APLICAÇÃO DA MINERAÇÃO DE DADOS NA MELHORIA DA QUALIDADE DE DADOS EM COLEÇÕES BOTÂNICAS

DATA MINING APPLICATION IN DATA QUALITY IMPROVEMENT IN BOTANICAL COLLECTIONS

Por:

Luís Alexandre Estevão da Silva
Caroline Corrêa Cabanillas

Resumo. *Bancos de dados de herbários auxiliam na geração de conhecimento sobre coleções científicas, provendo informações sobre taxonomia e biogeografia, entre outras áreas da pesquisa. Mas, por diversas razões, este tipo de banco de dados tem muitos problemas associados com a qualidade de dados. O presente trabalho descreve uma metodologia adotada para avaliar a qualidade de dados, identificando e classificando os tipos de erros encontrados nessa categoria de banco de dados. Ao final da metodologia um estudo foi realizado para indicar o método de limpeza de dados mais adequado de acordo com os erros encontrados. Foi então verificada a necessidade de uso dos recursos da mineração de dados, com a análise de associação, para o tratamento de alguns tipos de erros com um grau de dificuldade maior de identificação.*

Palavras-chave: qualidade de dados, dados primários da biodiversidade, bancos de dados de herbários, mineração de dados, coleções científicas.

Abstract. *Herbarium databases assist in the generation of knowledge about scientific collections, providing information on taxonomy and biogeography, among other areas of research. However, for various reasons, this type of database has many problems associated with data quality. This paper describes a methodology for assessing data quality, identifying and classifying the types of errors found in this database category. At the end of the methodology, was conducted a study to indicate the data most suitable cleaning method according to the errors found. It was then verified the need to use the capabilities of data mining, with the association analysis, for the treatment of some types of errors with a higher degree of difficulty of identification.*

Keywords: data quality, primary biodiversity data, herbarium database, data mining, scientific collections.

1. Introdução

Volume dos dados de coleções científicas têm aumentado consideravelmente nas últimas décadas, com destaque para as espécies vegetais. Instituições de pesquisas botânicas mantêm grandes acervos, tal como o Instituto de Pesquisas

Jardim Botânico do Rio de Janeiro com um total aproximado de 858.000 registros de espécimes (amostras) de plantas do Brasil e do exterior. Porém, o total armazenado em tais coleções é muito superior em outros países, com destaque para a Europa, sendo que o total mundial está na ordem de milhões de registros, tal como apresentado na Tabela 1, que exibe o volume de dados dos principais herbários do mundo e da América Latina, respectivamente. Além dessas fontes, outras instituições facilitam ainda mais o acesso, disponibilizando dados primários, como por exemplo, o GBIF (*Global Biodiversity Information Facility*) (GBIF, 2010), com aproximadamente 557.000.000 de registros on-line.

TABELA 1 – PRINCIPAIS HERBÁRIOS DO MUNDO E DA AMÉRICA LATINA.

HERBÁRIOS	TOTAL
Museum of Natural History, França - https://science.mnhn.fr/all/search/search/form	9.500.000
New York Botanical Garden, EUA - http://www.nybg.org	7.500.000
Royal Botanic Garden, Kew, Inglaterra - http://apps.kew.org/herbcat/	7.000.000
Herbario Nacional, México - http://www.ib.unam.mx/botanica/herbario/	1.300.000
Instituto de Pesquisas Jardim Botânico do Rio de Janeiro - http://aplicacoes.jbrj.gov.br/jabot/	858.000
Herbario Fanerogámico, Argentina - http://lillo.org.ar/	720.000

No Brasil, a coleção do herbário (RB) do Jardim Botânico do Rio de Janeiro, que é utilizada no estudo de caso neste trabalho, e seu respectivo banco de dados; é um elemento central dos sistemas de informação da biodiversidade brasileira, ele está conectado com três outros importantes sistemas: a Lista Oficial de Nomes da Flora Brasileira (FLORA, 2012), a Lista de Espécies Ameaçadas (CNCFLORA, 2012) e ao Herbário Virtual de Plantas Repatriadas (REFLORA, 2015).

O RB armazena informações de diferentes coleções (Figura 1), as principais coleções do Jardim Botânico em quantidade de registros são: Exsicatas ou *Vouchers,* que são testemunhos de espécimes coletados no campo; a Xiloteca que armazena informações sobre a estrutura anatômica de madeiras e contribui significativamente para o reconhecimento de árvores e arbustos para fins de pesquisas taxonômicas ou filogenéticas, principalmente quando o material reprodutivo (flores e frutos) é ausente ou escasso, sendo que nesse contexto, as madeiras representam uma importante fonte de informação para pesquisas, fornecendo possibilidades de identificação e

resgate de dados sobre procedência, coletores etc.; Coleção Viva (arboreto) com dados taxonômicos sobre as plantas vivas cultivadas no arboreto do parque, que possui exemplares de espécies de várias partes do mundo; a Coleção de DNA que é uma coleção de dados genéticos representativos da alta diversidade da flora brasileira, sendo um registro histórico da variação vegetal e uma base para a conservação e para a biotecnologia. O herbário RB possui 100% de seus dados e imagens armazenados e disponibilizados na internet, acessados por meio de seu sistema de informação Jabot (www.aplicacoes.jbrj.gov.br/jabot).

Total de amostras por tipo de coleção

Coleção	Quant.	%
Herbário do Jardim Botânico do Rio de Janeiro - RB	583717	67.68%
Fototeca - RBfoto	14001	1.62%
Xiloteca do Jardim Botânico do Rio de Janeiro - RBw	10122	1.17%
Coleção de fungos e líquens - RBfungo	9596	1.11%
Arboreto do Jardim Botânico do Rio de Janeiro - RBv	8644	1.00%
Carpoteca - RBcarpo	7549	0.88%
Banco de DNA - RBdna	5679	0.66%
Banco de Sementes - RBsem	2506	0.29%
Coleção em meio líquido - RBSpirit	2015	0.23%
Bromeliário do Jardim Botânico do Rio de Janeiro - RBvb	2006	0.23%
Viveiro Curadoria Coleções Vivas - RBvv	644	0.07%
Coleção de Sombra - RBvs	166	0.02%
Coleção de plantas insetívoras do Jardim Botânico do Rio de Janeiro - RBvi	57	0.01%
Cactário - RBvc	39	0.00%
Total	646741	74.97

FIGURA 1 – TOTAL DE REGISTROS POR TIPO DE COLEÇÃO NO JABOT.

O sistema foi totalmente desenvolvido na instituição com o uso de software livre, mas especificamente com Postgresql e PHP, entre 2003 e 2005. A carga dos dados no sistema foi trabalhada em dois grandes projetos de digitalização (Figura 2) que ocorreram entre 2005 e 2007 (GONZALEZ, 2009) e outro a partir de 2011.

FIGURA 2 - SALA DE DIGITAÇÃO DO HERBÁRIO DO JARDIM BOTÂNICO DO RIO DE JANEIRO.

As bases de dados de coleções científicas são utilizadas nos estudos de diversas linhas de pesquisas na Botânica, tais como a Ecologia, Taxonomia e Conservação. Porém, apesar da existência da grande quantidade de dados e das enormes possibilidades de estudos, os pesquisadores encontram dificuldades para a extração de conhecimento dessas fontes, principalmente pela qualidade dos dados obtidos.

Em 2010 uma nova versão do sistema de informação institucional começou a ser desenvolvida e, uma avaliação paralela dos problemas de qualidade de dados foi sendo aplicada. Avaliações prévias dos dados já haviam sido realizadas em (SILVA, 2010; PIPINO, 2002; STRONG, 1997) apresentado que o banco de dados JABOT tinha os principais problemas de qualidade de dados, distribuídos em diferentes dimensões de qualidade de dados e suas categorias. Os mais comuns tipos de erros identificados em coleções botânicas são relacionados com:

- Erros de classificação taxonômica;
- Digitação, campos parcialmente digitados ou incompletos – muitas vezes ocasionados por dificuldade de leitura de antigas etiquetas das exsicatas, como por exemplo na Figura 3;
- Erros de migração de dados;
- falta de padronização de dados e imprecisão de coordenadas geográficas (SILVA, 2010).

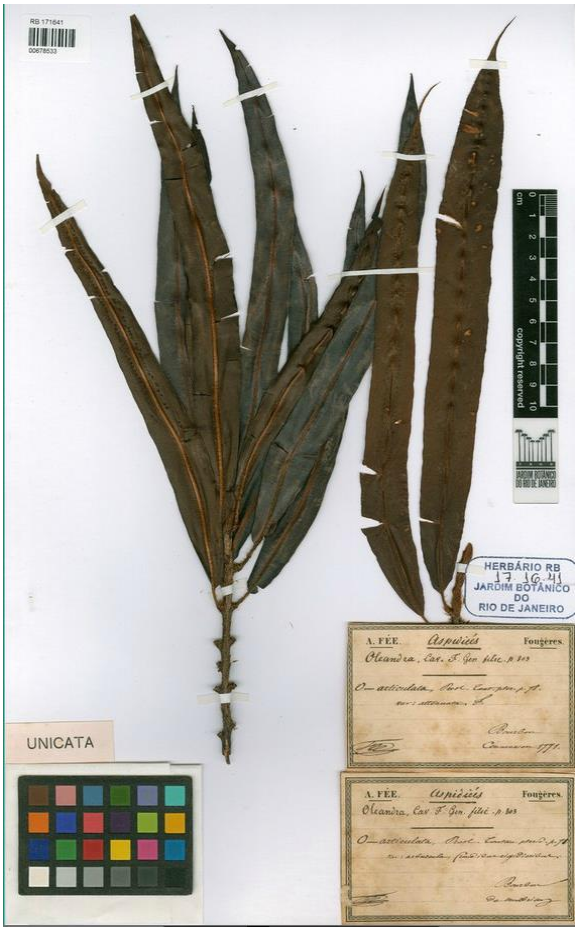

FIGURA 3 – EXSICATA ANTIGA DATADA DE 1771 COM ETIQUETA DE DIFÍCIL LEITURA.

A importância da análise de dados é valorizada muitas das vezes somente no momento em que erros são percebidos em outras pesquisas, como a modelagem de distribuição de espécies (CHAPMAN, 2005a).

A análise pela compreensão dos motivos dos erros no banco de dados indicou que muitos são consequência da ausência de validações na interface de entrada de dados no JABOT e também do conhecimento da equipe de digitadores usada. Isto pode ser averiguado pois o total de erros encontrados foram comparados em relação aos dois grandes movimentos de digitação de dados, o primeiro realizado por bolsistas não especializados na área botânica ou ainda por meio de importação de dados de planilhas enviadas por pesquisadores. Culminando em diversos erros na base de dados.

Diante do importante papel dos dados das coletas na biodiversidade, o controle da qualidade de dados é fundamental para a produção de pesquisas, quanto para

tomadores de decisão nos diversos níveis, como o governo em iniciativas de conservação e restauração de áreas degradadas. Assim, se faz necessário o desenvolvimento de metodologias que facilitem a extração de conhecimento de fontes de dados expressivas em termos da quantidade de registros. Porém, para que a extração possa ser realizada há a necessidade de um processo exaustivo de limpeza dos dados. O presente trabalho descreve a metodologia adotada para identificar erros e classificar a qualidade de dados; quantificando o número de registros afetados em cada categoria de erros, com o uso da mineração de dados.

2. Análise da Qualidade dos Dados

De acordo com Chapman (CHAPMAN, 2005b), a qualidade de dados e os erros encontrados são frequentemente ignorados em bancos de dados ambientais e na modelagem de distribuição potencial de espécies (SANTANA, 2008), sistemas de informações geográficas, sistemas de suporte à decisão, etc. Muito frequentemente, os dados são usados sem considerar os erros contidos, e isto pode acarretar em resultados errôneos, informação enganosa, decisões ambientais imprudentes e aumento dos custos.

Como ainda não existe um consenso ou uma forma padronizada para avaliar a qualidade de dados na botânica em (WANG, 2000), é encontrada uma divisão composta por 16 dimensões de qualidade agrupadas em 4 categorias:

- **Intrínseca**: características intrínsecas dos dados, independentes da sua aplicação;
- **Acessibilidade**: aspectos relativos ao acesso e a segurança dos dados;
- **Contextual**: características dependentes do contexto de utilização dos dados;
- **Representacional**: características derivadas da forma como a informação é apresentada.

Wang (2000) também clássica as 16 dimensões nas seguintes categorias:

TABELA 2 – CATEGORIAS E DIMENSÕES DE QUALIDADE DE DADOS

Categoria	Dimensões
Intrínseca	Acurácia, objetividade, credibilidade e reputação
Acessibilidade	Acessibilidade e segurança de acesso
Contextual	Relevância, valor agregado, validade temporal, Completude e quantidade de dados
Representacional	Interpretabilidade, facilidade de entendimento, representação concisa e representação consistente

Em (SILVA, 2010), foi realizado um levantamento mostrando os erros mais comuns de entrada de dados em aplicações com dados de coleções botânicas. A Figura 4 apresenta uma classificação em termos de importância e quantidade de erros para os dados analisados.

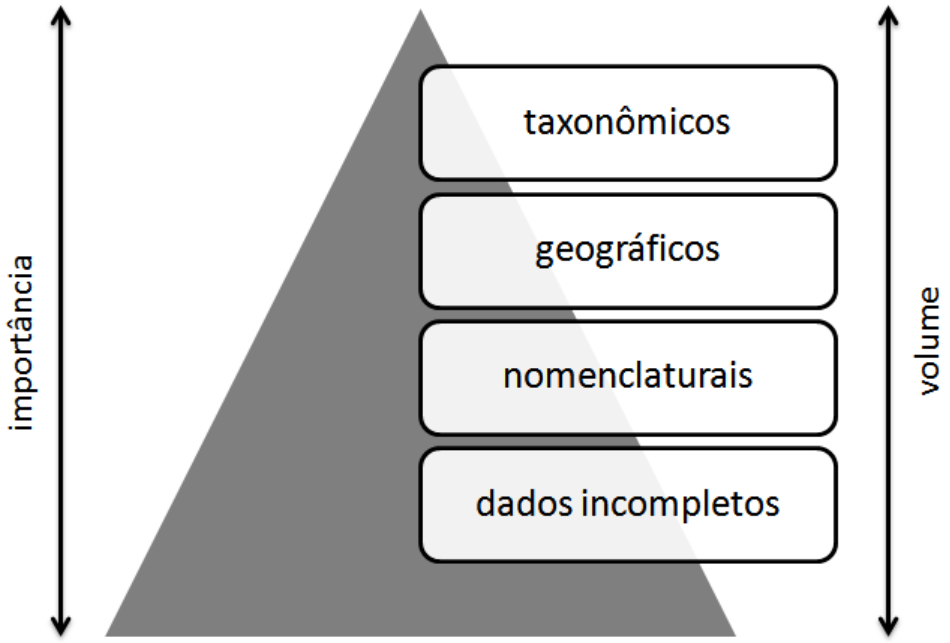

FIGURA 4 – METODOLOGIA USADA NO PROCESSO DE AVALIAÇÃO DA QUALIDADE DE DADOS.

Foram identificadas e classificadas 4 principais categorias de erros:

- **Taxonômicos** – subdivididos em *classificação incorreta* que são os erros ocasionados pela falta de conhecimentos taxonômicos e, os *erros de digitação* que são causados pela dificuldade de leitura de antigas etiquetas de identificação de alguns *vouchers*;
- **Geográficos** – relativos as coordenadas geográficas imprecisas. São muito comuns, pois muitas coletas foram feitas sem GPS e suas coordenadas

acrescentadas posteriormente, apenas com base no município onde foi realizada a coleta;

- **Nomenclaturais** – são aqueles ocorridos pela falta de padronização nos nomes e formas diversas de escrita para um mesmo valor. Um exemplo são os nomes dos coletores e localidades;
- **Dados incompletos** - alguns atributos foram parcialmente preenchidos, como é o caso dos dados de coleta da planta: ano, mês, dia, coordenadas geográficas.

Após a análise dos tipos de erros foi definida uma metodologia para a análise da qualidade de dados para esse tipo de aplicação

3. Metodologia

3.1. Metodologia aplicada

A metodologia proposta para a avaliação e melhoria da qualidade de dados nos dados das coleções botânicas foi dividida em 5 etapas, conforme a Figura 5. A metodologia considera as categorias e dimensões listadas em (SILVA, 2010).

Para a aplicação das etapas listadas na metodologia, parte-se do princípio que o banco de dados foi desenvolvido por meio de um projeto de banco de dados (TEOREY, 2011) onde regras de negócio foram incorporadas por meio dos relacionamentos das tabelas, restrições em atributos e funções escritas em linguagem procedural. Descrição das etapas:

- ***Escolha do padrão de metadados a ser usado no banco de dados*** – deve ser verificado, dentre os padrões utilizados na área científica da aplicação, qual o padrão metadado mais adequado. O uso de metadados facilita a organização dos dados por meio da definição precisa dos atributos, facilitando posteriores classificações de qualidade e também o intercâmbio de dados entre aplicações;

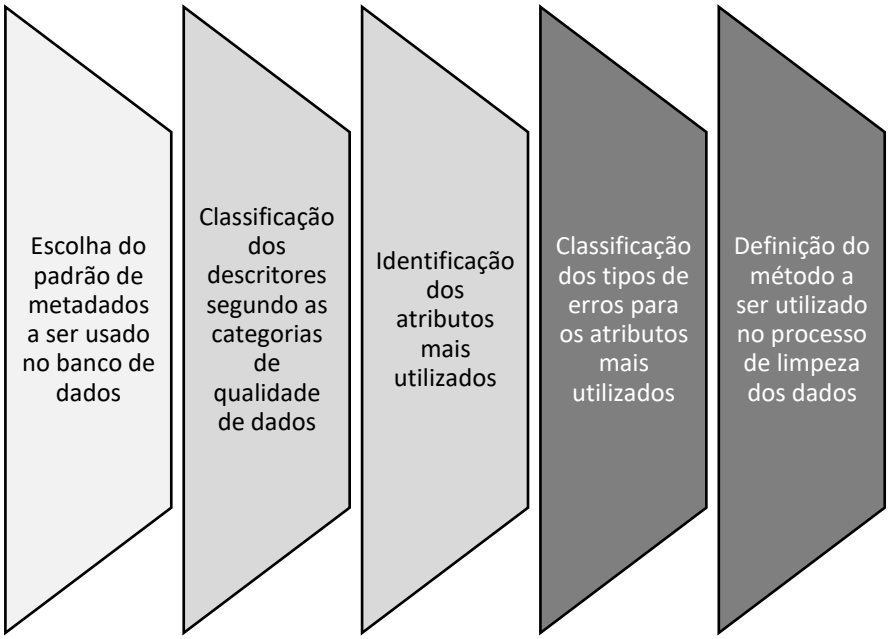

FIGURA. 5 – METODOLOGIA USADA NO PROCESSO DE AVALIAÇÃO DA QUALIDADE DE DADOS.

- ***Classificação dos descritores segundo as categorias de qualidade de dados*** - possibilita uma pré-avaliação dos atributos mais suscetíveis a erros. O critério refere-se à medida de quão é o dado é correto e confiável, e a reputação que corresponde à credibilidade da organização de origem;

- ***Identificação dos atributos mais utilizados –*** permite a definição da prioridade tanto para a tomadas das medidas preventivas, com o objetivo de evitar a entrada de erros no banco de dados. A avaliação quantitativa também é importante para a escolha do método a ser utilizado na atividade de limpeza dos dados, por conta da quantidade de registros;

- ***Classificação dos tipos de erros para os atributos mais utilizados*** – um estudo dos possíveis tipos de erros deve ser realizado para cada atributo usado. Isto somente pode ser realizado com base na análise dos dados presentes no banco de dados. Consultas de agrupamento podem ser utilizados para facilitar a identificação desses tipos, como por exemplo a verificação do total de coletas com o ano da coleta inválido;

- ***Definição do método a ser utilizado no processo de limpeza dos dados*** – diversos métodos podem ser utilizados no processo de limpeza e padronização

dos erros encontrados. Muitas ações podem ser realizadas em grandes operações com o desenvolvimento de rotinas na linguagem de manipulação de dados do próprio sistema gerenciador de banco de dados (SGBD) usado na instituição. Outras tarefas necessitam de métodos mais avançados, como a mineração de dados, que verificaremos neste trabalho.

4. Estudo de Caso

4.1. Material e métodos

Conforme mencionado anteriormente o banco de dados do Jardim Botânico do Rio de Janeiro é o estudo de caso utilizado para a aplicação da metodologia. O banco de dados utiliza o padrão de metadados Darwin Core, que estabelece um conjunto de atributos básicos relacionadas à taxonomia e ocorrência, para facilitar a troca de informações entre aplicações botânicas. O banco de dados é composto por 116 tabelas e implementado no Postgresql. As informações referentes à taxonomia e ocorrências das espécies são armazenadas nas tabelas "*arvoretaxon*", "*detacesso*", "*determinacao*" e "*testemunho*". Essas informações são importantes, não apenas para o acesso aos dados primários da coleção, mas também para a conservação e conhecimento da biodiversidade. Todas as demais tabelas servem de dicionários para essas e, justamente essas são as tabelas com a maior necessidade da melhoria da qualidade de dados. A seguir um maior detalhamento das tabelas:

- ***arvoretaxon*** – armazena os nomes científicos utilizados para a determinação das coletas e está organizada em uma estrutura hierárquica onde os níveis representam a própria classificação botânica;
- ***detacesso*** – apresenta os dados referentes a ocorrência dos testemunhos por meio do controle de seus acessos, tais como, local de coleta, coordenadas, altitude e data de coleta, coletor principal e outros coletores;
- ***testemunho*** – apresenta as informações dos materiais coletados das espécimes, como tipo de coleção, o número de tombamento no acervo e o código de barras dos *vouchers*;
- ***determinacao*** – apresenta os dados referentes ao histórico de determinações dos testemunhos. O armazenamento é relevante para permitir o controle das atualizações realizadas pelos taxonomistas no acervo.

4.2. Aplicação da metodologia no estudo de caso

A seguir são apresentadas as etapas da metodologia proposta:

- **Etapa 1** - Escolha do padrão de metadados a ser usado no banco de dados: No presente trabalho, o Darwin Core (DC, 2009) foi selecionado;
- **Etapa 2** - Classificação dos descritores segundo as categorias de qualidade de dados: uma avaliação quanto ao critério acurácia é apresentada na Tabela 3, que também apresenta para cada um dos atributos os tipos de erros encontrados juntamente com o critério;

TABELA 3 – TIPOS DE ERROS ENCONTRADOS EM DADOS DE COLEÇÕES CIENTÍFICAS.

Atributos	Tipos de erros	Critério
CatalogNumber	Data de coleta inválida Data de coleta não-informada Data de identificação não-informada Mês inválido Ano inválido Dia inválido	Acurácia
YearIdentified		
MonthIdentified		
DayIdentified		
TypeStatus		
CollectorNumber		
YearCollected		
MonthCollected		
DayCollected		
Country	Georreferenciado pela sede do município Altitude mínima e máxima invertida Georreferenciamento impreciso Altitude fora do limite Altitude não-numérica Coordenadas invertidas Longitude não-informada Latitude não-informada Coordenadas com o valor zero Altitude em pés Profundidade em pés	Acurácia
StateProvince		
Longitude-degree		
Latitude-degree		
Longitude-min		
latitude-min		
Longitude-sec		
Latitude-sec		
Altprof		
Minelevation		
Maxelevation		
Mindepth		
Maxdepth		
ScientificName	Táxon incorreto Identificação incorreta Ausência do valor do rank do táxon Reino não-informado Nome ambíguo	Acurácia / reputação
Family		
Genus		
Species		
Subspecies		
ScientificNameAuthor		
IdentifiedBy		

- **Etapa 3** - Identificação dos atributos mais utilizados: a utilização de cada um dos atributos é apresentada na Tabela 4;

- **Etapa 4** - Classificação dos tipos de erros para os atributos mais utilizados: é para os atributos mais utilizados (Tabela 4) foram observados os tipos de erros e classificados de acordo com (WANG, 2000; SILVA, 2010), a classificação é apresentada na Tabela 3;
- **Etapa 5** - Definição do método a ser utilizado no processo de limpeza dos dados: a partir do levantamento dos problemas identificados no banco de dados, foi realizado um estudo para a definição das técnicas a serem usadas no processo de limpeza, observando as categorias de erros anteriormente verificadas. Muitas das correções necessárias foram implementadas através de *scripts* desenvolvidos em SQL (*Structured Query Language*) (CHAMBERLIN, 1974).

TABELA 4 – PRINCIPAIS ATRIBUTOS USADOS NO BANCO DE DADOS DO HERBÁRIO RB.

Atributos	Preenchimento (%)
CatalogNumber	98,4
Collector	99,67
ScientificName	99,52
Family	99,52
Country	99,46
YearCollected	98,3
Locality	97,6
Altprof	96,61
Genus	96,21
MonthCollected	93,75
DayCollected	92,83
Notes	91,51
IdentifiedBy	89,61
Species	88,96
Maximum/MinimumElevation	85,78
MaximumDepth	85,78
ScientificNameAuthor	82,27
YearIdentified	72,1
Latitude/Longitude - min	23,37
Latitude/ Longitude - seg	23,37
Latitude /Longitude - grau	23,33

5. Resultados e soluções adotadas para a limpeza dos dados

5.1. Soluções tradicionais

Para os problemas de duplicidade de dados, foram desenvolvidos diversos *scripts* para consultar o banco de dados, realizando, por exemplo, verificações de duplicidades nos diversos níveis taxonômicos, com variação apenas no nome de

autor, acarretando duplicação do táxon. A diversidade de escritas de nomes de autores foi um dos mais difíceis de serem solucionados. Um exemplo deste caso é apresentado a seguir: *Abarema laeta (Benth.) Barneby & J.W.Grimes / Abarema laeta (Benth.) Barneby & J.W. Grimes*

De acordo com as regras taxonômicas, um nome de autor com ponto não deve possuir espaço após o mesmo. Portanto, o nome correto para os casos acima é o segundo. Para diminuir esse problema foi elaborada, uma expressão regular para validar nomes de autores, de acordo com as regras de taxonomia. Outro exemplo é o de uma função que faz as substituições dos táxons que possuem duplicações porém, com os nomes de autores diferentes, neste caso os nomes dos autores foram oriundos da Lista da Flora do Brasil (FLORA, 2012), que é adotada como padrão taxonômico. As principais atividades a serem realizadas com tal objetivo são apresentadas na Figura 6.

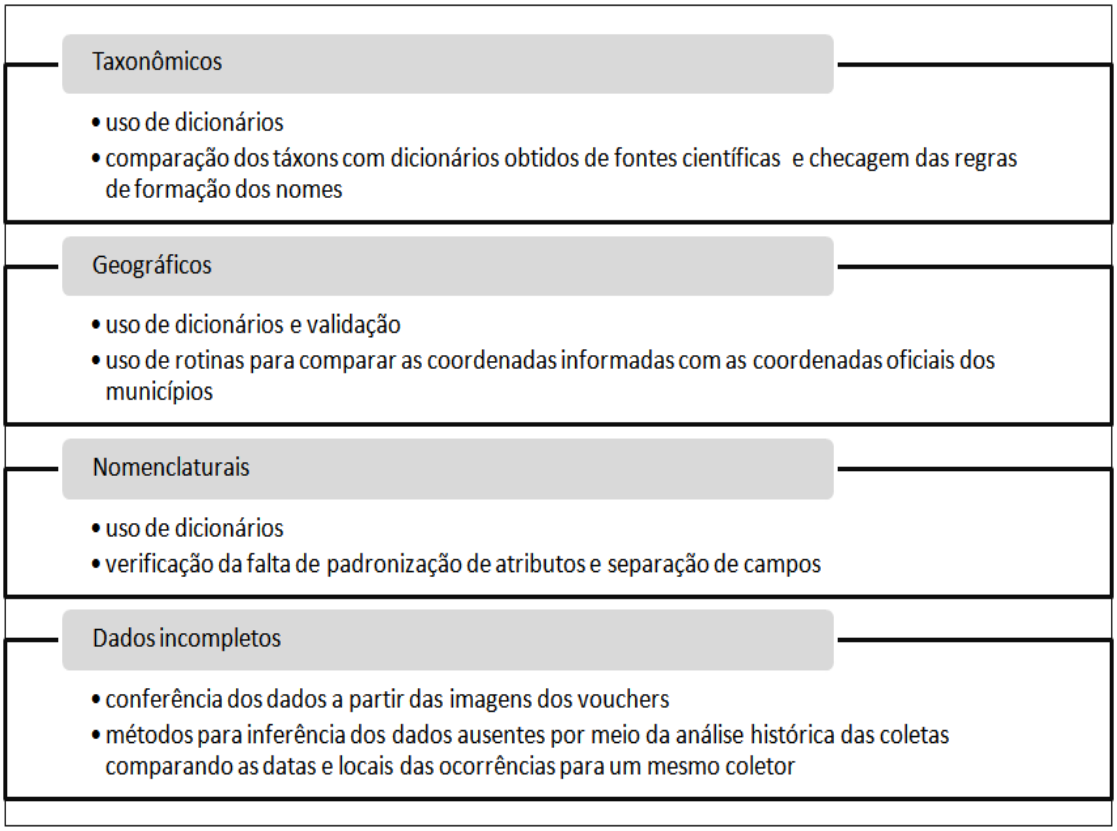

FIGURA 6 – ESTUDO DE MÉTODOS USADOS NA ATIVIDADE DE LIMPEZA DOS DADOS.

Dentre os métodos avaliados para a limpeza de dados, as Expressões Regulares – REGEX (MELTON, 2002) foram responsáveis por grande parte das operações de identificação de erros e correção dos mesmos, através de seus recursos

para a construção de complexas consultas. Expressões regulares permitem pesquisas de padrões em cadeias de caracteres por meio de sintaxes convencionadas e um conjunto de metacaracteres definidas em POSIX 1003.2 (POSIX, 1994) e amplamente usadas em aplicações de diversas áreas de pesquisas, principalmente em banco de dados.

A falta de padronização também se aplica ao nome dos coletores. Esses tipos de erros podem ser identificados de acordo com as regras taxonômicas. Nas Tabelas 5 e 6, outros exemplos de erros nomenclaturais são apresentados. Para cada tabela são exibidos os erros e solução adotada.

TABELA 5. TABELAS COM RESULTADOS DE DADOS TAXONÔMICOS.

Erro Encontrado	Atributo	Solução adotada
Duplicidade	Family, Genus, Species, Subspecies	Verificação se é da mesma família, gênero e espécie
Duplicidade	StateProvince, Country	Remoção de nomes duplicados
• Falta de padronização de espaços entre nomes e parênteses • Falta de pontuação em abreviações • Falta de padrão de letras maiúsculas e minúsculas • Ordem dos nomes • Nomes duplicados	ScientificNameAuthor, IdentifiedBy, Collector	• Padronização de espaços • Padronização de pontuação • Passar todos os caracteres para minúsculo e depois transformar somente os caracteres devidos para maiúsculo • Dar prioridade às abreviações • Remoção de nomes duplicados

TABELA 6. TABELA COM RESULTADOS DE DADOS DE OCORRÊNCIA.

Atributo	Erro Encontrado	Solução Adotada
Data coleta	• Anocoleta: data inválida (ano maior que atual e menor que 1700); • Diacoleta: datas inválidas (dia maior que 31) • Mescoleta: datas inválidas (mês maior que 12) • Datas inválidas (caracteres não numéricos. Ex.: algarismos romanos, palavras). • Datas invertidas (dia no lugar do mês)	• Quando o ano era maior que 12 e menor que 100 eram somados 1900 no campo ano. Quando o ano era maior que 500 e menor que 900 eram somados 1000 • Deixar o campo dia em branco • Realizar a conferência pela imagem do testemunho • Deixar o campo dia em branco • Realizar a conferência pela imagem do testemunho • Algarismos e textos foram convertidos para números • Seguiu a padronização para o formato dd/mm/aaaa
AltProf: altitude e profundidade	• Caracteres numéricos e textuais no mesmo campo • Sem unidade de medida • Unidade de medida inválida	• Separação de caracteres • Padronização de todos para metro

	• Valores muito altos (>2000m)	• As medidas que tinham **f** ou **t**, foram convertidas para *feets* e o restante para metro • Verificar coletas na mesma localidade e encontrar a altitude correta

5.2. Solução com uso da mineração de dados

Um dos métodos não-supervisionados mais populares da mineração de dados é o método destinado a encontrar conjuntos de itens frequentes a partir de transações registradas em banco de dados, com a extração de regras de associação entre os itens presentes nas transações, sem levar em consideração as implicações de causalidade (AGRAWAL, 1993; HAN, 2011; WU, 2007). A análise de associação visa apresentar regras que muitas vezes não são claras, como por exemplo, o famoso caso de compras de fraldas e cervejas revelando o padrão de pais que compravam fraldas e também cervejas nas sextas-feiras à noite para assistir a jogos pela TV. Assim, uma transação é o conjunto de itens encontrados em operações tais como os produtos comprados por um determinado cliente, tal como a análise de cestas de compras (BRIN, 1997).

Na análise de associação, itens frequentes são conjuntos de itens com frequência maior ou igual a um valor mínimo informado ou o valor de suporte. A manipulação desses itens pode ser uma tarefa não muito simples, em razão da complexidade ocasionada pela explosão combinatória que pode ser originada com base na quantidade de itens envolvidos na análise. De uma forma geral, o conjunto de itens frequentes pode ser encontrado por 2k – 1, excluindo o elemento nulo (TAN, 2009). Depois da obtenção dos itens frequentes, o próximo passo na análise é a aplicação do algoritmo para a identificação das regras de associação. Os valores de confiança juntamente com o valor de suporte serão usados para a seleção do conjunto de regras interessantes obtidas do conjunto de transações. O algoritmo então deve encontrar todas as regras que satisfaçam a condição na qual o valor do suporte deve ser maior ou igual ao valor mínimo de suporte e a confiança deve ser maior ou igual ao valor mínimo de confiança.

Considerando como unidade de aplicação da análise de associação o formato {*táxon, local_de_ocorrência*} foram geradas 3.880 transações. Na Figura 7 são apresentadas duas transações com os nomes de famílias que ocorrem nos municípios de Água Boa e Água Branca, respectivamente.

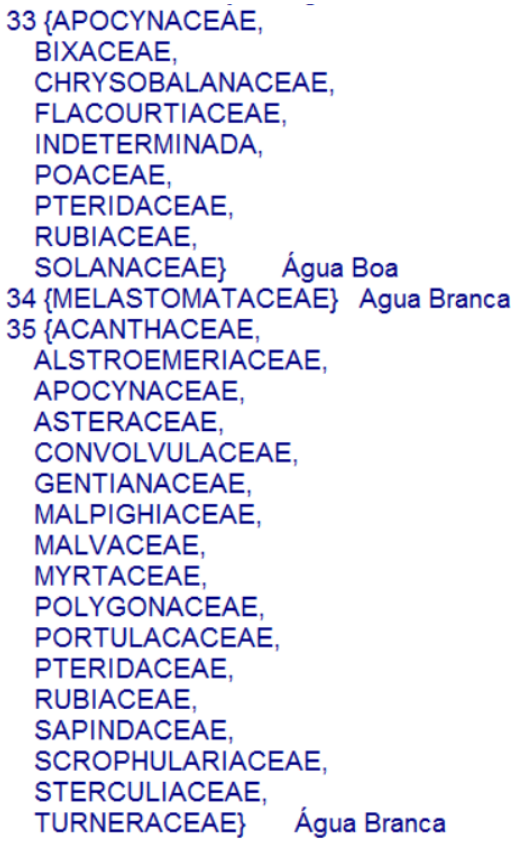

FIGURA 7 – EXEMPLOS DE TRANSAÇÕES NO FORMATO: FAMÍLIA, MUNICÍPIO.

Com a aplicação do algoritmo Apriori no *package* Arules (HAHSLER, 2005) no R (R DEVELOPMENT CORE TEAM, 2012); com o valor de suporte 0.2 e de confiança em 0.5, gerando 24 regras (Figura 8). As regras obtidas exibem as famílias que ocorrem em um mesmo local. Por exemplo, a regra 1 informa que em 20% (suporte) da base foram encontradas ocorrências das famílias MALPIGHIACEAE e ASTERACEAE em um local. O valor da confiança destaca que quando a primeira família (lhs - *left hand side*) foi encontrada, em 73% foi encontrada a segunda família (rhs – *right hand side*). A terceira coluna apresenta os valores do lift, que é uma medida objetiva de avaliação da qualidade de uma regra muito utilizada porque o uso somente da combinação suporte-confiança muitas vezes conduz a avaliações limitadas, tendo em vista que no cálculo da confiança o suporte do conjunto de itens do consequente da regra é ignorado (TAN, 2009). O *lift* identifica associações assumindo que os itens ocorrem inicialmente independentemente um dos outros (BRIN, 1997). Assim, as regras geradas podem ser armazenadas em tabelas para consultas comprobatórias de ocorrências de espécies em locais com semelhanças geográficas.

	lhs		rhs	support	confidence	lift
1	{MALPIGHIACEAE}	=>	{ASTERACEAE}	0.2028351	0.7334576	1.673025
2	{PIPERACEAE}	=>	{ASTERACEAE}	0.2087629	0.6852792	1.563129
3	{SOLANACEAE}	=>	{RUBIACEAE}	0.2038660	0.6510288	1.748091
4	{RUBIACEAE}	=>	{SOLANACEAE}	0.2038660	0.5474048	1.748091
5	{SOLANACEAE}	=>	{ASTERACEAE}	0.2231959	0.7127572	1.625807
6	{ASTERACEAE}	=>	{SOLANACEAE}	0.2231959	0.5091123	1.625807
7	{MELASTOMATACEAE}	=>	{MYRTACEAE}	0.2023196	0.6635672	2.003611
8	{MYRTACEAE}	=>	{MELASTOMATACEAE}	0.2023196	0.6108949	2.003611
9	{MELASTOMATACEAE}	=>	{RUBIACEAE}	0.2198454	0.7210482	1.936102
10	{RUBIACEAE}	=>	{MELASTOMATACEAE}	0.2198454	0.5903114	1.936102
11	{MELASTOMATACEAE}	=>	{ASTERACEAE}	0.2237113	0.7337278	1.673641
12	{ASTERACEAE}	=>	{MELASTOMATACEAE}	0.2237113	0.5102881	1.673641
13	{EUPHORBIACEAE}	=>	{MYRTACEAE}	0.2036082	0.6443719	1.945652
14	{MYRTACEAE}	=>	{EUPHORBIACEAE}	0.2036082	0.6147860	1.945652
15	{EUPHORBIACEAE}	=>	{RUBIACEAE}	0.2252577	0.7128874	1.914189
16	{RUBIACEAE}	=>	{EUPHORBIACEAE}	0.2252577	0.6048443	1.914189
17	{EUPHORBIACEAE}	=>	{ASTERACEAE}	0.2213918	0.7006525	1.598196
18	{ASTERACEAE}	=>	{EUPHORBIACEAE}	0.2213918	0.5049971	1.598196
19	{MYRTACEAE}	=>	{RUBIACEAE}	0.2322165	0.7011673	1.882719
20	{RUBIACEAE}	=>	{MYRTACEAE}	0.2322165	0.6235294	1.882719
21	{MYRTACEAE}	=>	{ASTERACEAE}	0.2360825	0.7128405	1.625997
22	{ASTERACEAE}	=>	{MYRTACEAE}	0.2360825	0.5385068	1.625997
23	{RUBIACEAE}	=>	{ASTERACEAE}	0.2551546	0.6851211	1.562769
24	{ASTERACEAE}	=>	{RUBIACEAE}	0.2551546	0.5820106	1.562769

FIGURA 8 – REGRAS GERADAS PELA APLICAÇÃO DO ALGORITMO APRIORI.

6. Conclusões

Banco de dados de coleções científicas possuem particularidades que facilitam a ocorrência de erros, pela própria natureza de nomes, dados históricos e, que podem ainda aumentar caso sejam utilizados no processo de inclusão de dados, usuários sem conhecimentos específicos da área. Apesar das dificuldades, o controle da qualidade de dados tem que ser contínuo e monitorado por especialistas para fomentar a produção segura de pesquisas a partir desses dados.

O presente trabalho avaliou o nível de qualidade de dados e propôs uma metodologia para melhoria da qualidade de dados na base de dados do Jardim Botânico do Rio de Janeiro. Na metodologia foi sugerido que qualquer trabalho de qualidade de dados deve ser desenvolvido após a definição do padrão de metadados a ser utilizado, pois isso facilita a definição e implementação de regras para a descoberta de erros. As outras etapas, mais relacionadas com o controle de qualidade de dados, caracterizam-se pelo levantamento dos atributos mais utilizados e o desenvolvimento das rotinas necessárias para identificação dos principais erros.

Outra parte da metodologia, avaliou a implementação de rotinas para a identificação de erros em grandes volumes de dados, como foi o caso de ocorrência de famílias em municípios. Nesse ponto, a análise de associação da mineração de dados foi usada com sucesso, auxiliando na comprovação da ocorrência de famílias em municípios. Diversas outras possibilidades de uso da mineração de dados e seus algoritmos estão em estudo para aplicações em dados de coleções científicas botânicas, contribuindo para o estudo de padrões em grandes volumes de dados.

Referências

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A., Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, v. 22, n. 2, p. 207–216, maio 1993.

BRIN, S.; MOTWANI, R.; SILVERSTEIN, C., Beyond market baskets. *ACM SIGMOD Record*, v. 26, n. 2, p. 265–276, jan. 1997.

CHAMBERLIN, D. D.; BOYCE, R., SEQUEL: A structured English query language. *Proceedings of the 1974 ACM SIGFIDET*, p. 249–264, 1974.

CHAPMAN, A.D. *Presentation: Principles of Data Quality*. . [S.l: s.n.], 2005a. Disponível em: <http://imsgbif.gbif.org/CMS_ORC/?doc_id=3135&download=1>.

CHAPMAN, A.D. *Principles and Methods of Data Cleaning : Primary Species and Species- Occurrence Data*. . [S.l: s.n.], 2005b. Disponível em: <http://www2.gbif.org/DataCleaning.pdf>.

CNCFLORA. *The Official Brazilian Plants Checklist*. Disponível em: <http://cncflora.jbrj.gov.br/?q=lista_vermelha/redlisting>.

DARWIN CORE TASK GROUP. *Darwin Core*. Disponível em: <http://rs.tdwg.org/dwc/>.

FLORA. *Lista de Espécies da Flora do Brasil*. Disponível em: <http://floradobrasil.jbrj.gov.br/2012>.

GBIF. Global Biodiversity Information Facility. *Natural History*, Scripta Botanica Belgica. v. 29, n. March, p. 1–2, 2010. Disponível em: <http://www.gbif.org/>.

GONZALEZ, M. Quantificação De Custo E Tempo No Processo De Informatização Das Coleções Biológicas Brasileiras: a Experiência Do Herbário Do Instituto De Pesquisas Jardim Botânico Do Rio De Janeiro. *Rodriguésia*, v. 60, p. 1–11, 2009. Disponível em: <http://rodriguesia.jbrj.gov.br/FASCICULOS/rodrig60_3/014-09a.pdf>.

HAHSLER, M; GRUEN, B; HORNIK, K. *arules: Mining Association Rules and Frequent Itemsets*. Disponível em: <http://cran.r-project.org/package=arules>. Acesso em: 2 maio 2013.

HAN, J.; KAMBER, M.; PEI, J., *Data Mining: Concepts and Techniques*. 3. ed. San Francisco: Morgan Kaufmann, 2011.

IEEE. *IEEE Standard for Information Technology--Portable Operating System Interface (POSIX) Part 1*. Disponível em: <http://standards.ieee.org/findstds/standard/1003.2d-1994.html>.

MELTON, J., SIMON, A. R. *SQL:1999: Understanding Relational Language Components*. [S.l.]: Morgan Kaufmann, 2002.

PIPINO, Leo L.; LEE, Yang W.; WANG, Richard Y. Data quality assessment. *Communications of the ACM*, v. 45, n. 4, p. 211, 2002. Disponível em: <http://dwquality.com/DQAssessment.pdf>.

R DEVELOPMENT CORE TEAM. R: A Language and Environment for Statistical Computing. Tertiary R: A language and environment for statistical computing, 2012, Vienna, Austria: R Foundation for Statistical Computing, 2012. Disponível em: <http://www.r-project.org>. Acesso em: 4 nov. 2013.

Reflora - Herbário Virtual. Disponível em: <http://reflora.jbrj.gov.br/jabot/herbarioVirtual>. Acesso em: 20 out. 2015.

SANTANA, F. S. *et al.* A reference business process for ecological niche modelling. *Ecological Informatics*, v. 3, n. 1, p. 75–86, jan. 2008. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S1574954107000933>. Acesso em: 9 mar. 2012.

SILVA, L. A. E; *et al.* Abordagem Colaborativa para a Melhoria da Qualidade de Dados em Bases de Dados Botânicas. 2010, Belo Horizonte: [s.n.], 2010.

STRONG, D. M.; LEE, Y. W.; WANG, R. Y. Data quality in context. *Communications of the ACM*, v. 40, n. 5, p. 103–110, 1997. Disponível em: <http://dl.acm.org/citation.cfm?id=253804>.

TAN, P., *Introduction to Data Mining*. 1st. ed. Boston, MA: Addison-Wesley Longman Publishing Co., Inc., 2007.

WU, X. *et al.* Top 10 algorithms in data mining. *Knowledge and Information Systems*, v. 14, n. 1, p. 1–37, dez. 2008.