

ALLAN KOCH VEIGA

UM ESTUDO SOBRE QUALIDADE DE DADOS EM BIODIVERSIDADE:  
APLICAÇÃO A UM SISTEMA DE DIGITALIZAÇÃO DE OCORRÊNCIAS  
DE ESPÉCIES

Dissertação apresentada à Escola  
Politécnica da Universidade de São Paulo  
para obtenção do título de Mestre em  
Ciências.

São Paulo  
2012



**Este exemplar foi revisado e alterado em relação à versão original, sob responsabilidade única do autor e com a anuênciā de seu orientador.**

**São Paulo, 27 de fevereiro de 2012.**

**Assinatura do autor** \_\_\_\_\_

**Assinatura do orientador** \_\_\_\_\_

## **FICHA CATALOGRÁFICA**

**Veiga, Allan Koch**

**Um estudo sobre qualidade de dados em biodiversidade:  
aplicação a um sistema de digitalização de ocorrências de  
espécies / A.K. Veiga. -- ed. rev -- São Paulo, 2012.**

**101 p.**

**Dissertação (Mestrado) - Escola Politécnica da Universidade  
de São Paulo. Departamento de Engenharia de Computação e  
Sistemas Digitais.**

**1. Informação (Qualidade) 2. Informática 3. Biodiversidade  
4. Sistemas de informação 5. Ocorrência de espécie I. Univer-  
sidade de São Paulo. Escola Politécnica. Departamento de  
Engenharia de Computação e Sistemas Digitais II. t.**

ALLAN KOCH VEIGA

UM ESTUDO SOBRE QUALIDADE DE DADOS EM BIODIVERSIDADE:  
APLICAÇÃO A UM SISTEMA DE DIGITALIZAÇÃO DE OCORRÊNCIAS  
DE ESPÉCIES

Dissertação apresentada à Escola  
Politécnica da Universidade de São Paulo  
para obtenção do título de Mestre em  
Ciências.

Área de Concentração: Sistemas Digitais

Orientador: Antonio Mauro Saraiva

São Paulo  
2012

*À minha família  
pelo incentivo e apoio  
à realização deste trabalho.*

## **AGRADECIMENTOS**

Ao meu orientador, Prof. Dr. Antonio Mauro Saraiva, pela orientação e dedicação na condução deste trabalho e, sobretudo, pela confiança, pulso firme, oportunidades e amizade que permitiram o meu crescimento profissional e pessoal durante estes dois anos de POLI.

À minha família que é meu ponto de confiança e que me deu incentivo, apoio e suporte para eu seguir em frente. Em especial, à minha mãe Leila, à minha madrasta Miriam, ao meu pai Jonas e ao meu irmão Felipe que contribuíram fortemente na minha formação pessoal e profissional.

À família LABI (Laboratório de Bioinformática da Universidade Estadual do Oeste do Paraná) pelos valorosos ensinamentos, conselhos e disciplinas que tiveram grande impacto sobre o meu modo de pensar, de agir e de ser. Em especial, aos coordenadores do LABI, Professores Dra. Huei Diana Lee, Dr. Wu Feng Chung e Msc. Renado Bobsin Machado, exemplos de pesquisadores e de pessoas para mim.

Ao Msc. Etienne Americo Cartolano Jr. pela contribuição, incentivo, apoio e pela experiência de dez anos de POLI compartilhada durante todo a minha permanência na USP.

Aos companheiros de pesquisa do LAA, em especial, João Victor, Lucas Dias, Lucas Isern, Luiz Saraiva, Guilhermo, Diogo, Jorge, Fernanda, Rafael, Willian, Raul, João Ferreira e Aline, que contribuíram diretamente e indiretamente para a realização deste trabalho.

Aos amigos, por contribuírem com os bons e breves momentos de “descanso” em Foz do Iguaçu e em Itajaí e com o enriquecimento da minha formação nas atividades extracurriculares.

À *Inter-American Biodiversity Information Network – Pollinators Thematic Network* (IABIN-PTN), cujo projeto apoiado pelo *Global Environmental Facility* (GEF) e Organização dos Estados Americanos (OEA) permitiu a concessão da Bolsa de Mestrado ao autor, por meio da Fundação de Apoio à USP (FUSP).

À Sra. Clélia e seus colaboradores da FUSP pelo excelente serviço no atendimento e no suporte as questões relativas as bolsas.

Este trabalho foi desenvolvido no âmbito do Núcleo de Apoio à Pesquisa em Biodiversidade e Computação da Universidade de São Paulo (NAP BioComp).

## RESUMO

Para o combate da atual crise de sustentabilidade ambiental, diversos estudos sobre a biodiversidade e o meio ambiente têm sido realizados com o propósito de embasar estratégias eficientes de conservação e uso de recursos naturais. Esses estudos são fundamentados em avaliações e monitoramentos da biodiversidade que ocorrem por meio da coleta, armazenamento, análise, simulação, modelagem, visualização e intercâmbio de um volume expressivo de dados sobre a biodiversidade em amplo escopo temporal e espacial. Dados sobre ocorrências de espécies são um tipo de dado de biodiversidade particularmente importante, pois são amplamente utilizados em diversos estudos. Contudo, para que as análises e os modelos gerados a partir desses dados sejam confiáveis, os dados utilizados devem ser de alta qualidade. Assim, para melhorar a Qualidade de Dados (QD) sobre ocorrências de espécies, o objetivo deste trabalho foi realizar um estudo sobre QD aplicado a dados de ocorrências de espécies que permitisse avaliar e melhorar a QD por meio de técnicas e recursos de prevenção a erros. O estudo foi aplicado a um Sistema de Informação (SI) de digitalização de dados de ocorrências de espécies, o *Biodiversity Data Digitizer* (BDD), desenvolvido no âmbito dos projetos da *Inter-American Biodiversity Information Network – Pollinators Thematic Network* (IABIN-PTN) e BioAbelha – FAPESP. Foi realizada uma revisão da literatura sobre dados de ocorrências de espécies e sobre os seus domínios de dados mais relevantes. Para os domínios de dados identificados como mais importantes (táxon, geoespacial e localização), foi realizado um estudo sobre a Avaliação da QD, no qual foi definido um conceito de QD em relação a cada domínio de dados por meio da identificação, definição e inter-relação de dimensões de QD (aspectos) importantes e de problemas que afetam essas dimensões. Embasado nesse estudo foram identificados recursos computacionais que permitissem melhorar a QD por meio da redução de erros. Utilizando uma abordagem de Gerenciamento da QD de prevenção a erros, foram identificados 13 recursos computacionais que auxiliam na prevenção de 8 problemas de QD, proporcionando, assim, uma melhoria da acurácia, precisão, completude, consistência, credibilidade da fonte e confiabilidade de dados taxonômicos, geoespaciais e de localização de ocorrências de espécies. Esses recursos foram implementados em duas ferramentas integradas ao BDD. A

primeira é a *BDD Taxon Tool*. Essa ferramenta facilita a entrada de dados taxonômicos de ocorrências livres de erros por meio de, entre outros recursos, técnicas de *fuzzy matching* e sugestões de nomes e de hierarquias taxonômicas baseados no *Catalog of Life*. A segunda ferramenta, a *BDD Geo Tool*, auxilia o preenchimento de dados geoespaciais e de localização de ocorrências de espécies livres de erros por meio de técnicas de georeferenciamento a partir de descrição em linguagem natural da localização, de georeferenciamento reverso e de mapas interativos do Google *Earth*, entre outros recursos. Este trabalho demonstrou que com a implementação de determinados recursos computacionais em SI, problemas de QD podem ser reduzidos por meio da prevenção a erros. Como consequência, a QD em domínios de dados específicos é melhorada em relação a determinadas dimensões de QD.

Palavras-chave: Qualidade de dados, informática para biodiversidade, sistemas de informação, ocorrências de espécies, biodiversidade.

## ABSTRACT

For fighting the current environment sustainability crisis, several studies on biodiversity and the environment have been conducted in order to support efficient strategies for conservation and sustainable use of natural resources. These studies are based on assessment and monitoring of biodiversity that occur by means of the collection, storage, analysis, simulation, modeling, visualization and sharing of a significant volume of biodiversity data in broad temporal and spatial scale. Species occurrences data are a particularly important type of biodiversity data because they are widely used in various studies. Nevertheless, for the analyzing and modeling obtained from these data to be reliable, the data used must be high quality. Thus, to improve the Data Quality (DQ) of species occurrences, the aim of this work was to conduct a study about DQ applied to species occurrences data that allowed assessing and improving the DQ using techniques and resources to prevent errors. This study was applied to an Information System (IS) designed to digitize species occurrences, the Biodiversity Data Digitizer (BDD), that was developed in the scope of the Inter-American Biodiversity Information Network – Pollinators Thematic Network (IABIN-PTN) and BioAbelha – FAPESP projects. A literature review about species occurrences data and about the most relevant data domains was conducted. For the most important data domains identified (taxon, geospatial and location), a study on the DQ Assessment was performed, in which important DQ dimensions (aspects) and problems that affect these dimensions were identified, defined and interrelated. Based upon this study, computational resources were identified that would allow improving the DQ by reducing errors. Using the errors preventing DQ Management approach, 13 computing resources to support the prevention of 8 DQ problems were identified, thus providing an improvement of accuracy, precision, completeness, consistency, credibility of source and believability of taxonomic, geospatial and location data of species occurrences. These resources were implemented in two tools integrated to the BDD IS. The first tool is the BDD Taxon Tool. This tool facilitates the entrance of error-free taxonomic data of occurrences by means of fuzzy matching techniques and suggestions for taxonomic names and hierarchies based on Catalog of Life, among other resources. The second tool, the BDD Geo Tool, helps to fill in error-free geospatial and location data about species

occurrence by means of georeferencing techniques from natural language description of location, reverse georeferencing and Google Earth interactive maps, among other resources. This work showed that with the development of certain computing resources integrated to an IS, DQ problems are reduced by preventing errors. As a result of reducing some problems in particular, the DQ in specific data domains is improved for certain DQ dimensions.

**Keywords:** Data quality, biodiversity informatics, information system, species occurrences, biodiversity.

## **LISTA DE FIGURAS**

Figura 1 - Cadeia de produção de informações de ocorrências de espécies.	28
Figura 2 - Modelo de requisitos de Kano. Baseado em Bolt & Mazur (1999).	34
Figura 3 - Um <i>framework</i> para Avaliação da QD.....	36
Figura 4 - Definição de dimensões de QD a partir de diferentes aspectos..	39
Figura 5 - Definição de problemas de QD.....	44
Figura 6 - Definição de dimensões de QD. Baseada em Dalcin (2005).....	46
Figura 7 - Metodologia de Avaliação da QD.....	47
Figura 8 - Relação entre precisão e acurácia em dados geoespaciais .....	56
Figura 9 - Modelo hierárquico de dimensões de QD .....	58
Figura 10 - Representação da Avaliação da QD de ocorrências.....	59
Figura 11 - Arquitetura de <i>software</i> do BDD.....	72
Figura 12 - <i>Autocomplete</i> de nomes de táxons.....	80
Figura 13 - Sugestões de nomes válidos.....	81
Figura 14 - Sugestões de hierarquias válidas e inválidas.....	82
Figura 15 - Sequência de uso da ferramenta BTT.....	83
Figura 16 - Primeira etapa da ferramenta BGT.....	85
Figura 17 - Georeferenciamento reverso.....	86
Figura 18 - Georeferenciamento utilizando um mapa interativo.....	87
Figura 19 - Georeferenciamento a partir da descrição “ <i>near sao paulo</i> ”.....	87
Figura 20 - Indicador de incerteza.....	88
Figura 21 - Sequência de uso da ferramenta BGT .....	89

## **LISTA DE ABREVIATURAS E SIGLAS**

ABCD – *Access to Biological Collections Data*

AJAX – Asynchronous Javascript and XML

ALA – *Atlas of Living Australia*

API – *Application Programming Interface*

BDD – *Biodiversity Data Digitizer*

BGT – BDD Geo Tool

BTT – BDD Taxon Tool

CDB – Convenção sobre Diversidade Biológica

CoL – *Catalog of Life*

DwC – Darwin Core

EoL – *Encyclopedia of Life*

GBIF – *Global Biodiversity Information Facility*

GPS – Global Positioning System

HTML – *Hypertext Markup Language*

IABIN – *Inter-American Biodiversity Information Network*

IABIN-PTN – *IABIN Pollinators Thematic Network*

IB – Informática para Biodiversidade

ITIS – *Integrated Taxonomic Information System*

JSON – Javascript Object Notation

LAA – Laboratório de Automação Agrícola

MVC – *Model-View-Controller*

PDD – *Pollinator Data Digitizer*

PHP - *Hypertext Preprocessor*

PTN – *Pollinator Thematic Network*

QD – Qualidade de Dados

SI – Sistema de Informação

SIG – Sistema de Informações Geográficas

TAPIR – *TDWG Access Protocol for Information Retrieval*

TDWG – *Biodiversity Information Standards (Taxonomic Database Working Group)*

TI – Tecnologia da Informação

XML – *eXtensible Markup Language*

## SUMÁRIO

1.	Introdução .....	13
1.1.	Objetivo .....	13
1.2.	Justificativa.....	13
1.3.	Metodologia.....	16
1.4.	Organização do texto .....	17
2.	Tópicos em Informática para Biodiversidade .....	19
2.1.	Introdução .....	19
2.1.1.	Iniciativas para a padronização de informações .....	20
2.2.	Ferramentas e serviços aplicados em IB .....	22
2.2.1.	Repositórios de informações taxonômicas.....	22
2.2.2.	Serviços de georeferenciamento.....	23
2.3.	Dados de ocorrências de espécies .....	25
2.4.	Cadeia de produção de informações sobre ocorrências de espécies	27
2.4.1.	Atores .....	27
2.5.	Considerações finais do capítulo .....	31
3.	Qualidade de Dados .....	33
3.1.	Abordagens de pesquisas sobre QD .....	35
3.1.1.	Avaliação da QD .....	35
3.1.2.	Gerenciamento da QD .....	40
3.2.	Considerações finais do capítulo .....	41
4.	Materiais e Métodos.....	42
4.1.	Estudo do domínio de aplicação .....	42
4.2.	Definição do escopo .....	42
4.3.	Estudo sobre a Avaliação da QD .....	43
4.3.1.	Identificar problemas de QD.....	43
4.3.2.	Definir problemas de QD .....	44
4.3.3.	Identificar dimensões de QD .....	44
4.3.4.	Definir dimensões de QD .....	45
4.3.5.	Definição de uma metodologia de avaliação da QD .....	46
4.4.	Estudo sobre o Gerenciamento da QD .....	47
4.5.	Estudo de caso de aplicação dos estudos de QD .....	47

4.5.1.	Análise e desenvolvimento de um SI .....	48
4.5.2.	Desenvolvimento de ferramentas de QD .....	48
5.	Resultados .....	49
5.1.	Estudo sobre QD de ocorrências de espécies.....	49
5.1.1.	Avaliação da QD .....	50
5.1.2.	Gerenciamento da QD .....	60
5.2.	Estudo de caso de aplicação dos estudos de QD .....	71
5.2.1.	Sistema de Informação: BDD .....	71
5.2.2.	Ferramentas de QD desenvolvidas .....	79
6.	Considerações Finais .....	91
6.1.	Contribuições .....	91
6.2.	Conclusões .....	93
6.3.	Trabalhos futuros .....	93

# **1. INTRODUÇÃO**

Neste capítulo são apresentados o objetivo do trabalho, a justificativa para a sua realização e a metodologia utilizada para a alcançar o objetivo proposto. Também é descrita a estrutura do texto.

## **1.1. Objetivo**

O objetivo deste trabalho é realizar um estudo sobre Qualidade de Dados no contexto de Informática para Biodiversidade que permita identificar causas que afetam a Qualidade de Dados de ocorrências de espécies em seus domínios de dados mais relevantes e identificar técnicas e recursos computacionais que permitam melhorar aspectos importantes da qualidade desses dados. Com base nesse estudo, implementar ferramentas que permitam melhorar a Qualidade de Dados de ocorrências de espécies em um Sistema de Informação de digitalização de dados de biodiversidade.

## **1.2. Justificativa**

O planeta tem passado por uma crise de sustentabilidade decorrente da dificuldade em se estabelecer um equilíbrio entre o desenvolvimento econômico e social e a conservação do meio ambiente e da biodiversidade (Brundtland, 1987). Para o combate dessa crise de magnitude ainda desconhecida, diversos estudos sobre sustentabilidade têm sido realizados com o propósito de embasar estratégias de conservação e de uso de recursos naturais de modo a não comprometer as futuras gerações, especialmente quando se trata de assuntos ambientais, nos quais os danos podem ser irreversíveis (Hill *et al.*, 2010).

A degradação da diversidade biológica, ou biodiversidade, pode afetar o planeta e a sociedade em diversos aspectos. Segundo Schnase *et al.* (2007), a biodiversidade e os ecossistemas são responsáveis por nos fornecer ar puro, água potável, alimento, vestuários, abrigos e medicamentos. Essa diversidade, em conjunto com os ecossistemas que a suporta, também contribui com trilhões de dólares para a economia mundial, de maneira direta, em setores como agricultura,

silvicultura, pesca e ecoturismo, e indireta, por meio de serviços biologicamente mediados, como polinização de plantas, dispersão de sementes, pastagens, remoção de dióxido de carbono, fixação de nitrogênio, controle de enchentes, degradação de resíduos e o controle biológico de pragas. Por esses motivos, tem havido uma crescente necessidade em se compreender e resolver complexos problemas relacionados a esse tema (MA, 2005).

Segundo Stockwell (2007, apud Cartolano, 2009, p. 21), a Ciência da Biodiversidade procura entender as tendências relacionadas à riqueza da diversidade dos ambientes biológicos. Essa riqueza é um dos fatores que mais influencia a estabilidade e a saúde do meio ambiente (Schnase *et al.*, 2007). Estudos demonstram que ações de conservação ou de degradação da biodiversidade têm impacto sobre conceitos complexos e interdependentes e de alto nível de organização, como biomas, ecossistemas, floras e faunas (Bisby, 2000).

Contudo, para que as ações de conservação e de uso sustentável da biodiversidade sejam efetivas, é necessário avaliar e monitorar de maneira integrada e contínua o *status* do risco de perda de recursos, por meio da coleta, armazenamento, análise, simulação, visualização e intercâmbio de um volume expressivo de informações de amplo escopo temporal e espacial. Para tanto, recursos computacionais e de comunicação modernas são necessárias (Saraiva, 2003). Como consequência dessa demanda, uma nova área da computação surgiu, a *Biodiversity Informatics*, ou Informática para Biodiversidade – IB (Canhos, 2003; Saraiva, 2003).

Nesse sentido, várias iniciativas foram criadas com a proposta de digitalizar, integrar e publicar dados de espécies e espécimes em portais na Internet (Canhos *et al.*, 2004). Dentre essas iniciativas destacam-se, pelo grande volume de dados publicados e pela abrangência geográfica global, o *Global Biodiversity Information Facility* – GBIF (GBIF, 2011), o *Encyclopedia of Life* – EoL (EoL, 2011) e o *Catalog of Life* – CoL (CoL, 2011). Há, também, iniciativas que se especializaram em determinados conjuntos de espécies ou regiões geográficas, ou em ambos, como a *Inter-American Biodiversity Information Network* – IABIN, a VertNet (VERTNET, 2011), a *Euro-Mediterranean Plant Diversity* e o *Atlas of Living Australia* – ALA (ALA, 2011), por exemplo.

A IABIN consiste em um importante fórum criado para promover a coordenação e colaboração técnica entre os países da América para realizar a digitalização, compartilhamento e uso de informações sobre biodiversidade (Cartolano, 2009). Essas ações são organizadas em cinco redes temáticas: Áreas Protegidas, Espécies e Espécimes, Ecossistemas, Espécies Invasoras e Polinizadores (IABIN, 2011). Cada rede temática possui um grupo de trabalho multidisciplinar dedicado ao seu tema.

A Rede Temática de Polinizadores (*Pollinators Thematics Network – PTN*) tem como principal objetivo coordenar esforços para realizar a digitalização, o compartilhamento e o uso de informações relacionadas a polinização e espécies polinizadoras. Para tanto, a Universidade de São Paulo, representada pelo Laboratório de Automação Agrícola – LAA da Escola Politécnica tem definido padrões técnicos e desenvolvido ferramentas de digitalização e de publicação de dados sobre espécies e espécimes. Uma dessas ferramentas é o *Biodiversity Data Digitizer – BDD* (Cartolano *et al.*, 2010; Saraiva *et al.*, 2011).

O BDD é um Sistema de Informação – SI baseado na *web* desenvolvido para facilitar a digitalização, manipulação e publicação de dados de biodiversidade. Por meio desse SI, dados sobre ocorrências de espécies, interações, déficit de polinização, monitoramento de polinizadores, recursos multimídia e recursos bibliográficos podem ser gerenciados e publicados em portais de dados, como os portais da IABIN-PTN, do projeto Polinizadores do Brasil, do GBIF, entre outros.

Dentre os diversos tipos de dados sobre a biodiversidade digitalizados por meio do BDD, destacam-se os dados sobre ocorrência de espécies. Esses dados são amplamente utilizados em muitos estudos sobre a gestão e o uso sustentável do meio ambiente (Chapman, 2005a). Uma ocorrência de espécie pode ser definida com uma observação ou uma coleta de um organismo biológico em um determinado espaço geográfico e tempo, ou seja, é um testemunho de um fato biológico.

Assim, o sistema BDD assume uma funcionalidade chave no contexto da IB: a digitalização. É por meio da digitalização que as informações sobre a biodiversidade são coletadas e armazenadas em formato digital para posteriormente serem utilizadas. A digitalização pode ser considerada uma questão crítica, pois os dados produzidos nesse processo são utilizados como insumo em análises, modelagens, simulações e visualizações, os quais são utilizados para melhorar a compreensão

sobre a natureza. Portanto, a digitalização constitui uma etapa fundamental para o cumprimento dos objetivos da IB e da IABIN-PTN.

Contudo, para que as análises e os modelos gerados a partir desses dados sejam confiáveis, é necessário que os dados utilizados sejam de alta qualidade. O uso indiscriminado de dados, sem considerar possíveis erros, pode levar a resultados incorretos, a informações enganosas e, por consequência, a tomadas de decisões que podem afetar negativamente a gestão e a manutenção do meio ambiente (Chapman, 2005b).

É consenso que a Qualidade de Dados – QD tem impacto nas tomadas de decisões, na credibilidade dos dados, na satisfação dos usuários, no custo de gerenciamento de banco de dados e no valor e no uso efetivo dos dados (Chapman, 2005c). A baixa QD pode ter um forte impacto sobre a eficácia global de uma organização (Wand & Wang, 1996).

Devido a esses fatores e com o rápido crescimento da disponibilidade e da troca de dados sobre biodiversidade global, os consumidores dos dados têm exigido um melhor detalhamento sobre a qualidade dos dados (Chapman, 2005b). Segundo o GBIF (Hill *et al.*, 2010), é essencial assegurar que a qualidade dos dados de ocorrências de espécies seja tão boa quanto reportada para que se possa determinar a sua adequação ao uso. Nesse sentido, a avaliação da QD assume um importante papel.

Portanto, a aplicação de métodos de Avaliação e de Gerenciamento da QD em SI de gestão de dados de ocorrências de espécies pode impactar significativamente a qualidade e a credibilidade de análises, simulações e modelos gerados a partir dos dados produzidos por meio desse SI e, por consequência, permitir que as ações estratégicas e de gestão da biodiversidade e do meio ambiente sejam mais efetivas, melhorando assim, a qualidade de vida dos seres humanos (Chapman, 2005b).

### **1.3. Metodologia**

Para atingir o objetivo deste trabalho, foi necessário realizar uma revisão da literatura sobre Informática para Biodiversidade, especialmente sobre dados de ocorrências de espécies, para compreender o domínio de aplicação. Dentro desse

domínio de aplicação, foi preciso definir um escopo. Esse escopo foi definido com base nos domínios de dados de ocorrências de espécies mais relevantes.

Posteriormente, foi necessário definir um conceito de QD para cada um dos domínios de dados identificados. Essa definição foi realizada com base no *framework* de Avaliação da QD proposto por Ge & Helfert (2007). Desse modo, foi realizada uma identificação de problemas de QD comuns em cada domínio de dados. Foi também preciso identificar as dimensões de QD e os seus significados em relação a cada domínio de dados. Por fim, foi feita uma análise para identificar quais dimensões de QD podem ser afetadas pelos problemas identificados em relação a cada domínio de dados.

Com base nesse estudo sobre Avaliação da QD, aplicado ao escopo definido neste trabalho, foi possível realizar um estudo sobre o Gerenciamento da QD de ocorrências de espécies, o qual consistiu em identificar mecanismos, técnicas e recursos para melhorar a QD. Como estudo de caso, esses recursos foram implementados em um SI, a fim de melhorar a QD por meio da prevenção a erros durante a digitalização de dados de ocorrências de espécies.

#### **1.4. Organização do texto**

Este texto é composto de seis capítulos, distribuídos da seguinte forma:

- O Capítulo 2 apresenta uma introdução sobre conceitos, ferramentas, importância e história do emergente campo da computação denominado de Informática para Biodiversidade.
- O Capítulo 3 introduz conceitos e abordagens metodológicas relacionados a Qualidade de Dados em Sistemas de Informação.
- O Capítulo 4 apresenta uma descrição dos materiais e métodos utilizados na pesquisa para alcançar o objetivo estabelecido neste trabalho.

- O Capítulo 5 traz os resultados obtidos quanto à aplicação de conceitos e métodos de Qualidade de Dados em um Sistema de Informação sobre Biodiversidade.
- As contribuições e as conclusões deste trabalho, e as propostas de trabalhos futuros são apresentados no Capítulo 6.

## **2. TÓPICOS EM INFORMÁTICA PARA BIODIVERSIDADE**

Para se definir um conceito de QD de biodiversidade é necessário entender o contexto em que esses dados são gerados, manipulados e utilizados. Isso é essencial para se definir diretrizes para realizar a Avaliação e o Gerenciamento da QD, bem como identificar fontes da degradação da qualidade. Portanto, nesse capítulo é apresentada uma breve revisão da literatura sobre Informática para Biodiversidade.

### **2.1. Introdução**

A Biodiversidade pode ser definida como “a variabilidade entre os organismos vivos de todas as fontes incluindo, entre outras, ecossistemas terrestres, marinhos e outros ecossistemas aquáticos, e os complexos ecológicos dos quais eles são parte; isso inclui diversidade dentro das espécies, entre espécies e de ecossistemas” (Steinhage, 2003 apud Saraiva, 2003). Em Hawksworth (1996), a Biodiversidade é descrita como: a diversidade de espécies em uma comunidade de organismos vivos e dos ecossistemas aos quais eles pertencem.

O uso de informações sobre a biodiversidade é crítico para a tomada de decisões em uma ampla gama de domínios (Canhos, 2004). Segundo Saraiva & Canhos (2011), há uma crescente necessidade de se entender e resolver complexos problemas de meio ambiente. Assim como tem-se desenvolvido a capacidade de predizer eventos climáticos, é necessário desenvolver a capacidade de predizer os resultados ecológicos do futuro.

Para isso, o uso integrado de informações de biodiversidade em amplo escopo geográfico e temporal, de diversas fontes de dados, associadas a informações sobre mudanças globais dos ecossistemas, dados sobre o ciclo de vida do carbono e de dados abióticos como precipitação, umidade, entre outros, é essencial para a realização de análises que produzam resultados comprehensíveis e úteis (Saraiva & Canhos, 2011; Canhos, 2004).

Assim, a forte demanda pela acessibilidade, integração e visualização de diferentes repositórios de dados de biodiversidade de diferentes campos de conhecimento, motivou o desenvolvimento de um novo campo de pesquisa, a IB.

Esse campo emergente tem o objetivo de usar e gerenciar de maneira eficiente informações globais sobre a biodiversidade por meio de ferramentas que auxiliam na análise e no entendimento dessas informações (Saraiva, 2003).

Entretanto, Schnase (2007) destaca dois fatores que mais afetam o trabalho nesse campo: a complexidade biológica e a complexidade sociológica. A complexidade biológica é resultante de mais de três bilhões de anos de evolução, o que causou uma alta variabilidade química, fisiológica, do ciclo de desenvolvimento e do comportamento das espécies. Existem milhões de espécies, cada uma apresenta variações em relação aos organismos individuais e a sua população. Há centenas, se não milhares, de ecossistemas, cada um compreendendo interações complexas entre diversas espécies e múltiplos fatores abióticos.

A complexidade sociológica inclui problemas de comunicação e de coordenação entre diferentes instituições, as quais podem ter interesses e pontos de vistas divergentes, possuírem diferentes níveis de experiência e conhecimento, além de estarem localizadas em regiões geográficas distintas. Os dados de biodiversidade e do ecossistema podem ser politicamente e comercialmente sensíveis, podendo haver conflitos de interesses. Além disso, os tipos de dados sobre os organismos coletados, os métodos de coleta, a acurácia e precisão, e a estrutura dos dados podem ser distintas. Por exemplo, importantes dados podem ser coletados por não-cientistas, como observadores amadores de pássaros, ou os dados podem ser obtidos de fontes de dados geográficos, meteorológicos, químicos e genéticos. Portanto, há uma necessidade de acomodar diferentes tipos de dados com diferentes níveis de qualidade dentro de uma infraestrutura democratizada de informação formal e informal (Schnase, 2007).

Assim, diversas iniciativas surgiram com a proposta de criar padrões, protocolos e ferramentas para facilitar o acesso e o uso de informações consistentes e confiáveis sobre a biodiversidade entre sistemas heterogêneos de diferentes organizações.

### **2.1.1. Iniciativas para a padronização de informações sobre biodiversidade**

Na Convenção sobre a Diversidade Biológica – CDB, comumente referida como Convenção da Biodiversidade, realizada no Rio de Janeiro em 1992, foi

firmado um acordo global que cobre todos os aspectos da diversidade biológica: conservação, uso sustentável e compartilhamento dos benefícios dos recursos genéticos (CBD, 2011).

Para alcançar os seus objetivos, a CDB passou a promover o surgimento de iniciativas para utilizar a Tecnologia da Informação – TI como suporte a Ciência da Biodiversidade. Essas iniciativas têm o papel de prover um conjunto de ferramentas para atender necessidades de: coleta de dados; registro e armazenamento dos dados; análise dos dados; acesso aos dados e sua divulgação; integração de dados (Saraiva, 2003).

Com os avanços decorrentes da CDB e dos esforços de governos e sociedade civil, houve uma enorme evolução da produção de informações associadas a amostras (espécimes) de material biológico depositadas em herbários e coleções zoológicas. Contudo, a base de conhecimento sobre a biodiversidade global continuava incipiente e desagregada (Canhos, 2003).

Portanto, para melhorar a interoperabilidade entre as informações mantidas pelas iniciativas, o *Taxonomic Database Working Group* – TDWG (TDWG, 2011), em parceria com outras organizações, passou a unir esforços para definir padrões para a troca de informações biológicas e protocolos para a interoperabilidade de sistemas de informação. O TDWG, atualmente chamado de *Biodiversity Information Standards*, promove as suas atividades em um ambiente colaborativo e internacional com uma equipe multidisciplinar que envolve: biólogos, zoólogos, entomologistas, ecologistas, geneticistas, cientistas da computação, engenheiros, etc. (Cartolano, 2009). Entre os principais padrões ratificados pelo TDWG, estão o *Access to Biological Collection Data* – ABCD, o *Darwin Core* – DwC e o *TDWG Access Protocol for Information Retrieval* – TAPIR.

#### **2.1.1.1. Padrão Darwin Core**

O DwC é um padrão de metadados baseado principalmente em táxons e suas ocorrências na natureza documentadas por observações ou coletas de espécies, e outras informações correlatas. Esse padrão é documentado por meio de um glossário de termos, os quais são utilizados para descrever ocorrências de espécies e permitir que essas informações sejam facilmente compartilhadas entre sistemas heterogêneos. O DwC disponibiliza documentos que descrevem como estes termos

são gerenciados, como podem ser utilizados e como podem ser estendidos, a fim de abranger novos domínios de dados.

Visando o compartilhamento de informações sobre ocorrências de espécies entre sistemas heterogêneos, o DwC foi implementado como um esquema de metadados em *Extensible Markup Language – XML*. Esse esquema é organizado em sete principais domínios de dados: nível de registro, ocorrência, evento, localização, contexto geológico, identificação e táxon (DwC, 2011). Além desses domínios de dados, o DwC admite o acoplamento de outros domínios específicos para determinadas aplicações, chamadas de extensões, como por exemplo, domínio de dados geoespaciais e de interações entre espécimes.

## **2.2. Ferramentas e serviços aplicados em IB**

Com o progresso da IB e os avanços da TI, diversas ferramentas e serviços foram desenvolvidos e disponibilizados na Internet para o uso da comunidade. Essas ferramentas e serviços podem ser utilizados como importantes recursos em SI sobre biodiversidade, facilitando a digitalização de dados de qualidade.

### **2.2.1. Repositórios de informações taxonômicas**

A padronização de nomenclaturas e de hierarquias taxonômicas é essencial para a manipulação computacional dessas informações, pois a ambiguidade e a incerteza que frequentemente estão associadas a dados taxonômicos podem representar obstáculos a sua computação. Portanto, a disponibilidade de informações taxonômicas consistentes, padronizadas e consideradas corretas pela comunidade, pode ser um importante recurso para melhorar a credibilidade, acurácia e a consistência de informações de biodiversidade. A seguir, são brevemente descritos dois importantes repositórios centralizados de informações taxonômicas disponíveis na Internet.

#### **2.2.1.1. Catalog of Life - COL**

O CoL é planejado para ser um catálogo de todas as espécies conhecidas de organismos do mundo. A décima primeira edição desse catálogo contém o registro

de 1.370.276 espécies. Esse catálogo é resultado da compilação de 101 bancos de dados de diversas partes do mundo. O CoL foi criado a partir da parceria do *Integrated Taxonomic Information System* – ITIS (ITIS, 2011) e do *Species 2000* (SPECIES2000, 2011), duas importantes iniciativas globais envolvidas com a criação de repositório de dados taxonômicos (CoL, 2011). Esses repositórios são considerados autoridades taxonômicas internacionais pela comunidade científica.

A consulta a esse catálogo pode ser feita por meio de uma aplicação web disponível no site do CoL (<http://www.catalogoflife.org>), por *web services*, *plug-ins* *web* ou por meio de aplicações *desktop* para os sistemas operacionais Windows, Linux ou MacOSX, as quais podem ser baixadas no *website* (<http://www.catalogueoflife.org/services>). O banco de dados do CoL também pode ser baixado (<http://www.catalogueoflife.org/services>) e usado em um banco de dados local.

#### **2.2.1.2. *Encyclopedia of Life - EoL***

O objetivo do EoL (<http://www.eol.org>) é aumentar a conscientização e a compreensão relacionadas à natureza por meio da produção e compartilhamento de conhecimento confiável sobre a biodiversidade em formatos digitais e de acesso livre. O EoL mantém um banco de dados que inclui informações taxonômicas, morfológicas, comportamentais, sobre hábitat, recursos multimídia e bibliográficos, etc. relacionados a espécies (EoL, 2011).

Como forma de compartilhamento dessas informações, o EoL disponibiliza uma *Application Programming Interface* – API (<http://eol.org/api>) que permite o uso das funcionalidades do *website* do EoL em outros SI ou ferramentas. Essa API permite, por meio de *web services*, realizar consultas para obter informações e imagens sobre táxons.

#### **2.2.2. Serviços de georeferenciamento de ocorrências**

Georeferenciamento pode ser definido como a atribuição de dados de coordenadas geoespaciais a um determinado objeto. O georeferenciamento pode ser realizado por meio de dispositivos de geoposicionamento ou (BioGeomancer, 2011) por meio de um processo de conversão de descrições de localizações para

informações geoespaciais legíveis computacionalmente por Sistemas de Informações Geográficas – SIG. A seguir são listados alguns recursos computacionais disponíveis na web que podem auxiliar o georeferenciamento de ocorrências de espécies.

#### **2.2.2.1. *BioGeomancer***

O Projeto BioGeomancer (<http://www.biogeomancer.org>) é resultado de uma colaboração entre especialistas em dados geoespaciais e em história natural. O principal objetivo do projeto é maximizar a qualidade e a quantidade de dados de biodiversidade georeferenciados, a fim de que esses dados sejam utilizados como suporte de pesquisas científicas, planejamentos, conservação e gestão da biodiversidade (BioGeomancer, 2011).

Dentre os resultados do projeto, está um *web service* (<http://bg.berkeley.edu:8080/ws>) que permite realizar o georeferenciamento a partir de uma descrição de uma localização em linguagem natural como “*5 miles north of São Paulo*”, por exemplo.

#### **2.2.2.2. *Google Maps APIs***

O *Google Maps* é um servidor de mapas que permite a visualização e a navegação em mapas interativos na web. O *Google Maps* possui um amplo conjunto de APIs (<http://code.google.com/apis/maps>) que permite a inclusão das funcionalidades do *Google Maps* em outras aplicações, websites e ferramentas (GoogleMaps, 2011). As APIs disponibilizadas pelo *Google Maps* são: *Google Earth API*, *Maps Javascript API*, *Maps Image APIs*, *web services* e *Maps API for Flash*.

#### **2.2.2.3. *GeoNames***

O banco de dados geográficos do *GeoNames* (<http://www.geonames.org>) cobre todos os países do planeta e contêm mais de oito milhões de nomes de lugares que estão disponíveis para serem baixados gratuitamente (*GeoNames*, 2011). Além de baixar essas informações, é possível acessá-las por meio de um

conjunto de *web services* disponíveis em <http://www.geonames.org/export/ws-overview.html>.

#### **2.2.2.1. GeoLocate**

O GeoLocate (<http://www.museum.tulane.edu/geolocate>) é uma plataforma para georeferenciamento de dados de coleções de história natural. O projeto GeoLocate coordena esforços para desenvolver *softwares* e serviços para traduzir descrições textuais de localidade, associados aos dados de biodiversidade, em coordenadas geográficas (GeoLocate, 2011). Como resultado do projeto, foi disponibilizado um conjunto de *web services* para auxiliar o georeferenciamento de dados de biodiversidade (<http://www.museum.tulane.edu/geolocate/developers>).

### **2.3. Dados de ocorrências de espécies**

As observações da natureza são fundamentais para a realização de estudos relacionados à ecologia e à conservação da biodiversidade. Informações relacionadas a essas observações são utilizadas como um importante recurso no desenvolvimento de planos de gestão da biodiversidade e na criação de políticas de conservação e uso sustentável do meio ambiente. Em muitos casos, para que complexas pesquisas sobre o meio ambiente e a biodiversidade possam ser realizadas, é necessária a disponibilidade de dados de observações de espécies em ampla escala geográfica e temporal (Chapman, 2005a). Para isso, diversas fontes de dados são utilizadas, para assim, integrar um banco de dados mais completo (Kelling, 2008).

Assim, para que haja interoperabilidade na integração de diferentes fontes de dados de ocorrências de espécies é necessário que haja uma padronização do esquema de metadados dos dados compartilhados. Um dos padrões mais utilizados para esse objetivo é o esquema de metadados DwC. Esse esquema é organizado em subconjuntos correlatos de elementos chamados, neste trabalho, de domínios de dados (Dalcin, 2005).

Dentre os domínios de dados disponíveis no DwC, três destacam-se em importância por serem essenciais para o uso em diversos domínios de aplicação, inclusive modelagem de distribuição de espécies biológicas:

- **Domínio de dados de localização:** Os dados desse domínio referem-se às informações geográficas das ocorrências. Entre os elementos desse domínio estão país, estado, município e localidade.
- **Domínio de dados geoespaciais:** Dados geoespaciais são relacionados ao georeferenciamento e detalham a localização da ocorrência. Os elementos desse domínio incluem latitude, longitude, altitude e *datum* geodésico.
- **Domínio de dados de taxonômicos:** Esse domínio comprehende informações sobre nomenclatura e hierarquia taxonômica de organismos. Reino, filo, classe, ordem, família, gênero e nome científico são exemplos de elementos que fazem parte desse domínio de dados.

Ambos os domínios de dados geoespaciais e de localização representam, de maneira diferente, o local aonde houve a ocorrência. Esses domínios são importantes componentes dos dados de ocorrências, pois são mandatórios para muitos estudos e modelos computacionais, como os modelos de distribuição de espécies (Chapman, 2005a). Esses modelos podem ser utilizados, por exemplo, para desenvolver cenários para uma espécie em relação às mudanças globais do meio ambiente (Hill *et al.*, 2010).

Taxonomia, ou Sistemática, é a teoria e a prática de classificação e nomenclatura de organismos (Chapman, 2005b; Dalcin, 2005). Classificação é o processo de criar e definir hierarquias sistemáticas de grupos de organismos de táxons conhecidos. A nomenclatura nesse contexto pode ser definida como atribuição de nomes únicos para cada grupo taxonômico (Dalcin, 2005). Todo organismo vivo conhecido faz parte de um táxon. Assim, toda ocorrência de espécie deve receber uma identificação taxonômica, ou seja, deve-se associar um nome de táxon à ocorrência. Dados de ocorrências sem táxons associados são inúteis em muitos estudos sobre biodiversidade (Chapman, 2005a). O domínio de dados de táxon, ou taxonômico, no DwC refere-se à classificação e nomenclatura de ocorrências de espécies.

Devido a importância desses dados, existem diversas soluções de SI utilizados para digitalizar dados de ocorrências de espécies, dos quais pode-se destacar o Specify 6<sup>1</sup>, o Brahms<sup>2</sup> e o SpeciesBase<sup>3</sup>.

## **2.4. Cadeia de produção de informações sobre ocorrências de espécies**

Segundo McGilvray (2008), problemas de QD podem ser causados principalmente por três elementos: usuários, processos ou sistemas. Portanto, para melhor entender e definir questões sobre a QD em SI sobre ocorrências de espécies é necessário compreender como esses elementos se relacionam e desempenham seus papéis nas atividades de criação, gestão, utilização e disseminação de dados sobre ocorrências de espécies.

### **2.4.1. Atores**

A Figura 1 apresenta um diagrama que ilustra a cadeia de produção de informações de ocorrências de espécies por meio do SI BDD. Esse diagrama está em concordância com uma representação do ciclo de informação taxonômica proposta por Dalcin (2005) e com uma representação do processo de descoberta e organização de informação sobre biodiversidade apresentado por Kelling (2008).

Essa cadeia de produção de informações sobre ocorrências de espécies é composta de seis atores:

- Produtor de dados;
- Taxonomista;
- Curador;
- Especialista em processamento de dados;
- Especialista em biodiversidade;
- Instituição.

---

<sup>1</sup> <http://specifysoftware.org/>

<sup>2</sup> <http://dps.plants.ox.ac.uk/bol/>

<sup>3</sup> <http://splink.cria.org.br/speciesbase?criaLANG=pt>

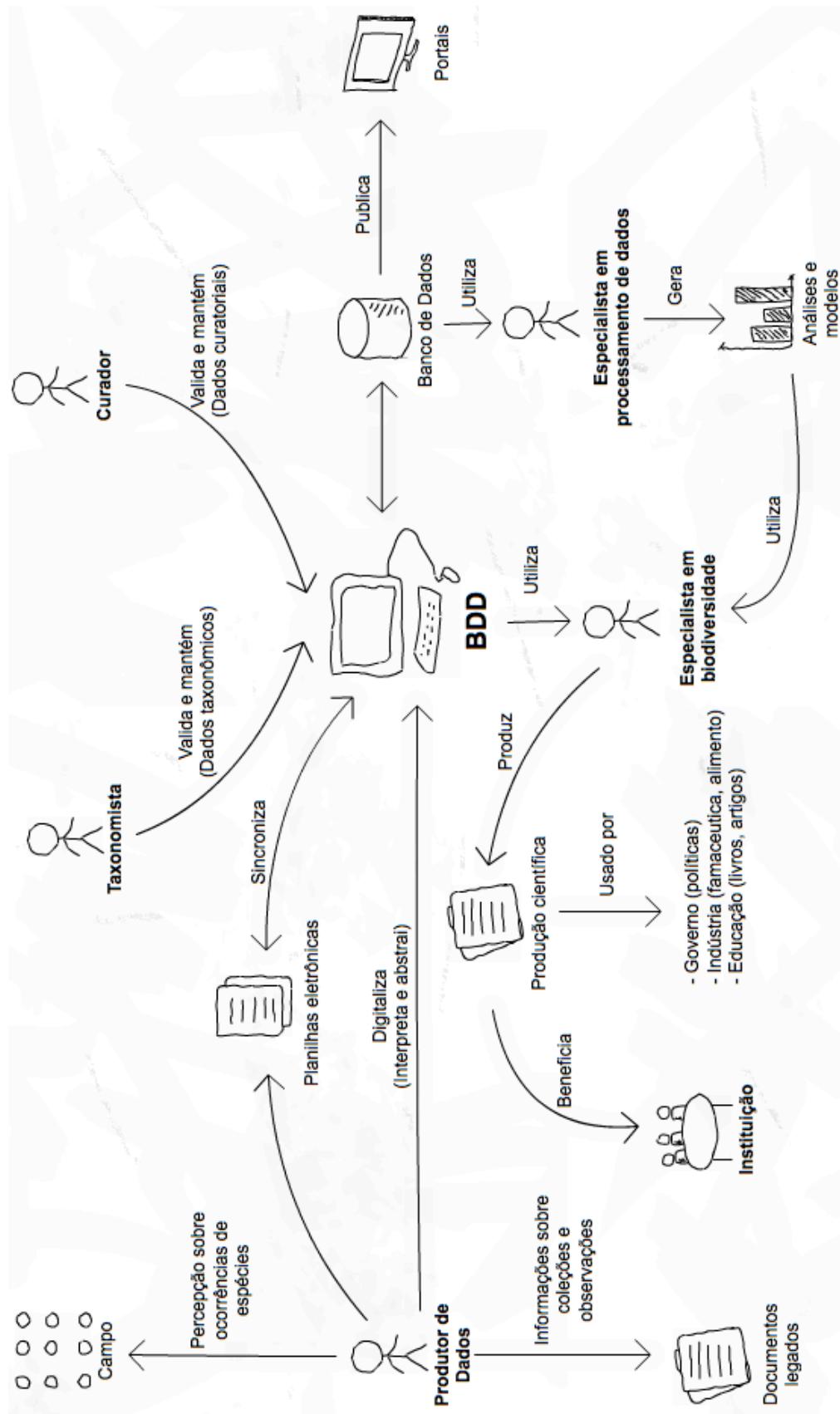


Figura 1 – Cadeia de produção de informações de ocorrências de espécies.

O processo de produção, gestão, utilização e disseminação de dados de ocorrências de espécies começa com os produtores de dados, que representam as pessoas que são responsáveis por coletar e digitalizar informações sobre ocorrências de espécies por meio do SI BDD ou de planilhas eletrônicas. Esses dados são, então, gerenciados e validados por taxonomistas e curadores. Os dados digitalizados são armazenados em um banco de dados que pode ser acessado por portais, como o do GBIF ou da IABIN, utilizando o protocolo TAPIR. O banco de dados também pode ser utilizado por especialistas em processamento de dados para gerar modelos e análises computacionais, os quais podem ser utilizados por especialistas em Biodiversidade, em conjunto com dados puros direto do BDD, para produzir conhecimento científico, o qual pode ser usado pela indústria, governo e na educação e, assim, beneficiar as Instituições mantenedoras.

Uma breve descrição das funções e das responsabilidades desses atores é apresentada a seguir.

#### **2.4.1.1. *Produtor de dados***

Os produtores de dados representam as pessoas responsáveis por coletar e digitalizar o insumo do SI: informações de ocorrências de espécies. Essa atividade normalmente é desenvolvida em duas partes, podendo, cada parte, ser executada por pessoas diferentes.

A primeira parte consiste em testemunhar o fato biológico, ou seja, observar ou coletar o organismo na natureza, e tomar nota desse acontecimento em meios não, necessariamente, digitais. Essa parte pode ser desenvolvidas em quatro etapas:

- **Percepção:** observação do organismo na natureza;
- **Interpretação:** entendimento das informações percebidas, baseado em conhecimentos adquiridos;
- **Abstração:** seleção de aspectos relevantes das informações interpretadas;
- **Anotação:** transcrição das informações abstraídas em meios não digitais.

A segunda parte está relacionada a digitalização das informações coletadas em campo. Essa parte também pode ser desenvolvida em quatro etapas:

- **Percepção**: observação de informações contidas nas anotações;
- **Interpretação**: entendimento das informações percebidas;
- **Abstração**: seleção de aspectos relevantes das informações interpretadas;
- **Digitalização**: transcrição das informações abstraídas em meios digitais.

Em cada etapa dessas atividades há a possibilidade de ocorrer erros ou anomalias que podem afetar a QD. Por exemplo, erros na etapa de interpretação podem causar baixa acurácia e inconsistência dos dados ou uma abstração inadequada pode levar a uma baixa completude de dados.

#### **2.4.1.2. *Taxonomista***

Os taxonomistas são especialistas em Taxonomia, ou Sistemática, e são responsáveis por realizar a identificação taxonômica dos organismos. Em relação ao SI, a sua função está relacionada a validação, correção e preenchimento de informações faltantes sobre a taxonomia das espécies observadas ou coletadas. O êxito da realização das tarefas desse ator é dependente do conhecimento adquirido e da experiência do especialista em relação a um determinado grupo taxonômico específico, como família ou gênero.

#### **2.4.1.3. *Curador***

Os curadores são responsáveis pelo zelo, manutenção e organização de coleções ou herbários. Esses usuários assumem uma função gerencial de supervisionar coleções de exemplares de organismos biológicos armazenados em museus, herbários e coleções biológicas de instituições. São também responsáveis pela gestão, manutenção e revisão dos dados relacionados aos espécimes, de modo a mantê-los consistentes com as amostras físicas armazenados na instituição.

#### **2.4.1.4. *Especialista em processamento de dados***

Esses especialistas são pessoas ligadas às áreas de Ciências Exatas com conhecimento em Computação, Engenharia ou Matemática. Esse ator representa pessoas responsáveis pela utilização de dados de ocorrências de espécies como insumo em métodos computacionais de análise inteligente de dados, na geração de modelos matemáticos ou computacionais e nas visualizações gráficas de um grande volume de dados, a fim de facilitar a interpretação dos dados e a descoberta de conhecimento.

#### **2.4.1.5. *Especialista em Biodiversidade***

Os especialistas em Biodiversidade são representados por pessoas que utilizam informações sobre ocorrências de espécies para gerar conhecimento e produção científica. A partir dos dados de ocorrências e de modelos e análises derivados desses dados, os especialistas em Biodiversidade podem realizar inferências que lhes permitirão responder questões sobre a biodiversidade, meio ambiente e sobre assuntos correlatos.

A partir desses estudos, poderão ser produzidos artigos científicos, livros, teses, relatórios, entre outras produções científicas. Esses produtos podem, então, ser utilizados pela indústria, pelo governo e pela própria academia para auxiliar a tomada de decisões estratégicas para o uso e a gestão sustentável da biodiversidade, além de servir como insumo para novas pesquisas.

#### **2.4.1.6. *Instituição***

As instituições representam as organizações responsáveis pelo fornecimento de recursos humanos e tecnológicos e de infraestrutura, necessários para manter e melhorar a produção científica.

### **2.5. Considerações finais do capítulo**

A IB desempenha um importante papel para o desenvolvimento sustentável do planeta. Para cumprir com os objetivos da IB, diversas organizações ao redor do

mundo têm unido esforços para, entre outras atividades, desenvolver padrões e ferramentas que auxiliem na tarefa de formar uma infraestrutura de compartilhamento, análise e uso de dados de biodiversidade, em escopo global. Um tipo particularmente importante de dados de biodiversidade são os dados de ocorrências de espécies, os quais são essenciais para o estudo relacionadas ao uso sustentável de recursos naturais. A produção e gestão desses dados é realizada por diversos atores especializados e faz uso de SI como o BDD. Para que esses dados possuam valor aceitável para serem usados com credibilidade e confiança, a sua qualidade deve ser avaliada e melhorada.

No Capítulo 3 é apresentada uma revisão da literatura que demonstra como a Avaliação e o Gerenciamento da QD podem ser realizados para medir e melhorar a QD.

### 3. QUALIDADE DE DADOS

De acordo com Crosby (1984, *apud* Wang et. al., 1993), é amplamente aceito que a qualidade pode ser definida como a conformidade com os requisitos. Isso implica que o conceito de qualidade muda à medida que os requisitos dos usuários mudam. De acordo com Rose (1994), para se definir um conceito de qualidade é necessário, além de entender quais são as reais exigências dos usuários, descobrir requisitos extras que não são esperados pelos usuários, mas que, quando apresentados, são percebidos como necessários.

Assim, para se definir um conceito de qualidade em um determinado contexto, é necessário entender e considerar os requisitos que afetam a satisfação dos usuários. Segundo o modelo de Kano, representado pela Figura 2, existem três tipos de requisitos, descritos a seguir (Mazur, 1993; Bolt & Mazur 1999):

- **Requisitos Normais (Normal Requirements)**: são requisitos declarados pelos usuários. A consideração desse tipo de requisito aumenta a satisfação do usuário na mesma proporção em que a desconsideração do requisito diminui a satisfação do usuário.
- **Requisitos Esperados (Expected Requirements)**: são requisitos tão básicos que normalmente os usuários não os declaram. A identificação e consideração de requisitos desse tipo, normalmente não aumenta a satisfação do usuário, contudo a desconsideração deles tende a diminuir muito a sua satisfação.
- **Requisitos Entusiasmantes (Exciting Requirements)**: são difíceis de identificar. Normalmente são requisitos implícitos e não declarados. Encontrar e considerar os requisitos desse tipo tende a aumentar muito a satisfação do usuário, pois excede a expectativa do usuário, contudo, a desconsideração deles não afeta a satisfação do usuário.

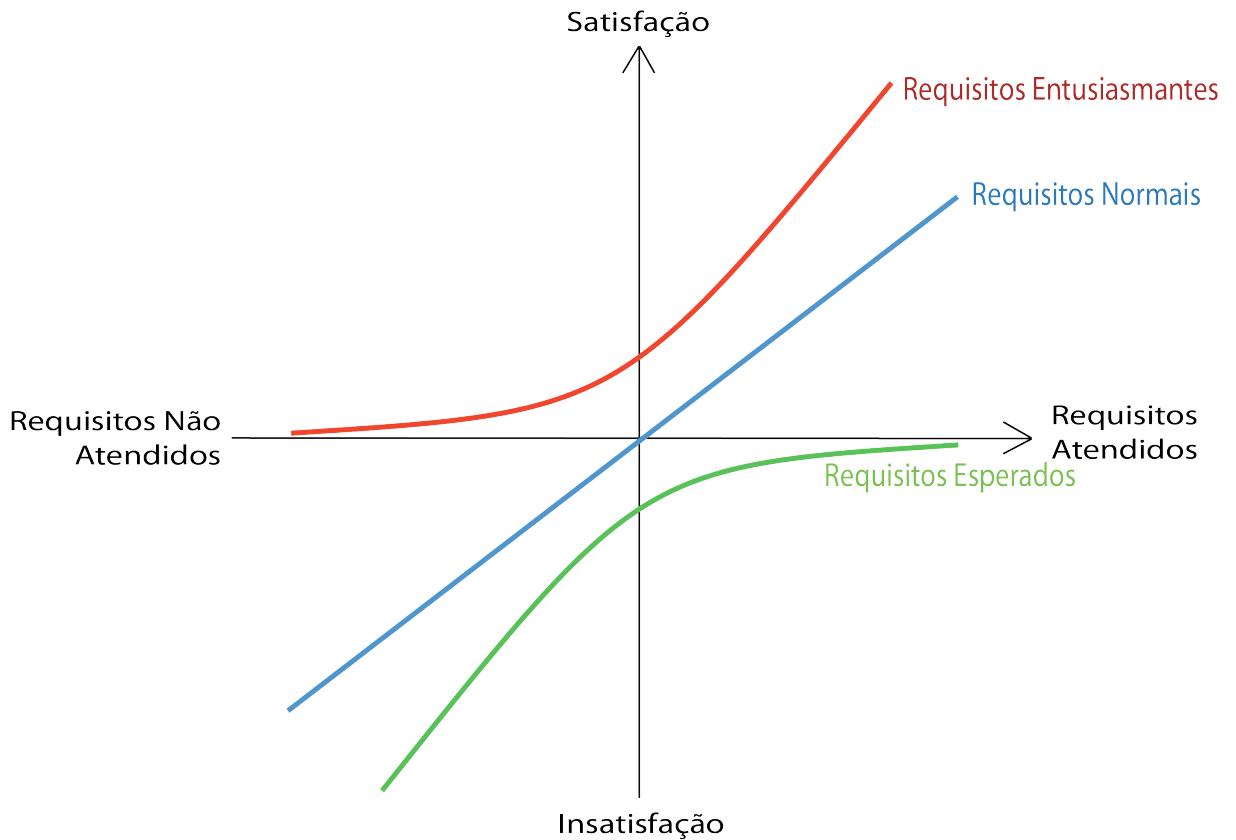


Figura 2 – Modelo de requisitos de Kano.

Baseado em Bolt & Mazur (1999).

De modo consonante, segundo Wang *et al.* (1993), esses conceitos de qualidade de produtos e serviços podem também ser aplicados ao conceito de QD, ou seja, a definição de QD depende da satisfação dos usuários em relação a um determinado conjunto de requisitos.

Uma definição frequentemente utilizada sobre QD é a de que dados de alta qualidade são dados adequados ao uso (Strong *et al.*, 1997); ou seja, os dados devem servir aos propósitos de quem os usa. De acordo com Wang *et al.* (1993), em geral, dados de alta qualidade são capazes de representar as condições do mundo real e de serem utilizados de maneira satisfatória pelos seus usuários.

Nesse sentido, em relação às várias definições, estudos e abordagens sobre QD, é consenso que para se definir um conceito de QD em um determinado domínio de aplicação, é necessário compreender o valor que os dados possuem ao serem utilizados para algum propósito (English, 1999 *apud* Dalcin, 2005). Strong *et al.* (1997) afirmam não ser possível definir um conceito de QD independentemente de

seus usuários. Nesse sentido, a qualidade pode ser definida como um conceito idiossincrático, ou seja, é definida em última instância por um indivíduo ou um grupo de indivíduos.

Portanto, dados de qualidade pode ser definidos como dados que são capazes de representar adequadamente as condições do “mundo real” e que atendem a um conjunto de requisitos relacionados a um contexto.

### **3.1. Abordagens de pesquisas sobre QD**

Na revisão da literatura sobre QD realizada por Ge & Helfert (2007) é sugerido que pesquisas em QD podem ser conduzidas conforme três abordagens: Avaliação, Gerenciamento e Contextualização da QD.

A Avaliação da QD pode ser definida como o processo de atribuir um valor numérico ou categórico a um aspecto da QD em um determinado contexto. A abordagem de Gerenciamento da QD está relacionada à aplicação de métodos e técnicas para aprimorar a qualidade, com base na Avaliação da QD. A abordagem de Contextualização está relacionada ao impacto da QD na organização gestora dos dados.

Assim, a seguir é apresentada uma revisão da literatura relacionada à Avaliação e ao Gerenciamento da QD, a qual pode auxiliar na concepção de metodologias de medição e aprimoramento da QD em um domínio de aplicação específico.

#### **3.1.1. Avaliação da QD**

Ge & Helfert (2007) afirmam que na literatura existem três componentes chaves utilizados na abordagem de Avaliação da QD:

- problemas de QD;
- dimensões de QD;
- metodologias de Avaliação da QD.

Esses componentes podem ser organizados em três camadas inter-relacionadas: problema, dimensão e metodologia, conforme a Figura 3.

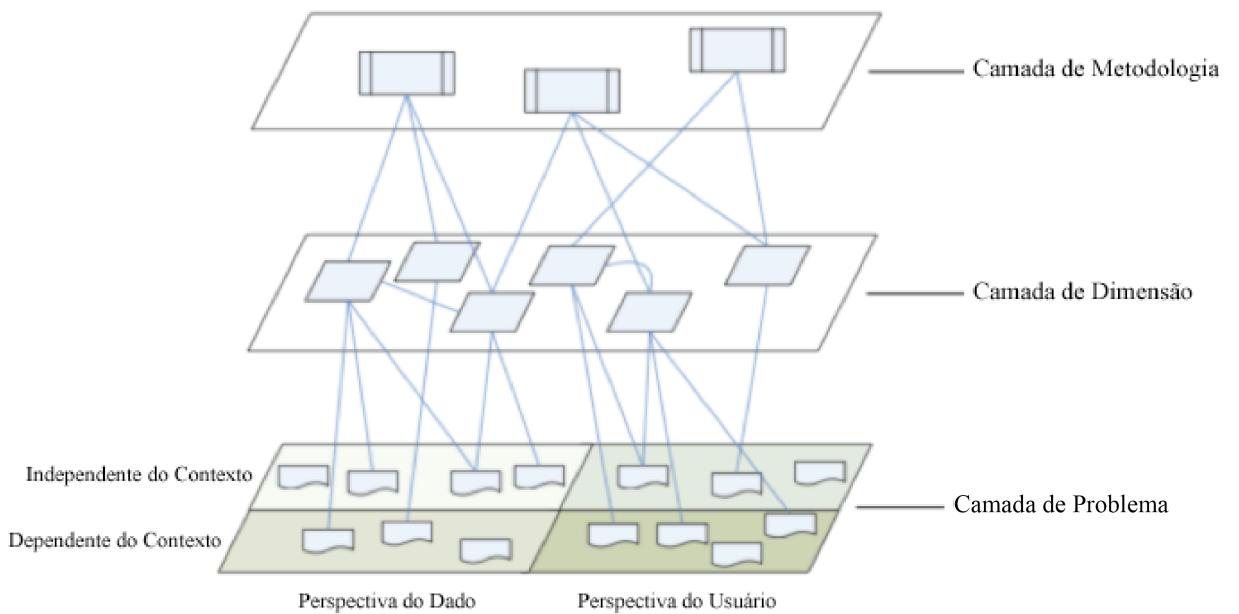


Figura 3 – Um *framework* para Avaliação da QD.

Adaptado de Ge & Helfert (2007).

Portanto, as pesquisas sobre Avaliação da QD podem ser conduzidas por meio do estudo da identificação, definição e relação dos elementos dessas camadas.

### 3.1.1.1. *Camada de problema*

Os elementos dessa camada representam os problemas mensuráveis de QD. Esses problemas, neste trabalho, podem ser definidos como classes de instâncias de erros, ou seja, são generalizações de erros que afetam a QD. Os problemas de QD podem ser classificados de acordo com o “contexto” e de acordo com a “perspectiva” (Chen *et al.*, 2009; Ge & Helfert, 2007), conforme exemplificado no Quadro 1, proposto por Ge & Helfert (2007).

Baseado nessa proposta de classificação, os problemas de QD podem ser, portanto, dependentes ou independentes de contexto. Problemas independentes do contexto são problemas que podem ser aplicados a qualquer conjunto de dados, sem considerar as regras de negócio. Enquanto que problemas dependentes do contexto são associados às regras de negócio (Ge & Helfert, 2007). Assim, para

realizar a identificação de problemas dependentes de contextos é necessário compreender as regras de negócio do domínio de aplicação.

	<b>Perspectiva Intrínseca aos Dados</b>	<b>Perspectiva do Consumidor</b>
<b>Independente do Contexto</b>	<ul style="list-style-type: none"> <li>▪ Erros de digitação</li> <li>▪ Dados faltantes</li> <li>▪ Dados duplicados</li> <li>▪ Valor incorreto</li> <li>▪ Formato do dado inconsistente</li> <li>▪ Dados desatualizados</li> <li>▪ Violação de sintaxe</li> <li>▪ Violação da restrição de integridade</li> <li>▪ Formatação de texto</li> </ul>	<ul style="list-style-type: none"> <li>▪ A informação está inacessível</li> <li>▪ A informação está insegura</li> <li>▪ A informação é dificilmente recuperável</li> <li>▪ A informação é difícil de agregar</li> <li>▪ Erros na transformação da informação</li> </ul>
<b>Dependente do Contexto</b>	<ul style="list-style-type: none"> <li>▪ Violação do domínio da restrição</li> <li>▪ Violação das regras de negócio da organização</li> <li>▪ Violação dos relacionamentos da companhia e do governo</li> <li>▪ Violações das restrições do banco de dados</li> </ul>	<ul style="list-style-type: none"> <li>▪ A informação não é baseada em fatos</li> <li>▪ A informação é de credibilidade questionável</li> <li>▪ A informação apresenta uma visão imparcial</li> <li>▪ A informação é irrelevante para o trabalho</li> <li>▪ A informação possui significados diferentes</li> <li>▪ A informação está incompleta</li> <li>▪ A informação está representada de maneira compacta</li> <li>▪ A informação é difícil de manipular</li> <li>▪ A informação é difícil de entender</li> </ul>

Quadro 1 – Classificação de problemas de QD (Ge & Helfert, 2007).

De acordo com Redman (2000) uma informação é considerada de alta qualidade se ela estiver livre de defeitos e possuir as características desejáveis. Nesse sentido, os problemas de QD também podem ser classificados em relação às seguintes perspectivas: problemas de qualidade intrínsecos aos dados e problemas de qualidade relacionados às expectativas do usuário.

Os problemas de QD intrínsecos aos dados normalmente podem ser resolvidos por processos autônomos como algoritmos de *data cleansing* e regras de *data mining*, por exemplo. Os problemas de QD sob a perspectiva do consumidor, por outro lado, normalmente não admitem o uso de processos autônomos para identificação e resolução de problemas. Análise do negócio e a reengenharia de

processos são exemplos de métodos que podem ser utilizados para identificar e resolver problemas sob a perspectiva do usuário (Ge & Helfert, 2007).

Portanto, a identificação de problemas de QD pode ser realizada por meio da observação das características intrínsecas aos dados, das regras de negócio e das necessidades dos consumidores dos dados.

### **3.1.1.2. Camada de dimensão**

Diversos autores afirmam que a QD é um conceito multidimensional (Dalcin, 2005; McGilvray, 2008; Wang *et al.*, 1995; Strong *et al.*, 1997). Uma dimensão de QD pode ser definida como um atributo que representa um aspecto da QD (Wang & Strong, 1996). As dimensões de QD podem possuir diferentes graus de relevância e diferentes significados dependendo do contexto. Portanto, é um fator importante a identificação de quais dimensões de QD devem ser utilizadas no Avaliação da QD e a definição do significado de cada dimensão nos diferentes contextos do domínio de aplicação (Dalcin, 2005; McGilvray, 2008).

Segundo Wang & Strong (1996), três abordagens podem ser utilizadas em estudos sobre QD: intuitiva, teórica e empírica. Essas abordagens podem ser utilizadas para realizar a identificação e a definição das dimensões de QD, conforme descrito a seguir (Wang & Strong, 1996; Ge & Helfert, 2007).

#### ***Identificação de dimensões***

A identificação de dimensões consiste em selecionar as dimensões de QD mais relevantes em um determinado contexto. Baseada nas abordagens propostas por Wang & Strong (1996), Ge & Helfert (2007) apresentaram três métodos de identificação de dimensões de QD, são elas: identificação intuitiva, teórica e empírica.

A identificação intuitiva de dimensões de QD apoia-se na experiência do pesquisador e no contexto de aplicação. Essa abordagem se baseia no conhecimento adquirido do pesquisador sobre o domínio de aplicação.

A abordagem teórica para a identificação baseia-se na observação das deficiências dos dados, causadas durante a produção de dados. Um exemplo de

utilização dessa abordagem é a observação da inconsistência entre o mundo real e o SI (Wang & Strong, 1996).

Na abordagem empírica, a identificação das dimensões de QD é realizada por meio de análises e coletas de atributos que determinem a adequação ao uso de dados com foco nos usuários (Ge & Helfert, 2007).

Após a identificação das dimensões de QD é necessário definir os significados de cada uma dessas dimensões.

### ***Definição de dimensões***

As dimensões de QD podem possuir diferentes significados ou se manifestarem de diferentes maneiras em relação aos domínios de dados. Por exemplo, a definição de completude em dados taxonômicos é diferente da definição de completude em dados geoespaciais.

Segundo Ge & Helfert (2007), a definição de dimensões de QD pode ser realizadas de acordo com três perspectivas, conforme ilustrado na Figura 4.



Figura 4 – Definição de dimensões de QD a partir de diferentes aspectos.

Adaptado de Ge & Helfert (2007).

Com a abordagem intuitiva, a definição de dimensões de QD é realizada a partir da perspectiva intrínseca aos dados. De acordo com essa abordagem, a dimensão de “completude” pode ser definida, por exemplo, como o preenchimento de todos os valores de uma determinada variável.

A abordagem teórica procura definir as dimensões de QD a partir da perspectiva do mundo real. Por exemplo, Wand & Wang (1996) definiram “completude” como a capacidade de um SI representar todos os estados significativos de sua representação do mundo real.

A abordagem empírica é utilizada para definir dimensões a partir das perspectivas dos usuários dos dados. Por exemplo, Wang & Strong (1996) definiram “completude” como a medida para o qual os dados sejam amplos e detalhados o suficiente para realizar uma determinada tarefa.

### **3.1.1.3. Camada de metodologia de avaliação**

Na camada de metodologia de avaliação, são propostos métodos para medir as dimensões de QD em relação ao contexto. Ou seja, nessa camada procura-se identificar como a QD, em cada dimensão, pode ser avaliada em relação a um domínio de dados.

Pipino *et al.* (2002) categorizaram a Avaliação da QD em objetiva e subjetiva. A avaliação objetiva identifica os problemas de QD de um conjunto de dados, e busca medir o quanto a informação está em concordância com a especificação de qualidade. A avaliação subjetiva da QD reflete as necessidades e expectativas dos usuários, e busca medir o quanto as informações estão adequadas para o uso (Ge & Helfert, 2007).

### **3.1.2. Gerenciamento da QD**

Segundo Dalcin (2005), a maioria dos especialistas em QD concordam que os princípios gerais do gerenciamento da qualidade de produtos podem também ser aplicados ao Gerenciamento da QD. Isso sugere que pode haver duas abordagens básicas para melhorar da QD: a prevenção a erros e a detecção e correção de erros (Embry, 2001 apud Dalcin, 2005; Chapman, 2005b).

Prevenção a erros é considerada superior à detecção e correção de erros, uma vez que a detecção e a correção é uma abordagem dispendiosa e não garante o total sucesso do procedimento (Dalcin, 2005; Chapman, 2005b). Contudo, não importa o quão eficiente seja o processo de digitalização, os dados estão inherentemente sujeitos a erros, e, portanto, a abordagem de detecção e correção de

erros não pode ser ignorada (Chapman, 2005c). Nesse sentido, a detecção de erros, validação e limpeza de dados tem um papel essencial, principalmente em dados legados, como por exemplo, dados de museus e herbários coletados há mais de 300 anos (Chapman, 2005c).

### **3.2. Considerações finais do capítulo**

A QD é um conceito idiossincrático e portanto não pode ser definida independentemente do domínio de aplicação. Portanto, pesquisas relacionadas à QD acarretam no estudo dos domínios de aplicação e de dados. A compreensão desses domínios, seus significados e como eles são utilizados é essencial para se estabelecer um conceito de QD e para avaliar o que são dados de qualidade e o que não são. Para esse propósito, um estudo sobre a Avaliação da QD pode ser realizado, o qual consiste em realizar a identificação, definição e inter-relação de elementos de três camadas: problema, dimensão e metodologia. Com base na Avaliação da QD é possível identificar e implementar recursos para a melhoria da QD, ou seja, realizar o Gerenciamento da QD. A identificação desses recursos de melhoria pode ser realizado baseado em duas abordagens: prevenção ou detecção e correção. Portanto, pesquisas em QD podem ser realizadas por meio de estudos sobre: o contexto (domínios de aplicação e de dados), a Avaliação da QD (conceito e medição de QD) e o Gerenciamento da QD (melhoria da QD).

## **4. MATERIAIS E MÉTODOS**

O método utilizado para atingir o objetivo deste trabalho foi organizado em cinco etapas principais:

- Estudo do domínio de aplicação;
- Definição de escopo;
- Estudo sobre Avaliação da QD;
- Estudo sobre Gerenciamento da QD;
- Estudo de caso de aplicação dos estudos de QD:
  - Análise e desenvolvimento de um SI;
  - Desenvolvimento de ferramentas de QD.

Essas etapas foram desenvolvidas de maneira iterativa e incremental, conforme descritas a seguir.

### **4.1. Estudo do domínio de aplicação**

Nesta etapa foi realizada uma ampla revisão da literatura sobre o domínio de aplicação, ou seja, sobre dados de ocorrências de espécies no contexto de SI. Para isso, foram estudados assuntos relacionados à IB, definições, importância e utilidade de dados de ocorrências de espécies, domínios de dados de ocorrências de espécies, padrões e iniciativas de padronização de informações sobre biodiversidade, ferramentas e protocolos relacionadas à IB, entre outros assuntos correlatos.

Também foi realizada uma análise dos atores envolvidos no processo de produção, gestão, utilização e disseminação de dados de ocorrências de espécies, por meio do SI BDD. Esse estudo foi descrito no Capítulo 2.

### **4.2. Definição do escopo**

Dados de ocorrências de espécies podem ser organizados em vários domínios de dados. O esquema de metadados DwC, utilizado neste trabalho para

fazer a modelagem dos dados de ocorrências de espécies, possui sete domínios de dados padrões. Segundo Dalcin (2005), a QD deve aplicada separadamente a cada domínio de dados e em relação a cada dimensão de QD.

Assim, o objetivo desta etapa foi realizar um estudo para identificar os domínios de dados mais relevantes e, assim, torna-los escopo do trabalho. Portanto, baseado no estudo apresentado na Seção 2.3 e na afirmação de Chapman (2005b) que diz que “erros em posições geoespaciais (georeferenciamento) e em identificações taxonômicas são duas das maiores causas de erros em ocorrências de espécies”, os domínios de dados de localização, geoespaciais e taxonômicos foram identificados como os domínios mais relevantes e, portanto, foram definidos como escopo deste trabalho.

### **4.3. Estudo sobre a Avaliação da QD**

Para melhorar a QD em um determinado domínio de dados é necessário compreender o que é QD nesse domínio e quais são os fatores que fazem essa qualidade variar. Portanto, para se definir um conceito de QD em relação aos domínio de dados identificados, foi necessário identificar os problemas de QD e como esses problemas se manifestam em cada domínio de dados. Também foi preciso identificar dimensões de QD relevantes e compreender os seus significados em relação aos domínios de dados. Posteriormente, foi realizada uma análise sobre como os elementos identificados e definidos (os problemas e as dimensões) se relacionam entre si. Com base nessa análise foi proposta uma metodologia que permitisse avaliar a QD de maneira mais objetiva em relação a cada domínio de dados. Esse processo se deu em cinco etapas secundárias, conforme descrito a seguir.

#### **4.3.1. Identificar problemas de QD**

A identificação dos problemas QD foi realizada por meio da observação das características intrínsecas aos dados de ocorrências de espécies, das regras de negócio do domínio de aplicação e das necessidades dos consumidores dos dados, conforme descrito na Subseção 3.1.1.1. Essa identificação foi realizada com base

em uma revisão da literatura sobre QD e sobre dados de biodiversidade, os quais permitiram listar um conjunto de erros comuns que afetam a QD.

Nesta etapa, portanto, foram identificados padrões de erros comuns em dados de biodiversidade e que podem ser aplicados aos domínios de dados de ocorrências de espécies.

#### 4.3.2. Definir problemas de QD

A definição de problemas de QD foi realizada por meio de uma análise de como os problemas identificados se manifestam em cada domínio de dados, conforme ilustrado na Figura 5.

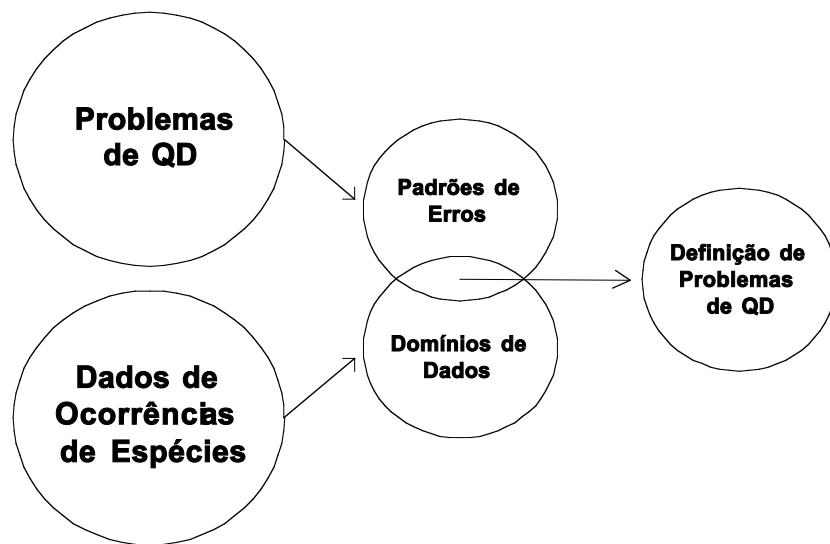


Figura 5 - Definição de problemas de QD.

#### 4.3.3. Identificar dimensões de QD

Para realizar a identificação de dimensões de QD relevantes no contexto de dados de ocorrências de espécies e do SI BDD, foram utilizadas três abordagens, apresentadas na Subseção 3.1.1.2 (Ge & Helfert, 2007), conforme descritas abaixo:

- **Identificação intuitiva:** com essa abordagem, a identificação das dimensões de QD foi realizada com base na experiência de

pesquisadores envolvidos com pesquisas em Informática para Biodiversidade. Foi também realizada uma ampla revisão bibliográfica sobre os assuntos de QD e IB, para auxiliar na identificação das dimensões de QD relevantes no domínio de aplicação.

- **Identificação teórica:** usando essa abordagem, a identificação das dimensões foi realizada por meio de observações dos bancos de dados do SI BDD e do seu antecessor *Pollinator Data Digitizer* – PDD, a fim de encontrar deficiências e padrões de erros que pudessem auxiliar na identificação de dimensões de QD relevantes.
- **Identificação empírica:** a identificação empírica foi realizada por meio de interações com biólogos. O objetivo dessas interações foi compreender as necessidades e expectativas dos usuários do SI. Essas interações ocorreram por meio de discussões e de debates com os biólogos em eventos, como o IABIN-PTN *Training Workshop* em 2010<sup>4</sup> e o TDWG *Annual Conference* em 2011<sup>5</sup>, e com os usuários do BDD e do PDD. Também foram utilizados meios digitais, como vídeo conferências, e-mails e *chats*, como recurso para a interação com biólogos e outros profissionais envolvidos com IB para obter informações relacionadas à QD e ao SI BDD.

Cada dimensão de QD identificada pode possuir significados diferentes e, portanto, é necessário definir as dimensões em relação aos domínios de dados, conforme descrito a seguir.

#### 4.3.4. Definir dimensões de QD

A principal abordagem utilizada para definir as dimensões de QD foi a abordagem empírica. Para entender o significado de cada uma das dimensões de QD foi realizada uma revisão bibliográfica sobre a utilidade dos dados de ocorrências de espécies em pesquisas no campo da IB. Com base nesse estudo

---

<sup>4</sup> Treinamento realizado para 40 pessoas do Brasil, Canadá, Chile, Colômbia, El Salvador, Equador, Estados Unidos, Guatemala, México, Paraguai e Peru, durante os dias 26 e 27 de Julho de 2010 em Ribeirão Preto – SP.

<sup>5</sup> Evento realizado em Nova Orleans nos Estados Unidos durante os dias 16 a 21 de Outubro de 2011.

foram definidas as dimensões de QD de acordo com os domínios de dados, conforme ilustrado na Figura 6.

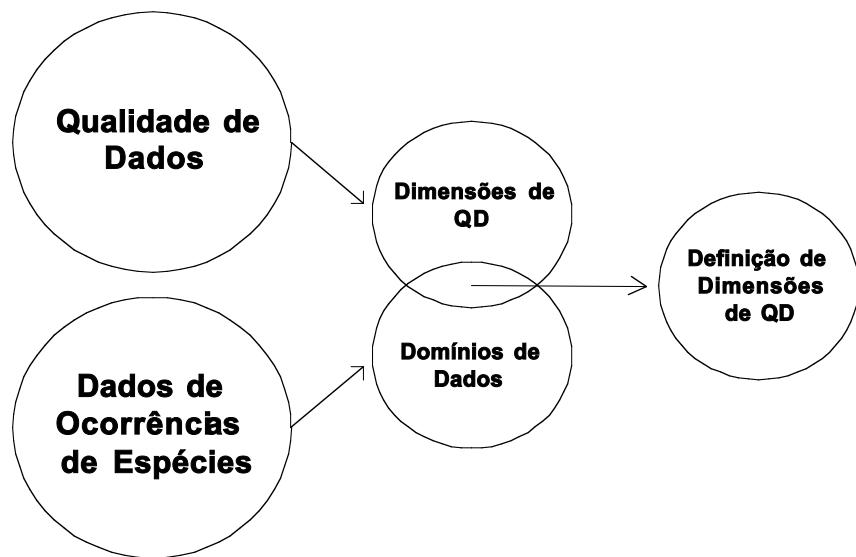


Figura 6 - Definição de dimensões de QD.

Baseada em Dalcin (2005).

#### 4.3.5. Definição de uma proposta de metodologia de avaliação da QD

Esta etapa está relacionada à camada de metodologia de avaliação do *framework* de Avaliação da QD, proposto por Ge & Helfert (2007). Nesta etapa são definido modos de se medir ou avaliar a QD em cada dimensão e de acordo com cada domínio de dados. Para isso foi realizada uma análise para identificar quais problemas afetam quais dimensões e em quais domínios de dados, conforme ilustrado na Figura 7. Assim, com base na presença de determinados problemas em determinados domínios de dados é possível inferir em quais dimensões a QD é degradada, e quais problemas devem ser reduzidos para melhorar a QD em determinadas dimensões.

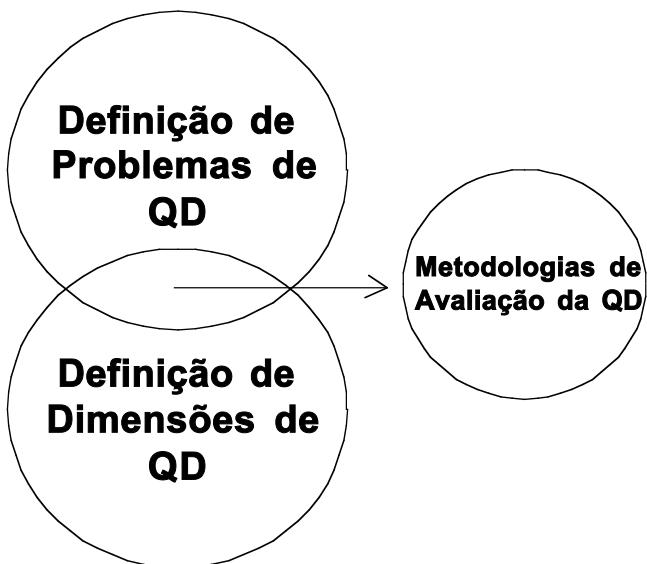


Figura 7 - Metodologia de Avaliação da QD.

#### **4.4. Estudo sobre o Gerenciamento da QD**

Para melhorar a QD, a abordagem de Gerenciamento da QD de prevenção a erros foi utilizada. Segundo Dalcin (2005), a prevenção é uma abordagem superior em relação à abordagem de detecção e correção de erros.

Portanto, baseado na metodologia de Avaliação da QD, definida neste trabalho, nesta etapa foi identificado um conjunto de recursos computacionais que, se integrados a um SI de digitalização de dados de ocorrências de espécies, pode provocar uma redução de erros durante o processo de digitalização, melhorando assim a QD.

#### **4.5. Estudo de caso de aplicação dos estudos de QD**

Nesta etapa, foram utilizados os resultados dos estudos sobre QD para projetar e desenvolver ferramentas de QD. Essas ferramentas foram implementadas para serem integradas a um SI de digitalização de dados de ocorrências de espécies em desenvolvimento. Para que esse SI pudesse suportar as novas ferramentas, foi necessário realizar uma análise do SI e readequá-lo quanto à arquitetura de software, banco de dados, Interface Humano-Máquina – IHM e codificação.

#### **4.5.1. Análise e desenvolvimento de um SI**

Os recursos identificados no estudo sobre Gerenciamento da QD foram implementados no SI BDD. Esse sistema é uma evolução do PDD, o qual foi projetado para a digitalização de dados de ocorrências de polinizadores. O BDD, na versão Beta, foi inicialmente implementado com base na antiga versão do DwC, o DwC 1.4.

Para a versão do BDD utilizada neste trabalho, o banco de dados do BDD passou por uma reestruturação para que ficasse adequada à nova versão do DwC (submetida ao TDWG dia 12 de Fevereiro de 2009 e modificada dia 08 de Outubro de 2009), adotado oficialmente como um padrão pelo TDWG.

Nesta etapa, também foi realizado uma análise de requisitos sob o ponto de vista dos usuários e uma restruturação do código e da arquitetura de *software*, para tornar o SI mais modular, manutenível, escalável e com uma melhor usabilidade. Essas mudanças foram necessárias para que o SI BDD pudesse suportar a implementação das novas ferramentas de QD e para que o SI fosse mais atrativo para os usuários. O SI e a análise de requisitos são descritos na Subseção 5.2.1.

#### **4.5.2. Desenvolvimento de ferramentas de QD**

As ferramentas de QD desenvolvidas nesta etapa (apresentadas na Subseção 5.2.2) foram baseadas nos recursos computacionais identificados no estudo sobre Gerenciamento da QD. Projetadas para serem acopladas ao formulário de cadastro de ocorrências de espécies do BDD, duas ferramentas *web* foram implementadas. Essas ferramentas foram projetadas com o objetivo de evitar que, inadvertidamente, os usuários cometessem erros durante a digitalização dos dados taxonômicos, geoespaciais e de localização de ocorrências de espécies.

## **5. RESULTADOS**

Neste capítulo são apresentados os resultados obtidos durante a pesquisa. Os resultados foram organizados em duas partes: estudo sobre QD de ocorrências de espécies e estudo de caso de aplicação desse estudos sobre QD em um SI de digitalização de ocorrências de espécies.

A primeira parte (Seção 5.1) apresenta um estudo sobre Avaliação e Gerenciamento da QD em SI sobre ocorrências de espécies. Nessa parte foi aplicada uma metodologia de Avaliação da QD para identificar problemas e dimensões de QD no contexto de dados de ocorrências de espécies a fim de propor uma forma de avaliar a qualidade desses dados. Com base nessa avaliação, no estudo sobre Gerenciamento da QD, foram identificados recursos computacionais que, se integrados a um SI, podem proporcionar uma melhora da QD por meio da prevenção a erros.

Na segunda parte dos Resultados (Seção 5.2) é apresentado o estudo de caso de aplicação do estudo de QD no SI BDD. Esse SI foi utilizado para implementar os recursos computacionais identificados no estudo sobre Gerenciamento da QD deste trabalho. Devido à abordagem de gerenciamento adotada, a abordagem de prevenção a erros, foi necessário realizar uma análise de requisitos do SI com foco nos usuários, a fim de identificar características do SI que possam melhorar a aceitação dos usuários ao sistema e às ferramentas de QD nele implementadas. Na Subseção 5.2.2, são descritas duas ferramentas de QD desenvolvidas e integradas ao SI BDD. Essas duas ferramentas implementam os recursos identificados no estudo sobre Gerenciamento da QD.

### **5.1. Estudo sobre QD de ocorrências de espécies**

Nesta seção são apresentados os resultados relacionados à QD aplicados a dados de ocorrências de espécies. Na primeira subseção é descrito o estudo sobre Avaliação da QD, no qual são identificados os problemas de QD e como esses problemas se manifestam em cada domínio de dados. Também são apresentadas as dimensões de QD identificadas e suas definições em relação a cada domínio de dados. Ainda na primeira seção, é descrita a metodologia de avaliação, que

consistiu em realizar uma análise para identificar, em relação a cada domínio de dados, quais problemas podem degradada a qualidade de cada dimensão de QD.

Com base nesse estudo da Avaliação da QD, na segunda subseção é apresentado um estudo sobre Gerenciamento da QD de ocorrências de espécies, o qual lista um conjunto de recursos computacionais que, se implementados em um SI, podem reduzir a ocorrência de problemas de QD por meio da prevenção a erros durante a digitalização.

### **5.1.1. Avaliação da QD**

Nesta subseção é apresentada a identificação e a definição de problemas e de dimensões de QD em relação os domínios de dados de localização, geoespaciais e taxonômicos. Com base nessas identificações e definições é apresentada uma análise sobre como esse elementos (problemas, dimensões e domínios de dados) se relacionam.

#### **5.1.1.1. Camada de problema**

Baseando-se em English (1999), Dalcin (2005) realizou um estudo sobre padrões de erros em dados sobre biodiversidade. Esses padrões de erros foram utilizados como base para a identificação de problemas de QD no contexto de ocorrências de espécies. Assim, os problemas de QD identificados neste trabalho são:

- ***Domain value redundancy*** (Redundância do valor de domínio): ocorre quando os valores dos dados não são padronizados ou são sinônimos. Ou seja, quando dois ou mais valores diferentes representam a mesma coisa no mundo real.
- ***Missing data value*** (Valor do dado faltante): ocorre quando há a ausência de dados necessários. Isso inclui campos obrigatórios e não obrigatórios, mas que são necessários para a realização de determinadas tarefas.
- ***Incorrect data values*** (Valores de dados incorretos): esses erros podem ser causados pela transposição de caracteres no momento da

digitação, por inserção de dados em campos incorretos, pela não compreensão do significado da informação ou, ainda, pela obrigatoriedade da inserção de algum dado que no momento não é conhecido.

- **Nonatomic data values** (Valores de dados não atômicos): ocorre quando um dado possui múltiplos valores, quando deveria possuir um único valor atômico.
- **Domain schizophrenia** (Esquizofrenia de domínio): ocorre quando campos são interpretados e utilizados de diferentes maneiras, dependendo do contexto.
- **Duplicate occurrences** (Ocorrências duplicadas): ocorre quando múltiplos registros com o mesmo valor representam uma única entidade no mundo real.
- **Inconsistent data values** (Valores de dados inconsistentes): as inconsistências podem ocorrer devido à heterogeneidade de padrões e de procedimentos adotados por diferentes instituições, coleções ou indivíduos. Esses erros são caracterizados por contradições em informações.
- **Information quality contamination** (Contaminação da qualidade da informação): a contaminação ocorre ao se utilizar dados incorretos combinados a dados corretos para a produção de novos dados.

Esses problemas podem ser manifestos de distintas maneiras em relação a cada domínio de dados, conforme descrito a seguir (Dalcin, 2005).

### ***Definição de problemas no domínio de dados de localização***

No domínio de dados de localização, os problemas de *domain value redundancy* podem estar relacionados ao idioma em que as informações foram digitalizadas. Por exemplo, Brasil (português) e *Brazil* (inglês); os dois são corretos e referem-se à mesma entidade no mundo real, contudo, são dados distintos. Os problemas de *duplicate occurrences* normalmente ocorrem quando não há uma restrição de unicidade em entidades de bancos de dados relacionais (Group, 2005).

O não uso dessa restrição permite, por exemplo, a inserção de dois ou mais registros com chaves primárias distintas, mas com os demais valores da entidade idênticos. Um exemplo desse erro seria o cadastro de um registro na entidade *country* (*ID, COUNTRY\_NAME*) com os valores “*ID = 1, COUNTRY\_NAME = Brazil*” e outro registro com os valores “*ID = 3, COUNTRY\_NAME = Brazil*”, em que o atributo “*ID*” é chave primária.

*Incorrect data values* no domínio de dados de localização são comumente causados por erros de digitação. *Nonatomic data values* podem ocorrer, por exemplo, quando no campo da cidade é digitado “*New York, NY*”, ou seja, nome da cidade e código do estado. *Information quality contamination* ocorre quando um dado com erro é reutilizado para produzir novos dados, por exemplo, reutilizar o nome da cidade “*New York, NY*” para cadastrar um novo registro.

### ***Definição de problemas no domínio de dados geoespaciais***

No domínio de dados geoespaciais o problema de *missing data value* está fortemente relacionado com os campos de latitude e longitude. A ausência de um desses valores, normalmente, tem o mesmo efeito da ausência de ambos os valores, visto que os dois valores em conjunto representam as coordenadas geoespaciais.

Erros de digitação também são comuns nesse domínio de dados. A transposição da vírgula ou a ausência de um sinal de menos (quando deveria haver) no campo de latitude ou de longitude decimal, por exemplo, podem ser considerados problemas de *incorrect data values*.

A inserção da latitude e da longitude em um mesmo campo, por exemplo, “Latitude: -23.834, -59.984”, é um problema de *nonatomic data values*. Problemas de *domain schizophrenia* e de *inconsistent data value* podem ocorrer quando coordenadas geoespaciais são preenchidas em formato de graus (minutos, segundos) em campos que deveriam ser preenchidos em formato decimal.

### ***Definição de problemas no domínio de dados taxonômicos***

No domínio de dados de táxon, os problemas de *domain value redundancy* e *duplicate values* podem ocorrer devido ao fato de a nomenclatura dos táxons

poderem mudar com o tempo, e assim sinônimos surgirem. Por exemplo, um filo pode receber o nome de Magnoliófita, Magnoliophyta ou Angiosperma. Essas três nomenclaturas são sinônimos e representam a mesma entidade no mundo real.

O problema de *missing value* é muito comum em táxons mais específicos da hierarquia taxonômica, como o nome científico da espécie ou o epíteto específico. Isso ocorre porque a identificação nos níveis mais específicos da hierarquia taxonômica pode ser uma tarefa mais difícil de ser realizada, pois essa tarefa pode exigir um grau elevado de experiência e de conhecimento específico sobre um determinado grupo taxonômico. Assim, esses dados são omitidos quando há dúvida em relação a sua corretude.

O erro de *incorrect data values* é muito comum nesse domínio de dados e é causado por erros de digitação. O fato de os nomes de táxons serem escritos em latim, pode contribuir para o aumento da quantidade de erros nesse domínio de dados.

*Nonatomic data values* podem ocorrer quando há a inserção de sinônimos de um táxon em um mesmo campo. Por exemplo, informar o nome de um filo como: “Angiosperma, Magnoliophyta”.

No domínio de dados de táxon, o problema de *domain schizophrenia* ocorre quando um campo é utilizado para um propósito ao qual ele não foi designado, por exemplo, utilizar o campo de nome científico da espécie para cadastrar “sp1”, que indica uma morfoespécie.

*Inconsistent data values* pode estar relacionado ao não uso de padrões de nomenclatura e de hierarquias taxonômicas ou à inadequação dos dados digitalizados ao padrão adotado.

O problema de *information quality contamination* ocorre quando parte de uma hierarquia taxonômica incorreta é utilizada para complementar outra hierarquia taxonômica mais completa, por exemplo.

Esses problemas identificados por English (1999), contextualizados para banco de dados taxonômicas por Dalcin (2005) e definidos no contexto de dados de localização, geoespaciais e taxonômicos de ocorrências de espécies (Veiga *et al.*, 2011a), foram utilizados na camada de problemas do *framework* de Avaliação da QD proposto por Ge & Helfert (2007). A vista desses padrões de erros, a seguir são apresentadas as dimensões de QD que foram utilizadas neste trabalho.

### **5.1.1.2. Camada de dimensão**

A QD é definida na literatura como um conceito multidimensional (Pipino et al., 2002), no qual as dimensões representam aspectos da qualidade dos dados. Essas dimensões permitem realizar a Avaliação e o Gerenciamento da QD de maneira mais objetiva e específica.

Nesse sentido, foram identificadas seis dimensões de QD importantes no contexto do SI BDD:

- **Completude** (*Completeness*) é uma dimensão gerenciável e mensurável que indica a suficiência de dados válidos para serem utilizadas na realização de uma determinada tarefa (Pipino et al., 2002; Dalcin, 2005);
- **Consistência** (*Consistency*) é utilizada para medir e gerenciar a ausência de contradições em banco de dados (McGilvray, 2008);
- **Credibilidade da fonte** (*Credibility of source*) está relacionada à medição de aspectos associados à reputação dos dados ou de sua fonte (Dalcin, 2005) e é utilizada para medir o quanto os dados merecem crédito para serem utilizados (Wang et al. 1995);
- **Acurácia** (*Accuracy*) é considerada em muitos estudos de QD como uma dimensão chave, e pode ser definida como a medida da corretude ou da veracidade dos dados (Pipino et al., 2002);
- **Precisão** (*Precision*) é frequentemente confundida com acurácia; contudo, acurácia está relacionada ao erro, enquanto que precisão está relacionada à resolução ou granularidade dos dados (Chapman, 2005b).
- **Confiabilidade** (*Believability*) indica o grau de confiança para que os dados possam ser utilizados. No campo de estudo da QD, essa dimensão é constituída pela composição das dimensões de completude, acurácia, consistência e credibilidade da fonte, de acordo com Wang et al. (1995).

Assim como os problemas discutidos anteriormente, as dimensões também podem assumir diferentes significados em relação aos domínios de dados, conforme descrito a seguir.

### ***Definição de dimensões no domínio de dados de localização e geoespaciais***

No domínio de dados geoespaciais e de localização, a completude de dados é considerada um fator importante, pois a ausência de alguns dados geoespaciais (como latitude ou longitude) ou de localização geográfica (como nome da cidade) de ocorrências de espécies limita o uso desses dados para muitas aplicações (Chapman, 2005b).

A causa da incompletude de dados geoespaciais pode ocorrer devido à indisponibilidade de recursos de geoposicionamento, como receptores GPS (*Global Positioning System*), no momento do registro da ocorrência. Uma técnica que pode ser utilizada para obter uma coordenada geoespacial, quando os dados de localização foram preenchidos corretamente, é utilizar a coordenada do centroide do município aonde houve a ocorrência. Contudo, essa técnica pode prejudicar a qualidade nas dimensões de acurácia e de precisão.

A consistência, nesses domínios de dados, pode ser interpretada como a ausência de contradições entre as coordenadas geoespaciais e os dados de localidade. Indicar que o local da ocorrência foi na cidade de São Paulo, Brasil, mas a coordenada geoespacial referir-se a uma posição no continente africano é um exemplo de inconsistência. Outra forma de inconsistência relacionada ao domínio de localização é indicar que a ocorrência foi registrada na cidade de Londres, estado de São Paulo e país Argentina, por exemplo.

No domínio de dados de localização, a acurácia pode estar relacionada à corretude ortográfica dos nomes das localizações geográficas. Enquanto que a precisão pode estar relacionada à presença ou ausência de dados de localidades mais específicas, como nome da cidade, por exemplo.

Em relação ao domínio de dados geoespaciais, as dimensões de acurácia e de precisão são fortemente correlatas, e suas definições são normalmente confundidas ou incompreendidas (Dalcin, 2005; Chapman, 2005b). A Figura 8 ilustra esses dois conceitos.



Figura 8 - Relação entre precisão e acurácia em dados geoespaciais (Dalcin, 2005).

A acurácia refere-se ao intervalo entre o valor real da posição e o valor informado. Precisão (ou resolução) pode ser dividida em duas abordagens principais: estatística e numérica. Precisão estatística refere-se à relativa conformidade das posições geoespaciais de um conjunto de ocorrências. Conforme demonstra a Figura 8, as posições geoespaciais das ocorrências podem ser precisas, mas não acuradas. Precisão numérica é relativa à quantidade de dígitos significativos utilizados para representar uma posição no espaço. Por exemplo, a latitude e a longitude decimal podem ser representadas com 10 casas decimais, ou seja, cerca 0,01 milímetros, contudo, a resolução real não é superior a 10 metros (Chapman, 2005b).

### ***Definição de dimensões no domínio de dados taxonômicos***

A definição de qualidade no domínio de dados de táxon difere consideravelmente dos domínios de dados geoespacial e de localização, pois normalmente dados taxonômicos são mais abstratos e mais difíceis de qualificar (Chapman, 2005b).

Esses dados são os principais identificadores das ocorrências de espécies, pois eles indicam a qual grupo taxonômico a espécie observada ou coletada pertence e, consequentemente, quais são suas características morfológicas, genéticas, ecológicas, fisiológicas, ambientais, entre outras. A tarefa de realizar uma identificação taxonômica, ou seja, associar uma nomenclatura taxonômica a um

determinado organismo exige experiência e conhecimento específico sobre determinados grupos taxonômicos. Muitas vezes, o pesquisador precisa consultar bibliografias, recursos multimídia e chaves taxonômicas para auxiliar na sua tomada de decisão em relação à identificação. Assim, a completude desses dados depende basicamente do conhecimento do pesquisador a cerca da espécie coletada ou observada.

A consistência de dados taxonômicos está relacionada à ausência de contradição em hierarquias taxonômicas e nas nomenclaturas. Contudo, cada instituição pode adotar uma nomenclatura ou hierarquia taxonômica própria (Kelling, 2008). Assim, ainda que haja essa “inconsistência” em um âmbito global, é necessário que, ao menos, cada instituição adote um padrão de hierarquia e de nomenclatura para ser usado em um âmbito mais restrito. Nesse contexto, existem as chamadas autoridades taxonômicas, como ITIS, Species 2000 ou CoL, por exemplo, que definem padrões de hierarquias e nomenclaturas. A adoção de um padrão de uma dessas autoridades pode influenciar na credibilidade da fonte dos dados.

A acurácia nesse domínio de dados está relacionada à corretude ortográfica dos nomes dos táxons. A precisão pode ser definida pela presença de dados de táxons mais específicos da hierarquia taxonômica, como gênero, subgênero, epíteto específico ou espécie, por exemplo. A acurácia também pode estar relacionada com a corretude da identificação taxonômica de um espécime, ou seja, informar corretamente que um determinado espécime pertence a um táxon X e não a um táxon Y.

A confiabilidade dos dados pode ser representada pelo modelo hierárquico de dimensões de QD proposto por Wang *et al.* (1995). Nesse modelo de representação, a confiabilidade é definida pela composição das dimensões de completude, consistência, credibilidade da fonte e acurácia, conforme a Figura 9.

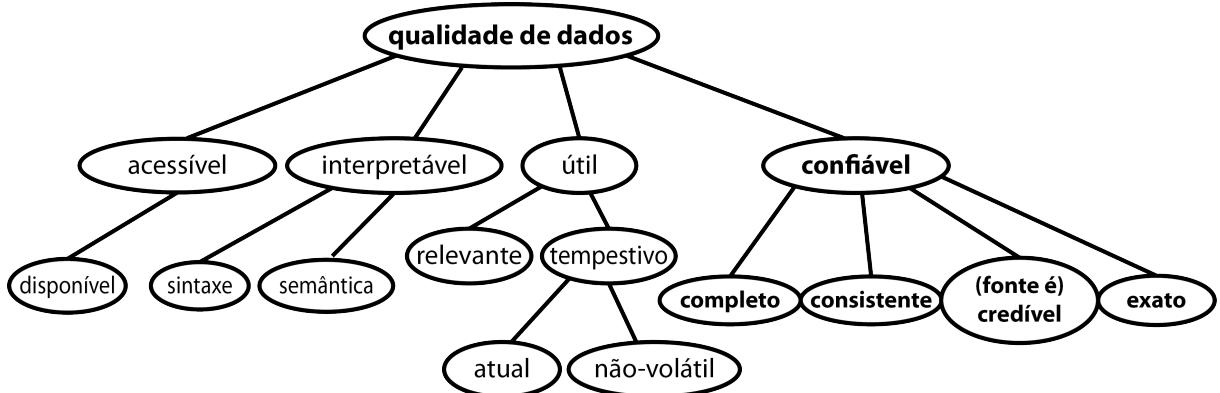


Figura 9 – Modelo hierárquico de dimensões de QD.

Adaptado de Wang et al (1995).

Assim, com base nesses conceitos, observa-se que a qualidade de um conjunto de dados depende de uma série de questões (Wang et al., 1995), e pode variar de acordo o domínio de dados. A seguir, será apresentada uma análise que identifica como a qualidade nessas dimensões de QD pode ser afetada.

#### **5.1.1.3. Camada de metodologia de avaliação**

Os elementos da camada de metodologia de avaliação estão relacionados aos domínios de dados, escopo deste trabalho. Desse modo, os elementos da camada de metodologia de avaliação são:

- Avaliação da QD Taxonômicos;
- Avaliação da QD Geoespaciais;
- Avaliação da QD de Localização.

Esses componentes estão relacionados com os elementos da camada de dimensão, os quais estão relacionados com os elementos da camada de problema, conforme exemplificado pela Figura 10.

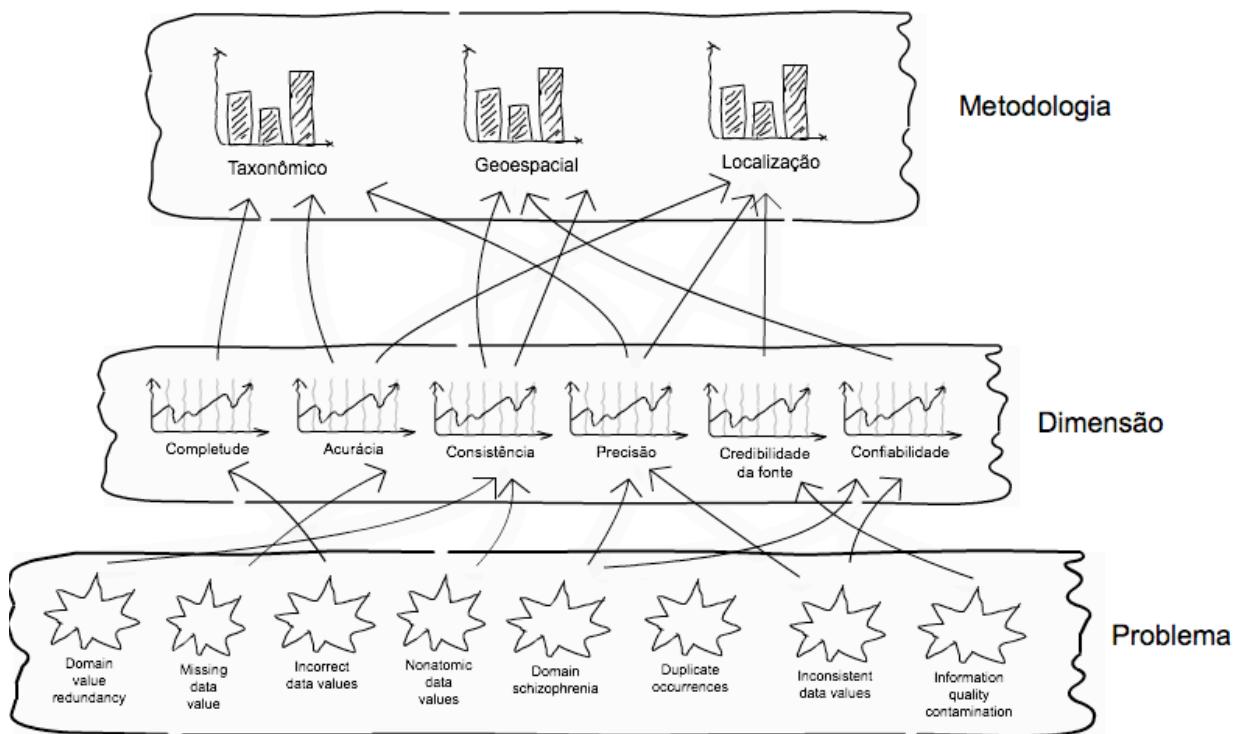


Figura 10 – Representação da Avaliação da QD de ocorrências de espécies.

Baseado em Ge & Helfert (2007).

A metodologia de Avaliação da QD proposta neste trabalho é baseada na premissa de que a qualidade em cada dimensão é afetada pela presença ou ausência de problemas. Além disso, como os problemas e as dimensões possuem diferentes significados em cada domínio de dados, os elementos da camada de metodologia de avaliação são baseados nesses domínios de dados. Assim, a avaliação da QD em determinada dimensão é definida com base na presença ou ausência de determinados erros em um determinado domínio de dados – Localização (L), Geoespacial (G) ou Taxonômico (T) – conforme o Quadro 2.

Problemas	Dimensões	Completude	Consistência	Acurácia	Precisão	Credibilidade da Fonte	Confiabilidade
<i>Domain value redundancy</i>	-	T   G   L	T   G   L	-	T   G   L	T   G   L	
<i>Missing data value</i>	T   G   L	-	-	T   -   L	T   G   L	T   G   L	
<i>Incorrect data values</i>	T   G   L	T   G   L	T   G   L	-   G   -	T   G   L	T   G   L	
<i>Nonatomic data values</i>	T   G   L	T   G   L	T   G   L	-	-	T   G   L	
<i>Domain schizophrenia</i>	T   G   L	T   G   L	T   G   L	-	T   G   L	T   G   L	
<i>Duplicate occurrences</i>	-	T   G   L	-	-	-	T   G   L	
<i>Inconsistent data values</i>	-	T   G   L	T   G   L	-	T   G   L	T   G   L	
<i>Information quality contamination</i>	-	T   -   L	T   -   L	-	T   -   L	T   -   L	

Quadro 2 - Impacto dos problemas nas dimensões de QD nos domínios de dados.

Domínio de dados de Localização (L), Geoespaciais (G) e Taxonômicos (T).

O Quadro 2 demonstra quais problemas (linhas) afetam, ou degradam, a qualidade nas dimensões de QD (colunas) em cada domínio de dados (L, G ou T). A presença do símbolo de um domínio de dados (L, G ou T) em uma célula indica que o erro naquela linha afeta a dimensão de QD da coluna correspondente à célula. Por exemplo, a célula destacada em cinza, que tem o valor “T | - | L” representa que o erro “*missing data value*” (linha) afeta a precisão (coluna) nos domínios de dados Taxonômicos (T) e de Localização (L), mas não no domínio de dados Geoespaciais (-).

Portanto, o resultado da camada de metodologia de avaliação, representada pelo Quadro 2, mostra como a QD pode ser melhorada em uma determinada dimensão e domínio de dados por meio da redução de determinados problemas.

### 5.1.2. Gerenciamento da QD

Com base na Avaliação da QD, a seguir são listadas algumas técnicas e recursos que podem ser implementados em SI para a prevenção a erros durante a digitalização de dados de ocorrências de espécies, e assim melhorar a qualidade desses dados.

### **5.1.2.1. Sugestões de nomenclaturas taxonômicas usando uma técnica de Fuzzy Matching**

Esse recurso pode ser implementado em um campo do tipo *autocomplete*. Com esse tipo de campo, à medida que o usuário começa a digitar caracteres, uma lista de sugestões é apresentada e atualizada à medida que caracteres vão sendo inseridos ou excluídos. Essa lista de sugestões pode ser gerada utilizando uma técnica de *Fuzzy Matching*. Essa técnica permite recuperar dados textuais ortograficamente similares. Por exemplo, se o usuário digitar *Apes mellifera*, o sistema pode sugerir *Apis mellifera*, caso o segundo nome exista no banco de dados consultado. Assim, se houver algum erro de digitação do nome taxonômico o sistema pode sugerir nomes similares e corretos.

Essa consulta pode ser realizada a um banco de dados de alguma autoridade taxonômica, como o CoL. Assim, caso haja dúvidas sobre a ortografia de algum nome taxonômico, sugestões de nomes ortograficamente similares e considerados internacionalmente corretos podem ser apresentadas ao usuário.

Contudo, as bases de dados dessas autoridades taxonômicas tendem a ser muito grandes, chegando a ter milhões de registros de nomes de todos os níveis da hierarquia taxonômica. Assim, consultas utilizando *Fuzzy Matching* a um volume muito grande de dados pode demandar muito processamento e memória para realizar os cálculos de similaridade para cada registro, podendo afetar o desempenho do SI e do banco de dados, aumentando, portanto, o tempo de resposta.

Para aumentar a produtividade do usuário, reduzindo o tempo de resposta desse recurso, consultas preliminares ao banco de dados local podem ser realizadas, visto que a quantidade de nomes taxonômicos distintos registrados localmente normalmente é menor que as registradas em banco de dados de autoridades taxonômicas. Caso o nome consultado não exista no banco de dados local, então, uma segunda consulta poderia ser feita ao banco de dados das autoridades taxonômicas. Com isso, o desempenho do sistema melhora para os casos em que o usuário procurar por um nome anteriormente utilizado no SI, aumentando, assim, a produtividade dos produtores de dados.

Em dados taxonômicos, esse recurso pode reduzir erros de *domain value redundancy*, visto que o sistema não irá sugerir sinônimos. *Incorrect data values*,

também pode ser reduzido pois, caso haja erros de digitação, o SI sugere uma “correção”. *Nonatomic data values* também são evitados, visto que o sistema sugere nomes atômicos. *Domain schizophrenia* pode se reduzido visto que não são sugeridos nomes de morfoespécies. O uso desse recurso pode melhorar, portanto, a QD nas dimensões de acurácia, consistência, credibilidade da fonte e confiabilidade.

#### **5.1.2.2. Sugestão de hierarquias taxonômicas**

Esse recurso consiste em preencher automaticamente a hierarquia taxonômica a partir da seleção de um nome de táxon mais específico. Ou seja, o usuário escolhe o nome de um táxon e baseado em fontes de dados de autoridades taxonômicas ou de um banco de dados local, o sistema sugere os demais nomes, menos específicos (mais altos) da hierarquia taxonômica. Por exemplo, se o usuário seleciona o nome de um gênero X, o sistema irá sugerir os nomes da família, ordem, classe, filo e reino relacionados a esse gênero com base nos registros das fontes de dados. Ao aceitar uma sugestão, o SI preenche automaticamente o formulário de cadastro de ocorrência de espécie com a hierarquia selecionada.

Esse recurso, além de agilizar o preenchimento dos dados taxonômicos, melhorando a produtividade dos usuários, permite também uma potencial redução de erros de *domain value redundancy*, *missing data value*, *incorrect data values*, *nonatomic data values*, *nonatomic data values*, *inconsistent data values* e *information quality contamination* no domínio de dados taxonômicos.

#### **5.1.2.3. Validação de nomenclaturas e hierarquias taxonômicas em relação a autoridades taxonômicas**

A validação de nomenclatura e de hierarquia taxonômica pode ser realizada por meio da verificação da conformidade desses dados a um padrão, norma ou amostra que seja considerado aceitável, correto ou válido pela comunidade científica. Essa validação é essencial, principalmente, para determinar ou avaliar a credibilidade dos dados.

Assim, a validação dos dados taxonômicos pode ser realizada por meio de consultas aos bancos de dados de autoridades taxonômicas, a fim de comparar os nomes e hierarquias taxonômicos digitalizados em relação aos nomes e hierarquias

considerados válidos de acordo com as autoridades taxonômicas. Caso o sistema encontre uma comparação que combine, então o SI associa aos dados taxonômicos digitalizados os nomes das autoridades taxonômicas que consideram tais dados válidos.

Desse modo, a credibilidade dos dados taxonômicos da ocorrência de espécie digitalizada pode ser avaliada com base na credibilidade da autoridade taxonômica validadora. Esse recurso pode reduzir, portanto, erros de *missing data value*, visto que uma informação importante sobre a credibilidade da fonte dos dados taxonômicos não é omitida.

#### **5.1.2.4. Consulta a recursos multimídia sobre táxons**

O uso de recursos multimídia, como fotografias, vídeos e sons, pode auxiliar taxonomistas na identificação taxonômica de espécies, caso haja dúvida sobre a classificação de um determinado organismo. Assim, consultas a recursos multimídia, implementadas no SI por meio de um dispositivo que apresente ao usuário fotos, vídeos ou sons relacionados a um determinado táxon, pode auxiliar na digitalização correta de dados taxonômicos de ocorrências de espécies. Visto que a atividade de identificação taxonômica depende do conhecimento sobre grupos taxonômicos específicos, imagens e sons podem ser úteis para melhorar a precisão, completude e acurácia de dados taxonômicos de ocorrências de espécies.

Esse recurso pode ser implementado no formulário de cadastro de ocorrências de espécies, de modo que permita ao usuário consultar um banco de dados de fotografias, vídeos e sons indexados por nome taxonômico. Assim, se o usuário não tiver certeza se uma determinada abelha é uma *Apis mellifera*, por exemplo, o usuário pode consultar imagens sobre essa espécie para comparar e tirar dúvidas, a fim de aumentar as chances de realizar uma identificação correta. Essa consulta pode ser realizada a um banco de dados local ou a repositório de informações taxonômicas, como ao banco de dados do EoL, que disponibiliza web services para a consulta de imagens sobre táxons.

Esse recurso pode minimizar erros de *missing data value*, *incorrect data value*, *inconsistent data values* e *information quality contamination* em dados taxonômicos.

#### **5.1.2.5. Consulta a recursos bibliográficos sobre táxons**

Durante a identificação taxonômica, recursos bibliográficos, como artigos e livros, podem ser consultados para ajudar a realizar a identificação taxonômica ou para validá-la.

Assim, a implementação de um dispositivo que permita fazer consultas a esse tipo de recurso durante a digitalização de ocorrências de espécies pode auxiliar o usuário a realizar uma identificação taxonômica mais precisa, completa e exata, diminuindo possíveis incertezas relacionadas à identificação. A consulta a esses materiais bibliográficos pode ser feita a um banco de dados local ou a fontes de dados externas.

A implementação desse recurso no SI pode auxiliar na diminuição de erros de *missing data value*, *incorrect data value*, *inconsistent data values* e *information quality contamination* no domínio de dados taxonômicos.

#### **5.1.2.6. Suporte a morfoespécies**

Existem casos em que há dúvidas em relação à classificação taxonômica de um conjunto de espécimes, contudo, sabe-se que todos os espécimes do conjunto pertencem a um mesmo táxon. Quando isso ocorre, esses espécimes podem receber um identificador temporário chamado de morfoespécie, como “sp1”, por exemplo. Isso indica que todos os espécimes identificados como “sp1” pertencem a um mesmo táxon. Assim, ao se realizar a identificação taxonômica de um indivíduo desse conjunto, todos os outros indivíduos, que receberam o mesmo identificador de morfoespécie serão, consequentemente, identificados como sendo do mesmo táxon.

Desse modo, o suporte a morfoespécie em campos de nomes taxonômicos pode reduzir alguns problemas. Esse recurso pode ser implementado no SI da seguinte maneira: quando o usuário digitar a sequência de caracteres “sp”, seguido por um número, por exemplo, “sp1”, “sp5”, em um campo de nome taxonômico, o sistema automaticamente identificará esse táxon como uma morfoespécie. Posteriormente, quando uma dessas morfoespécies for identificada, os outros registros com os mesmos identificadores serão, também, alterados para o mesmo táxon.

Esse recurso pode causar uma redução de erros de *incorrect data values*, *domain schizophrenia* e *information quality contamination* em dados taxonômicos.

#### **5.1.2.7. Indicador de incerteza da identificação taxonômica**

Esse recurso permite ao usuário reportar o grau de incerteza em relação a uma identificação taxonômica. Por exemplo, normalmente, se um usuário tem dúvidas sobre a identificação de um espécime ele pode tomar uma entre duas decisões: não cadastrar e, assim, diminuir a completude, ou cadastrar e, se a identificação estiver incorreta, diminuir a acurácia.

Nesse sentido, a disponibilidade de indicador de incerteza no SI permite ao usuário indicar que a informação cadastrada tem uma probabilidade de estar incorreta, necessitando assim de uma validação de um especialista. Assim, mesmo que o usuário não tenha certeza sobre a classificação do espécime, o dado pode ser inserido e a sua acurácia pode ser avaliada. Por meio desse indicador, pode-se avaliar a adequação ao uso dos dados, além de impactar nas dimensões de credibilidade da fonte e de completude de dados.

Portanto, a implementação desse recurso pode reduzir erros de *missing data value* em dados taxonômicos, visto que por meio desse recurso são fornecidas informações importantes para avaliar a credibilidade, a acurácia e a confiabilidade dos dados.

#### **5.1.2.8. Georeferenciamento – a partir de descrição da localização**

O georeferenciamento é processo de obter informações geoespaciais a partir da descrição de uma localização (BioGeomancer, 2011). A implementação desse recurso no SI pode ser realizada por meio de um campo de texto, no qual o usuário poderia digitar uma descrição da localização da ocorrência, como “*Bariloche, 25 km NNE via Ruta Nacional 40 (=Ruta 237)*”, por exemplo, e obter como resposta um conjunto de coordenadas geográficas correspondentes a essa descrição.

É comum a localização de uma ocorrência ser descrita em linguagem natural pela indisponibilidade de recursos geoposicionamento, como receptores GPS. Portanto, a implementação de um recurso de georeferenciamento no SI pode auxiliar no preenchimento de dados geoespaciais. Essa implementação pode ser realizada

por meio da API do Google *Maps* ou dos *web services* do BioGeomancer e do GeoLocate, por exemplo, os quais permitem georeferenciar ocorrências de espécies a partir da descrição de suas localizações.

Desse modo, com a implementação desse recurso, pode haver uma redução de erros de *missing data value*, *incorrect data values*, *inconsistent data values* e *information quality contamination* em dados geoespaciais.

#### **5.1.2.9. Georeferenciamento reverso – a partir das coordenadas geográficas**

O georeferenciamento reverso é o processo de obter informações geográficas de uma localização, como nome do país, do estado, da cidade e descrição da localização, a partir de dados geoespaciais, como coordenadas geoespaciais. A implementação desse recurso no SI de digitalização de ocorrências de espécies permite ao usuário preencher consistentemente os dados do domínio de localização a partir de coordenadas geoespaciais, como latitude e longitude decimais. Esse recurso pode ser implementado utilizando a API do Google *Maps* ou os *web services* do GeoNames e do GeoLocate.

Quando implementado no SI, esse recurso pode permitir uma redução de erros de *domain value redundancy*, *missing data value*, *nonatomic data values*, *domain schizophrenia*, *inconsistent data values* e *information quality contamination* no domínio de dados de localização.

#### **5.1.2.10. Georeferenciamento a partir de um mapa interativo**

Esse recurso consiste em permitir que o usuário utilize um mapa interativo para obter as coordenadas geoespaciais a partir de um clique sobre a localização desejada no mapa. Ao selecionar uma localização aproximada, o recurso de georeferenciamento reverso pode ser executado para obter informações mais completas sobre a localização.

A implementação desse recurso no SI pode ser feita utilizando a API do Google *Earth* e do Google *Maps*. Desse modo, pode haver uma potencial melhora da completude de dados, pois o usuário não precisa, necessariamente, ter as

coordenadas geoespaciais exatas da ocorrência da espécie para poder preencher os campos de latitude, longitude e altitude, por exemplo.

O usuário pode localizar alguma região conhecida no mapa, como um parque ou uma montanha, por exemplo, e utilizar uma referência mais específica, como um rio ou uma estrada, para obter os dados de localização e as coordenadas geoespaciais aproximadas. Em alguns casos, esse recurso pode contribuir, também, para a melhora da acurácia e da precisão no domínio de dados geoespaciais como, por exemplo, nos casos em que a coordenada geográfica é obtida a partir do centro de massa da cidade aonde houve a ocorrência.

Portanto, essa ferramenta pode reduzir erros de *missing data value*, *incorrect data value*, *domain schizophrenia* e *information quality contamination* em dados geoespaciais e de localização.

#### **5.1.2.11. Indicador de incerteza das coordenadas geográficas**

Um aspecto importante sobre a QD no domínio de dados geoespaciais é suscitado pelo GBIF em Hill *et al.* (2010). A acurácia e a precisão não precisam, necessariamente, ser perfeitas (Wang *et al.*, 1995). O uso dos dados geoespaciais em algumas aplicações admite baixa acurácia e baixa precisão. Contudo, há casos em que a acurácia e a precisão dos dados devem ser altas. Desse modo, a qualidade dos dados é definida pela adequação ao uso. Portanto, é necessário reportar o quanto preciso e acurado os dados são para avaliar a sua adequação ao uso (Hill *et al.*, 2010).

Visando essa necessidade, a implementação de um recurso que permita ao usuário indicar o grau de incerteza da exatidão dos valores informados é importante sob o ponto de vista de QD. Com esse recurso é possível reportar o quanto preciso ou exato os dados são. Assim, se um determinado usuário sabe que um espécime foi coletado em uma montanha específica, mas não sabe a posição geoespacial exata, é possível, por meio de um indicador de incerteza/erro, reportar que a posição geoespacial informada pode conter um erro de até 10 km, por exemplo. Desse modo, será possível avaliar adequação ao uso dos dados.

Esse recurso soluciona erros de *missing data value*, pois fornecem informações que pode melhorar a credibilidade da fonte em dados geoespaciais.

#### **5.1.2.12. Plotagem das coordenadas geoespaciais em um mapa**

Esse recurso permite ao usuário visualizar um mapa com as coordenadas geoespaciais plotadas. Desse modo, o usuário pode realizar uma validação visual das coordenadas geoespaciais digitalizadas.

É comum o usuário esquecer-se de colocar o sinal negativo nos campos de latitude e de longitude decimal, para localidades do hemisfério sul e para oeste do meridiano zero respectivamente, ocasionando uma plotagem incorreta das coordenadas. Com esse recurso, se o usuário cometer esse erro, um mapa será exibido com as coordenadas plotadas em uma região incorreta, facilitando a identificação e correção do erro.

Portanto, a implementação desse recurso no SI pode reduzir erros de *incorrect data values* em dados geoespaciais.

#### **5.1.2.13. Restrição de unicidade no banco de dados**

A restrição de unicidade garante que os valores contidos em uma coluna, ou no grupo de colunas, sejam únicos em relação aos valores de todas as outras linhas de uma tabela (PostgreSQL, 2011).

A implementação dessa restrição no banco de dados permite a prevenção de erros de *duplicate occurrences* em dados taxonômicos, geoespaciais e de localização.

#### **5.1.2.14. Consolidação dos recursos propostos**

Os recursos listados anteriormente, se implementados em um SI de digitalização de ocorrências de espécies, pode provocar uma potencial redução de erros e, consequentemente, uma melhora na QD nos domínios de dados de Localização (L), Geoespaciais (G) ou Taxonômicos (T).

O Quadro 3 demonstra quais recursos (linhas) previnem, ou reduzem, a ocorrência dos problemas de QD (colunas) em cada domínio de dados (L, G ou T). A presença do símbolo de um domínio de dados (L, G ou T) em uma célula indica que o recurso naquela linha pode prevenir o problema de QD da coluna correspondente à célula.

<b>Problemas</b>	<i>Domain value redundancy</i>	<i>Missing data value</i>	<i>Incorrect data values</i>	<i>Non-atomic data values</i>	<i>Domain schizophrenia</i>	<i>Duplicate occurrences</i>	<i>Inconsistent data values</i>	<i>Information quality contamination</i>
<b>Recursos</b>								
Sugestões de nomenclaturas taxonômicas usando uma técnica de Fuzzy Matching	T   -   -	-   -   -	T   -   -	T   -   -	T   -   -	-   -   -	-   -   -	-   -   -
Sugestões de hierarquias taxonômicas	T   -   -	T   -   -	T   -   -	T   -   -	-   -   -	-   -   -	T   -   -	T   -   -
Validação de nomenclaturas e hierarquias taxonômicas em relação a autoridades taxonômicas	T   -   -	T   -   -	T   -   -	T   -   -	T   -   -	-   -   -	T   -   -	T   -   -
Consulta a recursos multimídia sobre táxons	-   -   -	T   -   -	T   -   -	-   -   -	-   -   -	-   -   -	-   -   -	-   -   -
Consulta a recursos bibliográficos sobre táxons	T   -   -	T   -   -	T   -   -	-   -   -	-   -   -	-   -   -	T   -   -	T   -   -
Suporte a morfoespécies	-   -   -	T   -   -	T   -   -	-   -   -	T   -   -	-   -   -	T   -   -	-   -   -
Indicador de incerteza da identificação taxonômica	-   -   -	T   -   -	-   -   -	-   -   -	-   -   -	-   -   -	-   -   -	-   -   -
Georeferenciamento – a partir de descrição da localização	-   -   -	-   G   -	-   G   -	-   G   -	-   G   -	-   -   -	-   G   L	-   -   -
Georeferenciamento reverso – a partir das coordenadas geoespaciais	-   -   L	-   -   L	-   -   L	-   -   L	-   -   -	-   -   -	-   G   L	-   -   L
Georeferenciamento a partir de um mapa interativo	-   -   -	-   G   -	-   G   -	-   G   -	-   G   -	-   -   -	-   G   L	-   -   -
Indicador de incerteza das coordenadas geográficas	-   -   -	-   G   -	-   -   -	-   -   -	-   -   -	-   -   -	-   -   -	-   -   -
Plotagem das coordenadas geoespaciais em um mapa	-   -   -	-   -   -	-   G   L	-   -   -	-   -   -	-   -   -	-   -   -	-   -   -
Restrição de unicidade no banco de dados	-   -   -	-   -   -	-   -   -	-   -   -	-   -   -	T   G   L	-   -   -	-   -   -

Quadro 3 - Problemas potencialmente reduzidos com o uso dos recursos de Gerenciamento da QD.

Domínio de dados de Localização (L), Geoespaciais (G) e Taxonômicos (T).

O Quadro 4 demonstra quais recursos (linhas) melhoram, ou afetar, a qualidade nas dimensões de QD (colunas) em cada domínio de dados (L, G ou T). A presença do símbolo de um domínio de dados (L, G ou T) em uma célula indica que o recurso naquela linha pode melhorar a qualidade na dimensão de QD da coluna correspondente a célula.

Recursos \ Dimensões	Completude	Consistência	Acurácia	Precisão	Credibilidade da Fonte	Confiabilidade
Sugestões de nomenclaturas taxonômicas usando uma técnica de Fuzzy Matching	-   -   -	T   -   -	T   -   -	-   -   -	T   -   -	T   -   -
Sugestão de hierarquias taxonômicas	T   -   -	T   -   -	T   -   -	-   -   -	T   -   -	T   -   -
Validação de nomenclaturas e hierarquias taxonômicas em relação a autoridades taxonômicas	T   -   -	T   -   -	T   -   -	-   -   -	T   -   -	T   -   -
Consulta a recursos multimídia sobre táxons	T   -   -	-   -   -	T   -   -	T   -   -	T   -   -	T   -   -
Consulta a recursos bibliográficos sobre táxons	T   -   -	T   -   -	T   -   -	T   -   -	T   -   -	T   -   -
Suporte a morfoespécies	T   -   -	T   -   -	T   -   -	-   -   -	-   -   -	T   -   -
Indicador de incerteza da identificação taxonômica	T   -   -	-   -   -	T   -   -	T   -   -	T   -   -	T   -   -
Georeferenciamento – a partir de descrição da localização	-   G   -	-   G   L	-   G   -	-   G   -	-   G   -	-   G   L
Georeferenciamento reverso – a partir das coordenadas geoespaciais	-   -   L	-   G   L	-   -   L	-   -   L	-   -   L	-   G   L
Georeferenciamento a partir de um mapa interativo	-   G   -	-   G   L	-   G   -	-   G   -	-   -   L	-   G   L
Indicador de incerteza das coordenadas geográficas	-   G   -	-   -   -	-   G   -	-   G   -	-   G   -	-   G   -
Plotagem das coordenadas geoespaciais em um mapa	-   -   -	-   -   -	-   G   L	-   -   -	-   G   L	-   G   L
Restrição de unicidade no banco de dados	-   -   -	T   G   L	-   -   -	-   -   -	-   -   -	T   G   L

Quadro 4 - Dimensões de QD afetadas com o uso dos recursos de Gerenciamento da QD.

Domínio de dados de Localização (L), Geoespaciais (G) e Taxonômicos (T).

O resultado do estudo de Gerenciamento da QD deste trabalho, compilado por meio dos Quadros 3 e 4, são consistentes com os resultados do estudo da Avaliação da QD, representado pelo Quadro 2.

## **5.2. Estudo de caso de aplicação dos estudos de QD**

Os estudos de QD resultados deste trabalho, sobretudo os resultados do estudo sobre o Gerenciamento da QD, foram aplicados a um SI de digitalização de ocorrências de espécies. Para que isso pudesse ser feito, foi necessário considerar alguns aspectos relacionados ao SI, como arquitetura de *software*, manutenibilidade, escalabilidade, acoplamento, entre aspectos relacionados a engenharia de *software*, e aspectos relacionados aos usuários, como a cultura, expectativas e preferências, por exemplo. Baseado nessas considerações o SI existente, BDD versão Beta, passou por uma reestruturação de código e arquitetura de *software* para que pudesse suportar as ferramentas de QD desenvolvidas e descritas na Subseção 5.2.2.

### **5.2.1. Sistema de Informação: BDD**

O SI utilizado neste trabalho foi o BDD. Esse sistema *web* de código aberto (<http://code.google.com/p/laa-biodiversitydatadigitizer/>) foi projetado para permitir a fácil digitalização, manipulação e publicação de dados de biodiversidade, principalmente, dados de ocorrências de espécies.

O BDD foi projetado em uma arquitetura multi-módulos, no qual cada módulo tem o objetivo de digitalizar um tipo de dados diferente. São sete módulos desenvolvidos para manipular dados de ocorrências de espécies, espécies, interação entre espécies, monitoramento de polinizadores, déficit de polinização, recursos multimídia e recursos bibliográficos. Esses módulos permitem a digitalização (criação), manipulação (edição e exclusão) e consultas de registros.

A maioria desses módulos foram desenvolvidos baseados nos padrões publicados pelo TDWG, permitindo a publicação e o compartilhamento dos dados armazenados no BDD com outros sistemas, por meio do protocolo TAPIR.

Além dos módulos de digitalização, também fazem parte do BDD um módulo de análise estatística e visualização de dados e um módulo de sincronização de dados entre o banco de dados do BDD e planilhas eletrônicas.

Esse SI foi desenvolvido utilizando tecnologias *open source*, incluindo Javascript, PHP, Java, servidores Tomcat e Apache, biblioteca Javascript jQuery, framework Yii e banco de dados PostgreSQL. Visando a escalabilidade, manutenibilidade e baixo acoplamento, foi utilizada a arquitetura de *software Model-View-Controller – MVC*, conforme representado pela Figura 11.

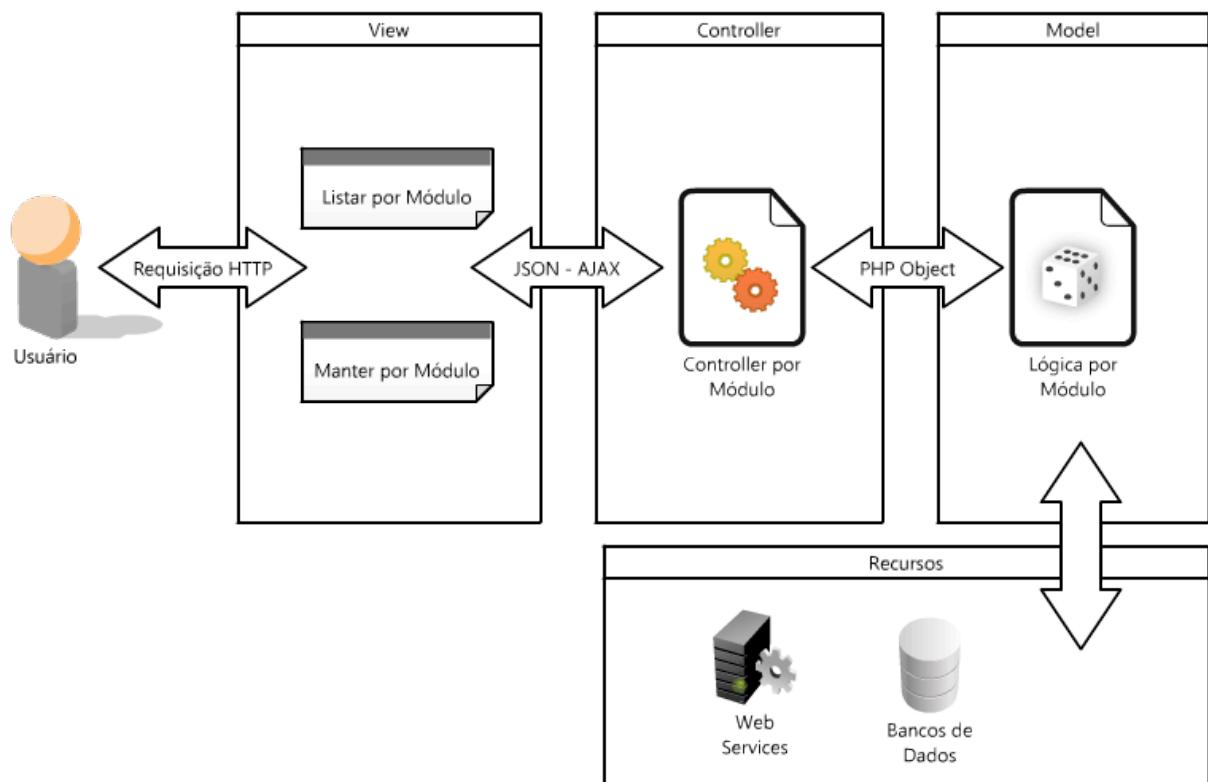


Figura 11 - Arquitetura de *software* do BDD.

Conforme representado na Figura 11, o usuário faz requisições HTTP por meio de um navegador *web* à camada *View* (Visão), na qual páginas escritas em PHP e interpretadas pelo servidor Apache geram documentos em HTML, CSS e Javascript, que são retornados como resposta ao usuário. Por meio dos documentos dessa camada, requisições assíncronas em *Asynchronous Javascript and XML* –

AJAX são realizadas usando o método POST e com parâmetros em formato Javascript *Object Notation* – JSON.

Essas requisições são feitas para camada *Controller* (Controlador), que interpreta e converte os dados passados por parâmetro para objetos PHP. Esses objetos são, então, enviados para a camada de *Model* (Modelo), onde fica a lógica de negócio do sistema. Nessa camada, toda a lógica para processar as requisições está implementada e, para isso, requisições aos bancos de dados e a *web services* externos podem ser realizadas.

Visto que a abordagem de Gerenciamento da QD neste trabalho é a prevenção a erros durante a digitalização e considerando alguns aspectos culturais dos usuários do BDD, uma análise de requisitos foi realizada para identificar características de um SI que possam aumentar a aceitação dos usuários ao BDD e as ferramentas de Gerenciamento da QD.

#### **5.2.1.1. Requisitos do sistema**

Produtores de dados de biodiversidade tendem a utilizar ferramentas que lhes são familiares, simples e fáceis de usar, pois permitem aos usuários realizarem suas atividades com maior agilidade. Essas ferramentas são, predominantemente, planilhas eletrônicas, conforme constatado no *Workshop de Treinamento da IABIN-PTN* em 2010 e no *TDWG Annual Conference* 2011. Contudo, erros que podem afetar negativamente a QD podem ocorrer durante a digitalização por meio de planilhas, os quais podem não ser despercebidos e, consequentemente, não corrigidos.

Assim, visto que a abordagem deste trabalho é a prevenção a erros por meio de ferramentas implementadas no SI BDD, a seguir é apresentada uma análise de requisitos de um SI para digitalização de ocorrências de espécies com suporte a QD, sob a perspectiva dos usuários do BDD, representados pelos atores listados na Subseção 2.4.1, com o objetivo de melhorar a aceitação de usuários de planilhas eletrônicas ao SI.

### **Produtividade**

Produtividade, no contexto de digitalização de dados de ocorrências de espécies, refere-se à quantidade de ocorrências digitalizadas por tempo. Esse requisito pode ser uma tônica em sistemas de digitalização, especialmente para os produtores de dados, visto que a produção de dados pode ser, com frequência, rotineira e dispendiosa. Talvez esse requisito seja o principal motivo de os produtores de dados preferirem utilizar planilhas eletrônicas, pois com elas é possível criar rapidamente um novo registro a partir da cópia de outro registro já existente e parecido.

Em muitos registros alguns campos têm valores iguais ou similares. Assim, o uso do recurso de copiar e colar uma linha na planilha eletrônica pode aumentar a produtividade do usuário. Sob essa perspectiva, uma solução compatível para aumentar a produtividade do usuário é desenvolver um recurso de *templates* de dados. Ou seja, o sistema permite ao usuário pré-cadastrar dados que são frequentemente utilizados por ele. Assim, quando for necessário cadastrar um novo registro, o usuário pode utilizar os dados pré-cadastrados no *template* para preencher o formulário de cadastro. Outro recurso que pode aumentar a produtividade consiste em o sistema permitir ao usuário reutilizar os dados do último registro cadastrado por ele.

### **Usabilidade**

A usabilidade é característica mensurável que indica o grau em que um sistema é fácil de usar. Desse modo, a seguir são listados alguns requisitos de usabilidade baseado em (FEC, 2003):

- O sistema deve guiar o usuário na execução de suas tarefas de maneira correta e eficiente;
- O sistema deve fornecer um nível adequado de orientação e *feedback* durante a realização de tarefas, permitindo ao usuário detectar e corrigir eventuais erros.
- O sistema deve oferecer uma interface intuitiva que facilite a navegação pelos recursos disponíveis no sistema.

- O usuário deve sentir-se emocionalmente confortável e confiante ao realizar suas tarefas por meio do sistema.

Sistemas fáceis de instalar e de aprender a usar têm maior aceitabilidade dos usuários (Rose, 1994). Assim, o processo de instalação do sistema não pode ser um problema para o usuário. A instalação deve ser conduzida com facilidade por usuários inexperientes. No entanto, pode haver questões técnicas que impeçam a fácil instalação. Nesses casos, um serviço de suporte à instalação deve ser oferecido aos usuários (Rose, 1994). Outra solução é disponibilizar o sistema remotamente por meio da web. Desse modo, o sistema pode ser mantido por uma equipe técnica especializada, dispensando assim, a instalação do sistema por usuários regulares.

### ***Beleza estética da interface gráfica***

A ideia de que a beleza não é apenas uma característica efêmera, mas que deve servir aos bons propósitos, já era defendida pelo filósofo Platão (Reber & Topolinski, 2010). Mas afinal, o que é beleza? Beleza, assim como qualidade, é um conceito idiossincrático, ou seja, é definido em última instância por um indivíduo ou um grupo de indivíduos (Rose, 1994; Reber *et al.*, 2004). Entretanto, existem alguns aspectos relativos à beleza que podem ser utilizados para definir diretrizes para a construção de interfaces gráficas relativamente belas.

Muitos teóricos veem a beleza como uma propriedade de um objeto que produz uma experiência agradável a um observador (Reber *et al.*, 2004). Esse conceito inspirou psicólogos a investigar quais são as características que contribuem para que as pessoas tenham a percepção de que algo é belo. Entre as características identificadas estão: equilíbrio e proporção, simetria, conteúdo informativo e complexidade, além de contraste e a nitidez (Reber *et al.*, 2004). Mas, como essas características são percebidas pelos usuários?

Uma teoria sobre a percepção estética considera que as pessoas consideram uma obra especialmente bela se ela é de fácil apreciação. Esse fenômeno é denominado pelos psicólogos de *processing fluency* (fluência do processamento) (Reber & Topolinski, 2010).

Pesquisas mostraram que objetos de fácil legibilidade, ou seja, que podem ser processados fluentemente e que se combinam de maneira harmoniosa criam uma predisposição mental de aceitação, despertando um sentimento positivo no observador (Reber & Topolinski, 2010).

Portanto, construir interfaces gráficas com cores e componentes harmoniosos, simétricos e de fácil apreciação pode contribuir para o aumento da aceitabilidade de novos SI de digitalização de ocorrências de espécies.

### ***Privacidade***

Alguns tipos de dados que podem ser sensíveis e, por motivos estratégicos, não podem ser publicados. Desse modo, é necessário permitir aos usuários indicarem quais dados podem ou não ser publicados.

Outro fator que deve ser considerado é que cada ação de criação, alteração e exclusão de registro deve ser registrada em um *log*. Visto que haverá vários usuários trabalhando com os mesmos dados, esse tipo de controle pode ser importante para resolver possíveis conflitos.

### ***Disponibilidade***

Disponibilidade, no contexto de SI de gestão de dados de ocorrência de espécies, refere-se a quando, onde e como o sistema estará disponível para o uso. Os produtores de dados, normalmente, coletam dados em campo, ou seja, na natureza. Para garantir a disponibilidade do sistema nesse contexto, uma solução é disponibilizar aos usuários um sistema para dispositivos móveis que permita a coleta de dados em campo e que, posteriormente, esses dados possam ser sincronizados com o SI principal.

Sistemas *web* acessíveis via Internet também pode aumentar a disponibilidade do SI para equipes geograficamente distantes.

### ***Suporte a sincronização de planilhas eletrônicas***

Suporte a sincronização de planilhas eletrônicas pode ser importante para os usuários por pelo menos dois aspectos:

- **Produtividade:** reaproveitar os dados anteriormente digitalizados em planilhas eletrônicas, visto que a redigitalização dessas informações pode ser muito custosa;
- **Disponibilidade:** permitir a manipulação dos dados em lugares onde não há acesso a Internet.

Apesar de o suporte a sincronização de planilhas eletrônicas propiciar uma provável redução da QD, visto que a digitalização ocorre sem ferramentas que dão suporte a prevenção a erros, esse recurso pode ser importante para que os usuários venham a ter um primeiro contato com um SI com suporte a QD. Sem esse recurso, a migração dos dados previamente digitalizados em planilhas teria que ser feita registro a registro, desperdiçando um tempo que poderia ser utilizado para produção de novos dados.

### ***Auxílio à tomada de decisões durante a digitalização***

Esse requisito é intrínseco à QD em SI de biodiversidade. Auxiliar os usuários a tomar decisões corretas no processo de digitalização e manipulação de dados de biodiversidade pode ajudar na prevenção de problemas de QD.

Esse requisito pode ser implementado de diversas maneiras, desde um simples recurso de *autocomplete* em determinados campos até complexas ferramentas de validação ou de georeferenciamento.

### ***Flexibilidade***

Com o objetivo de permitir o intercâmbio e a interoperabilidade de dados de biodiversidade num contexto global, têm sido propostos padrões de esquemas de metadados de biodiversidade, como *DwC*, ABCD entre outros. Naturalmente, o uso desses padrões em sistemas de gestão de dados de biodiversidade é recomendado e incentivado. Contudo, a delimitação de quais dados devem ser ou não digitalizados pode ter um impacto negativo para alguns pesquisadores. Algumas pesquisas podem exigir a digitalização de dados específicos que não estão

disponíveis nos esquemas de metadados de maneira explícita. Esse fator pode ser decisivo para a aceitação ou não do usuário a um determinado SI.

Assim, é necessário que o sistema permita ao usuário incrementar atributos específicos ao formulário de cadastro de maneira dinâmica e interativa, assim como ele faria em uma planilha eletrônica ao incrementar uma coluna. Para manter o sistema coerente com esquema de metadados, atributos como o *Dynamic Properties* do *Darwin Core*, por exemplo, podem ser utilizados na implementação desse recurso.

### **Adequação ao uso dos dados**

Assim como o requisito de auxílio à tomada de decisão, esse requisito também é intrínseco à QD. A QD pode ser definida como *fitness-for-use* (adequação ao uso), ou seja, os dados são de qualidade se eles forem úteis ao uso (Dalcin, 2005; Chapman, 2005b). Assim, é necessário haver indicadores que permitam identificar aspectos da qualidade, como completude, consistência, precisão, acurácia, entre outros. Desse modo, quando os usuários forem utilizar os dados eles poderão identificar se os dados são adequados, ou não, ao uso naquela aplicação específica (Hill *et al.*, 2010). Esse requisito afeta principalmente os consumidores dos dados, como os especialistas em processamento de dados e os especialistas em biodiversidade.

### **Suporte a impressão**

O recurso de geração e impressão de relatórios de produtividade e de QD é de grande importância para as instituições, pois permitem o acompanhamento do trabalho da equipe. Outro recurso importante nesse sentido é o suporte a impressão de etiquetas para espécimes preservadas em coleções. Em muitos casos, os espécimes são capturados e levados para fazer parte de uma coleção. Nesse contexto, o curador é responsável por manter a coleção organizada e consistente com o banco de dados. Para facilitar esse trabalho, o suporte a impressão de etiquetas para identificação dos espécimes coletados é de grande importância para os curadores.

### **5.2.2. Ferramentas de QD desenvolvidas**

Com base nos estudo de Avaliação e Gerenciamento da QD e na arquitetura do sistema do BDD, foram implementadas duas ferramentas para a prevenção a erros em dados de ocorrências de espécies. Cada ferramenta foi implementada como um módulo independente e acoplado ao formulário de cadastro de ocorrências de espécies.

#### **5.2.2.1. *BDD Taxon Tool***

A QD no domínio de dados taxonômicos está fortemente ligada à conformidade de nomes e de hierarquias de táxons a um determinado padrão, seja esse padrão reconhecido internacionalmente ou utilizado somente internamente por uma instituição. Portanto, para tentar melhorar a QD taxonômicos, está em desenvolvimento no SI BDD uma ferramenta *web*, denominado *BDD Taxon Tool – BTT* (Veiga *et al.*, 2011a, b, c). Essa ferramenta tem o objetivo de auxiliar os usuários a preencherem dados taxonômicos livres de erros.

Essa ferramenta é composta por sete recursos: (1) sugestões de nomenclaturas baseado no banco de dados local, (2) sugestões de hierarquias taxonômicas baseada no banco de dados local, (3) sugestões de nomenclaturas baseado em autoridades taxonômicas, (4) sugestões de hierarquias taxonômicas baseadas em autoridades taxonômicas, (5) auxílio à tomada de decisão na identificação taxonômica, (6) suporte a morfoespécies e (7) indicador de incerteza.

Ao utilizar o BTT, um campo texto é apresentado ao usuário. À medida que os caracteres vão sendo inseridos nesse campo, sugestões de nomes taxonômicos, de todos os níveis hierárquicos, são apresentadas, como ilustrado na Figura 12.

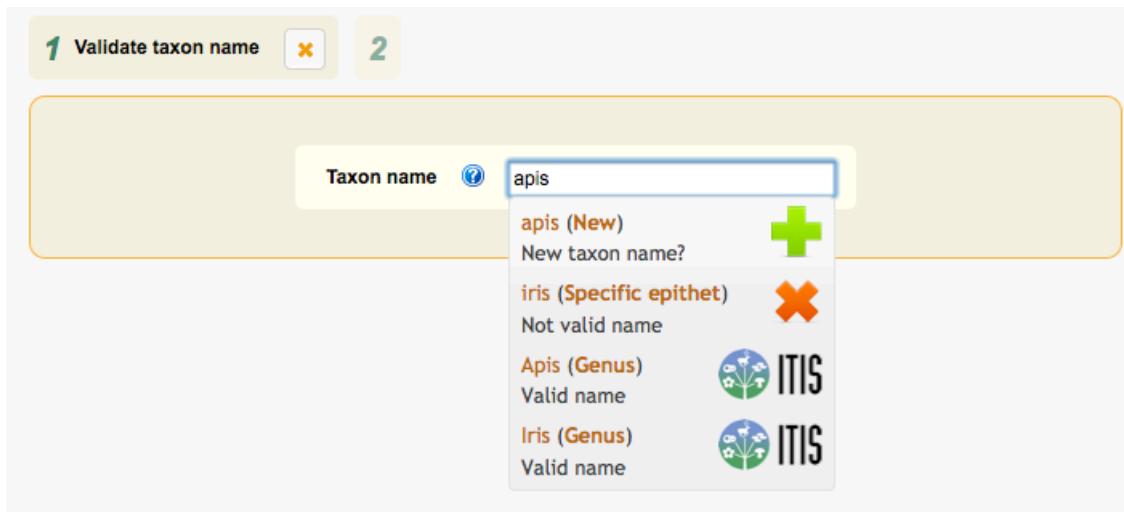


Figura 12 – *Autocomplete* de nomes de táxons.

Com esse recurso de *autocomplete*, as sugestões são recuperadas do banco de dados local da instituição utilizando uma implementação de *Fuzzy Matching* do PostgreSQL (Wagner & Fischer, 1974).

Os nomes recuperados do banco de dados local podem estar válidos ou inválidos em relação ao CoL. Se for selecionado um táxon inválido, uma segunda consulta é realizada ao banco de dados do CoL, conforme Figura 13. Também utilizando *Fuzzy Matching*, a ferramenta sugere nomes de táxons válidos de acordo com a autoridade taxonômica CoL.

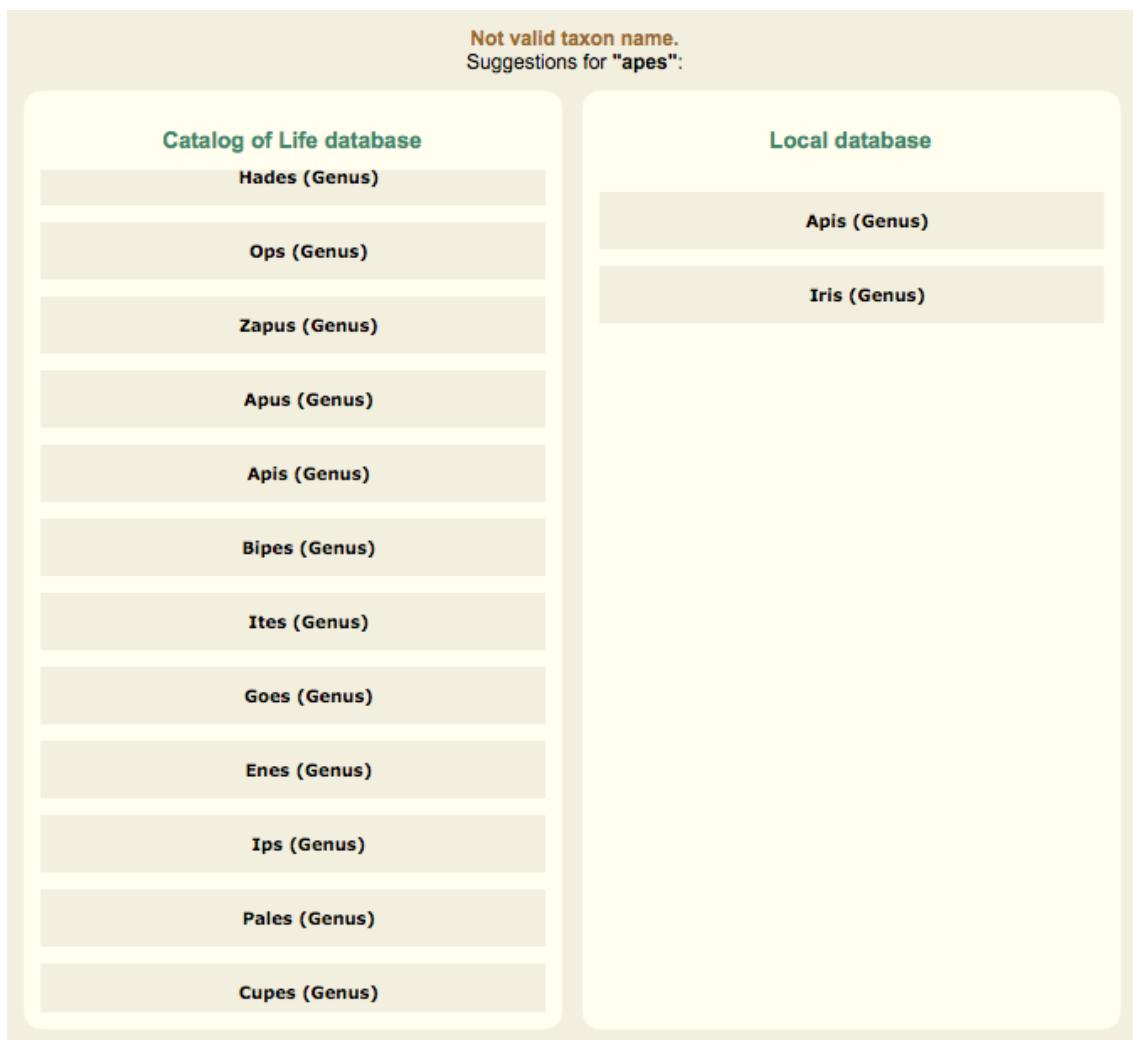


Figura 13 – Sugestões de nomes válidos.

Ao selecionar um nome, são apresentadas ao usuário todas as hierarquias taxonômicas distintas que possuem na sua composição o nome selecionado. Essas hierarquias são consultadas no banco de dados local e no banco de dados do CoL, como apresentado na Figura 14.

Catalog of Life database	Local database
<b>Valid according to Catalog of Life.</b> <b>Kingdom:</b> Animalia <b>Phylum:</b> Arthropoda <b>Class:</b> Insecta <b>Order:</b> Hymenoptera <b>Family:</b> Apidae <b>Genus:</b> Apis <b>Subgenus:</b> _____ <b>Specific epithet:</b> _____ <b>Infraspecific epithet:</b> _____ <b>Scientific name:</b> _____	<b>Not valid - 7 records using this hierarchy.</b> <b>Kingdom:</b> Animalia <b>Phylum:</b> Arthropoda <b>Class:</b> Insecta <b>Order:</b> Hymenoptera <b>Family:</b> Apoidea <b>Genus:</b> Apis <b>Subgenus:</b> _____ <b>Specific epithet:</b> _____ <b>Infraspecific epithet:</b> _____ <b>Scientific name:</b> _____
	<b>Not valid - 5 records using this hierarchy.</b> <b>Kingdom:</b> animalia <b>Phylum:</b> Arthropoda <b>Class:</b> Insecta <b>Order:</b> Hymenoptera <b>Family:</b> Apidae <b>Genus:</b> Apis <b>Subgenus:</b> _____ <b>Specific epithet:</b> _____ <b>Infraspecific epithet:</b> _____ <b>Scientific name:</b> _____
	<b>Not valid - 4 records using this hierarchy.</b> <b>Kingdom:</b> AniMalla <b>Phylum:</b> Arthropoda <b>Class:</b> Insecta <b>Order:</b> Hymenoptera <b>Family:</b> Apidae <b>Genus:</b> Apis <b>Subgenus:</b> _____

Figura 14 – Sugestões de hierarquias válidas e inválidas.

Esses recursos foram projetados para serem utilizados segundo o diagrama representado pela Figura 15.

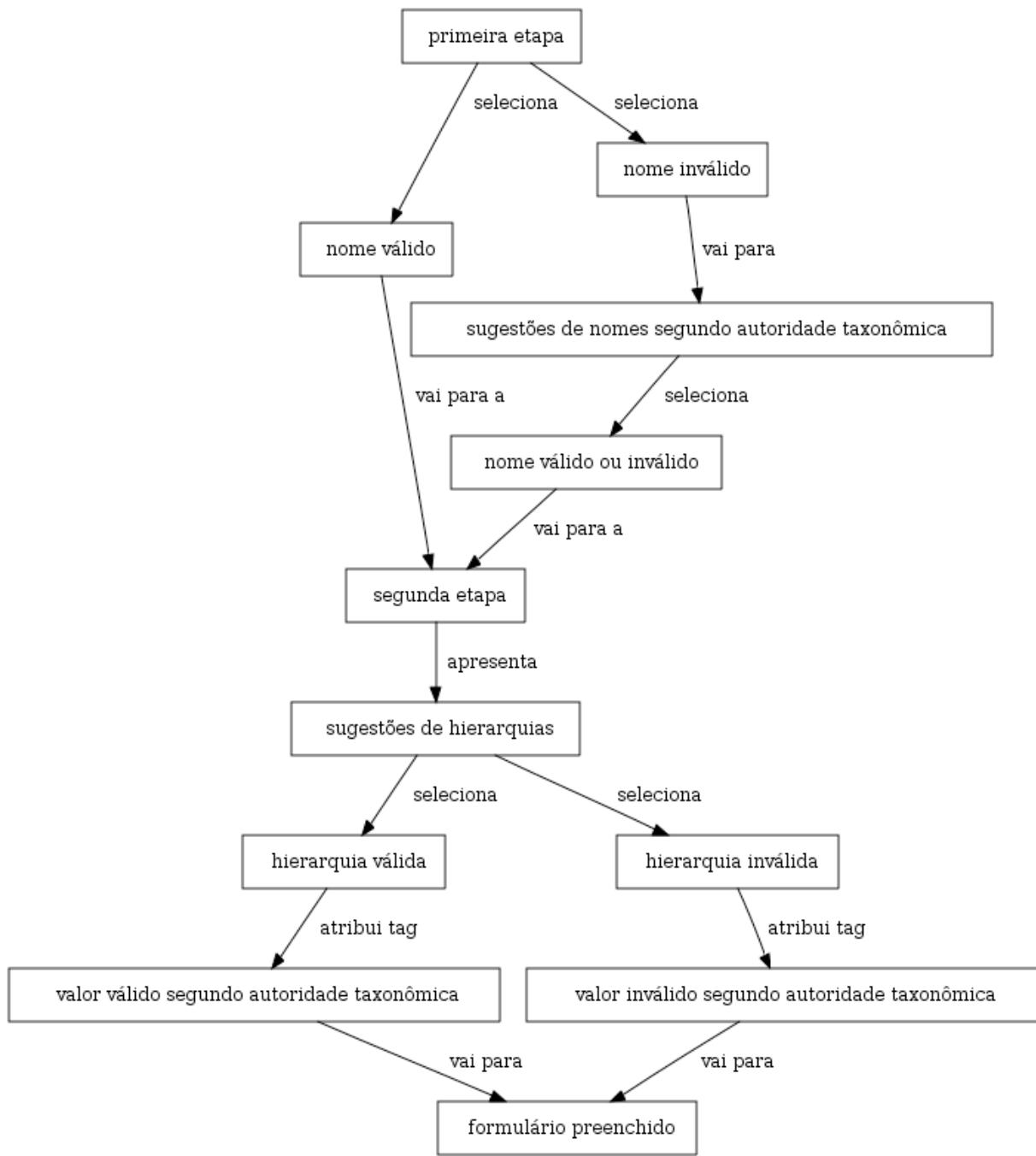


Figura 15 - Sequência de uso da ferramenta BTT.

Os próximos recursos projetados para o BTT estão em fase de desenvolvimento. Esse recursos estão relacionados à atividade de identificação de espécimes. Visto que essa atividade está fortemente ligada ao conhecimento sobre táxons específicos, a ferramenta permitirá obter informações úteis relativas a esses táxons, as quais podem auxiliar na identificação de um espécime.

Por exemplo, caso haja dúvidas sobre um espécime ser ou não da espécie *Tetragonisca angustula*, a BTT permitirá ao usuário obter informações bibliográficas e recursos multimídia, como fotos e chaves taxonômicas relacionadas ao táxon em questão. Essas informações são previamente cadastradas e relacionadas aos táxons no SI BDD. Também são consultados informações e recurso multimídias do banco de dados do EoL (EOL, 2011), por meio de *web services* disponibilizados na Internet.

Também foi projetado e está sendo implementado no BTT, um suporte a morfoespécies. Quando o usuário digitar a sequência de caracteres “sp” seguido por um número, por exemplo, “sp1”, “sp5”, no campo de táxon, o sistema automaticamente identificará esse táxon como uma morfoespécie. Posteriormente, quando uma dessas morfoespécies for identificada, os outros registros com os mesmos identificadores serão automaticamente alterados.

O último recurso da BTT permite ao usuário reportar o grau de incerteza em relação a uma identificação. Esse indicador de incerteza permite ao usuário indicar que a informação cadastrada tem uma probabilidade de estar incorreta, necessitando assim de uma validação de um especialista.

Para a implementação dessa ferramenta, foram utilizadas as linguagens de programação PHP (PHP, 2011) e Javascript (JAVASCRIPT, 2011). Os dados do CoL são atualizados periodicamente e mantidos em um banco de dados em um servidor dedicado. Os dados obtidos do EoL são recuperados via um *web service* disponibilizado por eles. Até o momento foram implementados no BDD os quatro primeiros recursos, (1) sugestões de nomenclaturas baseado no banco de dados local, (2) sugestões de hierarquias taxonômicas baseada no banco de dados local, (3) sugestões de nomenclaturas baseado em autoridades taxonômicas e (4) sugestões de hierarquias taxonômicas baseadas em autoridades taxonômicas.

#### **5.2.2.2. BDD Geo Tool**

Para a redução de erros em dado geoespaciais e de localização de ocorrências de espécies, o BDD implementa uma ferramenta denominado BDD Geo Tool – BGT (Veiga *et al.*, 2010; Veiga *et al.*, 2011a, b, c). Essa ferramenta está organizada em três etapas: (1) inserir dados primários, (2) selecionar informações e fontes de dados e (3) reportar incerteza.

Na primeira etapa, o usuário pode escolher entre três tipos de dados primários sobre o geoposicionamento da ocorrência. Conforme a Figura 16, o usuário pode inserir as coordenadas geoespaciais, usar um mapa interativo para obter uma localização aproximada conhecida ou utilizar uma descrição textual da localização da ocorrência.

The screenshot shows a user interface titled "Location Elements". On the left, a vertical sidebar lists categories: Record-level, Taxonomic, Location, Occurrence, Identification, Event, Media, and Reference. The "Location" category is selected. At the top, there are two buttons: "Tool BDD Georeferencing Tool" (highlighted in brown) and "Form". Below these are three numbered steps: 1. What kind of data do you have? (selected), 2. An approximate location, and 3. A recorded locality name. Step 1 contains fields for "Latitude" and "Longitude" with "Rev. Georeferencing" and "Georeferencing using Map" buttons. Step 2 features a globe icon. Step 3 contains fields for "Locality", "Context", "Country", and "State", with a "Georeferencing" button.

Figura 16 – Primeira etapa da ferramenta BGT.

Caso sejam conhecidas as coordenadas geoespaciais, o usuário pode inseri-las no local indicado e realizar um georeferenciamento reverso. Esse recurso permite obter os nomes da cidade, do estado, do país e a altitude relativa às coordenadas informadas. Essas informações são obtidas a partir de duas fontes de dados distintas: do Google Maps e do banco de dados geoespaciais do GeoNames (GeoNames, 2011). Na segunda etapa o usuário pode escolher qual fonte de dados será utilizada, conforme a Figura 17.

Please wait while data are retrieving from data sources.  
This may take a few seconds depending on the each data source.

 Google	 GeoNames	 Selected
<b>Country</b> Brazil	<b>GeoNames</b> Brazil	<b>Brazil</b>
<b>State or Province</b> Santa Catarina	<b>GeoNames</b> Santa Catarina	<b>Santa Catarina</b>
<b>Municipality</b> Itajaí	<b>GeoNames</b> Cabeçudas	<b>Itajaí</b>
<b>Geodetic Datum</b> WGS84	<b>GeoNames</b> WGS84	<b>WGS84</b>
<b>Select</b>	<b>Select</b>	<b>Valid according to:</b> <b>Google</b>

Figura 17 – Georeferenciamento reverso.

Caso as coordenadas geoespaciais não sejam conhecidas, o usuário pode utilizar um mapa interativo em três dimensões para obter a latitude e a longitude por meio de um clique sobre a localização desejada no mapa, conforme ilustrado na Figura 18. Ao selecionar uma localização aproximada, o recurso de georeferenciamento reverso é executado.

O usuário pode também utilizar uma descrição textual da localização, como “*Bariloche, 25 km NNE via Ruta Nacional 40 (=Ruta 237)*”, por exemplo, para obter as informações geoespaciais e de localização. Esse recurso de georeferenciamento utiliza um *web service* projetado pelo projeto BioGeomancer (BioGeomancer, 2011) e disponibilizado pela Universidade de Berkeley para obter um conjunto de possíveis coordenadas geoespaciais a partir da descrição da localidade, conforme ilustrado na Figura 19.

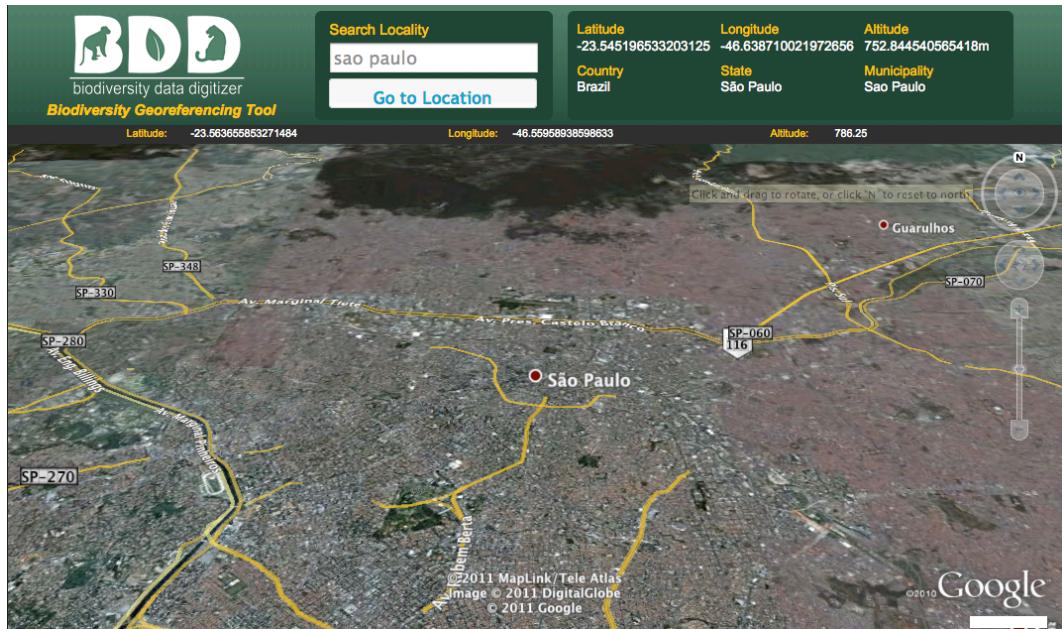


Figura 18 – Georeferenciamento utilizando um mapa interativo tridimensional.

**BioGeomancer**

**Biogeomancer**

Latitude	-23.6248414
Longitude	-46.5952706
Country	Brazil
State or Province	São Paulo
Municipality	Sao Paulo
Geodetic Datum	WGS84
Uncertainty in Meters	86363.0

**Select**

**BioGeomancer**

**Biogeomancer**

Latitude	-23.5333338
Longitude	-46.6166668
Country	Brazil
State or Province	São Paulo
Municipality	Sao Paulo
Geodetic Datum	WGS84
Uncertainty in Meters	4532.0

**Select**

Figura 19 – Georeferenciamento a partir da descrição “near sao paulo”.

Após a seleção das coordenadas geoespaciais e das informações geográficas, a ferramenta permite ao usuário reportar o nível de incerteza/erro em relação às informações geoespaciais, na terceira etapa. Essa incerteza é reportada em metros e pode ser visualizada em um mapa, de acordo com a ilustração na Figura 20. Por fim, os dados selecionados são utilizados para preencher o formulário de ocorrências de espécies do BDD.



Figura 20 – Indicador de incerteza.

Esses recursos foram projetados para serem utilizados segundo o diagrama representado pela Figura 21.

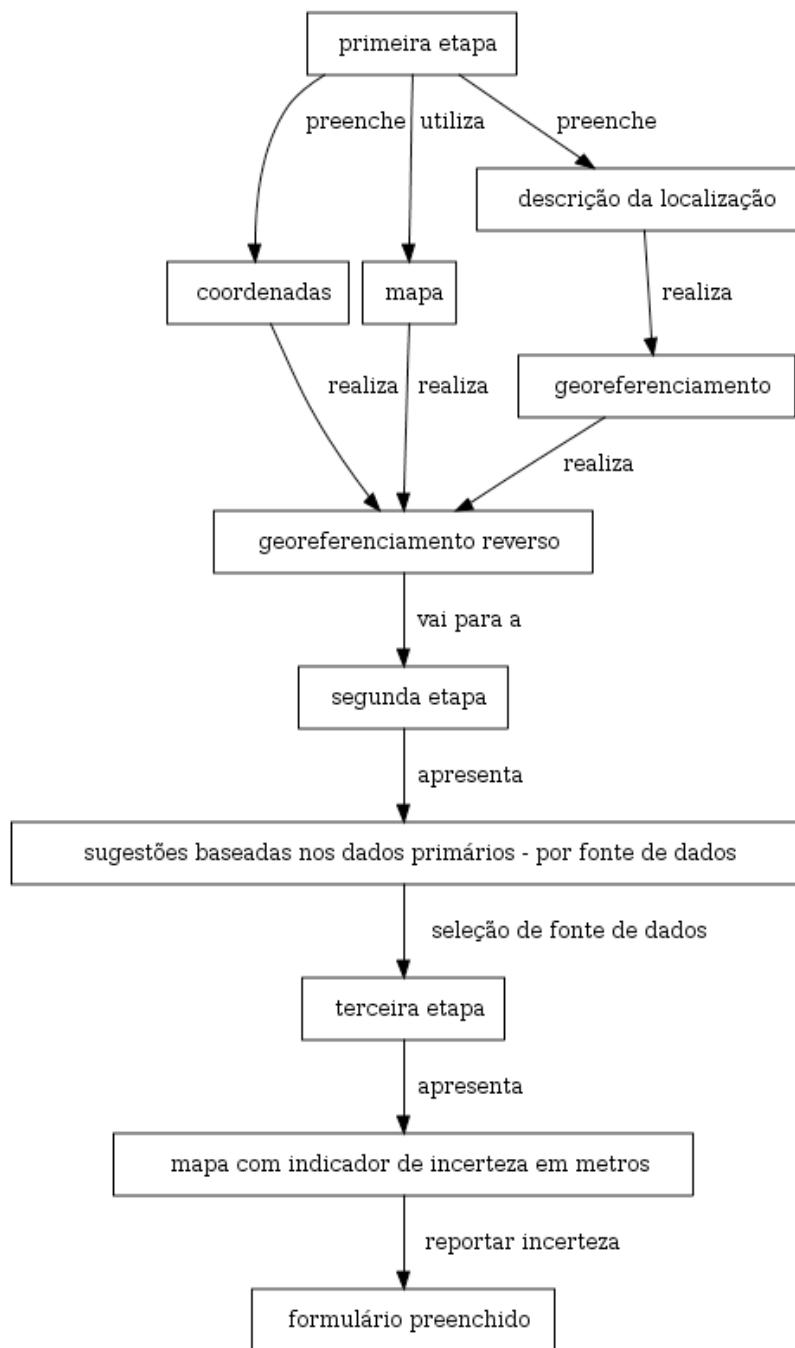


Figura 21 - Sequência de uso da ferramenta BGT.

Para a implementação dessa ferramenta foram utilizadas as linguagens de programação PHP (PHP, 2011) e Javascript (JAVASCRIPT, 2011). Para a renderização do mapa tridimensional foi utilizado a Google Earth API, sendo, desse

modo, necessária a instalação do plug-in do Google Earth no navegador para utilizar a ferramenta.

## **6. CONSIDERAÇÕES FINAIS**

Neste último capítulo são apresentadas as principais contribuições deste trabalho e os trabalhos futuros que podem ser realizados a partir dele.

### **6.1. Contribuições**

Como contribuição deste trabalho, destaca-se a metodologia utilizada para realizar o estudo da QD de ocorrências de espécies, pois ela pode ser usada em outras pesquisas sobre QD aplicada em outros domínios de aplicação, como em dados médicos, dados financeiros ou outros tipos de dados de biodiversidade, por exemplo.

Outra contribuição foi o resultado do estudo sobre QD aplicada a um importante tipo de dado sobre a biodiversidade: dados de ocorrências de espécies. Esse estudo foi dividido em duas partes: estudo sobre a Avaliação da QD e estudo sobre o Gerenciamento da QD.

Como resultado do estudo sobre a Avaliação da QD de ocorrências de espécies, foi identificado um conjunto de problemas de QD, os quais foram contextualizados em relação aos domínios de dados de localização, geoespaciais e taxonômicos. Também foi identificado um conjunto de dimensões de QD importantes no contexto de dados de ocorrência de espécies. Visto que cada dimensão pode possuir um significado diferente em relação aos domínios de dados (Dalcin, 2005), a definição dessas dimensões, em relação domínios de dados, escopo deste trabalho, foi realizada. Ainda no estudo da Avaliação da QD, foi identificado como os problemas, dimensões e domínios de dados se relacionam, a fim de definir um modo de avaliar a QD. Desse modo, o resultado desse estudo pode ser utilizado por outros pesquisadores para identificar quais problemas de QD devem ser reduzidos para melhorar a QD em determinadas dimensões e em determinados domínios de dados. Portanto, esse estudo contribui para o desenvolvimento de estratégias, políticas e ferramentas para melhorar a QD, por meio da identificação de quais tipos de erros devem ser evitados ou corrigidos para melhorar determinados aspectos da QD.

A segunda parte do estudo sobre QD está relacionada ao melhoramento da QD, ou seja, o Gerenciamento da QD. Nessa parte do estudo foram identificados

recursos computacionais que se implementados em um SI sobre ocorrências de espécies, pode auxiliar na redução de problemas por meio da prevenção a erros durante a digitalização informações sobre ocorrências de espécies. Esse estudo foi baseado na Avaliação da QD e demonstra como os recursos identificados se relacionam com os problemas e as dimensões de QD. Portanto, esse estudo também pode ser utilizado como diretriz para o desenvolvimento de SI sobre ocorrências de espécies com suporte a QD.

São contribuições também as duas ferramentas desenvolvidas com base no estudo sobre o Gerenciamento da QD, implementadas no SI BDD. Com o auxílio dessas ferramentas, BTT e BGT, a QD de ocorrências de espécies digitalizadas por meio do BDD pode ser melhorada, sendo, portanto, uma importante contribuição para os pesquisadores que utilizam os dados digitalizados.

É uma contribuição a evolução do SI que abriga as ferramentas citadas, o BDD. Com a reestruturação do SI, quanto a codificação, IHM, arquitetura de software e banco de dados, o BDD tornou-se uma ferramenta de digitalização de dados de ocorrências de espécies com suporte a QD com boa aceitação, conforme *feedback* em apresentações dos SI no TDWG *Annual Conference* 2011 e em reuniões com integrantes da IABIN-PTN. Por ser de código e de uso aberto, qualquer interessado pode utilizar o BDD para a digitalização de seus dados.

Por fim, a análise de requisitos de SI de ocorrências de espécies é uma importante contribuição. A identificação e análise desses requisitos teve por objetivo listar e descrever características que tornasse o SI mais atrativo aos usuários em relação as, frequentemente utilizadas, planilhas eletrônicas. Visto que a abordagem de Gerenciamento da QD utilizada neste trabalho foi a de prevenção a erros e, portanto, a melhora da QD só ocorre quando o SI é utilizado para a digitalização de informações sobre ocorrências de espécies, essa análise foi um estudo particularmente importante neste trabalho. Essa análise também pode ser utilizada por outros pesquisadores envolvidos com o projeto e o desenvolvimento de SI para digitalização de informações de ocorrências de espécies.

## **6.2. Conclusões**

O estudo sobre a Avaliação da QD, aplicado a dados de ocorrências de espécies, demonstrou que a QD, em determinadas dimensões e em relação aos domínios de dados de localização, geoespaciais e taxonômicos, é degradada pela incidência de determinados problemas e, portanto, a QD pode ser avaliada por meio da presença ou ausência desses problemas.

Com base no estudo da Avaliação da QD, conclui-se também que a QD pode ser melhorada em determinadas dimensões por meio da redução de erros específicos.

O estudo sobre o Gerenciamento da QD em SI sobre ocorrências de espécies demonstrou que determinados recursos computacionais, se implementados em um SI, são capazes de proporcionar uma redução de determinados problemas de QD por meio da prevenção a erros.

Assim, com a implementação no SI dos recursos identificados no estudo sobre o Gerenciamento da QD, possibilitou que houvesse uma redução de determinados erros e, conforme apresentado no estudo da Avaliação da QD, a QD de ocorrência de espécies é melhorada em determinadas dimensões nos domínio de dados de localização, geoespaciais e taxonômicos.

## **6.3. Trabalhos futuros**

Como trabalhos futuros sugere-se realizar o estudo de Avaliação e de Gerenciamento da QD aplicados aos demais domínios de dados do DwC, como os domínio de dados de evento e de nível de registro, por exemplo, e a outros tipos de dados de biodiversidade, como Interação entre Espécies, Espécies, Monitoramento de Espécies, entre outros.

É previsto, como trabalho futuro, finalizar e otimizar as ferramentas de QD propostas neste trabalho (BGT e BTT) e implementar no BDD todos os requisitos identificados para melhorar aceitabilidade dos usuários à BDD. Posteriormente, realizar uma pesquisa com usuários do BDD para avaliar estatisticamente a aceitação dos usuários ao SI e as ferramentas de QD.

Um importante trabalho futuro consiste em realizar um estudos sobre Gerenciamento da QD utilizando a abordagem de detecção e correção de erros. Ou

seja, com base no estudo da Avaliação da QD identificar técnicas e recursos que possam ser utilizados para detectar e corrigir erros em bases de dados de ocorrências de espécies.

A implementação de uma metodologia para a quantificação da QD em cada dimensão é um trabalho a ser implementado futuramente, o qual poderá trazer grandes contribuições para comunidade científica, pois permitirá realizar a Avaliação da QD de maneira mais objetiva.

## REFERÊNCIAS

ALA. **Atlas of Living Australia.** Disponível em: <http://www.ala.org.au>. Acesso em: 06 dez. 2011.

BIOGEOMANCER. **BioGeomancer.** Disponível em: <http://www.biogeomancer.org>. Acesso em: 06 ago. 2011.

BISBY, F. A. The quiet revolution: biodiversity informatics and the internet. *Science*, v. 289, n. 5488, p. 2309-2312, 2000.

BOLT, A.; MAZUR, G. H. Jurassic QFD: integration service and product quality function deployment. In: **The Eleventh Symposium on Quality Function Deployment**. Novi, Michigan, 1999.

BRUNDTLAND, G. H. Our Common Future. **Oxford University Press**, Oxford, p. 15-22, 1987.

CANHOS, V. P. Informática para biodiversidade: padrões, protocolos e ferramentas. **Ciência e Cultura**, 55, p. 45–47, 2003.

CANHOS, V. P.; SOUZA, R.; CANHOS, D. A. L. Global biodiversity informatics: setting the scene for a "New World" of ecological modeling. **Biodiversity Informatics**, v. 1, p. 1-13, 2004.

CARTOLANO, E. A. **Proposta de um sistema de informação orientado a serviços sobre a biodiversidade de abelhas.** 2009. Dissertação de Mestrado – Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica, Universidade de São Paulo, São Paulo, 2009.

CARTOLANO, E. A.; SARAIVA, A. M.; VEIGA, A. K.; KROBATH, D. B.; SARAIVA, L. G. P.; TAVARES, G. Biodiversity Data Digitizer. In: **The Proceedings**

**of TDWG: Provisional Abstracts of the 2010 Annual Conference of the Taxonomic Databases Working Group.** Woods Hole, USA, 2010.

**CBD. Convention on Biological Diversity.** Disponível em:  
<http://www.biodiv.org>. Acesso em: 06 dez. 2011.

**CHAPMAN, A. D.** Uses of Primary Species-Occurrence Data. **Report for the Global Biodiversity Information Facility**, v. 1.0, Copenhagen. 2005a.

\_\_\_\_\_. Principles and Methods of Data Cleaning – Primary Species and Species. **Report for the Global Biodiversity Information Facility**, v. 1.0, Copenhagen. 2005c.

\_\_\_\_\_. Principles of Data Quality. **Report for the Global Biodiversity Information Facility**, v. 1.0, Copenhagen. 2005b.

CHEN, B.; WANG, B.; ZHENG, C.; HU, X. Research and Implementation of Information Quality Improvement. In: **Proceedings** of Cooperation and Promotion of Information Resources in Science and Technology, p. 255-229, 2009.

**COL. Catalogo of Life.** Disponível em: <http://www.catalogoflife.org>. Acesso em: 06 ago. 2011.

CROSBY, P. B. Quality Without Tears. **McGraw-Hill Book Company**. New York. 1984.

DALCIN, E. C. **Data Quality Concepts and Techniques Applied to Taxonomic Databases**. 2005. Tese de Doutorado de Filosofia – School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton, Southampton, England, 2005.

**DWC. Darwin Core Terms: A quick reference guide.** Disponível em: <http://rs.tdwg.org/dwc/terms/>. Acesso em: 06 ago. 2011.

EMBURY, S. M. Data quality issues in information systems. Database Systems Cardiff University, School of Computer Science, Cardiff, p. 41, 2001.

ENGLISH, L. P. Improving data warehouse and business information quality: methods for reducing costs and increasing profits. **John Wiley & Sons, Inc.**, New York, 1999.

EOL. **Encyclopedia of Life**. Disponível em: <http://www.eol.org>. Acesso em: 06 ago. 2011.

FEC. Developing a user-centered voting system. Technical report. **Federal Election Commission**. 2003.

GBIF. **Global Biodiversity Information Facility**. Disponível em: <http://www.gbif.org>. Acesso em: 06 ago. 2011.

GE, M.; HELFERT, M. A review of information quality research-develop a research agenda. In: **Proceedings** of the 12th International Conference on Information Quality. 2007.

GEOLOCATE. **GeLocate: A platform for georeferencing natural history collections data**. Disponível em: <http://www.museum.tulane.edu/geolocate>. Acesso em: 06 dez. 2011.

GEONAMES. **GeoNames – Geographical Database**. Disponível em: <http://www.geonames.org>. Acesso em: 06 ago. 2011.

SARAIVA A.M., CANHOS, D.A.L. Sistemas de informação e ferramentas computacionais para pesquisa, educação e disseminação do conhecimento sobre polinizadores. In: **Polinizadores no Brasil - contribuição e perspectivas para a biodiversidade, uso sustentável, conservação e serviços ambientais**.

(IMPERATRIZ-FONSECA V.L., CANHOS D.A.L., ALVES D.A., SARAIVA A.M., eds), São Paulo, SP: EDUSP.

GOOGLEMAPS. **Família da Google Maps API.** Disponível em: <http://code.google.com/intl/pt-BR/apis/maps/index.html>. Acesso em: 06 dez. 2011.

GROUP, I. Introduction to Database Management Systems. **McGraw-Hill Education (India) Pvt Ltd**, 2005.

HAWKSWORTH, D. L. Biodiversity: measurement and estimation. **Chapman & Hall in association with The Royal Society**, 1996.

HILL, A. W.; OTEGUI, J.; ARIÑO, A. H.; GURALNICK, R. P. GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for- Use Across the GBIF Network. **Report for the Global Biodiversity Information Facility**, v 1.0, Copenhagen, 2010.

IABIN. **Inter-American Biobiversity Network.** Disponível em: <http://www.iabin.net> . Acesso em: 03 ago. 2011.

ITIS. **Integrated Taxonomic Information System.** Disponível em: <http://www.itis.gov>. Acesso em: 06 ago. 2011.

JAVASCRIPT. **Javascript.** Disponível em: <http://www.w3schools.com/js/default.asp>. Acesso em: 06 ago. 2011.

KELLING, S. Significance of organism observations: Data discovery and access in biodiversity research. **Report for the Global Biodiversity Information Facility**, 2008.

MA. Ecosystems and human well-beings: Biodiversity Synthesis. **Millennium Ecosystem Assessment Report**. World Resources Institute. Washington, DC. 2005.

MAZUR, G. H. QFD for service industries: from voice of customer to task deployment. In: **The Fifth Symposium on Quality Function Deployment**. Novi, Michigan. 1993.

MCGILVRAY, D. Executing data quality projects: ten steps to quality data and trusted information. **Morgan Kaufmann, Elsevier**, 2008.

PHP. **PHP**. Disponível em: <http://www.php.net> . Acesso em: 18 mar. 2011.

PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. Data quality assessment. **Transactions on Communications of the ACM**. 45, n. 4, p. 211-218, 2002.

POSTGRESQL. **Documentação do PostgreSQL**. Disponível em: <http://pgdocptbr.sourceforge.net/pg74/ddl-constraints.html>. Acesso em: 06 dez. 2011.

REBER, R.; SCHWARZ, N.; WINKIELMAN, P. Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? **Personality and Social Psychology Review**. p. 364-382. 2004.

REBER, R.; Topolinski, S. Simples + belo = correto: sera? **Mente e Cérebro, Scientific American**. p. 60-65. 2010.

REDMAN, T. C. Data quality: the field guide. **Digital Press**. Newton, MA, USA, 2000.

ROSE, P. Quality in services and services in quality. **Customer Driven Quality in Product Design, IEEE Colloquium on**. p. 1–6, 1994.

SARAIVA, A. M. **Tecnologia da informação na agricultura de precisão e biodiversidade: estudos e proposta de utilização de Web Services para desenvolvimento e integração de sistemas**. 2003. Tese (Livre Docência) –

Departamento de Engenharia de Computação e Sistemas Digitais, Escola Politécnica da Universidade de São Paulo, São Paulo, 2003.

SCHNASE, J. L.; CUSHING, J.; SMITH, J. A. Biodiversity and ecosystem informatics. **Journal of Intelligent Information Systems**, v. 29, n. 1, p. 1-6, 2007.

SPECIES2000. **Species 2000**. Disponível em: <http://www.sp2000.org>. Acesso em: 06 ago. 2011.

STEINHAGE, V. **Automated identification of Bee Species in biodiversity information system**. Disponível em: [http://www.informatik.uni-bonn.de/~steinhag/stelleabis/abis\\_ui\\_200.pdf](http://www.informatik.uni-bonn.de/~steinhag/stelleabis/abis_ui_200.pdf). Acesso em: 22 jan. 2003.

STOCKWELL, D.R.B. **Overview of computational biodiversity research**. Publicado em 03/02/97. Disponível em: <http://biodi.sdsc.edu/doc/bis/overview.html> Acesso em 22 jan. 2007.

STRONG, D. M.; LEE, Y. W.; WANG, R. Y. Data quality in context. **Transactions on Communications of the ACM**. 40, n. 5, p. 103–110, 1997.

TDWG. **Biodiversity Information Standards**. Disponível em: <http://www.tdwg.org> . Acesso em 30 jul. 2011.

VEIGA, A. K.; CARTOLANO, E. A; SARAIVA, A. M. Data quality resources in Species occurrences digitization. In: **The Proceedings of TDWG: Provisional Abstracts of the 2011 Annual Conference of the Taxonomic Databases Working Group**. New Orleans, USA, 2011c.

VEIGA, A. K.; SARAIVA, A. M.; CARTOLANO, E. A. A georeferencing tool to improve biodiversity data quality. In: **The Proceedings of TDWG: Provisional Abstracts of the 2010 Annual Conference of the Taxonomic Databases Working Group**. Woods Hole, USA, 2010.

VEIGA, A. K.; SARAIVA, A. M.; CARTOLANO, E. A. Methods and tools to improve data quality in biodiversity specimens-occurrence data. In: **Proceedings of the World Congress on Computers In Agriculture of the European Federation for Information Technology in Agriculture, Food and the Environment**. Prague, Czech Republic, 2011a.

VEIGA, A. K.; SARAIVA, A. M.; CARTOLANO, E. A. Métodos e ferramentas de prevenção a erros em dados de ocorrências de espécies. In: **Proceedings of VIII Congresso Brasileiro de Agroinformática da Sociedade Brasileira de Agroinformática**. Bento Gonçalves, RS, Brasil, 2011b.

VERTNET. **VertNet**. Disponível em: <http://vertnet.org>. Acesso em: 06 dez. 2011.

WAGNER, R. A.; FISCHER, M. J. The String-to-String Correction Problem. **Journal of the ACM**. 21, n. 1, p. 168–176, 1974.

WAND, Y.; WANG, R. Y. Anchoring data quality dimensions in ontological foundations. **Transactions on Communications of the ACM**. 39, n. 11, p. 86–95, 1996.

WANG, R. Y.; KON, H. B.; MADNICK, S. E. Data quality requirements analysis and modeling. In: Data Engineering, 1993. **Proceedings**. Ninth International Conference on. p. 670–677, 1993.

WANG, R. Y.; REDDY, M. P.; KON, H. B. Toward quality data: An attribute-based approach. **Decision Support Systems**. 13, 3-4, p. 349-372, 1995.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: What data quality means to data consumers. **Journal of Management Information Systems**. 12, n. 4, p. 5–33, 1996.