Computacional de Sistemas

Daniela Mascarenhas de Queiroz Trevisan

Filhote - Ferramenta de Suporte à Análise e Interpretação de Dados Biológicos

Palmas - TO, Brasil

2015

Daniela Mascarenhas de Queiroz Trevisan

Filhote - Ferramenta de Suporte à Análise e Interpretação de Dados Biológicos

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional de Sistemas da Universidade Federal do Tocantins, como requisito parcial para obtenção do título de Mestre em Modelagem Computacional de Sistemas

Universidade Federal do Tocantins – UFT

Programa de Pós-Graduação em Modelagem Computacional de Sistemas

Orientadores: Prof. Dr. David Nadler Prata e Prof.ª Dra. Elineide Eugênio Marques

Palmas - TO, Brasil

2015

Daniela Mascarenhas de Queiroz Trevisan
Filhote - Ferramenta de Suporte à Análise e Interpretação de Dados Biológicos/ Daniela Mascarenhas de Queiroz Trevisan. – Palmas - TO, Brasil, 2015-
92 páginas.

Orientadores: Prof. Dr. David Nadler Prata e Prof.ª Dra. Elineide Eugênio Marques

Dissertação (Mestrado) – Universidade Federal do Tocantins – UFT
Programa de Pós-Graduação em Modelagem Computacional de Sistemas, 2015.

1. Banco de Dados. 2. Modelo de Dados. 3. Mineração de Dados. 4. Peixes. 5. Ictiofauna. I. Prof. Dr. David Nadler Prata e Prof.ª Dr.ª Elineide Eugênio Marques. II. Universidade Federal do Tocantins - UFT. III. Programa de Pós-Graduação em Modelagem Computacional de Sistemas. IV. Filhote - Ferramenta de suporte à análise e interpretação de dados biológicos

SERVIÇO PÚBLICO FEDERAL
UNIVERSIDADE FEDERAL DO TOCANTINS
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO
MODELAGEM COMPUTACIONAL DE SISTEMAS

Palmas, 01 de Dezembro de 2015.

Ao 1º (primeiro) dia do mês de Dezembro de 2015, realizou-se a defesa da Dissertação de Mestrado da aluna **DANIELA MASCARENHAS DE QUEIROZ TREVISAN**, do Curso de Mestrado em Modelagem Computacional de Sistemas, da Universidade Federal do Tocantins (UFT), intitulado: "***Filhote - Ferramenta de Suporte à Análise e Interpretação de Dados Biológicos***", realizado sob a Orientação dos Professores **Dr. David Nadler Prata** e **Dra. Elineide Eugênio Marques**, tendo como banca avaliadora, os professores abaixo relacionados.

Atribuíram a Nota Final A (Aprovado) pelo trabalho, tendo sido considerado Aprovado. Nada mais tendo a constar, assinam esta Ata os professores componentes da banca.
Observações:_____

_____

_____

_____
Professor David Nadler Prata, Dr.
PPGMCS – Orientador

_____
Professora Elineide Eugênio Marques, Dra.
PPGMCS – Orientadora

_____
Professor Marcelo Lisboa Rocha, Dr.
PPGMCS – Membro

_____
Professor Gustavo Enrique de A. P. Alves Batista, Dr.
PPGMCS – Membro Externo

Este trabalho é dedicado às minhas duas filhas queridas, Camila e Laís, que, mesmo escutando a frase "agora não posso, estou estudando" várias vezes ao dia, continuam procurando uma maneira de ficar comigo.

Agradecimentos

Agradeço imensamente e carinhosamente aos meus orientadores David Nadler Prata e Elineide Eugênio Marques por me acompanharem e me direcionarem todos esses anos. Muito obrigada pela paciência, pela motivação, pelo pensamento positivo e por contribuírem tanto na minha vida profissional quanto pessoal.

Ao meu esposo, Márcio Trevisan, pelo carinho e cuidado com a manutenção da minha felicidade e do meu equilíbrio.

A minha filha Camila que é um exemplo de doçura, calma e companheirismo, sempre ajudando a mamãe, mesmo ainda sendo um nenê, e à minha filha Laís por ter participado junto comigo deste curso, desde que estava na minha barriga. Vocês são a alegria da nossa casa!

Aos meus pais pela presença em minha vida e, especialmente à minha mãe pelos dias dedicados a cuidar das crianças para que este trabalho pudesse ser desenvolvido;

Ao meu parceiro de pesquisa Michel Almeida. Considero-me realmente privilegiada por ter tido a honra de trabalhar com alguém tão bem preparado tecnicamente e tão companheiro. Este trabalho também é seu.

Aos amigos Aislan Max, Emilio Mario, Douglas Chagas e Célia Kondo, meus colegas de trabalho. Definitivamente, vocês fizeram toda a diferença na minha pesquisa. Obrigada por terem me socorrido e me acalmado todas as vezes em que eu já estava no limite da minha força de vontade, por todas as conversas e revisões. Pessoas como vocês provam que ainda podemos acreditar na existência de gentileza.

Aos meus colegas de sala na UFT, pela torcida e companheirismo, especialmente ao Ricardo Egídio, pela boa vontade em dividir as tarefas do dia a dia. Sem o suporte de vocês este trabalho levaria muitos meses mais.

Muito obrigada!

"...And you ask "What if I fall?"
Oh, but my darling,
What if you fly?"
(Erin Hanson)

Resumo

Dissertação de Mestrado

Trevisan, D. M. Q.; Prata, D. N.; Marques, E. E.; Filhote: Ferramenta de Suporte à Análise e Interpretação de Dados Biológicos. Dissertação de Mestrado. Modelagem Computacional de Sistemas. Universidade Federal do Tocantins. Palmas - TO. 2015

Este trabalho apresenta uma proposta para a estruturação dos dados de peixes coletados na região da Usina Hidrelétrica Luís Eduardo Magalhães (ou Usina de Lajeado), no período de 1999 a 2012, e o desenvolvimento de uma ferramenta, chamada Filhote, para a administração destes dados. O principal objetivo é oferecer um meio de manipulação e armazenamento eficiente aos dados obtidos possibilitando a construção de séries históricas com a agregação de resultados de futuras coletas. Para isto, foi desenvolvido um modelo de dados para o armazenamento estruturado desta série, visando servir de alicerce aos estudos de monitoramento da fauna de peixes em ambientes com e sem reservatório. Tomando este modelo como base, a ferramenta Filhote foi integrada à aplicação de Mineração de Dados *WEKA* com o intuito de prover ao pesquisador um meio de análise de dados através da geração de regras de associação. O modelo de dados e a ferramenta desenvolvida são viáveis para o tratamento dos dados existentes e se apresentam como uma boa alternativa para projetos que coletam dados neste mesmo sentido, possibilitando a expansão dos módulos de armazenamento, bem como com a inclusão de novos algoritmos de mineração de dados.

Palavras-chaves: Banco de Dados. Modelo de Dados. Mineração de Dados. Peixes. Ictiofauna.

Abstract

Master's thesis

Trevisan, D. M. Q.; Prata, D. N.; Marques, E. E.; Filhote - Ferramenta de Suporte à Análise e Interpretação de Dados Biológicos. Master's thesis. Modelagem Computacional de Sistemas. Universidade Federal do Tocantins. Palmas - TO. 2015

This work presents a proposal to structure the data of fishes collected in the region of hydroelectric plant Luís Eduardo Magalhães (or hydroelectric plant of Lajeado), during the 1999-2012 period and also a tool development, called *Filhote*, for the administration of these data. The main purpose is to provide an efficient way to manipulate and store the obtained data, enabling the construction of time series aggregating results from future collections. To this intent, it was developed a data model for the structured storage of this set, aiming to provide the basis for studies on the monitoring of fish fauna in environments with and without reservoir. Taking this model as the basis, the *Filhote* tool has been integrated into the application of Data Mining *WEKA* in order to provide the researcher a means of data analysis through the generation of association rules. The data model and the developed tool are viable to treatment of existing data and they are presented as a good alternative for projects that collect data in this same direction, enabling the expansion of the storage modules, as well as the inclusion of new data mining algorithms.

Key-words: Database. Data Model. Data Mining. Fishes. Ichthyofauna.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

BSD	Berkeley Software Distribution
CSV	Comma-separated Values
DDL	Data Definition Language
DER	Diagrama Entidade Relacionamento
JSP	JavaServer Pages
KDD	Knowledge Discovery in Databases
MER	Modelo Entidade Relacionamento
NEAMB	Núcleo de Estudos Ambientais
NSA	National Security Agency - Estados Unidos
SHA	Secure Hash Algorithm
SGBD	Sistema de Gerenciamento de Banco de Dados
SQL	Structured Query Language
TDWI	The Data Warehousing Institute
UML	Unified Modeling Language
USGS	United States Geological Survey
WEKA	Waikato Environment for Knowledge Analysis
XLS	MS Excel file extension

Sumário

1 Introdução

1.1 Visão geral do problema

A espécie humana é uma dentre as 8,7 milhões de espécies estimadas para o planeta Terra, com mais de 7 bilhões de indivíduos atualmente. O aumento do número de indivíduos e a intensificação de suas atividades têm modificado a paisagem terrestre em um curto espaço de tempo (HEILIG, 2012).

Contudo, segundo Mora et al. (2011) os resultados recaem sobre a própria população humana à medida que um grande número de espécies e de ambientes é extinto antes mesmo de conhecê-los.

Neste sentido, a Ciência tem buscado descrever os fenômenos observados na natureza e construir modelos que possibilitem a previsão e o manejo direcionado para a sustentação da própria população humana e conservação das espécies (HILL et al., 2010).

O desafio é ainda maior em regiões tropicais, onde a diversidade de espécies e de ambientes é alta e pouco conhecida. Em se tratando de ambientes aquáticos a dificuldade é ainda maior devido às dificuldades metodológicas de obtenção das informações.

No caso da fauna de peixes na região amazônica, o levantamento das informações apresenta um valor agregado ainda maior em função das distâncias, das dimensões espaciais, das dificuldades em se realizar os trabalhos de campo e da velocidade de alteração do ambiente, acelerada pela expansão do agronegócio e pelos grandes projetos aprovados para a região.

Neste cenário diverso, o cuidado com as séries de dados obtidas nos levantamentos de biodiversidade, independente de sua finalidade, é fundamental para o acompanhamento do histórico de modificação em escalas local, regional, continental e mesmo planetária, como os estudos relacionados às mudanças climáticas, por exemplo.

Se o foco for direcionado para os investimentos em recursos humanos, financeiros e de tempo para o levantamento de dados biológicos relativos às comunidades aquáticas, pode-se verificar que o valor agregado a cada dado obtido é alto. A possibilidade de comparação dos conjuntos de dados obtidos com diferentes metodologias, em diferentes locais, condições climáticas e períodos de tempo é o principal caminho para a análise das alterações ocorridas ao longo do tempo nos ecossistemas tropicais.

Apesar da legislação ambiental no Brasil exigir a elaboração dos estudos de

impacto ambiental antes da execução de empreendimentos que alterem o ambiente (BRASIL, 1986), muitos dados têm se perdido ou encontram-se inacessíveis.

Neste contexto, a formação de recursos humanos que percebam e valorizem as informações obtidas e que deem suporte à elaboração e ao armazenamento de dados é um dos primeiros passos necessários para os avanços nos modelos robustos de previsão na área ambiental.

Visando entender esta questão focamos este estudo no caso dos dados coletados durante a execução dos Programas Básicos Ambientais [1] da Ictiofauna na região da Usina de Lajeado (ou Usina Hidrelétrica Luís Eduardo Magalhães), pelo Núcleo de Estudos Ambientais da Universidade Federal do Tocantins (Neamb/UFT).

A construção da Usina de Lajeado foi iniciada em 1997 e finalizada em 2001. As informações sobre a fauna de peixes foram obtidas por meio de amostragem experimental padronizada, utilizando redes de espera com malhas entre 2,4 e 16 cm entre nós opostos. As coletas foram realizadas no período entre outubro de 1999 e março de 2012, com diferentes intervalos de tempo e de locais de amostragem.

1.2 Motivação

Neste período foi realizado o levantamento de espécies e registrado o número de indivíduos, o comprimento, o peso e outras informações sobre cada peixe coletado em ficha de papel juntamente com as informações sobre as condições do ambiente aferidas no local. As fichas estão armazenadas fisicamente no Neamb/UFT.

Os dados foram digitados e disponibilizados para este estudo em arquivos de planilha eletrônica, no formato *xls*. Um dos arquivos *xls* contem dados de peixes (dados bióticos) e o outro arquivo armazena dados do ambiente (dados abióticos).

Por se tratar de uma grande quantidade de registros, estes dados armazenados em planilhas eletrônicas eram de difícil manipulação e frequentemente eram constatadas lentidão e falhas no aplicativo editor. Além disso, os dados dispostos apenas neste padrão estavam sob o risco de perdas de integridade, quando um ou mais registros eram submetidos a atualizações pelo pesquisador. Ademais, os dados também estavam vulneráveis à perda definitiva, considerando que, sendo o dispositivo de armazenamento um computador pessoal, considera-se o risco do equipamento estar mais suscetível a danos.

Perdas de dados como estes significariam a perda de parte dos resultados de

[1] Plano Básico Ambiental (PBA) é um documento técnico que contém a orientação e a especificação das ações mitigatórias dos programas ambientais propostos no Estudo Prévio e Relatório de Impacto Ambiental.

anos de pesquisa, principalmente considerando que o ambiente em que foram coletados foi alterado com a construção da Usina e a fauna de peixes se adapta a esta alteração de modo singular. Perceber como as populações naturais se adaptam às novas condições ambientais ao longo do tempo contribui para a elaboração de prognósticos sobre as mudanças populacionais neste e em outros locais sob impactos semelhantes.

1.3 Justificativa

Ciente da relevância destes dados biológicos no contexto da conservação e acompanhamento histórico de espécies de peixes, especialmente por causa da condição de vulnerabilidade em que se encontram, os dados foram estruturados e foi planejada uma ferramenta *web* que permite que o pesquisador gerencie este conjunto de dados existentes de forma segura, possibilitando que dados futuros possam ser agregados a este banco.

Além disto, nesta aplicação o pesquisador terá mais um método de explorar os dados armazenados em sua base, através da análise de regras de associação e a descoberta de padrões que poderiam, inclusive, estar implícitos nos dados. Isto amplia o leque de opções de análise de dados e fortalece a integração de áreas de estudo.

O nome "Filhote" foi eleito para esta ferramenta em homenagem a um dos maiores bagres registrados na bacia (*Brachyplathystoma Filamentosum*) e também considerando o significado do termo, uma ferramenta que pode se desenvolver e crescer muito. Este peixe está atualmente presente na lista de animais ameaçados de extinção, principalmente, por causa da construção de novas usinas hidrelétricas e, além disso, faz parte do conjunto de registros que compõem este banco de dados.

1.4 Objetivos

1.4.1 Geral

Estruturar os dados da fauna de peixes coletados na usina de Lajeado e prover uma ferramenta de armazenamento e de manutenção de dados biológicos e ambientais relativos à fauna de peixes coletados em ambientes naturais que possibilite o acesso amigável aos dados, além de oferecer um meio de análise de dados através da geração de regras de associação.

Hipótese:

Os dados biológicos coletados compõem a base para o diagnóstico da condição inicial de um ambiente, por isto, o cuidado com a manutenção dos conjuntos é extre-

mamente relevante, especialmente na Amazônia, onde a biodiversidade é elevada e as alterações ambientais têm ocorrido rapidamente.

Planilhas eletrônicas não são o meio adequado para persistir os dados biológicos e faz-se necessário o uso de ferramenta planejada para esse tipo de dado, que ofereça recursos para manter, organizar e exportar séries de dados cruciais para o desenvolvimento de pesquisas.

1.4.2 Específicos

- Desenvolver um modelo de dados adequado aos dados registrados durante as coletas;
- Criar chaves para identificar individualmente os registros;
- Importar os dados para o modelo de dados desenvolvido;
- Desenvolver uma ferramenta que dê suporte à entrada, edição e exportação de dados;
- Integrar a ferramenta desenvolvida com o *WEKA (Waikato Environment for Knowledge Analysis)*[2] para geração de regras de associação

1.5 Organização da dissertação

Os demais capítulos desta dissertação estão organizados da seguinte forma:

O Capítulo 2 apresenta o referencial teórico fazendo uma abordagem sobre a importância dos dados e de sua manutenção de forma eficiente, discorre sobre a problemática da perda de dados, faz apontamentos sobre o tratamento de dados biológicos, descreve conceitos de banco de dados e faz um apanhado geral sobre as características de mineração de dados.

O Capítulo 3 descreve a metodologia deste trabalho e está dividido em duas partes. Na primeira parte está descrita a estruturação e o mapeamento dos dados deste estudo e na segunda parte está a especificação da ferramenta proposta, apresentando os detalhes de seu funcionamento e implementação.

No Capítulo 4 estão as discussões referentes ao desenvolvimento do trabalho, a importância dos resultados obtidos e a análise das dificuldades e contribuições alcançadas.

[2] Disponível em: http://www.cs.waikato.ac.nz/ml/index.html

Este trabalho é concluído no Capítulo 5. O Apêndice A apresenta as especificações dos casos de uso; O Apêndice B apresenta o dicionário de dados; No Anexo A e no Anexo B estão as imagens amostrais dos dados abióticos e bióticos, respectivamente.

2 Referencial Teórico

Este capítulo está dividido em três partes que têm como principal objetivo discorrer sobre a importância da estruturação dos dados, realizar um breve apanhado sobre o problema de perda de dados relacionados às pesquisas, fazer apontamentos sobre o tratamento dos dados biológicos, apresentar os fundamentos de banco de dados e fazer uma abordagem geral sobre mineração de dados. A revisão sobre estes tópicos, são fundamentais à compreensão deste trabalho.

2.1 A importância dos dados

O processo de informatização nas instituições iniciou-se com a busca da automatização de processos rotineiros e repetitivos com o objetivo de trazer celeridade aos trabalhos nas mais diversas áreas de atuação. O crescimento e a disseminação dos recursos computacionais resultaram no constante aumento das séries de dados.

De acordo com o relatório do TI BPO Book - 2013/2014, 2,5 quintilhões de bytes de dados estão sendo criados por dia e a quantidade de informação no mundo dobra a cada ano (BRASSCOM, 2014).

Dados são resultados de observações documentadas, uma representação de fatos, conceitos e/ou instruções. A simples existência dos dados pode não ser suficiente para a obtenção de informações, para isto é necessário que eles estejam disponíveis aos usuários e sejam processados (DAVENPORT, 1998). Eles são elementos base sobre os quais o computador efetua as operações necessárias à tarefa em questão (COADIC; GOMES, 1996).

Para os autores (REZENDE; ABREU, 2013) um dado é transformado em uma informação compreensível por seus usuários quando são processados, o que os torna úteis e com valor agregado para auxiliar nas tomadas de decisão.

As informações, resultado do processamento dos dados, por sua vez, torna-se um dos recursos mais importantes de uma instituição, contribuindo decisivamente no resultados de suas atividades (FERNANDES, 2005).

Segundo Coadic e Gomes (1996), "Sem informação, a ciência não pode se desenvolver e viver. Sem informação a pesquisa seria inútil e o conhecimento não existiria". Mas, não basta que exista a informação, é necessário que os dados que a compõem sejam de qualidade.

Por isso, o processo de coleta dos dados primários é de fundamental importância

para geração de informações de qualidade, o que significa que os instrumentos de registro (fichas, planilhas, sistemas de informação, entre outros) devem ser adequadamente preenchidos e que os dados registrados sejam seguramente armazenados.

Quando se fala de qualidade dos dados, em geral se trata de sua corretude, que é a adequação para o uso esperado destes dados, e da ausência de erros. Isto implica na possibilidade de se avaliar o dado em qualquer tempo e, comparando-o com um valor real autenticado, corrigi-lo caso seja necessário (CAPPIELLO et al., 2013).

Segundo Felix (2003), a qualidade da informação pode ser medida por meio da avaliação de suas dimensões e seus atributos, conforme sintetizado na Tabela 1.

Tabela 1 – Qualidade da informação (FELIX, 2003)

Tempo	Prontidão	A informação deve ser fornecida quando necessária
	Aceitação	A informação deve estar atualizada quando for fornecida
	Frequência	A informação deve ser fornecida todas as vezes que forem necessárias
	Período	A informação pode ser sobre períodos e instantes do presente, passado ou futuro
Conteúdo	Precisão	A informação deve estar isenta de erros
	Relevância	A informação deve relacionada às necessidades do seu receptor específico, para uma situação específica
	Integridade	Toda informação que for necessária deve ser fornecida
	Concisão	Apenas a informação que for necessária deve ser fornecida
	Amplitude	A informação pode ter um alcance amplo ou reduzido, um foco externo ou interno
	Atualização	A informação é continuamente atualizada para garantir que as pessoas utilizem o que há de melhor
Forma	Clareza	A informação deve ser fornecida de uma forma fácil de ser compreendida
	Detalhe	A informação deve ser fornecida na forma normal, detalhada ou resumida
	Ordem	A informação deve ser organizada em uma sequência predeterminante
	Apresentação	A informação deve ser apresentada na forma narrativa, numérica, gráfica ou outras

Neste mesmo sentido, Batini e Scannapieco (2006) discutem que a baixa qualidade dos dados pode prejudicar seriamente a eficiência e a eficácia das organizações. Segundo estimativas do *The Data Warehousing Institute* (TWDI, 2006), dados de má qualidade custam, só nos Estados Unidos, 611 bilhões de dólares por ano, em postagem, impressão e repetição de trabalhos de funcionários. Outro exemplo aconteceu em Tokyo, no Japão, onde um dado incorreto armazenado no sistema fez o segundo maior banco daquele país tivesse prejuízo de 27 bilhões de yen (equivalente a US$

224.000 na época). Um usuário deveria colocar uma ação à venda por 610.000 yen, mas erroneamente informou que 610.000 ações custavam 1 yen (HYUGA, 2005).

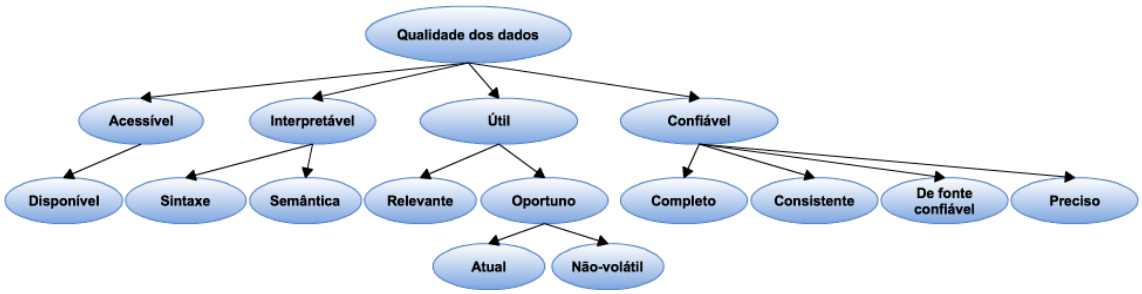
São inúmeros os casos ao redor do mundo semelhantes a estes. Por constatações como estas, pesquisas vêm sendo desenvolvidas com foco no custo provocado pela baixa qualidade dos dados.

O autor (VEIGA, 2012) realizou um estudo onde foram identificados seis aspectos de confiabilidade que influenciam na qualidade dos dados:

- Completude: indicador da suficiência de dados úteis para um determinado domínio;
- Consistência: indicador de ausência de contradições em banco de dados;
- Credibilidade da fonte: indicador de reputação dos dados ou de sua fonte. Faz a mensuração se os dados merecem crédito para serem considerados úteis;
- Acurácia: indicador de qualidade dos dados, define a medida ou a veracidade dos dados;
- Precisão: frequentemente confundida com acurácia porém, diferente da acurácia que está relacionada ao erro, a precisão está relacionada à granularidade dos dados;
- Confiabilidade: indicador de confiança dos dados. É definida através da análise dos resultados de completude, acurácia, consistência e credibilidade da fonte.

A Figura 1 é uma adaptação de (WANG; REDDY; KON, 1995) que representa as características de qualidade dos dados listadas por este autor.

Figura 1 – Modelo hierárquico de dimensões de qualidade dos dados - Fonte: (WANG; REDDY; KON, 1995) - adaptado



Dada a importância das informações, focando em sua qualidade, quanto melhor for o seu processo de geração e sua matéria prima - os dados, melhores serão os resultados obtidos a partir de sua boa administração. E, ainda, para sua geração, quanto

mais estruturado for o processo, mais indicado é o uso de sistemas de informação para a sua gerência (GUIMARÃES; ÉVORA, 2004).

2.2 O problema com a manutenção de dados

Esta seção apresenta um resumo das ideias expostas pelos autores Vines et al. (2014), onde é discutida a problemática em torno da condição de sobrevida dos dados utilizados em pesquisas.

O trabalho supracitado se refere aos dados como partes distintas de informações, geralmente formatadas de uma forma especial. Os autores consideram que em pesquisas, os dados podem ser gerados para fins diferentes e por processos diferentes. Porém, um dos problemas atuais observados com os resultados das pesquisas está relacionado diretamente com os dados. Segundo eles, na maioria das vezes os dados estão armazenados de forma ineficaz e suscetíveis a perdas.

O resultado do referido trabalho mostrou que as principais causas de indisponibilidade de dados de artigos científicos antigos foram a perda da mídia de armazenamento. O termo *perda* se refere a várias maneiras de indisponibilidade dos dados como um roubo de computador pessoal, mídias antigas onde o hardware necessário para leitura já não está mais disponível ou e-mails em que os autores das pesquisas já perderam o acesso.

É notório que a conservação dos dados é diferente nas mais diversas comunidades acadêmicas e que alguns dados são mais fáceis de manter do que outros, mas muitos conjuntos de dados produzidos em pesquisas cientificas são únicos para o seu tempo e localização e, uma vez perdidos não podem ser substituídos.

É consenso e bastante aceito entre pesquisadores que se dedicam ao estudo das relações entre as pesquisas e os dados referentes a elas que é extremamente necessária uma coleta eficaz e, além disso, é indispensável a manutenção desses dados oriundos de pesquisas (VINES et al., 2014).

No caso específico da biodiversidade os dados costumam ser recolhidos de uma mesma maneira por muitos anos, então esse aspecto sobre a conservação dos dados se torna ainda mais crítico.

2.3 Tratamento de dados biológicos

No Brasil, a legislação requer que sejam realizadas avaliações ambientais no momento que precede impactos, por exemplo, a construção de usinas hidrelétricas, criação de aterros sanitários, ferrovias, dentre outros empreendimentos (BRASIL, 1986).

O conjunto de dados coletados são fator base para o diagnóstico da condição inicial do ambiente e das comunidades biológicas. Porém, embora observado o valor deste trabalho para a conservação do meio ambiente, muitas séries estão sendo perdidas.

Além da perda das séries de dados ainda existem outros problemas mais específicos à área biológica. Esses problemas incluem a quantidade de dados biológicos, o grande número de bases de dados biológicos, a rápida taxa de crescimento destas bases, o excesso de tipos e formatos de dados, a variedade de técnicas de acesso aos dados, a heterogeneidade dos conjuntos de dados, erros nos dados e, até mesmo, a natureza interdisciplinar do campo da bioinformática (ERIKSSON, 2015).

Estes dados são muito mais heterogêneos do que dados da física, por exemplo. São oriundos de uma ampla gama de experiências e resultam em diversos tipos de informações, tais como sequências genéticas, interações de proteínas, registros médicos, séries temporais de ambientes e comunidades biológicas (MARX, 2013).

Os pesquisadores que lidam com dados biológicos precisam armazenar grandes conjuntos, analisar, comparar e compartilhar, o que não se configura uma tarefa fácil com o uso de softwares não planejados para esses fins, como é o caso de planilhas eletrônicas. O uso de planilhas eletrônicas é comum no meio acadêmico e apresentam alguns problemas como dificuldade em eliminar redundância, a falta de consistência e a dificuldade da manutenção de recursos de programação que, na maioria das vezes, como o software não foi projetado especificamente para os tipos de dados inseridos, são em macros de difícil manutenção (OIKAWA, 2012).

Em países mais desenvolvidos, a probabilidade de erros tem sido minimizada a partir do gerenciamento seguro dos dados que compõem a elaboração de modelos de previsão. Há casos, inclusive, em que os dados são de responsabilidade de instituições específicas, como é o caso da USGS (*United States Geological Survey*) nos Estados Unidos onde o sistema *BioData Retrieval* disponibiliza publicamente dados de mais de 15.000 peixes, macroinvertebrados aquáticos e amostras da comunidade de algas. Além disso, o sistema disponibiliza mais de 5.000 conjuntos de dados físicos (amostras), tais como dados relativos ao *habitat* e da presença de luz, que foram recolhidos para apoiar as pesquisas. O sistema contém dados de amostra que foram recolhidos e tratados, desde 1991, usando os protocolos específicos para esse fim (USGS, 2015).

2.4 Armazenamento de dados

2.4.1 Modelo de dados

A busca por informações de qualidade e, em consequência, por dados estruturados, armazenados de forma segura e com acesso ágil, nas mais diversas áreas de atuação, impulsionou os estudos e o desenvolvimento de ferramentas que dessem suporte ao armazenamento e à manipulação de dados. Este crescimento acelerado, também contribuiu para a evolução do *hardware* e para a disseminação da utilização de Bancos de Dados colocando-os como centralizador de dados dos sistemas de informação desenvolvidos (FANDERUFF, 2003).

Segundo Date (2004), "Um banco de dados é uma coleção de dados persistentes, usada pelos sistemas de aplicação de uma determinada organização". Os sistemas informatizados armazenam grandes quantidades de dados em seus bancos, o que faz com cresça proporcionalmente o número de organizações que precisam despender esforços no planejamento, controle e gestão mais eficiente dos dados coletados.

Como parte do planejamento, o desenvolvimento do modelo de dados é considerado um valioso instrumento de especificação. É através dele que a arquitetura de um banco de dados é projetada e descrita (ELMASRI; NAVATHE, 2011). Se o modelo for bem especificado, com chaves primárias coerentes com os relacionamentos, suas generalizações e agregações identificadas corretamente, a reutilizações de dados avaliadas, as regras de normalização aplicadas, entre outros aspectos, então, espera-se construir um projeto que irá oferecer bons resultados.

Assim, Silberschatz, Korth e Sudarshan (2010) definiram um modelo de dados como um conjunto de ferramentas conceituais usadas para descrição, semântica de dados e regras de consistência. Ele está sob a estrutura do banco de dados. Portanto, o modelo de dados é considerado a base estrutural sobre o qual um sistema de banco de dados é desenvolvido. Ele é ferramenta essencial ao desenvolvimento de sistema de informação.

Segundo (SILBERSCHATZ; KORTH; SUDARSHAN, 2010), a modelagem de dados está dividida em:

2.4.1.1 Modelo conceitual

O modelo conceitual descreve uma visão do mundo real, resultando em uma descrição geral dos principais dados e relacionamentos entre eles sem se preocupar com regras de implementação. É a etapa inicial de um projeto de banco de dados e o seu o objetivo é descrever como o sistema deve funcionar, baseando-se no processo real.

2.4.1.2 Modelo lógico

O modelo lógico é criado após o desenvolvimento do esquema conceitual para descrever as estruturas que formarão o banco de dados. Nesta etapa são elencados os atributos, as entidades e os relacionamentos do modelo.

2.4.1.3 Modelo físico

O modelo físico, criado após o desenvolvimento do modelo lógico, tem a função de descrever como será a estrutura de armazenamento de dados em um SGBD (Sistema Gerenciador de Banco de Dados). Nele são informados os campos, seus tipos e tamanhos, além de índices e qualquer outra informação física definida.

Esta é a última fase da modelagem de dados e é realizada através da Linguagem de Definição de Dados (DDL), gerando o banco de dados físico.

Baseando em Reingruber e Gregory (1994), é possível dizer de modo sucinto que um modelo de dados de qualidade deve ter as seguintes características:

- Incorporar os planos de negócio da instituição, suas políticas e estratégias: é preciso compreender os processos da organização;
- Utilizar regras claras: é necessário que a descrição das regras sejam uniformemente compreendidas não apenas por quem as escreveu, mas por outras equipes também;
- Possibilitar a geração de uma estrutura de dados de qualidade: deve ser base consistente para a estrutura física;
- Observar o contexto de elementos relacionados: o modelo deve ser criado considerando seus possíveis relacionamentos externos;
- Observar o contexto de toda instituição: o modelo deve ser aderente ao relacionamento entre departamentos e/ou processos;
- Considerar a infra-estrutura: ferramentas, formação, procedimentos de gestão;
- Envolver participantes relevantes: é necessária a participação direta de especialistas do domínio, profissionais de gestão, facilitadores e modeladores de dados, profissionais de gerenciamento de dados, desenvolvedores de sistemas de informação e administradores de banco de dados.

Após o desenvolvimento do modelo de dados, a estrutura do modelo físico para o armazenamento dos dados está apta a ser implementada em um sistema de banco de dados.

2.4.2 Sistema Gerenciador de Banco de Dados

A implementação do modelo físico de dados envolve um ambiente com softwares, operações de sistemas, *hardwares*, ferramentas de administração e usuários. Esta arquitetura compõe a estrutura para suporte de uma coleção de dados armazenados de forma que fiquem persistentes e possam ser inseridos, excluídos, recuperados e atualizados em um sistema que protege a integridade das operações e dos dados (GRAVES; GOLDFARB, 2001).

Os primeiros bancos de dados tiveram início a partir de sistemas de gerenciamento de arquivos, buscando suprir a necessidade de gerenciamento dos dados. Esses sistemas guardavam os registros em diversos arquivos e tinham programas aplicativos para extrair e adicionar registros nos arquivos apropriados (SILBERSCHATZ; KORTH; SUDARSHAN, 2010). Porém, com desvantagens no que diz respeito ao armazenamento, redundância, múltiplos formatos de arquivos, recuperação e segurança dos dados.

Para suprir essas limitações, surgiram estudos com o intuito de desenvolver os Sistemas Gerenciadores de Banco de Dados (SGBD's) que, segundo (ELMASRI; NAVATHE, 2011), são constituídos por um conjunto de dados associados a um conjunto de programas para acesso a esses dados. Os SGBD's obedecem as noções de Atomicidade, Consistência, Isolamento e Durabilidade - propriedades ACID e atividades básicas de banco de dados como, por exemplo, o *backup* e a restauração de bases de dados.

Estes sistemas proporcionam um ambiente eficiente e eficaz para a recuperação dos dados, além de permitir um gerenciamento adequado sobre eles, incluindo a criação de estruturas de bases de dados, a atualização de dados, a geração de relatórios de dados, além de outras atividades de acordo à necessidade do usuário como, por exemplo, o acesso aos dados de forma concorrente.

Elmasri e Navathe (2011) destacam que, um SGBD apresenta os dados em uma visão independente dos aplicativos e garantem três características importantes: a eficiência, com a manutenção de um grande volume de dados, a integridade, através do controle de concorrência e a persistência, através do controle de falhas e recuperação concorrentemente.

Sumariamente, Silberschatz, Korth e Sudarshan (2010) definem os principais objetivos de um SGBD:

- a garantia da privacidade dos dados através do isolamento;
- o compartilhamento de dados de forma segura;
- a garantia de integridade dos dados através de suas chaves;

- a abstração dos dados, de forma que as informações são disponibilizadas ao grupo de usuários que realmente tem necessidade de acessá-las.

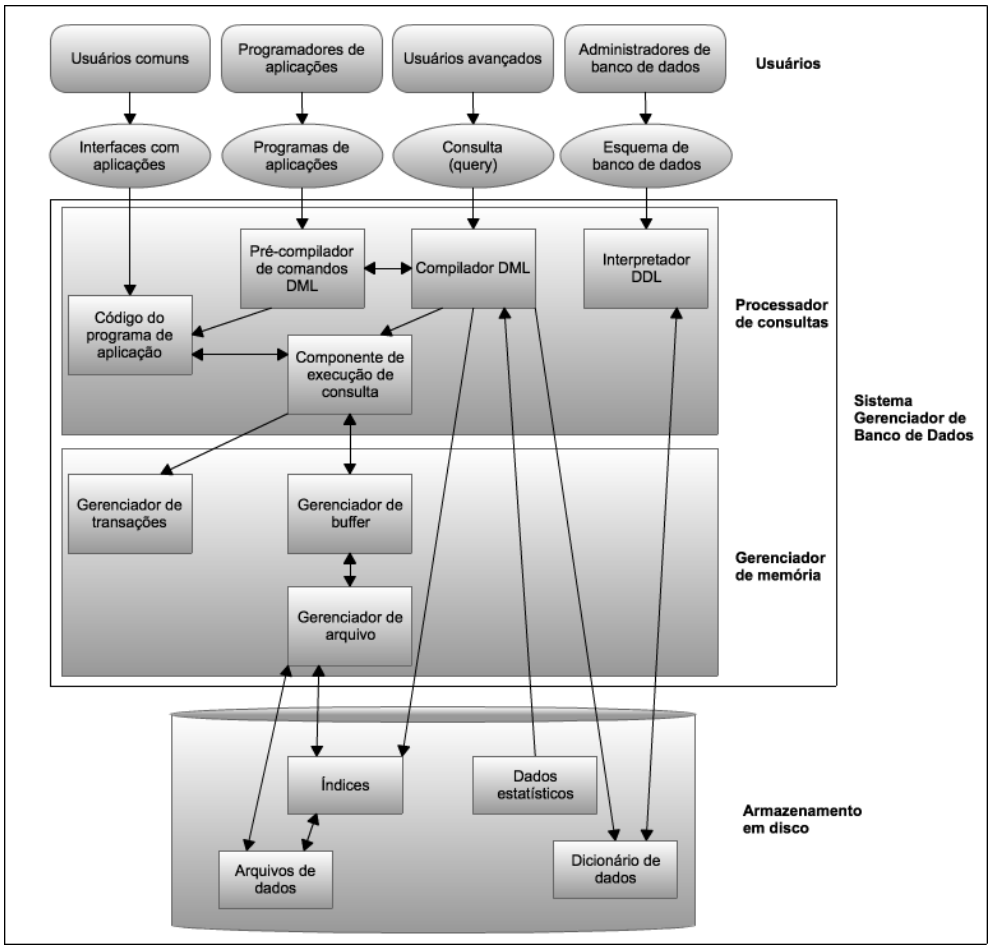
A Figura 2 representa a estrutura geral do sistema, adaptado de (SILBERSCHATZ; KORTH; SUDARSHAN, 2010), onde os usuários se conectam aos dados através de aplicações de administração dos dados ou ferramentas de desenvolvimento e administração da estrutura do sistema, cada um utilizando a visão dos dados a que tem permissão de acesso.

Os acessos dessas aplicações são realizados através de módulos internos de interpretação de linguagens específicas do SGBD, mecanismos de avaliação de consulta e de definição de dados. Este módulo é conhecido como processador de consulta.

O módulo de gerenciamento de armazenamento é responsável pelo controle de transações, que garante que o banco de dados continue consistente em relação às transações concorrentes; pela gerência de autorizações e integridade, que garante a efetivação das regras de integridade e a concessão de permissão ao usuário ao dado correto; pela administração de arquivos, que irá gerenciar a alocação do espaço físico de armazenamento; bem como gerenciamento de buffer, que é responsável pela recuperação de dados para a memória.

O armazenamento em disco gerencia o arquivo de dados, que contém o banco de dados propriamente dito; o dicionário de dados, que contém metadados referentes à estrutura do banco de dados; os índices, que melhoram o desempenho no acesso aos dados; e as estatísticas coletadas sobre os objetos do sistema.

Figura 2 – Estrutura geral do SGBD - Fonte: (SILBERSCHATZ; KORTH; SUDARSHAN, 2010) - adaptado



Um SGBD segue um determinado modelo, que define a forma como os dados estarão organizados internamente. Cronologicamente, os modelos de banco de dados desenvolvidos são: em redes, hierárquicos, relacionais, objeto-relacionais e orientados a objetos (SILBERSCHATZ; KORTH; SUDARSHAN, 2010). Atualmente, a estrutura dos SGBD's comerciais mais utilizados são do tipo Relacional (MASSINO; ROLAND, 2015), onde ela é representada como um conjunto de relações. Os SGBD's implementados sob este modelo de dados, dão suporte ao tipo de dados básicos compostos por números, letras e caracteres.

Apresentado publicamente em um artigo em 1970 por Edgar Frank Codd, a teoria do SGBD relacional foi implementada apenas nos anos 80 (ELMASRI; NAVATHE, 2011). O modelo relacional se mostrou como o mais flexível e adequado ao solucionar os problemas que se colocam no nível da concepção e implementação da base de dados.

A estrutura fundamental do modelo relacional é a relação, constituída por um ou mais atributos, onde cada atributo tem um tipo de dado. Cada linha da tabela é conhecida como tupla ou registro, representa uma coleção de valores de dados relacio-

nados e deve seguir restrições de integridade referencial, chaves e integridade de junções de relações (SILBERSCHATZ; KORTH; SUDARSHAN, 2010).

A ordem física dos registros das tabelas são irrelevantes para o usuário, pois cada um deles é identificado por um valor único, conhecido como chave. Então, para a recuperação de dados, o usuário necessita conhecer apenas a estrutura do banco de dados e seus relacionamentos. A relação entre duas tabelas é estabelecida através de um campo compartilhado, que determinam a unicidade de cada registro. Este campo é conhecido como chave primária (*primary key*) na tabela originária e chave estrangeira (*foreing key*) na tabela que relacionada (ELMASRI; NAVATHE, 2011). Este modelo atende especialmente sistemas de informação que terão dados não-complexos, como é o caso dos dados trabalhados nesta pesquisa.

2.5 Mineração de Dados

Mineração de Dados é um campo que permeia diversos outros como aprendizado de máquina e reconhecimento de padrões, além de fazer uso de diversos conceitos de estatística e de computação. Sendo assim possui uma natureza interdisciplinar podendo se relacionar com as mais diversas áreas de estudo, contribuindo para pesquisas de diferentes assuntos. O principal processo da mineração de dados é a extração de padrões sobre grande quantidade de dados, fato que motivou o acréscimo de uma etapa de mineração ao projeto.

Apesar de Mineração de Dados ser o termo mais conhecido, ele é na verdade parte de um processo maior denominado Descoberta do Conhecimento, do inglês *Knowledge Discovery in Databases* (KDD), que segundo Fayyad et al. (1996) é o processo não trivial de identificar padrões em dados que sejam válidos, novos, potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou um processo de tomada de decisão.

(FAYYAD et al., 1996) define a Mineração de Dados como a etapa do processo de descoberta de conhecimento que realiza a análise dos dados e aplica os algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões sobre os dados.

2.5.1 Tipos de aplicação de Mineração de Dados

Uma das fases do processo KDD, a mineração de dados é geralmente reconhecida pela capacidade de aplicar métodos que buscam padrões válidos sobre os dados Fayyad et al. (1996). Segundo Larose (2014), as aplicações mais comuns são:

Classificação: uma das técnicas mais comuns da mineração de dados, a classificação, tem o objetivo de identificar a qual classe um determinado registro pertence. Este método pode ser sintetizado por um processo de discriminação de unidades em categorias, considerando a atribuição de uma classe pré-definida pelo analista. Na aplicação deste método, o modelo analisa o conjunto de dados a fim de "aprender" como classificar um novo registro, conhecido por isto como aprendizado supervisionado.

Estimativa: tem o objetivo de realizar a estimativa sobre um índice é determinar o valor mais provável encontrado considerando dados obtidos no passado ou dados de índices semelhantes. Sumariamente, pode se estimar o valor de uma determinada variável através da análise dos valores das demais.

Predição: a tarefa de predição tem características similares às tarefas de classificação e de estimação, mas com o objetivo de descobrir o valor futuro de um determinado atributo.

Descrição: é o método utilizado para descrever os padrões e tendências revelados pelos dados. Esta técnica usualmente oferece uma interpretação para os resultados obtidos e é muito utilizada junto com as técnicas de análise exploratória de dados, para comprovar a influência de determinadas variáveis no resultado obtido.

Agrupamento: também conhecida como clusterização, a técnica de agrupamento tem o objetivo de identificar e aproximar dados de coleção de registros similares entre si, avaliando suas diferenças em relação aos outros registros nos demais agrupamentos. Este método se difere da classificação por não necessitar que os registros sejam previamente categorizados sendo, portanto, considerado aprendizado não-supervisionado. Ademais, ela não tem o objetivo de classificar, estimar ou predizer o valor de uma variável, mas apenas de formar grupos de elementos homogêneos entre si.

Associação: o método de associação pretende determinar que fatos ocorrem simultaneamente com probabilidade razoável ou quais itens de um conjunto de dados estão presentes juntos com uma certa chance, buscando sua correlação. Apresentam a forma: SE atributo X ENTÃO atributo Y. É uma técnica bastante conhecida devido aos resultados de sucesso obtidos, especialmente nas análises da "Cestas de Compras" (*Market Basket*).

2.5.2 Regras de associação

Os algoritmos capazes de encontrar relacionamentos entre os dados são chamados de algoritmos de regras de associação e trabalham com a extração de conjuntos de atributos frequentes inseridos em um conjunto maior. Esses algoritmos variam bastante em relação à geração de subconjuntos e em como os conjuntos de atributos escolhidos

são suportados durante a geração das regras de associação. Segundo os autores Agarwal, Imielinski e Swami (1993), uma regra de associação tem o formato A $\longrightarrow$ B, onde A é chamado de antecedente, e B é chamado de consequente. A e B são conjuntos de itens ou transações, e a regra pode ser lida como: o atributo A frequentemente implica no atributo B. Para avaliar as regras geradas são utilizadas algumas medidas de interesse, onde as mais utilizadas são *suporte* e *confiança*.

Os autores Geng e Hamilton (2006) realizaram uma pesquisa e sugeriram estratégias para selecionar medidas adequadas para determinados domínios e exigências. As seguintes medidas são utilizadas para mensurar as regras geradas:

- Suporte: $P(AB)$. O suporte de uma regra é definido como sendo a fração de itens I que satisfazem o conjunto A e B da regra. Se o suporte não é grande o suficiente, isso significa que a regra não é digna de consideração ou que é simplesmente preterida e pode ser considerada mais tarde;
- Confiança: $P(A/B)$. É uma medida da força de suporte às regras e corresponde à significância estatística. A probabilidade de encontrar B da regra nas transações sobre a condição que essas transações também contenham A;
- Interesse: $P(B|A)/P(B)$ ou $P(AB)/P(A) * P(B)$. Utilizada para encontrar dependências, ela indica o quanto mais frequente torna-se B quando A ocorre.

Encontrar conjuntos de itens frequentes, com frequência maior ou igual à especificada pelo usuário como sendo o suporte mínimo, não é trivial, devido à explosão combinatória ocorrida ao gerar os subconjuntos de itens. Mas, uma vez que os conjuntos de itens frequentes são obtidos, é muito simples gerar regras de associação com confiança maior ou igual à especificada pelo usuário como sendo o valor mínimo (WU et al., 2008).

2.5.3 Algoritmo apriori

Um dos algoritmos mais utilizados para a mineração de dados, o Apriori foi introduzido por Agarwal, Imielinski e Swami (1993) como uma maneira de gerar regras de associação de dados. Basicamente, o algoritmo é composto por duas etapas:

- Encontrar todos os conjuntos de itens frequentes, ou seja, com valor de suporte superior ou igual ao suporte mínimo especificado pelo usuário. Em termos de custo computacional, esta é a etapa com o custo mais elevado;

- Usar os conjuntos de itens frequentes para induzir as regras de associação, com valores de suporte e de confiança superior ou igual ao suporte e a confiança mínimos especificados pelo usuário.

Empregando busca em profundidade o algoritmo é capaz de gerar conjuntos de itens candidatos (reconhecidos como o padrão) com k elementos a partir de conjuntos de itens de $k \neq 1$ elementos. A varredura só termina no último elemento da base de dados e padrões não frequentes são descartados. O pseudo-código apresentado a seguir demonstra o funcionamento do Apriori (SILVA, 2014).

Algoritmo 1: Pseudo código do algoritmo *Apriori*

```
Input: (T,minSupport)
// T é o conjunto de dados e minSupport é o suporte mínimo
begin
    L1={itens frequentes};
    for (k=2; L_{k-1} != ∅; k++ ) do
        C_k = candidatos gerados a partir de L_{k-1};
        // é o produto cartesiano L_{k-1} x L_{k-1} e elimina
        // qualquer conjunto de itens de tamanho k − 1 que não seja frequente
        foreach transação t ∈ T do
            #incrementa o contador de todos candidatos em C_k que estão ⊂ em t;
            L_k = candidatos em C_k com minSupport;
        end
        // fim do for each
    end
    // fim do for
end
return U_k L_k;
```

Como é comum na mineração de regra de associação, dado um conjunto de conjuntos de itens (por exemplo, das operações de varejo de moda, cada listagem individual itens comprados), o algoritmo tenta encontrar subconjuntos, que são comuns a pelo menos um número mínimo C dos conjuntos de itens. O Apriori usa uma abordagem *bottom-up*, onde subconjuntos frequentes são estendidos um item por vez (uma etapa conhecida como geração de candidatos), e grupos de candidatos são testados contra os dados.

2.5.3.1 Parâmetros do algoritmo apriori

Para encontrar relacionamentos ocultos entre as regras alguns parâmetros podem ser definidos para o algoritmo Apriori. O *suporte* é o parâmetro que indica o

percentual de vezes que o conjunto A aparece no conjunto de transações, o que também é chamado de cobertura da regra. Sua precisão, chamada de Confiança, é o número de instâncias que a regra prevê corretamente, expressa como uma percentagem de todas as instâncias a que se aplica.

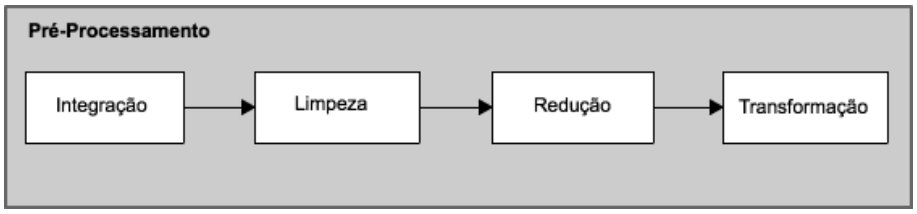
Os parâmetros confiança e suporte são essenciais para o funcionamento do algoritmo. Eles determinam diretamente tanto a quantidade como a qualidade das regras geradas e, por isto, é de fundamental importância que eles sejam avaliados no momento da configuração do algoritmo para a geração de regras de associação significativas (WITTEN; FRANK; HALL, 2011).

2.5.4 Pré-processamento

Os bancos de dados reais, ou que não são construídos apenas para testes, são altamente suscetíveis a ruídos, ausência de dados e inconsistentes devido a seu tamanho, geralmente grande (muitas vezes de vários terabytes ou mais), e sua provável origem de múltiplas fontes heterogêneas (SILVA, 2014).

Consequentemente, é possível que a qualidade desses dados esteja comprometida, e dados de baixa qualidade irão levar a resultados de mineração de dados de baixa qualidade. Portanto, necessita-se de uma fase anterior, que consiste na preparação dos dados para que possam ser aplicados os algoritmos de mineração, essa fase é denominada pré-processamento (HAN; KAMBER; PEI, 2006).

Figura 3 – Etapas do pré-processamento. Fonte: (SILVA, 2014)



Conforme a Figura 3 as principais etapas envolvidas no pré-processamento são:

- **Integração dos dados**: ao trabalhar com dados não é raro se deparar com dados provenientes de fontes heterogêneas como banco de dados, arquivos textos, planilhas, vídeos, imagens, etc. Por este motivo, percebe-se a necessidade da integração destes dados com o objetivo de obter uma base única e consistente. Em busca deste resultado uma análise aprofundada dos dados deve ser realizada, observando redundâncias, dependências entre as variáveis e valores conflitantes. Exemplo destas características são categorias diferentes para os mesmos valores, chaves divergentes, regras diferentes para os mesmos dados, entre outros.

- **Limpeza dos dados**: esta etapa tem o objetivo de eliminar os problemas como dados faltantes, valores errôneos e dados inconsistentes, de modo que eles não influenciem nos resultados. As técnicas usadas nesta atividade podem incluir a eliminação de registros com problemas, atribuição de valores padrões ou nulos, e a aplicação de técnicas de agrupamento (ou discretização de atributos) para auxiliar na obtenção de melhores resultados.

- **Redução dos dados**: embora a quantidade de dados usado na mineração seja grande na maioria das vezes, em alguns casos este volume é tão grande que torna o processo de análise dos dados e da própria mineração impraticável. Desta forma, técnicas de redução de dados devem ser aplicadas para que a massa de dados original seja convertida em uma massa de dados menor, mas com a mesma representatividade. Esta decisão sugere que os algoritmos de mineração sejam executados com mais eficiência, mantendo a qualidade do resultado. As estratégias adotadas nesta etapa podem ser a criação de estruturas otimizadas para os dados, a seleção de um subconjunto dos atributos, a redução da dimensionalidade e a discretização.

- **Transformação dos dados**: alguns algoritmos trabalham apenas com valores numéricos e outros apenas com valores nominais, nestes casos é necessário transformar os valores de acordo ao algoritmo que se deseja executar. Não existe um critério único para transformação dos dados e diversas técnicas podem ser usadas de acordo com os objetivos pretendidos. Algumas das técnicas empregadas nesta etapa são o agrupamento, a generalização, a normalização e a criação de novos atributos a partir de outros já preexistente.

A realização de uma ou de várias etapas técnicas de pré-processamento pode ser a chave para resolver o problema. Pode ser o fator decisivo entre uma modelagem algorítmica realizada com sucesso e a fracassada. Desta forma, é possível perceber que se o pré-processamento não for realizado corretamente, os dados que serão a base da mineração de dados poderão ser inviabilizados.

3 Metodologia

Este capítulo está dividido em duas partes. A primeira seção descreve os detalhes de todos os procedimentos efetuados sobre os dados de peixes da Usina de Lajeado, bem como a modelagem e os métodos seguidos para a importação dos dados primários. Na segunda seção serão apresentados os detalhes da arquitetura da ferramenta Filhote, com o propósito de detalhar o desenvolvimento da sua construção.

3.1 Estruturação e mapeamento dos dados

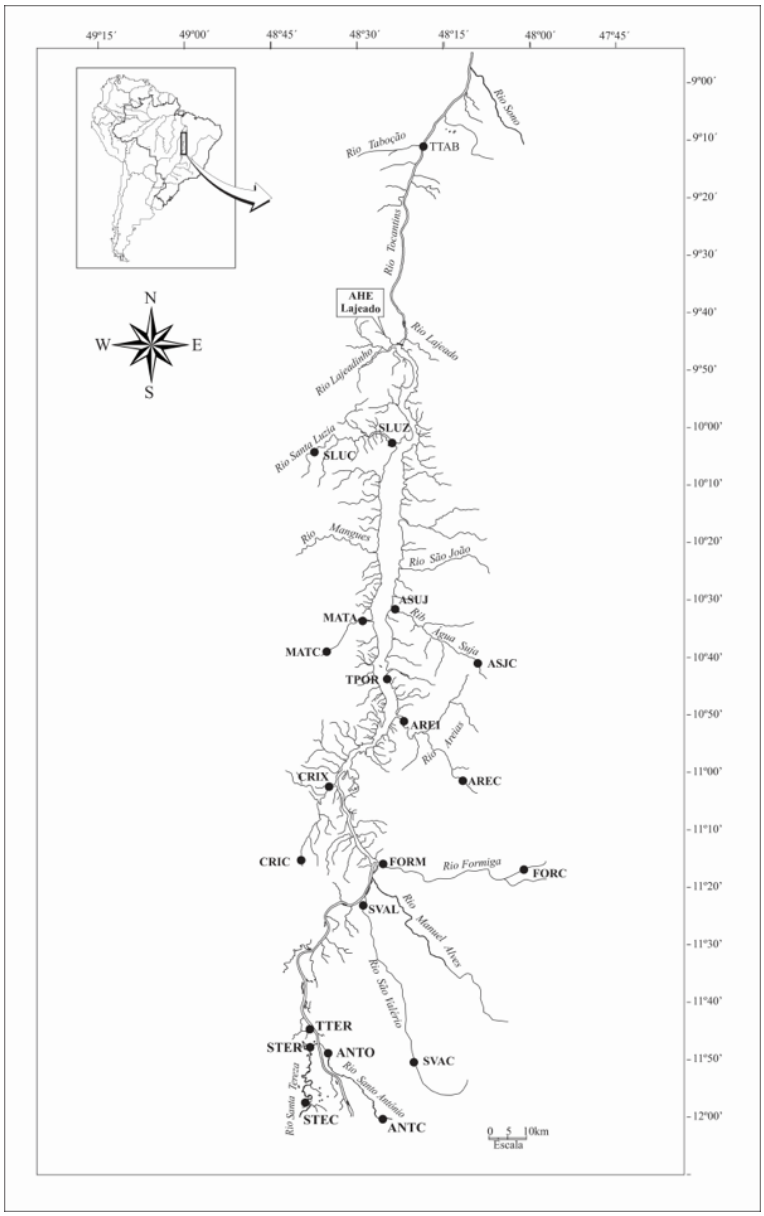
Esta seção descreve os detalhes das atividades realizadas sobre os dados primários, compreendendo a modelagem, a estruturação e a importação da base de dados de peixes da Usina de Lajeado. Para isto, buscou-se o entendimento de como foram realizadas as coletas, a identificação da região geográfica onde foram coletados os dados, a estrutura primária da base e a dinâmica utilizada para o levantamento deste conjunto.

A partir deste mapeamento inicial, realizou-se estudos sobre os dados visando a compreensão dos campos que fazem parte das planilhas, dentre outros aspectos que serão discutidos detalhadamente nesta seção.

3.1.1 Região geográfica das coletas de dados

Neste estudo foram trabalhados os dados de peixes coletados no reservatório da Usina de Lajeado, compreendida entre os municípios Miracema do Tocantins e Ipueiras, Figura 4 (AGOSTINHO et al., 2007), no período de outubro de 1999 a março de 2012, contemplando amostragens realizadas antes, durante e após a construção da barragem da Usina de Lajeado.

Figura 4 – Locais de coleta dos dados de peixes



O enchimento do reservatório foi iniciado no final de setembro de 2001 e concluído em fevereiro de 2002 (AGOSTINHO et al., 2007).

O conjunto de dados resultantes deste acompanhamento da construção da usina foi registrado e armazenado em fichas de papel durante a rotina de coleta. Após os registros *in loco*, dados foram digitados em planilhas eletrônicas

3.1.2 Tratamento dos dados

Para melhorar a compreensão e o entendimento sobre os dados objeto deste trabalho, foi necessário um período de estudo sobre eles e a realização de entrevistas com pesquisadores do Neamb, no intuito de colher detalhes dos dados armazenados. Além

disso, para que fosse possível o completo entendimento dos campos que compunham as planilhas, realizou-se uma visita *in loco* para observar pré-amostras aleatorizada de um conjunto de peixes, em 2014.

As tarefas que compreendem o estudo das planilhas recebidas do grupo de pesquisadores responsáveis pelos dados foram divididas em 2 fases.

3.1.2.1 Fase 1 – Descrição dos campos

Nesta fase buscou-se detalhar os campos que fazem parte das duas planilhas eletrônicas em formato *xls*, sendo uma planilha com dados abióticos e outra com dados bióticos de peixes, onde todos os tipos de dados eram considerados como texto.

A planilha com dados abióticos (imagem no Anexo A) contém 10.967 registros e tem os campos detalhados na Tabela 2.

Tabela 2 – Descrição dos dados abióticos

Campo	**Significado**
Projeto	Nome do projeto
Amostra	Número da amostra real armazenada
Local	Denominação do local em que a amostra foi coletada
Data	Hora em que as redes foram revistadas
Hora	Hora em que a coleta foi realizada
Margem	Margem em que foi jogado o aparelho de coleta
SMF	Indica se a coleta foi realizada na “superfície”, “meio”, ou “fundo”
Ativ	Informação sobre a atividade da água, indicando qual das amostras das 24 horas estava sendo realizada
Vento	Informação sobre a presença ou ausência de vento
Nebulosidade	Grau de cobertura de nuvem no céu
Chuva	Informação sobre a presença ou ausência de chuva
Tar	Temperatura do ar
Tagua	Temperatura da água
Profund	Profundidade do rio no local
ODmg	Oxigênio dissolvido
Transp	Transparência da água
PH	PH da água
Cond	Condutividade da água

E a planilha com dados bióticos (imagem no Anexo B), contém 433.641 registros e tem os campos dispostos na Tabela 3.

Tabela 3 – Descrição dos dados bióticos

Campo	**Significado**
Número	Número de registro do espécime coletado, utilizado como chave para as demais análises
Projeto	Nome do projeto
Ambiente	Ambiente da coleta (canal do rio principal, tributário, lagoa)
Ponto	Ponto de realização da coleta no canal do rio (superfície, fundo, margem)
Local	Denominação do local em que a amostra foi coletada
Data	Data em que a coleta foi realizada
Hora	Hora em que as redes foram revistadas
Espécie	Espécie do peixe
SpGrupo	Agrupamento das espécies com problemas taxonômicos por gênero
Ctrof	Categorização trófica das espécies em relação a dieta
Migrad	Categorização das espécies em relação as características reprodutivas
Aparelho	Aparelho utilizado na coleta
Apar	Especificação dos aparelhos de pesca
Lt	Comprimento total
Ls	Comprimento padrão
Wt	Peso total
Sexo	Sexo do peixe
Estádio	Estádio de desenvolvimento gonadal
Estad	Estádio de desenvolvimento gonadal agrupado
Wg	Peso da gônada
Rgs	Relação gônadossomática
Gre	Grau de repleção do estômago
Gri	Grau de repleção do intestino
Gv	Grau de gordura visceral
We	Peso do estômago
Wv	Peso da víscera
Ordem	Ordem
Família	Família
Gnespecie	Nome da espécie

Vale frisar que o campo *numero* da tabela *abiotico* é considerado chave para algumas análises internas do grupo de pesquisa, mas não é chave primária para este conjunto de dados não sendo, inclusive, preenchido para todos os registros. As tarefas envolvidas na descrição dos campos foram de fundamental importância para que fosse possível reconhecer todos os campos que fazem parte da metodologia de coleta de dados de peixes utilizada pelos especialistas.

3.1.2.2 Fase 2 – Pré-processamento e validação primária

De maneira a elevar a qualidade dos dados, no que se refere à verificação de ruídos, redundâncias e a existência de relacionamentos entre eles, foi aplicado o pré-processamento seguindo as atividades de limpeza, transformação e integração descritas no Capítulo 2. A tarefa de redução dos dados, que compõe o pré-processamento não foi aplicada neste conjunto de dados, visto que este momento estava relacionado à importação dos dados e todos os dados referentes ao conjunto foram importados. Os tópicos a seguir apresentam os detalhes das atividades realizadas nesta fase.

Limpeza

Na atividade que compreende a limpeza dos dados, foram substituídos por nulo os campos que continham caracteres especiais considerados fora do padrão pelo pesquisador responsável. Por exemplo, o valor de um campo que armazena a temperatura do ar deveria ser numérico, porém havia um caracter *'n'*. Em casos como este, o valor do campo foi considerado inválido e considerado nulo.

Ainda sobre esta atividade chegou-se ao entendimento de que havia campos que eram derivados de outros campos, e eles deveriam ser removidos, já que para o modelo de banco de dados isto seria considerado redundância.

Além disso, foi necessário ajustar datas que eram posteriores ao ano 2000, pois elas estavam com o ano incorreto (exemplo: 2001 = 1901; 2002 = 1902; 2009 = 1909; 2010 = 1910).

Nas duas planilhas haviam sub-planilhas com identificação ou descrição de campos e valores da planilha principal. Por exemplo, o valor armazenado no campo local, tanto da planilha de dados abióticos quanto da planilha de dados bióticos, continha apenas a sigla do local de coleta, sem o nome do local ou o nome do município. Na sub-planilha tinha um campo com o nome do local, outro campo com a sigla e outro com o Município. Estes campos foram analisados e considerados no momento da inserção para que na entidade com dados sobre o local da coleta tivesse o nome completo do local, bem como a sigla e o nome do município.

Transformação

Os dados passaram por um processo de transformação onde eles puderam ser mapeados para um conjunto de tabelas, considerando os seus tipos de acordo à sua utilização. Na planilha eletrônica todos os campos tinham o mesmo tipo de dados, que era texto. No mapeamento, foi necessário considerar um tipo *numérico* para campos com valores numerais, por exemplo as chaves, e o tipo *string* para campos com textos.

Integração

Considerando a disposição dos dados primários, foi necessário que estes dados passassem por um processo de integração, de relacionamento, para estivessem em um só repositório. Para isto, foi necessária uma análise aprofundada sobre os dados, onde foram observadas suas características para posterior estruturação e relacionamento entre eles.

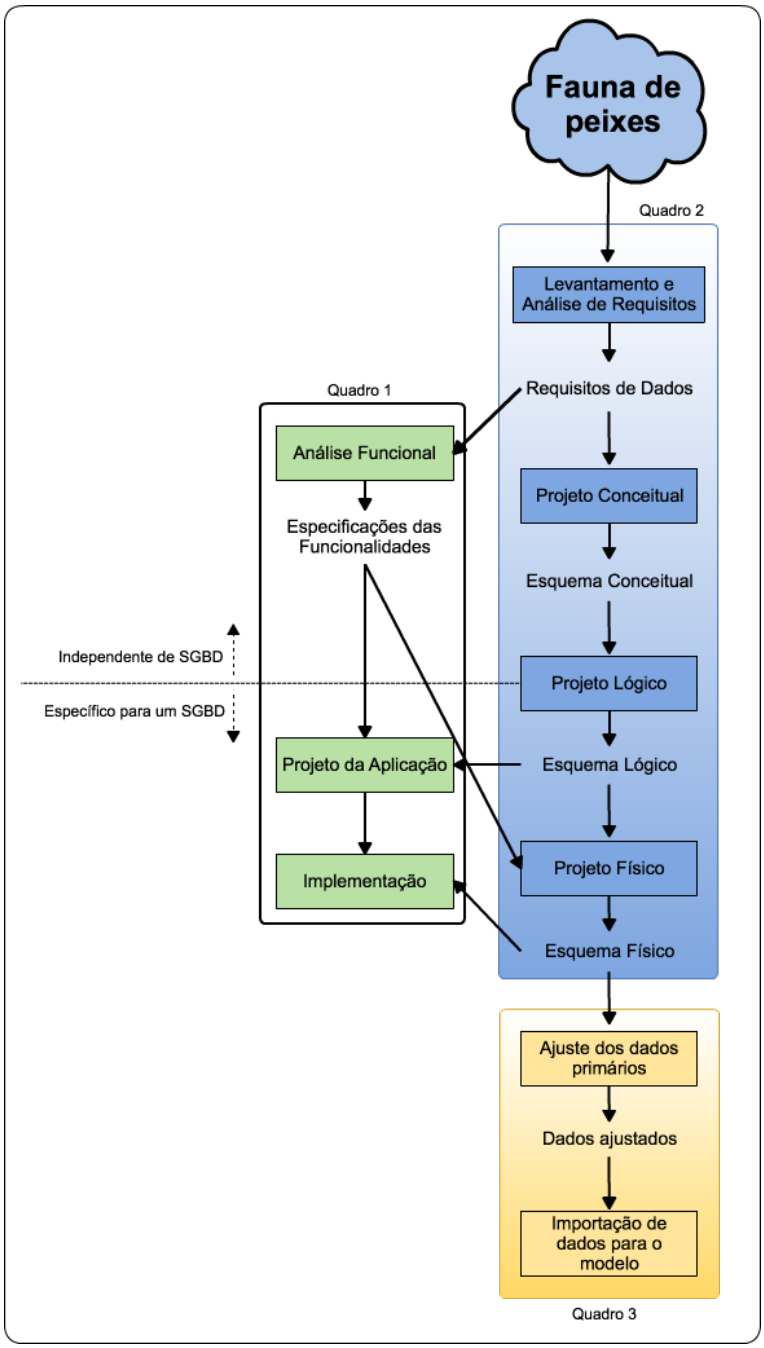
O resultado do pré-processamento contribuiu sobremaneira com o desenvolvimento do projeto de banco de dados, bem como para a atividade de importação para a base construída. Detalhes deste projeto estão descritas na subseção 3.1.3.5.

3.1.3 Modelagem e Estruturação do banco de dados

As atividades da estruturação do banco de dados, seguiram os passos sugeridos por Elmasri e Navathe (2011), ilustrados na Figura 5 no quadro 2. Estas fases envolvem o levantamento e análise de requisitos, o projeto conceitual, o projeto lógico e o projeto físico.

Extensivamente a estas tarefas, em adição ao modelo proposto pelos autores e, especialmente por haver dados primários, neste trabalho foram realizadas as atividades descritas no quadro 3 da Figura 5, envolvendo o ajuste dos dados primários e a sua importação para o modelo.

Figura 5 – Projeto do banco de dados - adaptada de (ELMASRI; NAVATHE, 2011)



Os detalhes de cada um dos processos que compõem o projeto do banco de dados serão apresentados nas seções subsequentes.

3.1.3.1 Levantamento e análise de requisitos

A atividade de levantamento e análise de requisitos iniciou-se juntamente com a etapa de descrição dos dados, realizando entrevistas com especialistas, gerando anotações sobre os requisitos desejados tanto para a estruturação dos dados quanto para a

ferramenta, bem como sobre os requisitos implícitos necessários ao bom funcionamento do projeto.

Esta etapa resultou nos requisitos constantes nos tópicos seguintes:

R1: A ferramenta deverá atender a um grupo de pesquisadores. Cada usuário deste grupo deve ser registrado e ter sua senha pessoal de acesso à ferramenta. Todos os usuários podem ter a mesma visão e controle dos dados;

R2: Este grupo coleta e administra uma série de dados de peixes e do meio ambiente em que vivem;

R3: Deseja-se armazenar dados de peixes (bióticos) e do meio ambiente (abióticos) de forma relacionada entre eles, mas considerando que podem haver coletas apenas de dados bióticos e outras apenas de dados abióticos; Para um dado biótico pode haver um dado abiótico; um dado abiótico pode estar relacionado a nenhum ou a vários dados bióticos;

R4: É necessário que os dados primários de peixes pertencentes ao grupo Neamb/UFT possam se adequar ao modelo e serem importados;

R5: É necessário que os dados primários registrados em duas planilhas eletrônicas – uma com dados bióticos e outra com dados abióticos – sejam relacionados entre si, considerando os atributos *data*, *hora*, *local*, *smf* e *ponto*, existentes para todos o conjunto de dados;

R6: Novos projetos podem ser executados e novos dados podem estar relacionados a um projeto;

R7: Deseja-se ter acesso aos dados para consulta, atualização e exclusão. No caso de todas as *exclusões* será considerada a função de deixar o dado inativo no banco de dados, invisível ao usuário, porém sem exclusões definitivas;

R8: Para exclusão (tratatada aqui como *inativação*) deve-se considerar as chaves de relacionamento – não deve ser permitido inativar um registro utilizado por outra entidade;

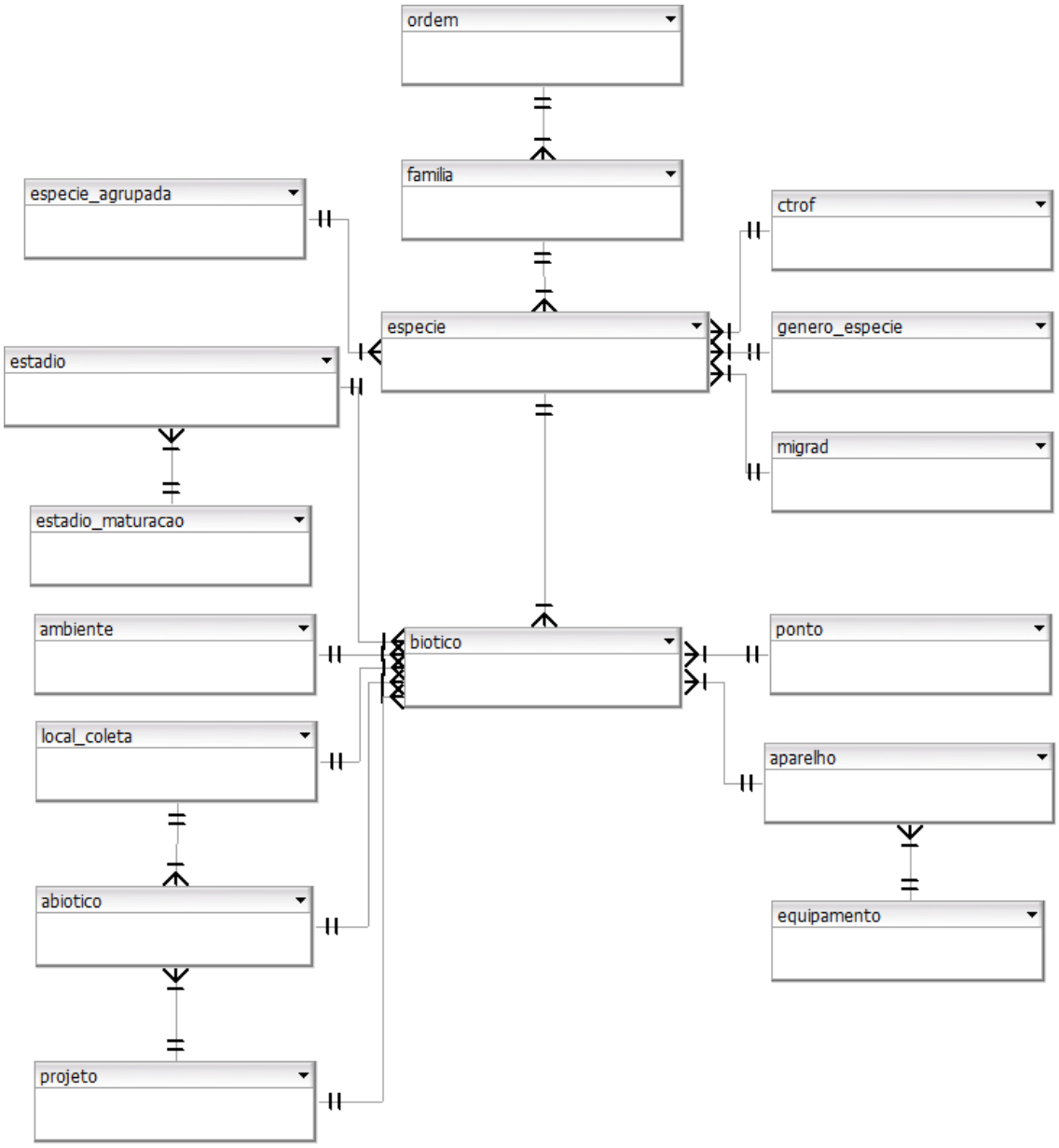
R9: Deseja-se exportar os dados armazenados em formato de arquivo texto;

R10: Pretende-se poder agregar um meio de análise de dados à ferramenta.

3.1.3.2 Projeto conceitual

Na etapa do projeto conceitual, portanto ainda sem considerar detalhes de armazenamento e implementação, foi modelado o diagrama entidade-relacionamento (DER), ilustrado na Figura 6, gerando um esquema conceitual a partir dos requisitos de dados elicitados.

Figura 6 – Modelo entidade-relacionamento



Observe que o DER apresenta as entidades *biotico* e *abiotico* e as entidades que se relacionam com elas. Além disso, neste diagrama estão representadas as cardinalidades que descrevem a relação as entidades.

3.1.3.3 Projeto lógico

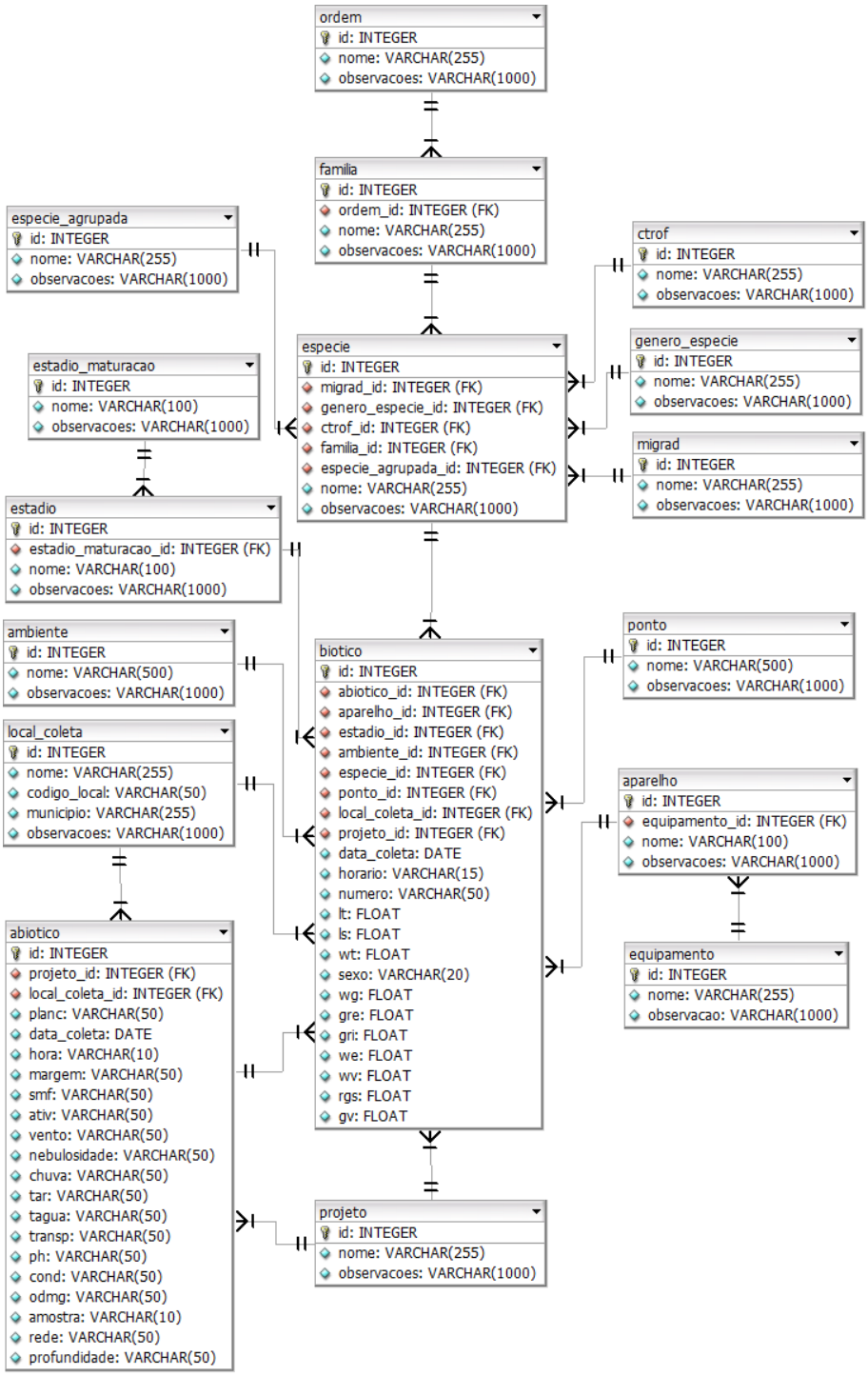
A partir da modelagem conceitual realizada, iniciou-se a etapa da criação da estrutura do banco de dado, realizando o mapeamento do esquema conceitual para o esquema lógico.

Nesta etapa verificou-se que o modelo relacional se mostrou adequado à estru-

tura deste projeto, especialmente por se tratar de dados não-complexos e por incluírem garantias de integridade dos dados, gerenciamento de concorrência, recuperação de falhas, segurança e otimização de consultas.

Desta forma, foi criado o esquema lógico que descreve os campos, as chaves e os relacionamentos constantes no modelo. O resultado desta fase está ilustrado na Figura 7.

Figura 7 – Esquema lógico de banco de dados



3.1.3.4 Projeto físico

Sabendo que a etapa do projeto físico corresponde ao mapeamento do modelo conceitual para as rotinas de implementação da estrutura do banco de dados, esta atividade foi realizada seguindo os três passos descritos a seguir, todos eles utilizando a linguagem padrão universal para manipulação de dados SQL (*Structured Query Language*), através dos comandos pertencente ao conjunto DDL (*Data Definition Language*):

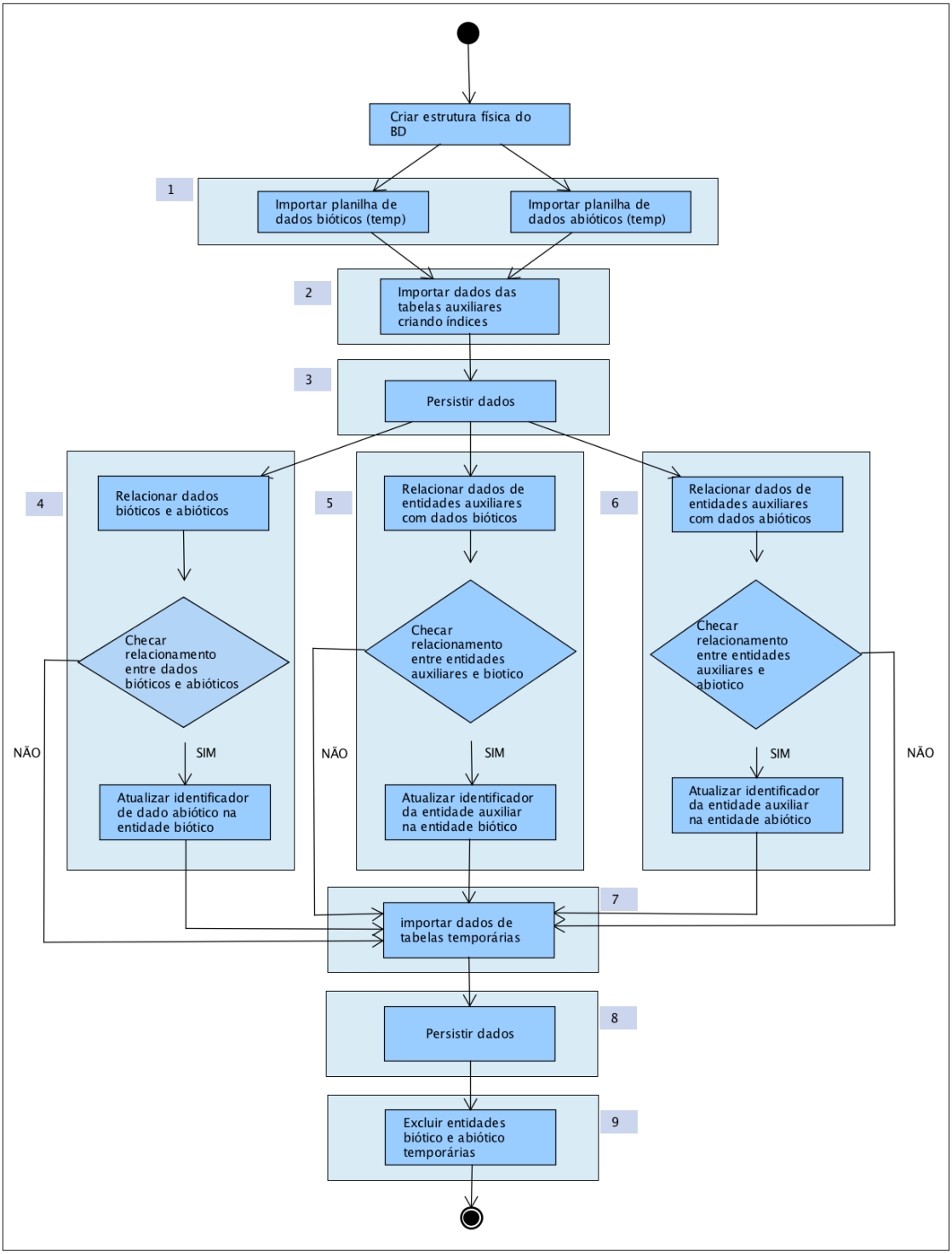
- Criação do esquema de banco de dados: foi gerada uma base de dados;
- Criação do conjunto de tabelas e relacionamentos, entre elas: dezessete tabelas foram criadas na base de dados;
- Criação de duas tabelas temporárias: foi criada uma tabela temporária para armazenar os dados bióticos e outra para armazenar os dados abióticos primários das planilhas, exatamente como estavam após o pré-processamento. Estas tabelas foram extremamente necessárias para a manipulação de dados no momento da importação, descrito na subseção 3.1.3.5.

3.1.3.5 Importação dos dados

Com a estrutura do banco de dados criada fisicamente, conforme ilustrado no quadro 3 da Figura 5, iniciou-se a importação dos dados contidos nas planilhas eletrônicas para o modelo de banco de dados.

A Figura 8 ilustra as atividades realizadas para o processo de relacionar e persistir os dados nas tabelas.

Figura 8 – Detalhamento da importação dos dados



A seguir são detalhados os processos constantes no fluxograma, para melhor entendimento.

Quadro 1

Como pode ser observado no quadro 1 da Figura 8, a tarefa de importação das

tabelas temporárias consistiu no processo de persistir os dados contidos nas planilhas na base exatamente como eles estavam, com os tipos criados automaticamente pelo SGBD. Estas relações populadas permitiram o início do processo de persistência dos dados no novo modelo.

Quadro 2

A importação dos dados para as entidades auxiliares considerou cada um dos campos que são chave estrangeira nas tabelas *biotico* e *abiotico* individualmente. Para esta atividade foi desenvolvida uma rotina que agrupa os dados dos campos a serem importados nas tabelas auxiliares, ordena-os e insere em sua respectiva tabela.

Esta operação pode ser exemplificada com o preenchimento da tabela *especie*, na qual foi feita uma consulta na tabela *biotico temporária*, selecionando apenas o campo *especie* e todos os seus valores agrupando-os e ordenando-os, obtendo, portanto, um retorno de todos os nomes de espécies da planilha de dados bióticos primária sem redundâncias e em ordem alfabética.

Após esta seleção, foi desenvolvida uma rotina de inserção para que os dados fossem inseridos na tabela *especie* definitiva. Esta operação implementou a criação de chaves auto-incremento para a indexação dos dados inseridos.

Este mesmo processo foi realizado sobre as entidades que tem relacionamento com as duas principais tabelas do modelo, *biotico* e *abiotico*. Para algumas delas havia a necessidade de relacionamento com outras tabelas e, então, a prioridade de inserção teve que ser avaliada. Um destes casos é a relação entre as entidades *ordem* e *familia* que, por se tratar de tabelas relacionadas (cada família tem uma ordem), a entidade ordem teve seus dados inseridos primeiro, concluindo a tarefa respeitando o relacionamento previsto.

Quadro 3

O quadro 3 descreve o processo de persistir os dados que foram importados para as entidades, permitindo que eles possam ser recuperados posteriormente.

Quadro 4

No processo ilustrado no quadro 4, foi realizada a busca pelo o relacionamento entre as entidades *biotico* e *abiotico*. Para isto, o primeiro passo foi a checagem da existência do relacionamento de cada registro biótico com um registro da entidade *abiotico*, considerando os campos chave *data*, *hora*, *local*, *ponto* e *smf*.

Neste procedimento, a existência do relacionamento entre estas tabelas implicou na atualização do campo *abiotico_id* da relação *biotico* com o valor referente a esta chave estrangeira. Vale frisar que nem todo dado biótico tem um registro de dado

abiótico. Também, nem todo dado abiótico armazenado neste banco de dados tem relação com um dado biótico. Isto pode acontecer, por exemplo, quando em uma coleta os aparelhos para registros de um tipo de dados tiverem problema, ou mesmo quando o objetivo da coleta não contempla a coleta dos dois conjuntos.

Quadro 5

O quadro 5 descreve a busca pelo relacionamento entre as chaves estrangeiras constantes na relação *biotico* e suas tabelas de origem. Neste caso, o valor era constante na tabela temporária *biotico* e comparado com o campo nome da tabela mapeada no quadro 2. Assim, quando havia relacionamento, o campo com a chave estrangeira recebia o valor do campo *id* da tabela auxiliar.

Quadro 6

O processo ilustrado no quadro 6 descreve o processo semelhante ao descrito no quadro 5, porém atualizando os campos com chaves estrangeiras da entidade *abiotico* às suas tabelas de origem.

Quadro 7

Após a finalização do processo de relacionamento entre as entidades do modelo, os dados das relações *biotico* e *abiotico* temporárias foram inseridos nas tabelas definitivas, onde seus campos continham tipos de dados de acordo à sua função e descrição do modelo. Este processo está ilustrado no quadro 7.

Quadro 8

Desta forma, o último passo para conclusão da estruturação do banco de dados foi a exclusão das tabelas temporárias. Sendo assim, o modelo permaneceu com 17 entidades relacionadas entre si, com os dados primários persistidos e indexados no SGBD.

Neste capítulo foram descritas as atividades realizadas com o objetivo de transformar os dados armazenados nas duas planilhas eletrônicas em um conjunto estruturado de dados sob o modelo relacional.

3.2 Especificação da ferramenta proposta

3.2.1 Introdução

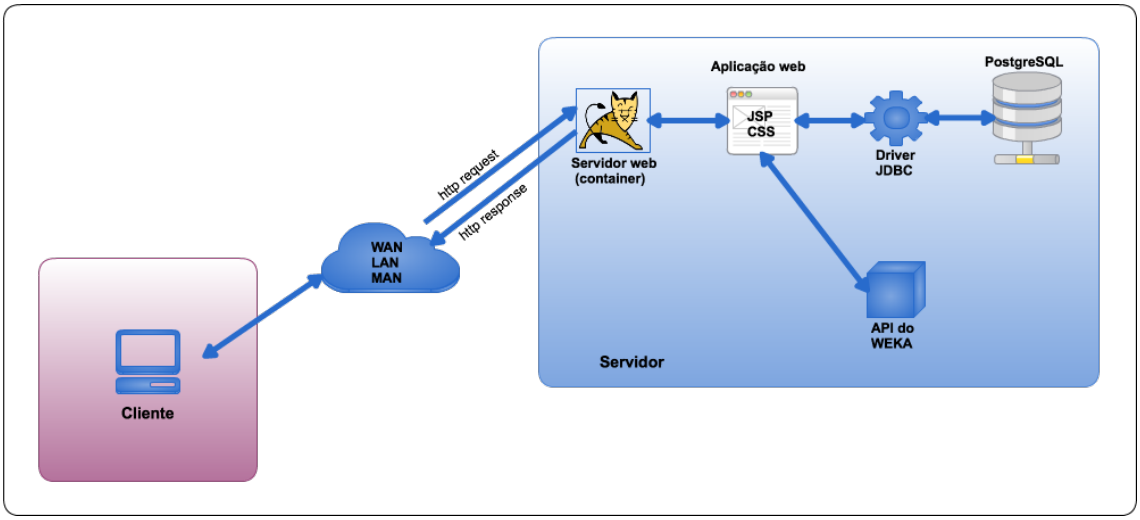
Esta seção descreve a proposta da construção da ferramenta Filhote, que terá o papel de servir de instrumento de administração dos dados deste projeto. Além disso, é apresentado o detalhamento dos componentes, a especificação do software e as tecnologias utilizadas no desenvolvimento.

3.2.2 Visão Geral da Proposta

Com o intuito de atender os objetivos deste trabalho, paralelamente à estruturação dos dados primários detalhados na seção 3.1, foi realizada a construção do projeto da ferramenta. Esta tarefa, teve o objetivo de atender as necessidades apresentadas pelo grupo de trabalho do Neamb/UFT, que apresentou as especificações sobre como os dados são coletados, armazenados e recuperados.

A Figura 9 apresenta a estrutura completa da ferramenta e as subseções a seguir detalham cada um dos seus componentes.

Figura 9 – Arquitetura proposta



- **Cliente**

  A entidade cliente representa os usuários que acessarão a aplicação, realizando atividades de consulta e manipulação dos dados, através de um *browser* em um dispositivo conectado à internet ou a uma rede local;

- **Conectividade**

  Representa o acesso aos dados através da conexão à rede local ou por meio de acesso à internet. Dessa forma a aplicação não precisa ser instalada no computador do cliente e pode ser acessada de qualquer dispositivo desde que haja uma conexão de dados.

- **Servidor web**

  O servidor *web* utilizado é o *Apache Tomcat*[1], na versão 8.0.15, que pode ser descrito de maneira simplificada como sendo um *container web* de código aberto

[1] Disponível em: http://tomcat.apache.org

implementado na linguagem *Java*, na qual é utilizado para o desenvolvimento de *servlets*, sites dinâmicos e *webservices*, dentre outros.

- **Aplicação web**

  Este item da arquitetura proposta é a ferramenta propriamente dita, com todas suas funcionalidades. A linguagem de programação escolhida para o projeto foi o *Java* versão 8 por ser multiplataforma e principalmente por facilitar a ligação com a *API* do *WEKA*. Para criação do conteúdo dinâmico dentro das classes foi utilizado a tecnologia *JavaServer Pages* (JSP). Os estilos das páginas ficam a cargo do *Cascading Style Sheets* (CSS) separando assim o formato e o conteúdo da aplicação.

  Esta aplicação *web* se baseou no padrão de arquitetura de *software* MVC (*Model-View-Controller*), visando separar as regras de negócio da interface com a qual o usuário interage com o principal objetivo de facilitar a reusabilidade do código.

  O *hardware* utilizado para o desenvolvimento foi um MacBook Pro, 2.8GHz Dual-core Intel Core i7, Turbo Boost up to 3.3GHz, 16GB 1600MHz DDR3L SDRAM, com o sistema OS X *Yosemite* na versão 10.10.2. Os *softwares* utilizados para a modelagem foram o Astah *professional*[2] e o DBDesigner[3], e para o desenvolvimento foram o NetBeans[4], na versão 8.0.2 e o pgAdmin III[5], na versão 1.20.

  Para a proteção da senha do usuário da aplicação foi utilizado o conjunto de funções *hash* criptográficas *sha-256* (algoritmo de hash seguro) projetadas pela NSA (Agência de Segurança Nacional dos EUA).

  O projeto da aplicação descrevendo suas funcionalidades está detalhado na subseção 3.2.3 deste capítulo.

- **Driver JDBC**

  Este módulo da ferramenta permite que a aplicação *Java* possa interagir com o banco de dados. O driver JDBC (*Java Database Connectivity*) fornece a conexão ao banco de dados através de um conjunto de classes e implementa o protocolo para transferir a consulta e o resultado entre cliente e o banco de dados.

- **PostgreSQL**

  Este módulo representa o armazenamento dos dados no SGBD que, neste caso, é o PostgreSQL[6] na versão 9.4.5.

[2] Disponível em: http://astah.net/editions/professional
[3] Disponível em: http://www.fabforce.net/dbdesigner4/
[4] Disponível em: https://netbeans.org
[5] Disponível em: http://www.pgadmin.org
[6] Disponível em: http://www.postgresql.org

Este SGBD é distribuído sob a licença BSD (*Berkeley Software Distribution*), oferece suporte a consultas complexas, transações, controle de concorrência multi-versão, tipos de objetos definidos pelo usuário e *views* (BORGES, 2014). Estas características subsidiaram a escolha por este banco de dados.

Convém salientar que o detalhamento do modelo relacional, bem como a estruturação e a especificação dos componentes do modelo são apresentados na subseção 3.2.3.

- ***API* do WEKA**

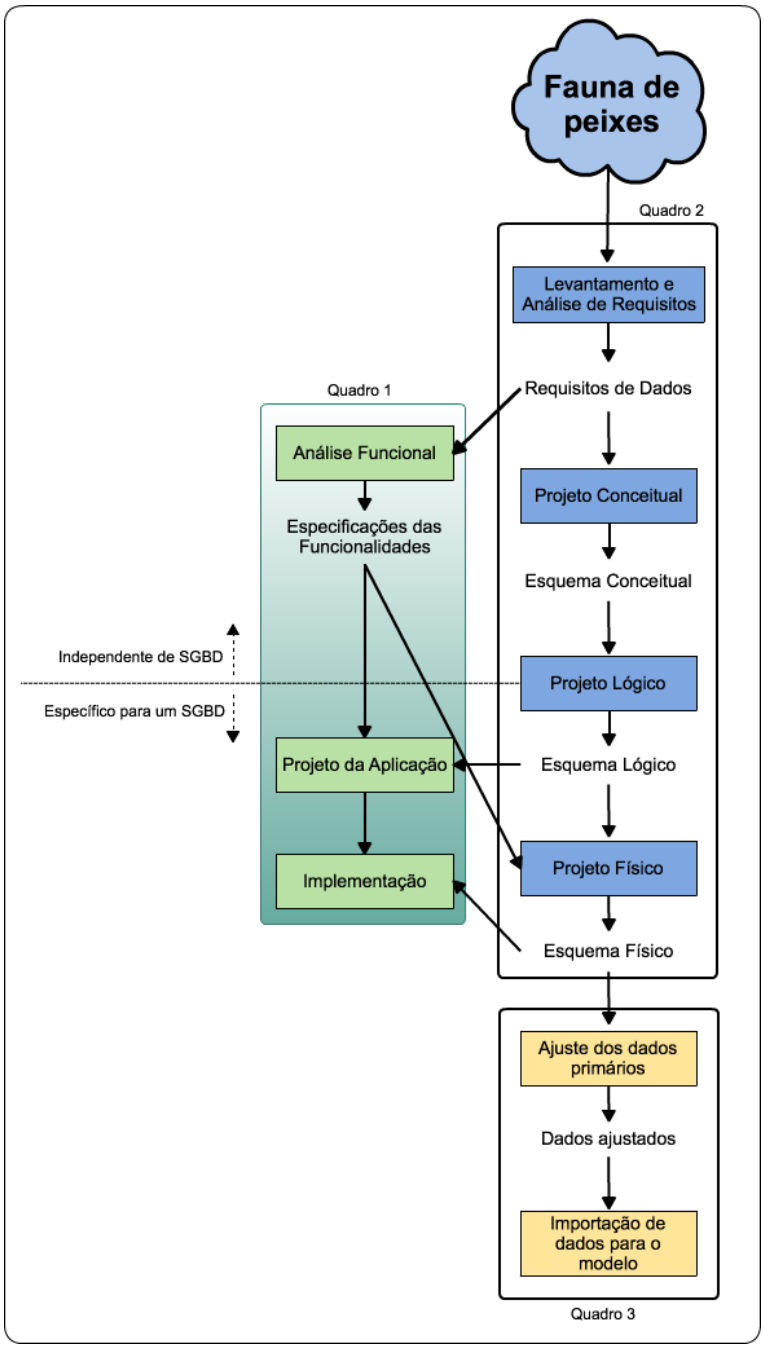
  A *API* é utilizada para acesso das funcionalidades do *framework WEKA* pela aplicação *web* Filhote. Seu uso foi motivado pelas características de facilidade de aplicação de algoritmos de Mineração observadas em Silva et al. (2013) e pela disponibilidade de uma *API* pública. A subseção 3.2.4 deste capítulo demonstra a descrição da utilização, bem como os detalhes da integração com a aplicação web.

3.2.3 Modelagem da aplicação

A modelagem desta aplicação inicia-se com a análise funcional, elicitando os requisitos descritos pelo usuário. Além disso, abrange o projeto do banco de dados, descreve a estrutura a ser implementada, bem como detalhes internos de funcionamento desta ferramenta.

Nesta etapa, foram seguidos os passos apresentados na Figura 10 constantes no quadro 1.

Figura 10 – Projeto da aplicação. Fonte: (ELMASRI; NAVATHE, 2011) - adaptada



Na primeira etapa, análise funcional, foram analisadas as operações definidas pelos usuários que serão aplicadas no banco de dados, incluindo recuperações e atualizações, obtendo os Requisitos Funcionais da aplicação. Estendeu-se um pouco mais a tarefa de análise funcional e foram levantados alguns Requisitos não Funcionais fundamentais ao bom funcionamento da aplicação.

Abaixo estão listados os Requisitos Funcionais e os Requisitos não Funcionais, respectivamente.

3.2.3.1 Requisitos funcionais

RF1: Permitir a inclusão, consulta, alteração e inativação dos dados referentes ao projeto;

RF2: Permitir a inclusão, consulta, alteração e inativação dos dados referentes local de coleta;

RF3: Permitir a inclusão, consulta, alteração e inativação dos dados referentes ao equipamento utilizado na coleta;

RF4: Permitir a inclusão, consulta, alteração e inativação dos dados referentes ao aparelho utilizado na coleta;

RF5: Permitir a inclusão, consulta, alteração e inativação dos dados referentes à ordem;

RF6: Permitir a inclusão, consulta, alteração e inativação dos dados referentes à família;

RF7: Permitir a inclusão, consulta, alteração e inativação dos dados referentes ao estádio da maturação;

RF8: Permitir a inclusão, consulta, alteração e inativação dos dados referentes ao estádio;

RF9: Permitir a inclusão, consulta, alteração e inativação dos dados referentes à espécie agrupada;

RF10: Permitir a inclusão, consulta, alteração e inativação dos dados referentes à espécie;

RF11: Permitir a inclusão, consulta, alteração e inativação dos dados referentes ao ponto da coleta;

RF12: Permitir a inclusão, consulta, alteração e inativação dos dados referentes ao ambiente da coleta;

RF13: Permitir a inclusão, consulta, alteração e inativação dos dados referentes às características de migrad;

RF14: Permitir a inclusão, consulta, alteração e inativação dos dados referentes às características de gênero/espécie;

RF15: Permitir a inclusão, consulta, alteração e inativação dos dados referentes às características de categoria trófica;

RF16: Permitir a inclusão, consulta, alteração e inativação dos dados referentes a um item biótico com seus relacionamentos;

RF17: Permitir a inclusão, consulta, alteração e inativação dos dados referentes a um item abiótico com seus relacionamentos;

RF18: Permitir a exportação de dados para o formato *csv* (*Comma-separated values*);

RF19: Permitir a mineração de dados através da geração das regras de associação.

3.2.3.2 Requisitos não funcionais

RNF1: Estar preparada para ser acessada de qualquer lugar do planeta que tenha conexão com a internet;

RNF2: Utilizar ferramentas gratuitas;

RNF3: Deve executar ao menos nos navegadores mais populares (*chrome, firefox, safari, internet explorer*);

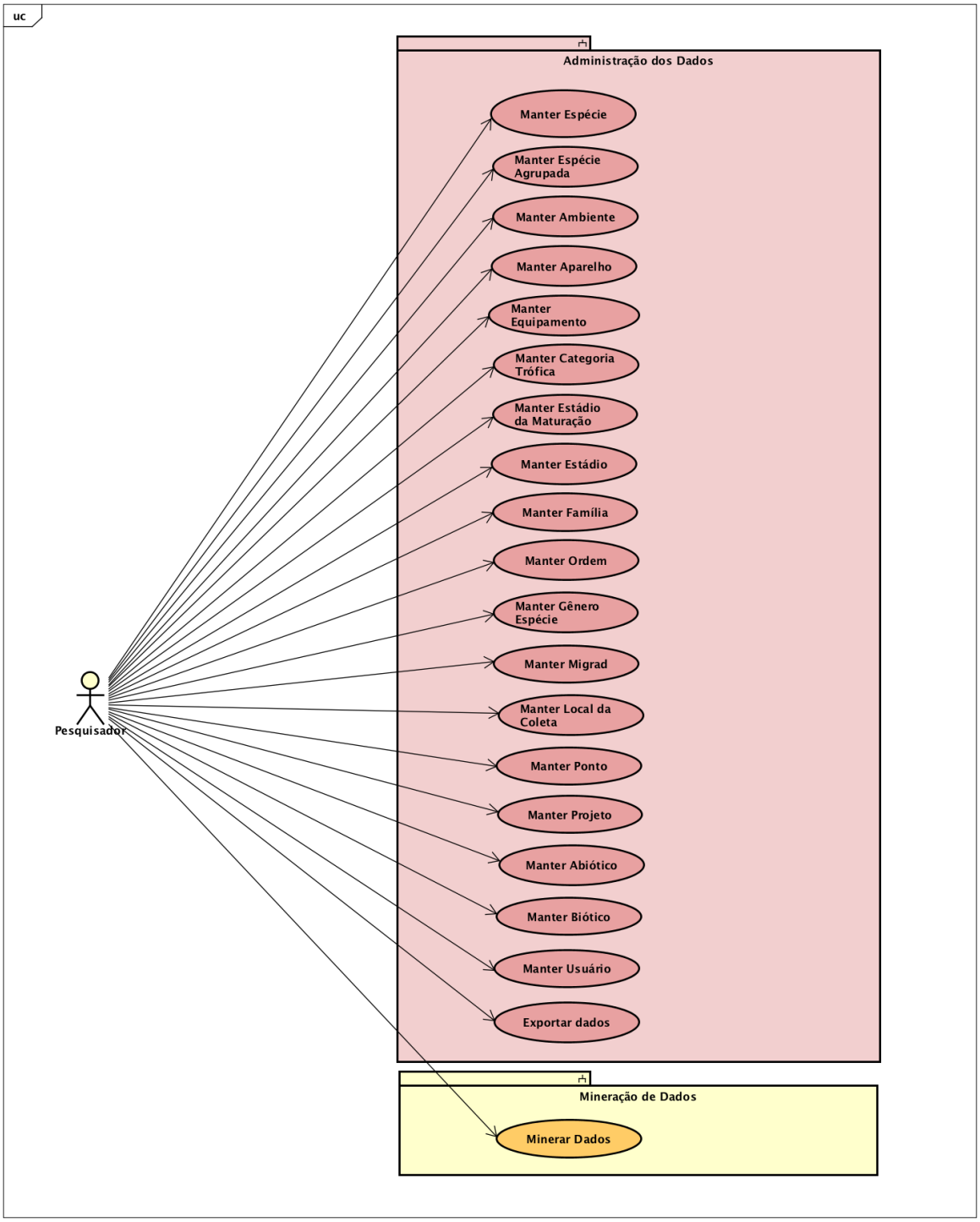
RNF4: Desenvolvimento utilizando linguagem Java;

RNF5: Integração com o *Weka*;

RNF6: Deve disponibilizar acesso aos dados apenas a usuários registrados.

Com o intuito de ilustrar os requisitos funcionais, a Figura 11 apresenta o diagrama de caso de uso.

Figura 11 – Diagrama de caso de uso



Este diagrama foi separado em duas partes para facilitar o entendimento, conforme ilustrado na Figura 11. Observe que o conjunto de operações que compõe a *administração dos dados* representa o módulo que descreve as funcionalidades de inclusão, seleção, atualização e exclusão de dados, que também serão chamadas neste trabalho de CRUD's (*Create, Read, Update, Delete*). Conforme já mencionado neste

trabalho, a funcionalidade *exclusão* tratará os dados na forma de inativação, deixando o dado invisível ao usuário, mas presente no banco de dados.

Além disto, faz parte deste conjunto a funcionalidade de exportar dados, onde os usuários podem selecionar quais grupos da base de dados desejam exportar, criando um arquivo no formato *csv*.

A aplicação permitirá ainda a mineração dos dados existentes em sua base, conforme o caso de uso ilustrado no bloco *minerar dados*, descrito na Figura 11. É através desta funcionalidade que os usuários poderão gerar regras de associação de acordo com as configurações parâmetros informados.

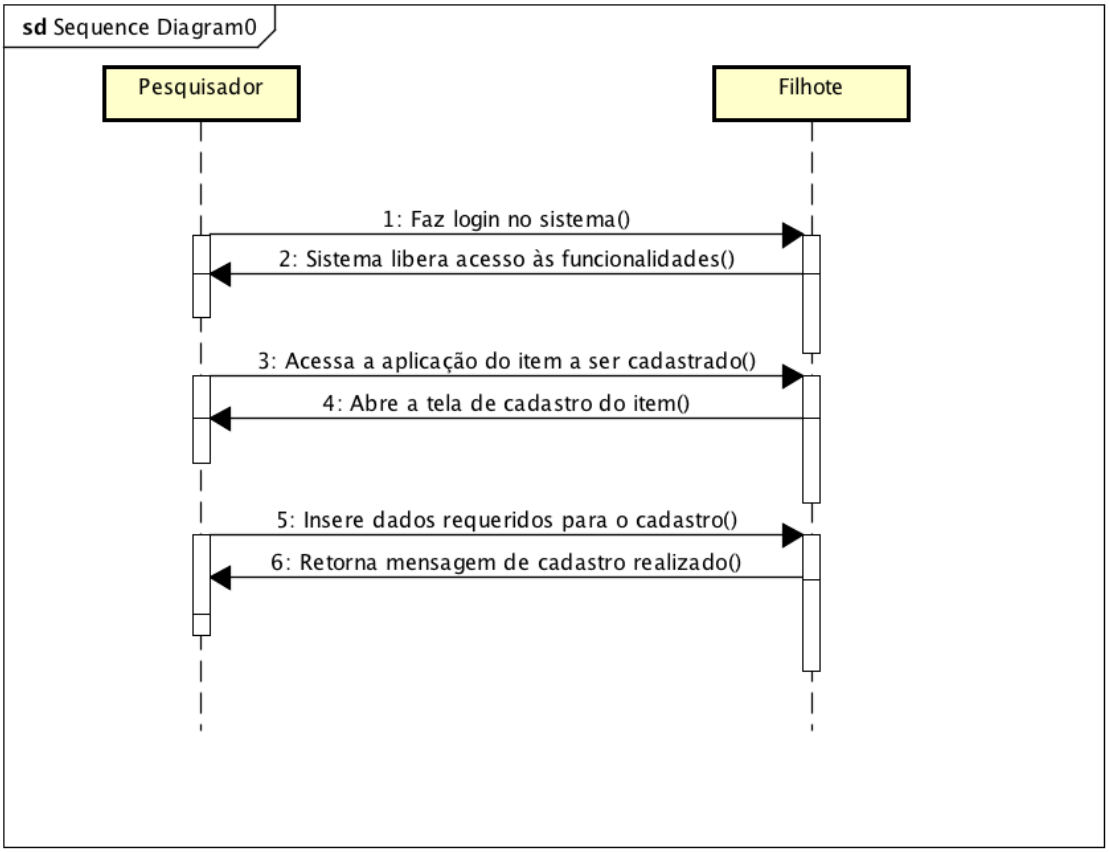
Para o atendimento dos requisitos citados, conforme ilustrado no diagrama dos casos de uso, ficou definido que a aplicação deveria ser *web*, isto aumenta a facilidade de acesso dos pesquisadores à ferramenta desenvolvida, visto que eles podem realizar a administração ou a mineração de dados através de um dispositivo conectado a uma rede, sem necessidade de ter uma estrutura mais robusta instalada em uma máquina específica.

A especificação dos casos de uso elicitados neste capítulo está no Apêndice A.

Em seguida, com o intuito de representar a sequência de processos entre as entidades que compõem a aplicação, são apresentados os diagramas de sequência. Por convenção e objetividade, os diagramas que representam os *CRUD's* foram divididos em: Diagrama de sequência das funcionalidades de cadastro, Diagrama de sequência das funcionalidades de consulta a dados, Diagrama de sequência das funcionalidades de alteração de dados, Diagrama de sequência das funcionalidades de excluir (ou inativar) dado.

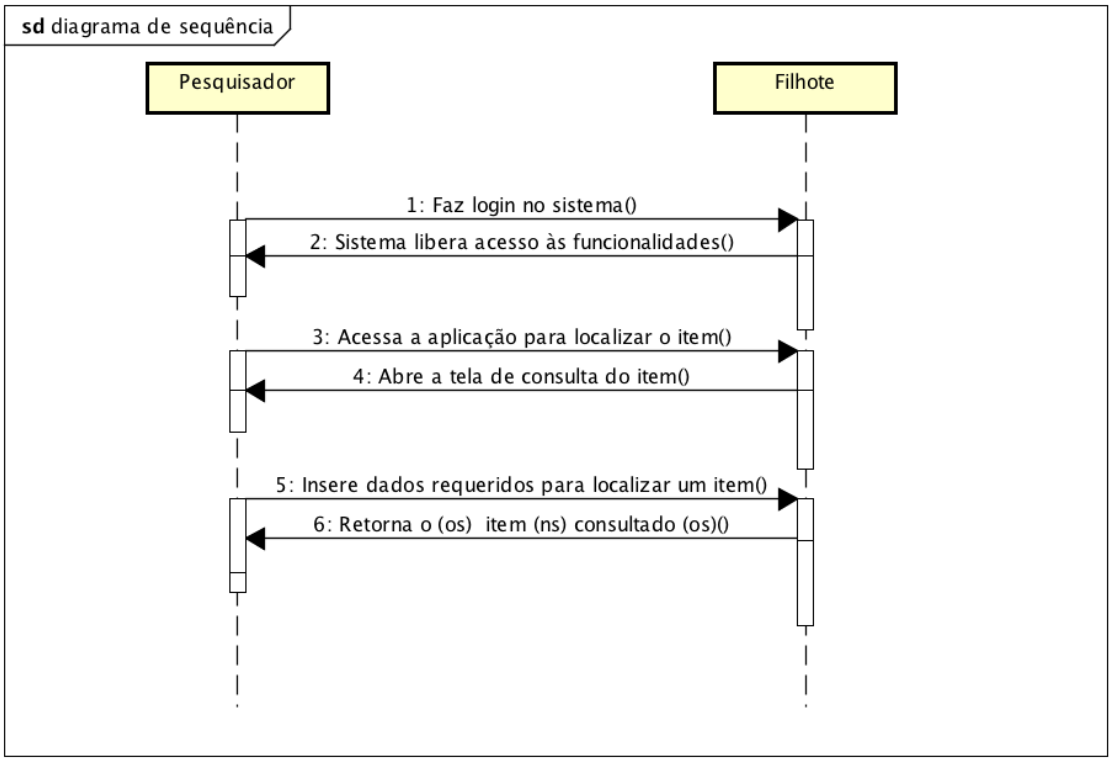
A partir da análise da Figura 12, observa-se as principais interações realizadas entre as entidades *pesquisador* e *aplicação* (Filhote), que implementam as operações de cadastro.

Figura 12 – Diagrama de sequência das funcionalidades de cadastro



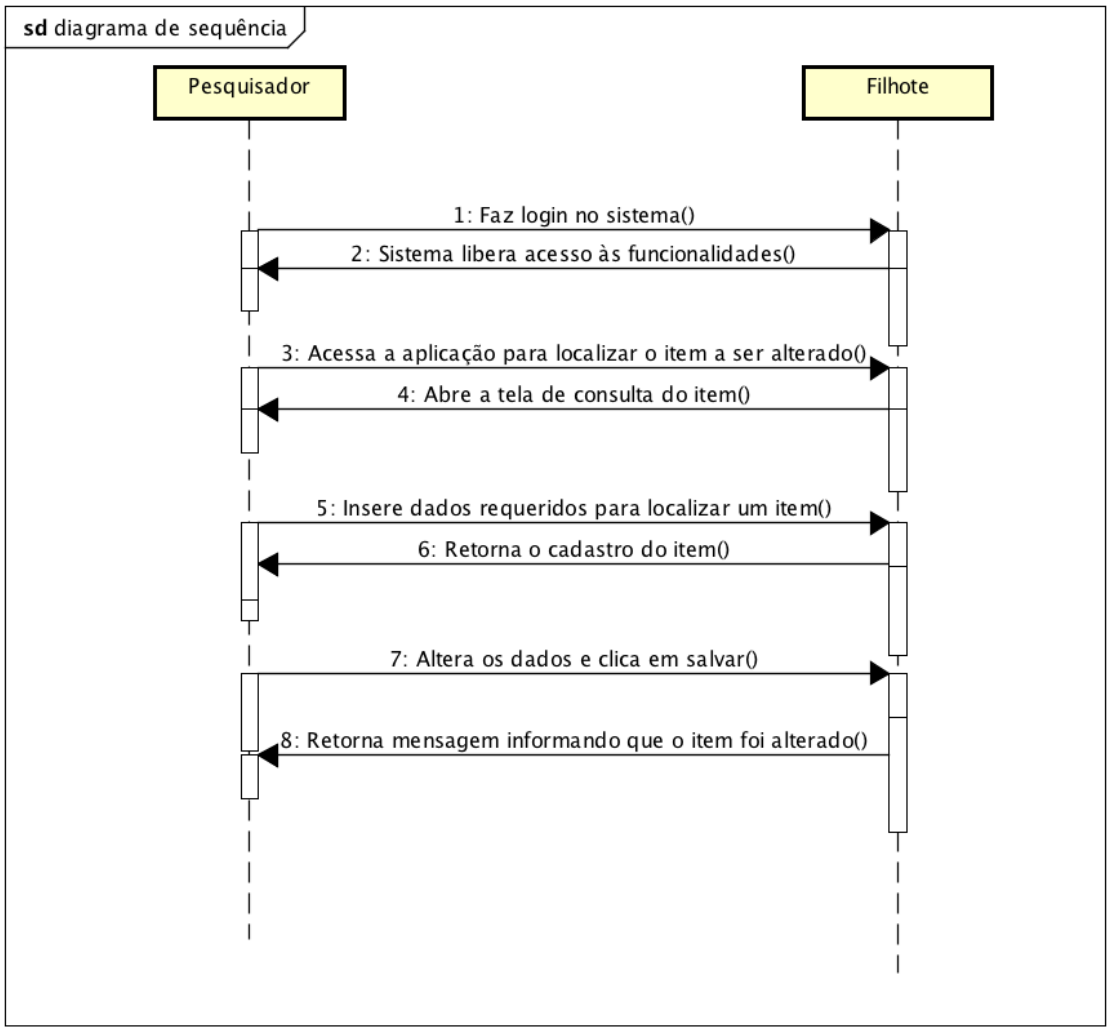
No que tange à operação de consulta dos dados, o diagrama ilustrado na Figura 13 descreve as interações para localizar um registro na base, de acordo com o(s) parâmetro(s) informado pelo *pesquisador*.

Figura 13 – Diagrama de sequência das funcionalidades de consulta a dados



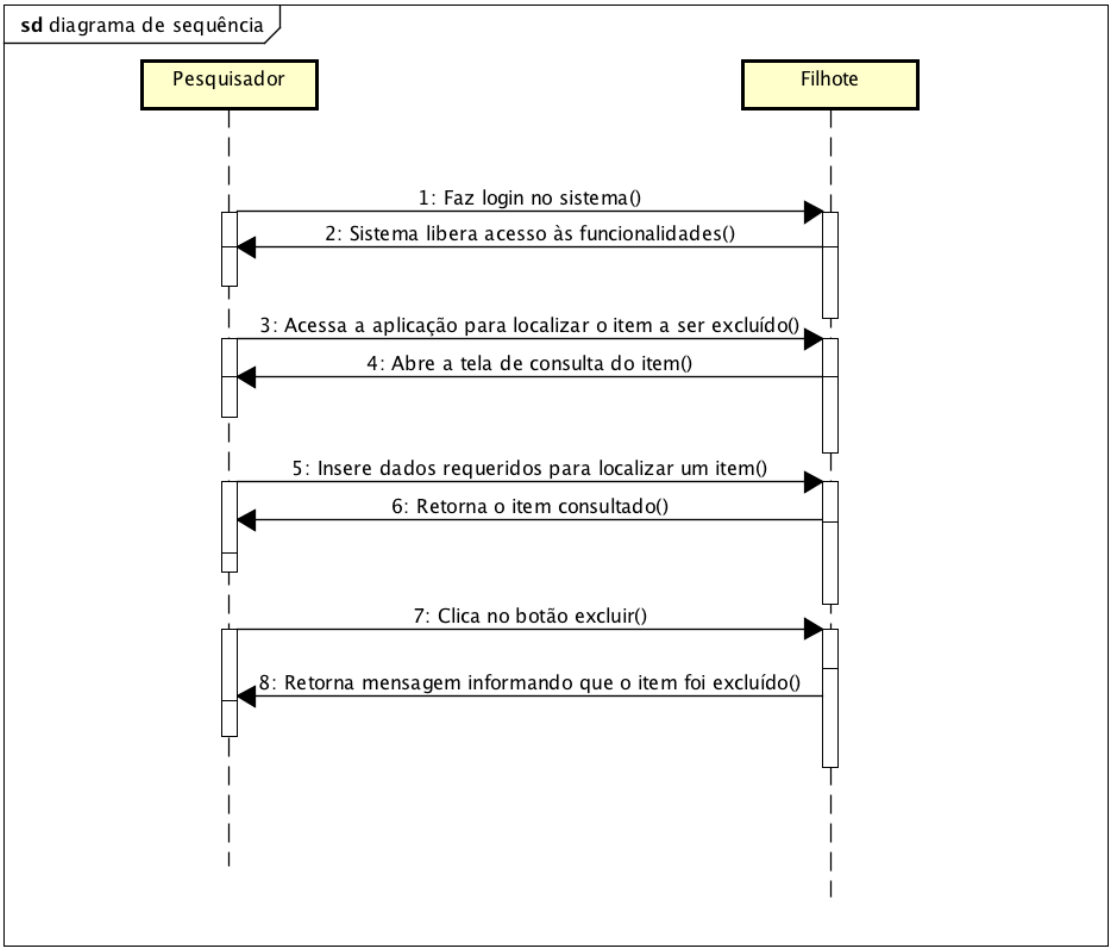
Em relação à funcionalidade ilustrada na Figura 14, que define a operação de alteração de um ou mais campos retornado(s) a partir de uma consulta.

Figura 14 – Diagrama de sequência das funcionalidades de alteração de dados



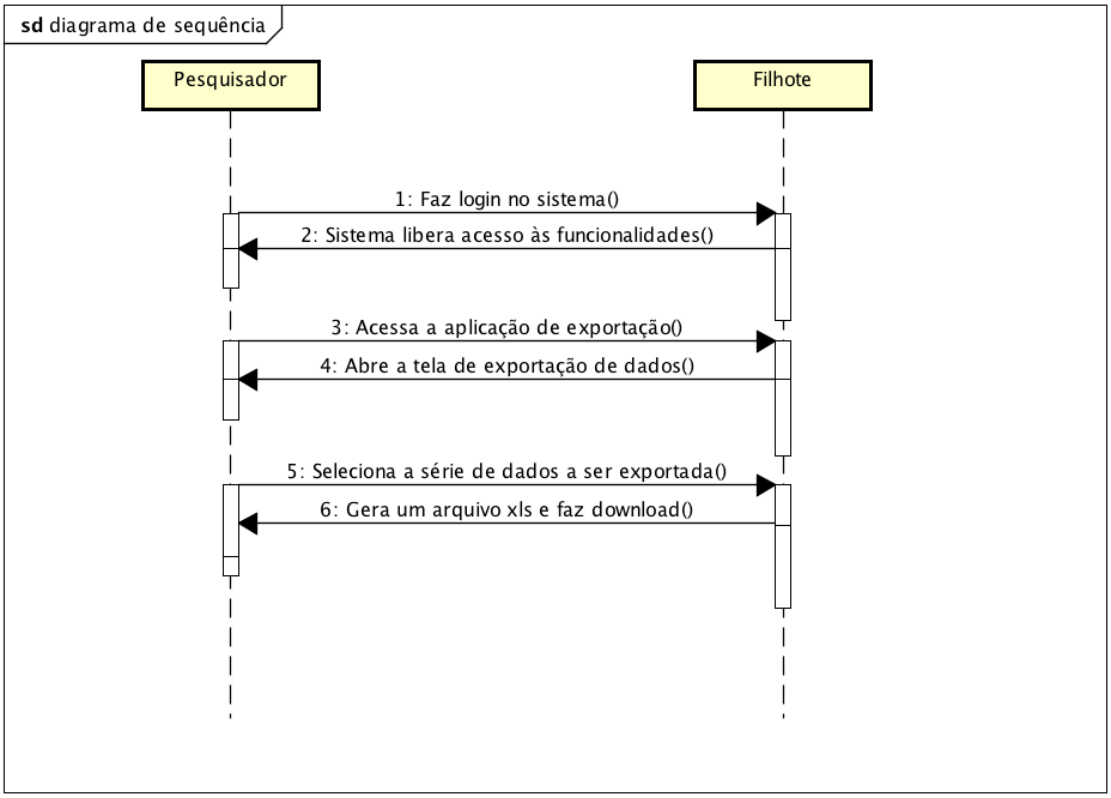
A Figura 15 descreve a interação do pesquisador com a aplicação demonstrando a operação de exclusão de um registro na base de dados. Com o objetivo de manter os dados sempre presentes no banco de dados, nesta aplicação um dado marcado como excluído pelo usuário será apenas inativado.

Figura 15 – Diagrama de sequência das funcionalidades de excluir dados



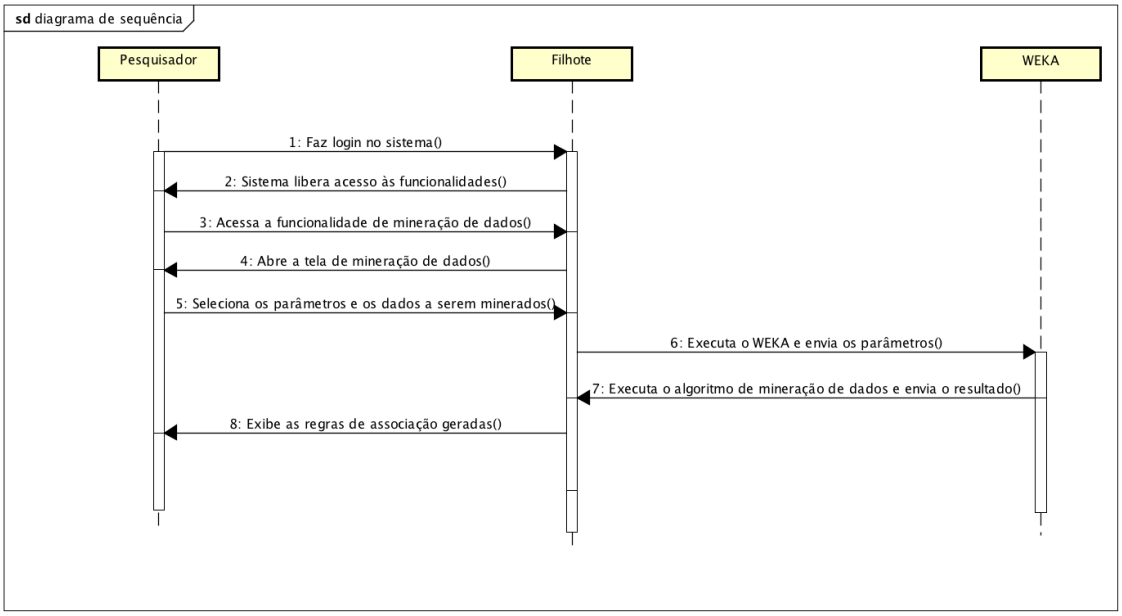
O diagrama de sequência apresentado na Figura 16 apresenta a funcionalidade de exportação de dados. Note que é necessário que o usuário informe quais conjuntos de dados (bióticos e/ou abióticos) devem fazer parte do arquivo *csv*. Este processo permite ainda que sejam selecionados dados de um ou de vários locais registrados, bem como o período de coleta.

Figura 16 – Diagrama de sequência da funcionalidade de exportar dados



Por fim, a Figura 17 exibe o Diagrama de sequência da funcionalidade de mineração de dados. Neste conjunto de interações o pesquisador seleciona os campos que deseja remover do processo de mineração, assim como define o período e o local de coleta (seleção dos dados). Além disto, informa os parâmetros necessários para o acionamento do *framework WEKA*, que gera as regras de associação e retorna para a aplicação Filhote. A aplicação apresenta as regras geradas ao usuário.

Figura 17 – Diagrama de sequência da funcionalidade de minerar dado

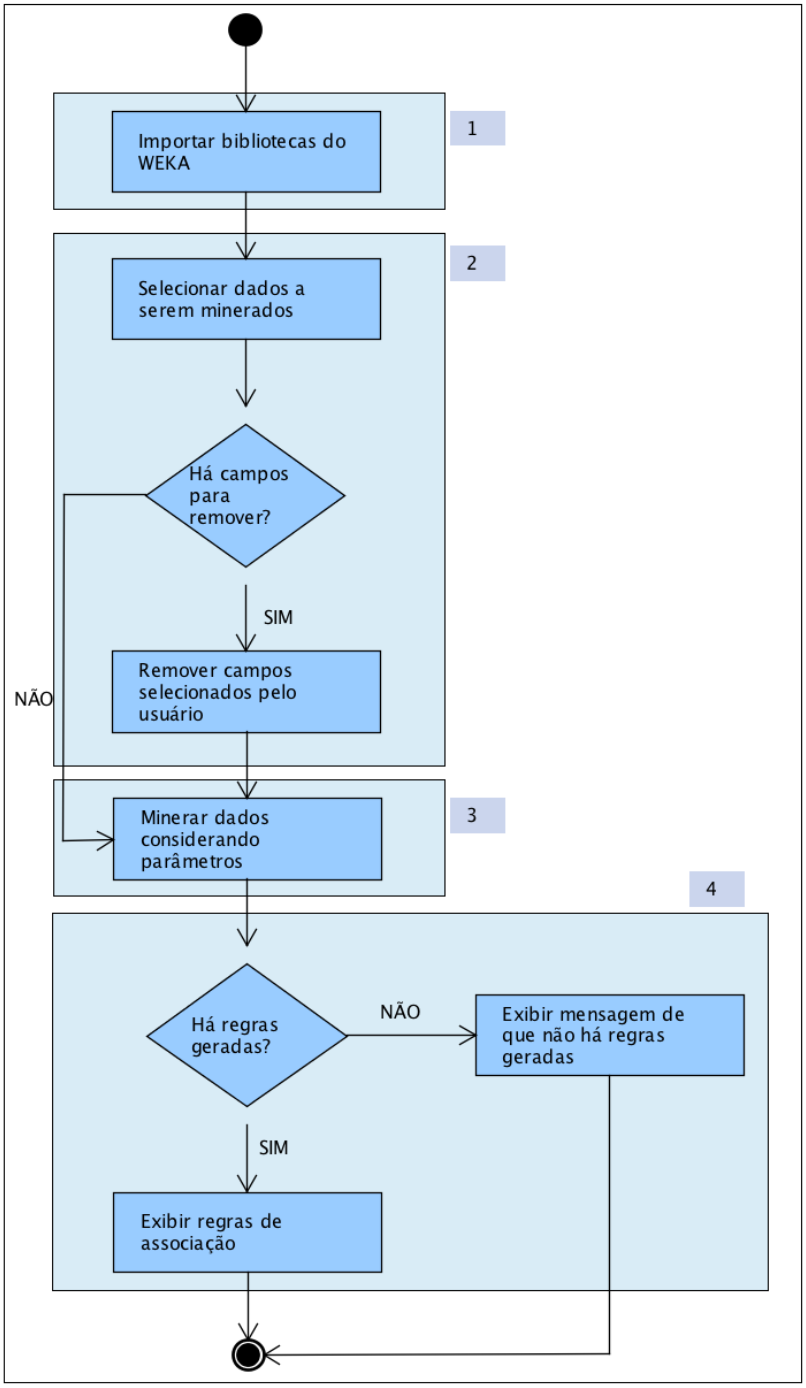


Esta seção descreveu a estrutura da aplicação desenvolvida, suas funcionalidades, a modelagem UML (*Unified Modeling Language*) baseada em Larman (2004), as tecnologias utilizadas, bem como a interação das principais entidades envolvidas. O detalhamento da estruturação dos dados, a construção e a implementação dos modelos e a integração com a *framework WEKA* é apresentada na subseção 3.2.4.

3.2.4 Estrutura e funcionamento da funcionalidade minerar dados na aplicação Filhote

Entendendo que na aplicação desenvolvida, o módulo menos comum do ponto de vista do desenvolvimento de *softwares* é referente à integração com a aplicação *WEKA*, nesta seção será apresentado o detalhamento do processo de integração, bem como do funcionamento, ilustrados na Figura 18, com o objetivo de facilitar o entendimento, onde cada retângulo representa um processo para a obtenção de dados minerados.

Figura 18 – Detalhamento da mineração de dados



Conforme ilustrado no quadro 1 da Figura 18, para a integração da aplicação Filhote com a ferramenta de mineração de dados *WEKA*, foi necessário obter o arquivo *weka.jar* do site do da Universidade de Waikato – Nova Zelândia. Esta biblioteca foi inserida no diretório do projeto para a disponibilização dos recursos de mineração através de sua API.

A importação do pacote *associations*, através da utilização de suas classes, per-

mitiu a utilização do algoritmo *Apriori*, que é um dos algoritmos responsáveis pela geração de regras de associação. Além disso, este algoritmo está presente no *ranking* dos 10 algoritmos mais influentes em Mineração de Dados identificados e eleitos na IEEE *International Conference on Data Mining* (ICDM) (WU et al., 2008).

A ideia de utilizar nesta ferramenta a técnica de geração de regras de associação foi motivada por possibilitar a sequência aos trabalhos iniciados em (SILVA et al., 2013), onde realizou-se a mineração de dados gerando regras de associação sobre ovos e larvas que, embora as tarefas tenham sido aplicadas sobre uma outra base de dados, este conjunto segue os mesmos princípios da base trabalhada neste projeto.

Outro pacote importado nesta aplicação é chamado *remove*. Este pacote permite que os dados selecionados sejam filtrados, retirando campos que, por ventura, não são desejados na análise a ser executada.

O quadro 2 da Figura 18 apresenta o processo de seleção do conjunto de dados que serão minerados. A seleção de campos deve ser realizada pelo usuário da aplicação de acordo com os critérios definidos para atender o que se delimitou como problema.

Nesta aplicação uma rotina para seleção de dados foi criada, utilizando o pacote *remove* do *WEKA*. Para isto, dois campos fazem parte da seleção de dados, apresentando os valores do campo *local da coleta* agrupados, oferecendo a possibilidade de que um ou vários locais possam ser adicionados pelo usuário ao conjunto de dados que serão minerados, e o campo *período da coleta* permite que o usuário defina a data de início e a data de fim em que os dados foram coletados.

Sendo assim, para cada conjunto de dados a ser minerado, dispostos na aplicação Filhote no item *minerar dados*, o usuário tem a opção de selecionar vários campos que serão removidos deste grupo e, além disso, definir um recorte da base de dados delimitando o período e o local de coleta.

O quadro 3 da Figura 18 ilustra a execução do algoritmo para minerar dados. É neste momento que o conjunto de dados selecionado fica à disposição do algoritmo para buscar regras de associação sobre eles.

O processo de mineração de dados requer a configuração de parâmetros do algoritmo, que neste caso, para o *Apriori*, são: métrica, que pode ser escolhida pelo usuário entre *Confiança*, *Interesse*, *Influência* e *Convicção*; o valor mínimo a ser utilizado para a medida selecionada, no campo *valor mínimo da métrica*; o *número de regras a serem retornadas*; o *suporte mínimo*; e o *suporte máximo*. Para todos eles, a aplicação já traz valores preestabelecidos, com os mesmos valores padrões do *WEKA*, sendo a métrica igual a "*Confiança*", o valor da métrica igual a "0.9", o número de regras igual a "10", o suporte mínimo igual a "0.1" e suporte máximo igual a "1.0".

No quadro 4 da Figura 18 está representado o retorno do algoritmo de mineração para o usuário. No caso de não haver regras geradas, uma mensagem é apresentada informando o resultado. Caso contrário, as regras são apresentadas na tela para o usuário, na forma como segue:

- Nome do algoritmo;
- Suporte mínimo encontrado;
- Valor mínimo da métrica;
- Número de ciclos realizados;
- Tamanho dos conjuntos de item;
- Listagem das melhores regras de associação encontradas.

Com os resultados listados, assim como normalmente acontece nos trabalhos de mineração de dados, é desejado que o pesquisador especialista nos dados possa analisar as regras de associação retornadas e, se necessário, reinicie o processo com a alteração das configurações para que novas regras sejam geradas.

4 Discussões

Os resultados de uma pesquisa têm seu suporte firmado em uma rede de processos que se articulam para gerar respostas baseadas em informações e, por isto, há bastante tempo a informação tem ganhado seu título de importância, sendo considerada um dos principais ativos de uma instituição.

Este fator tem feito com que os pesquisadores percebam que, para que haja bons resultados em um trabalho, a simples obtenção dos dados pode ser insuficiente. Em se tratando de bancos de dados com conteúdos de caráter biológicos, grandes quantidades de dados são coletadas e o valor destes conjuntos de dados estão muitas vezes relacionados a pesquisas sobre ambientes que passaram por alterações definitivas.

Em adição a isto, sabe-se que não basta que a informação seja gerada, é de fundamental importância ter ciência da qualidade dos dados que a compõem. A informação não contribuiria de forma positiva nos resultados de uma pesquisa caso sua base fosse frágil.

Fundamentado nestas características, este trabalho identificou que o desenvolvimento da informação tem seu pilar principal na concepção e tratamento empregado no processo de coleta e na manutenção dos dados primários. Isto significa afirmar que o planejamento e os recursos utilizados para a gestão das bases de dados devem receber atenção especial no que diz respeito a estruturação e a preservação destes conjuntos, bem como no correto preenchimento dos campos inerentes ao modelo.

Avaliando os dados de peixes tratados neste trabalho, foi possível perceber que a coleta de dados primários para a caracterização do ambiente e das espécies de peixes requer investimentos volumosos de recursos financeiros e humanos, algumas vezes envolvendo a segurança da equipe de campo, especialmente quando as amostragens são realizadas durante 24 horas, incluindo a navegação noturna.

A série de dados que precedem os impactos é o diagnóstico da condição inicial do ambiente e das comunidades biológicas. Este diagnóstico inicial é o balizador das mudanças que ocorrem tanto nas escalas espaciais quanto temporais, sendo únicas para cada área. Em muitos casos estas séries estão sendo perdidas, sendo utilizadas apenas para a elaboração dos relatórios requeridos pela legislação para a obtenção das licenças.

Em países mais desenvolvidos, onde o planejamento, gestão e manejo seguros da biodiversidade e da qualidade do ambiente passam pela análise de séries seguras de dados, a probabilidade de erros tem sido minimizada a partir da elaboração de modelos de previsão. Em alguns casos, instituições específicas são responsáveis pela guarda e

disponibilização desses dados considerados estratégicos como, por exemplo, a USGS nos Estados Unidos.

Isto justifica a busca por melhorias de gestão dos dados primários através de aplicação de sistemas computacionais e da inserção de aplicações que possam prover recursos facilitadores para gerir os dados primários de forma estruturada e acessível. Ciente disto, a primeira etapa deste trabalho teve seu foco direcionado à estruturação da base de dados de peixes coletada na região da Usina de Lajeado visando atender os objetivos específicos desta pesquisa no que se refere aos dados.

Nesta etapa, a implementação do banco de dados a partir de um conjunto de dados dispostos em planilhas apresentou as dificuldades da estruturação de banco de dados a posteriori. O estudo da base de dados de Lajeado mostrou claramente este problema e a necessidade de preocupação com esta questão.

Foi possível constatar que projetos como este ainda necessitam ser incentivados a se preocupar com a estrutura dos dados antes mesmo do início dos trabalhos de coleta, na tentativa de evitar maiores custos para o tratamento dos mesmos, trabalhando para diminuir riscos de perdas e facilitando a manipulação e a análise de dados no futuro.

A fase de projeto de banco de dados considerou o requisito de modelar os dados para serem independentes da aplicação, ou seja, mesmo que a aplicação venha a mudar futuramente, os dados devem permanecer em seu estado de integridade. Desta forma, o modelo relacional se mostrou adequado para armazenar este conjunto de dados atendendo a este requisito e, especialmente, pela característica deste projeto de tratar de dados não–complexos, além de ser o modelo mais adotado comercialmente.

O resultado do projeto do banco de dados mostrou que as principais redundâncias foram resolvidas através da criação do modelo entidade-relacionamento. Durante a análise dos dados, optou-se por permitir que os campos *data, hora, local e projeto* estivessem presentes nas duas entidades principais, considerando que pode haver coleta apenas de dados bióticos em um momento, ou coletas apenas de dados abióticos, dependendo do objetivo da pesquisa.

Desta forma, analisando o modelo de dados gerado, vide Figura 7, é possível dizer que este conjunto de dados encontra-se na Quarta Forma Normal, visto que os atributos das tabelas não contem grupos de repetição (tabelas aninhadas) atendendo à primeira forma normal, os atributo não-chave são dependentes da chave primária de acordo à segunda forma normal, não há dependências funcionais transitivas seguindo a terceira forma normal e as tabelas não possuem mais de uma dependência multivalorada, em conformidade com a quarta forma normal.

Convém salientar que o processo de importação dos dados bióticos, teve a com-

plexidade das rotinas desenvolvidas consideradas bastante alta. Foi possível chegar a este entendimento não apenas pela grande quantidade de registros pertencentes à entidade mas, especialmente, pelo fato de haver várias chaves estrangeiras que deveriam receber o valor do campo *id* da tabela relacionada e, também, pelas comparações de *strings* de muitos campos para a seleção dos valores destas chaves. Este procedimento teve custo de tempo elevado para ser realizado, mas foi concluído de forma satisfatória para 10 das 11 chaves estrangeiras pertencentes à entidade *biotico*.

Embora esta atividade tenha sido realizada sob alto grau de dificuldade, a decisão por não criar estas restrições poderia acarretar em inserção de valores diferentes para campos que se referenciam ao mesmo dado, gerando a perda de integridade e/ou redundância. Com a implementação dos relacionamentos e a utilização de chaves estrangeiras este problema é evitado. Ademais, este custo observado se refere apenas à fase de mapeamento dos dados primários e não afetará o usuário no dia a dia.

A última verificação de chave estrangeira foi um dos passos mais importantes e também mais complexos deste trabalho. Neste momento foi checada a existência de relacionamento entre as tabelas *biotico* e *abiotico*, considerando os campos *data, hora e local* porém, porém apenas 20% dos campos eram passíveis de relacionamento através destas chaves.

Isto demandou nova análise por parte do especialista nos dados para verificar a correta execução desta atividade. Assim, foi constatado nesta verificação que, além destes campos, o atributo *ponto* da tabela *biotico* deveria ser relacionado com o campo *smf* da tabela *abiotico*. Desta forma, 61,60% dos registros da entidade *biotico* se relacionava com a entidade *abiotico* e o campo com a chave estrangeira *abiotico_id* foi atualizado, recebendo o valor do atributo *id* da entidade *abiotico*. A informação de que 38,4% dos dados bióticos não puderam ser relacionados a um dado abiótico foi considerado consistente pelo especialista dos dados, visto que nem todo dado biótico tem um dado abiótico.

O processo de execução da rotina de atualização de chaves estrangeiras na tabela *biotico* levou cerca 2 horas para a finalização, momento em que foi considerada encerrada a primeira etapa deste trabalho.

Desta maneira é possível afirmar que o modelo de dados gerado dará suporte aos dados de peixes existentes no conjunto das coletas de Lajeado e poderá receber novos dados com as mesmas características, fazendo uso dos mesmos recursos já implementados, de forma transparente ao usuário. Isto também facilita novas implementações de módulos que permitam a disponibilização para consulta pública por outros grupos de pesquisas, fazendo o uso de tecnologias do tipo *webservices*, bem como o desenvolvimento de aplicações móveis para uso em ambientes de coleta (campo), evitando assim

o preenchimento de fichas manuais ou planilhas eletrônicas.

A segunda fase deste trabalho tratou da disponibilização de uma ferramenta para manutenção destes dados, atendendo os objetivos específicos desta pesquisa se referem à aplicação para manipulação e análise de dados.

Conforme já discutido nesta dissertação, sabe-se que os dados que estavam mantidos em planilhas eletrônicas tinham limitações de manipulação e estavam sob risco de perda de integridade.

Embora os programas de criação e de manutenção de planilhas eletrônicas sejam amplamente difundidos para visualização e manipulação de dados, permitam a criação de gráficos, automatização de cálculos e, além disso, sejam de fácil aprendizagem para operações básicas, o tratamento de um conjunto de dados como o de peixes da usina de Lajeado poderia ser comprometido por algumas limitações.

Neste tipo de aplicação os mecanismos de segurança são restritos, há dificuldades de verificação de tipos de dados, o registro histórico de modificações é trabalhoso, além de requerer muito esforço na integração de dados. Especialmente analisando sua utilização, verificou-se a dificuldade de eliminar redundâncias e a dificuldade de manter a consistência dos dados. Além disto, é praticamente impossível a interoperabilidade com outras aplicações, confinando quase que exclusivamente a operação ao programa editor, por exemplo, *BrCalc*, *MS Excel*®, *Numbers*, dentre outros.

Embasado nos riscos e dificuldades encontradas no armazenamento e manipulação de dados e, principalmente, por conhecer a importância desta série, bem como a vulnerabilidade em que ela se encontrava, a ferramenta concebida se comportará como um facilitador ao acesso e à manipulação da base constituída. Este foi de fato o principal motivador para o desenvolvimento da proposta apresentada e discutida neste trabalho.

Notoriamente, dados de peixes de pesquisas futuras teriam os mesmos riscos citados anteriormente, o que levaria a perda de recursos importantes. Na tentativa de suprir estas dificuldades, a aplicação Filhote buscou cumprir os requisitos primordiais levantados pelos especialistas nos dados e discutidos amplamente no Capítulo 3. Assim, garante a integridade dos dados e oferece a facilidade de acesso ao pesquisador, que poderá ter acesso aos dados em qualquer lugar do planeta onde haja conexão com a internet, visto que é uma aplicação *web*.

Além disso, a aplicação oferece uma forma colaborativa de se trabalhar com os dados, considerando que uma alteração realizada sobre eles é persistida sobre a mesma base acessada pelos demais pesquisadores, tendo todas as vantagens de controles implementados por um SGBD. A ausência da ferramenta implicava na necessidade de

se manter o controle de versões de forma manual, o que oferece grande riscos de perdas (de dados, de registros, de integridade).

Adicionalmente, outra funcionalidade atendida pela ferramenta é a exportação de um conjunto de dados selecionados pelo usuário para um arquivo *csv*. Justificou-se este requisito o fato de que os dados podem ser requeridos para uma determinada análise e, por isto, devem compor um arquivo texto passível de ser importado em ferramentas de análises estatísticas tais como os *softwares R*, *octave*, *matlab* ou aplicações desenvolvidas por terceiros.

Outra vantagem refere-se à utilização da aplicação Filhote, no que tange à exportação dos dados retornados relacionados entre si. Assim, um registro de dado biótico que tenha relacionamento com um dado abiótico será retornado em uma mesma tupla (linha). Na manipulação dos dados sobre planilhas eletrônicas este relacionamento era custoso, bastante suscetível a erros e estava atrelado exclusivamente ao conhecimento, à atenção e ao trabalho manual do especialista.

O desenvolvimento do trabalho até este ponto atendeu os objetivos específicos de projetar um banco de dados adequado ao conjunto de dados coletados na usina de Lajeado e que desse suporte às próximas coletas de projetos similares, bem como a importação dos dados para o modelo gerado. Foi atendido, também, o objetivo de criação de uma ferramenta para facilitar o acesso e a manipulação de dados.

Além do processo de estruturação e de desenvolvimento da aplicação para manipulação e exportação dos dados, objetivou-se oferecer um mecanismo de análise de dados através da geração de regras de associação. Esta característica foi definida no momento em que foi realizado o estudo e a descrição do conjunto de dados. No início desta fase pretendia-se realizar a mineração de dados sobre este conjunto, mas verificou-se que a maior urgência era a busca por estruturar esta base e alterar o estado de vulnerabilidade a que estava exposta, submetê-la à estruturação e trabalhar para oferecer organização e segurança permitindo que novas pesquisas possam ser realizadas sobre ela a partir de agora.

Desta forma, o objetivo de trilhar o caminho da análise desta base através da mineração de dados, ainda que inicialmente, teve seu processo vinculado à integração da ferramenta de mineração de dados *WEKA*. Espera-se com isso, integrar áreas de estudo, oferecendo um meio de análise utilizado mais amplamente dentro da área de tecnologia da informação, mas que certamente trará bons resultados quando utilizados por pesquisadores da área ambiental.

Convém ressaltar que este não é o único *framework* que poderia se integrar à aplicação e, da mesma forma, a geração de regras de associação não é a única técnica

que poderia ser oferecida. Outros métodos de mineração de dados poderiam consumir os dados estruturados como, por exemplo, a classificação, através da execução de algoritmos como o *J48, ID3*, dentre outros.

Além de agrupar especialistas de diferentes áreas, com distintas especialidades, a interdisciplinaridade possibilita a evolução da aplicação de métodos em áreas diversas. Para o contexto descrito neste trabalho, esta interdisciplinaridade foi alcançada através da possibilidade de integrar pesquisadores de outras áreas do conhecimento ao estudo de técnicas de mineração.

Os esforços dispendidos para a realização desta integração de ferramentas mostraram que a elaboração de hipóteses de pesquisa e análises ambientais complexas envolvendo questões ambientais em ambientes tropicais, incluindo a Amazônia, somente serão possíveis a partir da estruturação de bancos de dados e de ferramentas de análise que agreguem informações obtidas em estudos primários com amplitudes espaciais e temporais consistentes.

Este é um filão estratégico para a Ciência como um todo, inclusive para os profissionais e instituições brasileiras ligadas à tecnologia, desenvolvendo projetos de estruturação e armazenamento seguros de informações. As informações existentes sobre a biodiversidade brasileira poderiam nos dizer muito mais hoje se estivessem estruturadas.

5 Conclusão

Este trabalho concentrou esforços no tratamento dos dados coletados na usina de Lajeado, com o intuito de oferecer melhoria na sua estruturação e no seu armazenamento. Esta pesquisa teve seu início nas tentativas de utilizar este conjunto de dados para a geração de regras de associação, através da mineração de dados.

Constatou-se que estes dados, da forma que se encontravam, poderiam oferecer resultados frágeis quando submetidos a esta técnica, considerando que não havia relacionamento claro entre eles.

Foi possível observar que os dados registrados em planilhas eletrônica, estavam sob risco de perda de integridade ou, até mesmo, perda total do conjunto. Ademais, o relacionamento dependia exclusivamente do conhecimento, do trabalho manual e da atenção do especialista nos dados. Atividades deste tipo se tornam muito dispendiosas, inclusive por causa da manipulacao de um grande número de linhas e variáveis. Estas características tornaram o trabalho de mineração de dados inviável.

Considerando estas tentativas realizadas anteriormente a este trabalho e a importância dos dados referentes a um ambiente que foi alterado de modo definitivo, foi necessário repensar a estruturação do trabalho.

Desta forma, iniciou-se um levantamento de informações e conceitos não apenas sobre os dados mas, sobre como eles eram armazenados e mantidos, e também sobre o contexto em que estão inseridos.

Observou-se que já existem discussões que abordam a importância da estruturação dos dados de biodiversidade e, especialmente, a necessidade de que a sua manutenção e sua segurança sejam considerados antes mesmo do início das coletas. É necessário que haja preocupação com a garantia de integridade e acessos facilitados ao conjunto, no intuito de que os dados estejam corretos, oportunos, relevantes e consistentes para a geração de informações de qualidade.

Neste sentido, este trabalho teve o principal objetivo voltado para dispor este conjunto de dados sob um modelo estruturado, relacionando os dados contidos em planilhas eletrônicas para futuro acesso e, também, para possibilitar que novo dados possam ser agregados a este conjunto.

As tarefas realizadas para atender este objetivo mostrou de forma evidente as dificuldades do tratamento de dados a posteriori. O relacionamento de registros que não têm um campo com chave exclusiva é de difícil solução e, neste caso, demandou

várias análises para definições de critérios e outras verificações para a confirmação de que os relacionamentos realizados estavam corretos.

O modelo de dados desenvolvido para persistir os dados, procurou incorporar os objetivos e requisitos do grupo de pesquisa do Neamb/UFT, que descreveu as regras para o desenvolvimento do modelo, cumprindo os requisitos de R1 a R10 listados na subseção 3.1.3.1.

Com a preocupação do crescimento das séries de dados, este trabalho atendeu o objetivo de desenvolvimento de uma ferramenta para o suporte à entrada, edição e exportação de dados através da criação da aplicação Filhote, que cumpre a função de facilitar e trazer melhorias à gestão dos dados. O seu desenvolvimento contemplou os requisitos funcionais de RF1 a RF17 e os requisitos não funcionais de RFN1 a RFN6, listados na subseção 3.2.3.

Em adição à administração dos dados, com o intuito de trilhar o caminho da integração de áreas do conhecimento e dispor de técnicas de mineração de dados para ecólogos, ainda que inicialmente, a aplicação Filhote se apresentou como uma forma alternativa para a mineração de dados, através da geração de regras de associação, com a utilização da *API* do *WEKA*.

Além disto, este trabalho procurou sensibilizar os estudantes, pesquisadores e gestores sobre a importância da estruturação dos dados, bem com da integração de áreas, buscando a modernização do tratamento dos dados e métodos de análise, especialmente considerando que recursos computacionais são essenciais para o tratamento e intercâmbio de grandes volumes de dados.

Tabela 4 – Objetivos específicos alcançados

Objetivo	**Como foi alcançado**
Desenvolver um modelo de dados adequado aos dados registrados durante as coletas	A análise de requisitos realizada proporcionou meio de esclarecimento para que fosse projetado um modelo de dados;
Criar chaves para identificar individualmente os registros	Foram criadas chaves primárias para cada entidade importada para o modelo
Importar os dados para o modelo de dados desenvolvido	Os dados foram relacionados e inseridos em cada uma das entidades do modelo
Desenvolver uma ferramenta que dê suporte à entrada, edição e exportação de dados	Os requisitos funcionais e não funcionais foram elicitados, tendo servido de guia para o desenvolvimento da arquitetura proposta
Integrar a ferramenta desenvolvida com o *WEKA* para geração de regras de associação	A aplicação Filhote foi integrada à ferramenta de mineração de dados através de sua *API java* disponibilizada para este fim

5.1 Contribuições e publicações relacionadas

SILVA, M. d. A.; TREVISAN, D. Q.; PRATA, D. N.; MARQUES, E. E.; LISBOA, M.; PRATA, M. Exploring an ichthyoplankton database from a freshwater reservoir in legal amazon. In: Advanced Data Mining and Applications. [S.l.]: Springer, 2013. p. 384–395

SILVA, M. d. A.; TREVISAN, D. Q.; PRATA, D. N.; MARQUES, E. E. Aplicação do algoritmo apriori para uma base de dados de ictioplâncton em um reservatório de água doce da amazônia legal. X Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). 2013.

PRATA, D. N.; SOARES, K. P.; SILVA, M. A.; TREVISAN, D. Q.; LETOUZE, P. Social data analysis of brazilian's mood from twitter. International Journal of Social Science and Humanity, Vol. 6, No. 3, March 2016

5.2 Perspectivas

Como trabalhos futuros pretende-se realizar a mineração de dados utilizando a ferramenta *Filhote* analisando as regras de associação geradas na tentativa de encontrar regras úteis e que possam responder a questões ambientais ainda ocultas.

Planeja-se, também, agregar novos módulos de mineração de dados para aplicações de técnicas de classificação, para a categorização dos elementos da base, e de agrupamento, para a formação de grupos similares dentre o conjunto de dados existente. A disponibilidade de novos métodos aumenta a possibilidade do desenvolvimento de novas pesquisas através da busca por padrões.

Além disso, a aplicação *Filhote* está preparada para que sejam incluídos módulos de consulta pública de dados a partir do uso de *webservices* do tipo *RESTFull*, que possibilitarão a utilização do conjunto de dados por outros grupos de pesquisas que desejam realizar análises pontuais sobre determinadas características da base.

Para a expansão da base de dados gerida pelo SGBD deseja-se ainda realizar o desenvolvimento de um aplicativo *mobile* contendo as características atuais da base, possibilitando que os pesquisadores possam inserir os dados no momento da coleta *in loco*. A validação das consistências dos campos contidos na aplicação poderia ser realizada a nível de aplicação ou a nível de banco de dados, minimizando erros no armazenamento dos dados da coleta.

Tal aplicação permitira ainda a definição da área de coleta a partir de técnicas de georreferenciamento, a partir da utilização da *API*do Google Maps, ampliando a base já existente.

Deseja-se realizar análises temporáis sobre a base de dados de Lajeado, considerando as coletas anteriores ao enchimento do reservatório, as coletas durante este período e os dados coletados após a finalização do enchimento.

Pretende-se desenvolver um algoritmo para tratar os resultados retornados pela mineração de dados. Este algoritmo deve tratar as regras de associação geradas utilizando os conceitos da teoria de conjuntos, elencando as regras pertencentes a um mesmo grupo, de acordo às suas característidas, bem como listando as regras raras, que não contém as características mais frequentes.

Referências

AGARWAL, R.; IMIELINSKI, T.; SWAMI, A. Database mining: A performance perspective. *Knowledge and Data Engineering, IEEE Transactions on*, IEEE, v. 5, n. 6, p. 914–925, 1993. Citado na página 31.

AGOSTINHO, C. S.; PEREIRA, C. R.; OLIVEIRA, R. J. d.; FREITAS, I. S.; MARQUES, E. E. Movements through a fish ladder: temporal patterns and motivations to move upstream. *Neotropical Ichthyology*, SciELO Brasil, v. 5, n. 2, p. 161–167, 2007. Citado 2 vezes nas páginas 35 e 36.

BATINI, C.; SCANNAPIECO, M. *Data Quality Concepts, Methodologies and Techniques. 2006*. [S.l.]: Springer-Verlag, 2006. Citado na página 20.

BORGES, L. E. *Python para desenvolvedores*. [S.l.]: Novatec Editora, 2014. Citado na página 51.

BRASIL. Estabelece as definições, as responsabilidades, os critérios básicos e as diretrizes gerais para uso e implementação da avaliação de impacto ambiental como um dos instrumentos da política nacional do meio ambiente. *Resolução Conama nº 001, de 23 de janeiro de 1986*, 1986. Citado 2 vezes nas páginas 15 e 22.

BRASSCOM. Relatório brasscom brasil ti-bpo - 2013-2014. 2014. Citado na página 19.

CAPPIELLO, C.; CARO, A.; RODRIGUEZ, A.; CABALLERO, I. An approach to design business processes addressing data quality issues. In: *ECIS*. [S.l.: s.n.], 2013. p. 216. Citado na página 20.

COADIC, Y.-F. L.; GOMES, M. Y. F. *A ciência da informação*. [S.l.]: Briquet de lemos Livros, 1996. Citado na página 19.

DATE, C. J. *Introdução a sistemas de bancos de dados*. [S.l.]: Elsevier Brasil, 2004. Citado na página 24.

DAVENPORT, T. *Ecologia da informação: por que só a tecnologia não basta para o sucesso na era da informação*. Futura, 1998. ISBN 9788586082726. Disponível em: <https://books.google.com.br/books?id=JGAxAAAACAAJ>. Citado na página 19.

ELMASRI, R.; NAVATHE, S. *Sistemas de Banco de Dados–Fundamentos e aplicações, tradução da 6a. ed.[por] Daniel Vieira*. [S.l.]: São Paulo, Pearson Addison Wesley, 2011. Citado 8 vezes nas páginas 9, 24, 26, 28, 29, 40, 41 e 52.

ERIKSSON, D. *Database Integration - Data Integration Issues*. http://www.avajava.com/tutorials/lessons/biological-database-integration—data-integration-issues.html?page=, 2015. Citado na página 23.

FANDERUFF, D. *Dominando o oracle 9i: modelagem e desenvolvimento*. [S.l.]: MAKRON, 2003. ISBN 9788534615136. Citado na página 24.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. Advances in knowledge discovery and data mining. the MIT Press, 1996. Citado na página 29.

FELIX, W. *Introdução à gestão da informação*. [S.l.]: Alínea, 2003. (Administração & sociedade). ISBN 9788575160435. Citado 2 vezes nas páginas 10 e 20.

FERNANDES, A. C. *Gesão de Sistemas de Informação*. Tese (Doutorado) — Universidade Técnica de Lisboa, 2005. Citado na página 19.

GENG, L.; HAMILTON, H. J. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, ACM, v. 38, n. 3, p. 9, 2006. Citado na página 31.

GRAVES, M.; GOLDFARB, C. F. *Designing XML databases*. [S.l.]: Prentice Hall PTR, 2001. Citado na página 26.

GUIMARÃES, E. M. P.; ÉVORA, Y. D. M. Sistema de informação: instrumento para tomada de decisão no exercício da gerência. *Ciência da Informação, Brasília*, SciELO Brasil, v. 33, n. 1, p. 72–80, 2004. Citado na página 22.

HAN, J.; KAMBER, M.; PEI, J. *Data mining: concepts and techniques*. [S.l.]: Morgan kaufmann, 2006. Citado na página 33.

HEILIG, G. K. World urbanization prospects: the 2011 revision. *United Nations, Department of Economic and Social Affairs (DESA), Population Division, Population Estimates and Projections Section, New York*, 2012. Citado na página 14.

HILL, A.; OTEGUI, J.; ARIÑO, A.; GURALNICK, R. Gbif position paper on future directions and recommendations for enhancing fitness-for-use across the gbif network, version 1.0. *Copenhagen: Global Biodiversity Information Facility*, v. 25, 2010. Citado na página 14.

HYUGA, T. Mizuho says typing error sparked $3.5 billion in j-com trades. 2005. Citado na página 21.

LARMAN, C. *Utilizando UML e padrões: uma introdução à análise e ao projeto orientados a objetos e ao processo unificado*. Bookman, 2004. ISBN 9788536303581. Disponível em: <https://books.google.com.br/books?id=FFo8uAAACAAJ>. Citado na página 62.

LAROSE, D. T. *Discovering knowledge in data: an introduction to data mining*. [S.l.]: John Wiley & Sons, 2014. Citado na página 29.

MARX, V. Biology: The big challenges of big data. *Nature*, Nature Publishing Group, v. 498, n. 7453, p. 255–260, 2013. Citado na página 23.

MASSINO, E. G.; ROLAND, C. E. de F. Banco de dados objeto-relacional para aplicações web. *Revista Eletrônica de Sistemas de Informação e de Gestão Tecnológica*, v. 5, n. 1, 2015. Citado na página 28.

MORA, C.; TITTENSOR, D. P.; ADL, S.; SIMPSON, A. G.; WORM, B. How many species are there on earth and in the ocean? 2011. Citado na página 14.

OIKAWA, M. K. *Bancos de Dados Biológicos*. 2012. Citado na página 23.

REINGRUBER, M. C.; GREGORY, W. W. *The Data Modeling Handbook: A Best-Practice Approach to Building Quality Data Models*. [S.l.]: John Wiley & Sons, Inc., 1994. Citado na página 25.

REZENDE, D.; ABREU, A. D. *Tecnologia da Informaçao: Aplicada a Sistemas de Informaçao Empresariais*. [S.l.]: ATLAS, 2013. ISBN 9788522475483. Citado na página 19.

SILBERSCHATZ, A.; KORTH, H.; SUDARSHAN, S. *Database System Concepts*. McGraw-Hill Education, 2010. ISBN 9780073523323. Disponível em: <https://books-.google.com.br/books?id=re4YQAAACAAJ>. Citado 6 vezes nas páginas 9, 24, 26, 27, 28 e 29.

SILVA, M. A. *O Pré-Processamento em Mineração de Dados como suporte a modelagem algorítmica*. Tese (Dissertação de mestrado) — Dissertação de Mestrado em Modelagem Computacional de Sistemas - Fundação Universidade Federal do Tocantins - UFT, 2014. Citado 3 vezes nas páginas 9, 32 e 33.

SILVA, M. d. A.; TREVISAN, D. Q.; PRATA, D. N.; MARQUES, E. E.; LISBOA, M.; PRATA, M. Exploring an ichthyoplankton database from a freshwater reservoir in legal amazon. In: *Advanced Data Mining and Applications*. [S.l.]: Springer, 2013. p. 384–395. Citado 2 vezes nas páginas 51 e 64.

TWDI. *The data warehousing institute*. 2006. Citado na página 20.

USGS. *BioData - Aquatic Bioassessment Data for the Nation*. 2015. Citado na página 23.

VEIGA, A. K. *Um estudo sobre qualidade de dados em biodiversidade: aplicação a um sistema de digitalização de ocorrências de espécies*. Tese (Doutorado), 2012. Citado na página 21.

VINES, T. H.; ALBERT, A. Y.; ANDREW, R. L.; DÉBARRE, F.; BOCK, D. G.; FRANKLIN, M. T.; GILBERT, K. J.; MOORE, J.-S.; RENAUT, S.; RENNISON, D. J. The availability of research data declines rapidly with article age. *Current Biology*, Elsevier, v. 24, n. 1, p. 94–97, 2014. Citado na página 22.

WANG, R. Y.; REDDY, M. P.; KON, H. B. Toward quality data: An attribute-based approach. *Decision Support Systems*, Elsevier, v. 13, n. 3, p. 349–372, 1995. Citado 2 vezes nas páginas 9 e 21.

WITTEN, I.; FRANK, E.; HALL, M. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. [S.l.]: Elsevier Science, 2011. (The Morgan Kaufmann Series in Data Management Systems). ISBN 9780080890364. Citado na página 33.

WU, X.; KUMAR, V.; QUINLAN, J. R.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G. J.; NG, A.; LIU, B.; PHILIP, S. Y. et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, Springer, v. 14, n. 1, p. 1–37, 2008. Citado 2 vezes nas páginas 31 e 64.

Apêndices

APÊNDICE A – Especificação dos casos de uso

Este apêndice apresenta o detalhamento dos casos de uso de alto nível.

Cada caso de uso do tipo CRUD (*create, read, update and delete*) ou cadastrar, localizar, alterar e excluir, foi nominado como "manter" com o objetivo de evitar redundância.

Para os casos de uso Manter Aparelho, Manter Espécie, Manter Estádio, Manter Família, Manter Biótico e Manter Abiótico, deve ser observado que existem campos que listam dados de outra tabela relacionada e, então uma pós-condição é que estes dados estejam previamente cadastrados.

Tabela 5 – Especificação dos casos de uso de cadastro de dados

Casos de Uso	Manter Ambiente, Manter Categoria Trófica, Manter Equipamento, Manter Espécie Agrupada, Manter Estádio da Maturação, Manter Gênero/Espécie, Manter Local, Manter Migrad, Manter Ordem, Manter Ponto, Manter Projeto, Manter Aparelho, Manter Espécie, Manter Estádio, Manter Família, Manter Biótico, Manter Abiótico
Ação	Cadastrar
Descrição	Estes casos de uso especificam a ação de cadastrar um item no banco de dados. O pesquisador fornece os dados para cada campo, sendo que alguns deles são obrigatórios, e clica no botão "Salvar"
Atores	Pesquisador
Precondições	O usuário deve estar cadastrado e logado no sistema Caso haja campos que são recuperados de outra tabela (chave estrangeira), eles devem estar previamente cadastrados
Pós-condições	Um novo registro é criado no banco de dados
Restrições	Impedir acesso ao sistema a usuários não cadastrados
Cenário de insucesso	Uma mensagem de erro/alerta é retornada pelo sistema e nenhum registro será criado

Tabela 6 – Especificação dos casos de uso de consulta de dados

Casos de Uso	Manter Ambiente, Manter Categoria Trófica, Manter Equipamento, Manter Espécie Agrupada, Manter Estádio da Maturação, Manter Gênero/Espécie, Manter Local, Manter Migrad, Manter Ordem, Manter Ponto, Manter Projeto, Manter Aparelho, Manter Espécie, Manter Estádio, Manter Família, Manter Biótico, Manter Abiótico
Ação	Consultar
Descrição	Estes casos de uso especificam a ação de localizar um item no banco de dados. O pesquisador fornece os dados de um ou mais campos e clica no botão "Pesquisar"
Atores	Pesquisador
Precondições	O usuário deve estar cadastrado e logado no sistema
Pós-condições	Não há
Restrições	Impedir acesso ao sistema a usuários não cadastrados
Cenário de insucesso	Uma mensagem de alerta é retornada pelo sistema informando que nenhum registro será retornado

Tabela 7 – Especificação dos casos de uso de alteração de dados

Casos de Uso	Manter Ambiente, Manter Categoria Trófica, Manter Equipamento, Manter Espécie Agrupada, Manter Estádio da Maturação, Manter Gênero/Espécie, Manter Local, Manter Migrad, Manter Ordem, Manter Ponto, Manter Projeto, Manter Aparelho, Manter Espécie, Manter Estádio, Manter Família, Manter Biótico, Manter Abiótico
Ação	Alterar
Descrição	Estes casos de uso especificam a ação de alterar um item existente no banco de dados. Após consultar um item, o pesquisador clica no botão "Alterar". Uma tela com todos os dados referentes a este item é retornada para o usuário. O usuário altera os dados desejados. Ao clicar no botão "Alterar" o item é atualizado no banco de dados
Atores	Pesquisador
Precondições	O usuário deve estar cadastrado e logado no sistema; O item a ser alterado deve estar armazenado no banco de dados
Pós-condições	O registro selecionado é atualizado no banco de dados
Restrições	Impedir acesso ao sistema a usuários não cadastrados
Cenário de insucesso	Uma mensagem de alerta é retornada pelo sistema informando que nenhum registro será alterado

Tabela 8 – Especificação dos casos de uso de exclusão de dados

Casos de Uso	Manter Ambiente, Manter Categoria Trófica, Manter Equipamento, Manter Espécie Agrupada, Manter Estádio da Maturação, Manter Gênero/Espécie, Manter Local, Manter Migrad, Manter Ordem, Manter Ponto, Manter Projeto, Manter Aparelho, Manter Espécie, Manter Estádio, Manter Família, Manter Biótico, Manter Abiótico
Ação	Excluir
Descrição	Estes casos de uso especificam a ação de excluir um item do banco de dados. Após consultar um item, o pesquisador clica no botão "Excluir". Uma tela com todos os dados referentes a este item é retornada para o usuário. Ao clicar no botão "Excluir" o item é inativado no banco de dados
Atores	Pesquisador
Precondições	O usuário deve estar cadastrado e logado no sistema; O item a ser excluído deve estar armazenado no banco de dados
Pós-condições	O registro selecionado é atualizado no banco de dados
Restrições	Impedir acesso ao sistema a usuários não cadastrados; Impedir exclusão de dados que são usados por outros itens (chaves estrangeiras)
Cenário de insucesso	Uma mensagem de erro/alerta é retornada pelo sistema e nenhum registro será inativado

Tabela 9 – Especificação do caso de uso exportar dados

Casos de Uso	Exportar Dados
Ação	Extrair dados e gerar arquivo *csv*
Descrição	Este caso de uso especifica a ação de extrair dados bióticos e/ou abióticos do banco de dados e exportar para um arquivo *csv*. O pesquisador escolhe entre as opções biótico e/ou abiótico e o período de início e fim das coletas para a geração do arquivo. Ao clicar no botão "Exportar", é iniciado o processo de geração e download do arquivo gerado
Atores	Pesquisador
Precondições	O usuário deve estar cadastrado e logado no sistema; Ao menos um campo do formulário deve ser preenchido pelo usuário
Pós-condições	Um novo arquivo é gerado com os dados selecionados
Restrições	Impedir acesso ao sistema a usuários não cadastrados
Cenário de insucesso	Uma mensagem de erro/alerta é retornada pelo sistema e nenhum arquivo é criado

Tabela 10 – Especificação do caso de uso minerar dados

Casos de Uso	Minerar Dados
Ação	Gerar regras de associação *csv*
Descrição	Este caso de uso especifica a ação de realizar a mineração de dados armazenados no banco. O usuário escolhe um ou mais dos campos listados, decide entre o preenchimento ou não do período de início e fim da coleta e seleciona um ou mais locais para enviar para a mineração. O usuário configura os parâmetros da mineração preenchendo os campos "Métrica", "Valor mínimo da métrica", "Nº de Regras", "Suporte Mínimo", "Suporte Máximo", considerando que os campos de parâmetro podem manter os valores padrão já preenchidos pela ferramenta. Ao clicar no botão "Minerar" a ferramenta irá processar e retornar regras de associação
Atores	Pesquisador
Precondições	O usuário deve estar cadastrado e logado no sistema
Pós-condições	São listadas regras de associação
Restrições	Impedir acesso ao sistema a usuários não cadastrados
Cenário de insucesso	Uma mensagem de erro/alerta é retornada pelo sistema e nenhuma regra de associação é gerada

APÊNDICE B – Dicionário de dados

Este capítulo apresenta o dicionário de dados da aplicação Filhote.

Dicionário de dados - Filhote

abiotico

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	**INTEGER**	PK	NN	Sim
local_coleta_id	INTEGER		NN	
projeto_id	INTEGER		NN	
planc	VARCHAR(50)			
data_coleta	DATE			
hora	VARCHAR(10)			
margem	VARCHAR(50)			
smf	VARCHAR(50)			
ativ	VARCHAR(50)			
vento	VARCHAR(50)			
nebulosidade	VARCHAR(50)			
chuva	VARCHAR(50)			
tar	VARCHAR(50)			
tagua	VARCHAR(50)			
transp	VARCHAR(50)			
ph	VARCHAR(50)			
cond	VARCHAR(50)			
odmg	VARCHAR(50)			
amostra	VARCHAR(10)			
rede	VARCHAR(50)			
profundidade	VARCHAR(50)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id
abiotico_FKIndex2	FOREING_KEY	projeto_id
abiotico_FKIndex	FOREING_KEY	local_coleta_id

ambiente

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	**INTEGER**	PK	NN	Sim
nome	VARCHAR(500)			
observacoes	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id

aparelho

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	**INTEGER**	PK	NN	Sim
equipamento_id	INTEGER		NN	
nome	VARCHAR(100)		NN	
observacoes	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id
aparelho_FKIndex1	FOREING_KEY	equipamento_id

biotico

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	**INTEGER**	PK	NN	Sim
projeto_id	INTEGER		NN	
abiotico_id	INTEGER		NN	
aparelho_id	INTEGER			
estadio_id	INTEGER			
ambiente_id	INTEGER			
especie_id	INTEGER			
ponto_id	INTEGER			
local_coleta_id	INTEGER			
data_coleta	DATE			
horario	VARCHAR(15)			
numero	VARCHAR(50)			
lt	FLOAT			
ls	FLOAT			
wt	FLOAT			
sexo	VARCHAR(20)			
wg	FLOAT			
gre	FLOAT			
gri	FLOAT			
we	FLOAT			
wv	FLOAT			
rgs	FLOAT			
gv	FLOAT			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id
exemplar_FKIndex3	FOREING_KEY	local_coleta_id
exemplar_FKIndex5	FOREING_KEY	ponto_id
exemplar_FKIndex10	FOREING_KEY	especie_id
exemplar_FKIndex18	FOREING_KEY	ambiente_id
exemplar_FKIndex13	FOREING_KEY	estadio_id
exemplar_FKIndex13	FOREING_KEY	aparelho_id
exemplar_FKIndex18	FOREING_KEY	abiotico_id
biotico_FKIndex8	FOREING_KEY	projeto_id

ctrof

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	**INTEGER**	PK	NN	Sim
nome	VARCHAR(255)		NN	
observacoes	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id

equipamento

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	**INTEGER**	PK	NN	Sim
nome	VARCHAR(255)			
observacao	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id

especie

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	**INTEGER**	PK	NN	Sim
migrad_id	INTEGER		NN	
genero_especie_id	INTEGER		NN	
ctrof_id	INTEGER		NN	
familia_id	INTEGER		NN	
especie_agrupada_id	INTEGER			
nome	VARCHAR(255)		NN	
observacoes	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id
especie_FKIndex1	FOREING_KEY	especie_agrupada_id
especie_FKIndex2	FOREING_KEY	familia_id
especie_FKIndex3	FOREING_KEY	ctrof_id
especie_FKIndex4	FOREING_KEY	genero_especie_id
especie_FKIndex5	FOREING_KEY	migrad_id

especie_agrupada

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	**INTEGER**	PK	NN	Sim
nome	VARCHAR(255)		NN	
observacoes	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id

estadio

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	**INTEGER**	PK	NN	Sim
estadio_maturacao_id	INTEGER		NN	
nome	VARCHAR(100)		NN	
observacoes	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id
estadio_FKIndex1	FOREING_KEY	estadio_maturacao_id

estadio_maturacao

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	**INTEGER**	PK	NN	Sim
nome	VARCHAR(100)		NN	
observacoes	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id

familia

ColumnName	DataType	PrimaryKey	NotNull	AutoInc

id	INTEGER	PK	NN	Sim
ordem_id	INTEGER		NN	
nome	VARCHAR(255)		NN	
observacoes	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id
familia_FKIndex1	FOREING_KEY	ordem_id

genero_especie

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	INTEGER	PK	NN	Sim
nome	VARCHAR(255)		NN	
observacoes	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id

local_coleta

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	INTEGER	PK	NN	Sim
nome	VARCHAR(255)		NN	
codigo_local	VARCHAR(50)			
municipio	VARCHAR(255)			
observacoes	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id

migrad

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	INTEGER	PK	NN	Sim
nome	VARCHAR(255)		NN	
observacoes	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id

ordem

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	INTEGER	PK	NN	Sim
nome	VARCHAR(255)		NN	
observacoes	VARCHAR(1000)			

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id

ponto

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	INTEGER	PK	NN	Sim
nome	VARCHAR(500)		NN	

observacoes	VARCHAR(1000)

IndexName	IndexType	Columns
PRIMARY	PRIMARY	id

projeto

ColumnName	DataType	PrimaryKey	NotNull	AutoInc
id	**INTEGER**	PK	NN	Sim
nome	VARCHAR(255)		NN	
observacoes	VARCHAR(1000)			

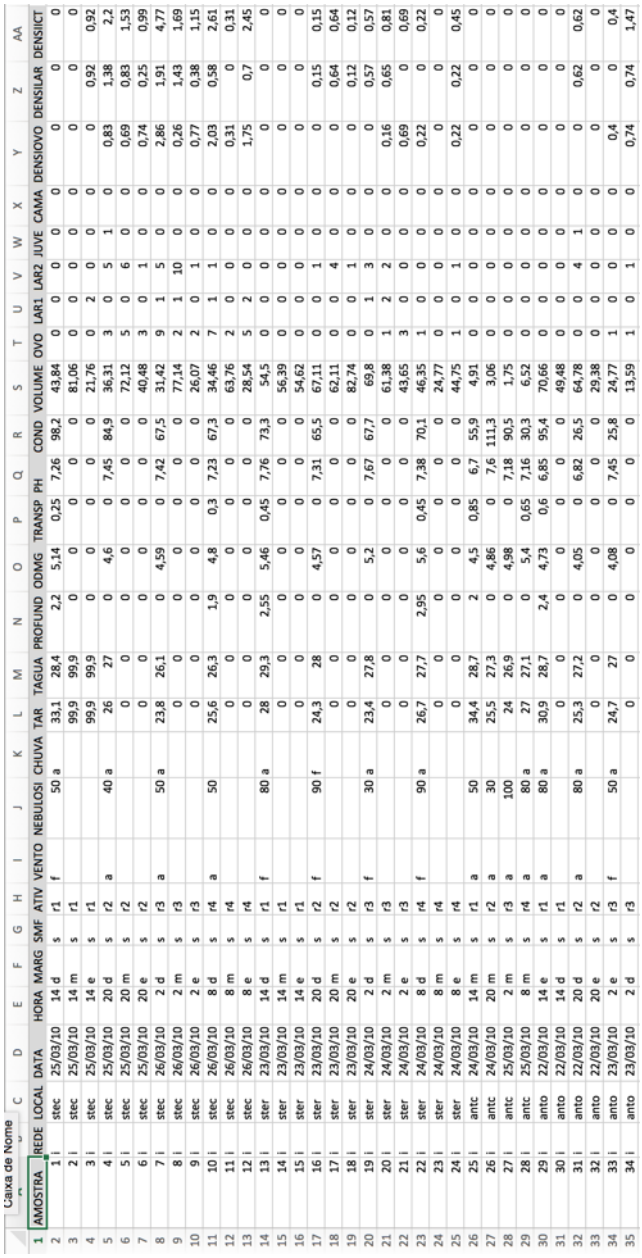
IndexName	IndexType	Columns
PRIMARY	PRIMARY	id

Anexos

ANEXO A – Amostra de dados abióticos na planilha inicial

A Figura 19 ilustra o conjunto de dados abióticos na planilha inicial.

Figura 19 – Dados abióticos primários



Caixa de Nome		C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	AMOSTRA	REDE	LOCAL	DATA	HORA	MARG	SMF	ATIV	VENTO	NEBULOSI	CHUVA	TAR	TAGUA	PROFUND	ODMG	TRANSP	PH	COND	VOLUME	OVO	LAR1	LAR2	JUVE	CAMA	DENSIOVO	DENSILAR	DENSIICT
2	1	i	stec	25/03/10	14 d	s	r1	f		50 a		33,1	28,4	2,2	5,14	0,25	7,26	98,2	43,84	0	0	0	0	0	0	0	0
3	2	i	stec	25/03/10	14 m	s	r1					99,9	99,9	0	0	0	0	0	81,06	0	0	0	0	0	0	0	0
4	3	i	stec	25/03/10	14 e	s	r1					99,9	99,9	0	0	0	0	0	21,76	0	2	0	0	0	0	0,92	0,92
5	4	i	stec	25/03/10	20 d	s	r2	a		40 a		26	27	0	4,6	0	7,45	84,9	36,31	3	0	5	1	0	0,83	1,38	2,2
6	5	i	stec	25/03/10	20 m	s	r2					0	0	0	0	0	0	0	72,12	5	0	6	0	0	0,69	0,83	1,53
7	6	i	stec	25/03/10	20 e	s	r2					0	0	0	0	0	0	0	40,48	3	0	1	0	0	0,74	0,25	0,99
8	7	i	stec	26/03/10	2 d	s	r3	a		50 a		23,8	26,1	0	4,59	0	7,42	67,5	31,42	9	1	5	0	0	2,86	1,91	4,77
9	8	i	stec	26/03/10	2 m	s	r3					0	0	0	0	0	0	0	77,14	2	1	10	0	0	0,26	1,43	1,69
10	9	i	stec	26/03/10	2 e	s	r3					0	0	0	0	0	0	0	26,07	2	0	1	0	0	0,77	0,38	1,15
11	10	i	stec	26/03/10	8 d	s	r4	a		50		25,6	26,3	1,9	4,8	0,3	7,23	67,3	34,46	7	1	1	0	0	2,03	0,58	2,61
12	11	i	stec	26/03/10	8 m	s	r4					0	0	0	0	0	0	0	63,76	2	0	0	0	0	0,31	0	0,31
13	12	i	stec	26/03/10	8 e	s	r4					0	0	0	0	0	0	0	28,54	5	2	0	0	0	1,75	0,7	2,45
14	13	i	ster	23/03/10	14 d	s	r1	f		80 a		28	29,3	2,55	5,46	0,45	7,76	73,3	54,5	0	0	0	0	0	0	0	0
15	14	i	ster	23/03/10	14 m	s	r1					0	0	0	0	0	0	0	56,39	0	0	0	0	0	0	0	0
16	15	i	ster	23/03/10	14 e	s	r1					0	0	0	0	0	0	0	54,62	0	0	0	0	0	0	0	0
17	16	i	ster	23/03/10	20 d	s	r2	f		90 f		24,3	28	0	4,57	0	7,31	65,5	67,11	0	0	1	0	0	0	0,15	0,15
18	17	i	ster	23/03/10	20 m	s	r2					0	0	0	0	0	0	0	62,11	0	0	4	0	0	0	0,64	0,64
19	18	i	ster	23/03/10	20 e	s	r2					0	0	0	0	0	0	0	82,74	0	0	1	0	0	0	0,12	0,12
20	19	i	ster	24/03/10	2 d	s	r3	f		30 a		23,4	27,8	0	5,2	0	7,67	67,7	69,8	0	1	3	0	0	0	0,57	0,57
21	20	i	ster	24/03/10	2 m	s	r3					0	0	0	0	0	0	0	61,38	1	2	2	0	0	0,16	0,65	0,81
22	21	i	ster	24/03/10	2 e	s	r3					0	0	0	0	0	0	0	43,65	3	0	0	0	0	0,69	0	0,69
23	22	i	ster	24/03/10	8 d	s	r4	f		90 a		26,7	27,7	2,95	5,6	0,45	7,38	70,1	46,35	1	0	0	0	0	0,22	0	0,22
24	23	i	ster	24/03/10	8 m	s	r4					0	0	0	0	0	0	0	24,77	0	0	0	0	0	0	0	0
25	24	i	ster	24/03/10	8 e	s	r4					0	0	0	0	0	0	0	44,75	1	0	1	0	0	0,22	0,22	0,45
26	25	i	antc	24/03/10	14 m	s	r1	a		50		34,4	28,7	2	4,5	0,85	6,7	55,9	4,91	0	0	0	0	0	0	0	0
27	26	i	antc	24/03/10	20 m	s	r2	a		30		25,5	27,3	0	4,86	0	7,6	111,3	3,06	0	0	0	0	0	0	0	0
28	27	i	antc	25/03/10	2 m	s	r3	a		100		24	26,9	0	4,98	0	7,18	90,5	1,75	0	0	0	0	0	0	0	0
29	28	i	antc	25/03/10	8 m	s	r4	a		80 a		27	27,1	0	5,4	0,65	7,16	30,3	6,52	0	0	0	0	0	0	0	0
30	29	i	anto	22/03/10	14 e	s	r1	a		80 a		30,9	28,7	2,4	4,73	0,6	6,85	95,4	70,66	0	0	0	0	0	0	0	0
31	30	i	anto	22/03/10	14 d	s	r1					0	0	0	0	0	0	0	49,48	0	0	0	0	0	0	0	0
32	31	i	anto	22/03/10	20 d	s	r2	a		80 a		25,3	27,2	0	4,05	0	6,82	26,5	64,78	0	0	4	1	0	0	0,62	0,62
33	32	i	anto	22/03/10	20 e	s	r2					0	0	0	0	0	0	0	29,38	0	0	0	0	0	0	0	0
34	33	i	anto	23/03/10	2 e	s	r3	f		50 a		24,7	27	0	4,08	0	7,45	25,8	24,77	1	0	0	0	0	0,4	0	0,4
35	34	i	anto	23/03/10	2 d	s	r3					0	0	0	0	0	0	0	13,59	1	0	1	0	0	0,74	0,74	1,47

ANEXO B – Amostra de dados bióticos na planilha inicial

A Figura 20 ilustra o conjunto de dados bióticos na planilha inicial.

Figura 20 – Dados bióticos primários