

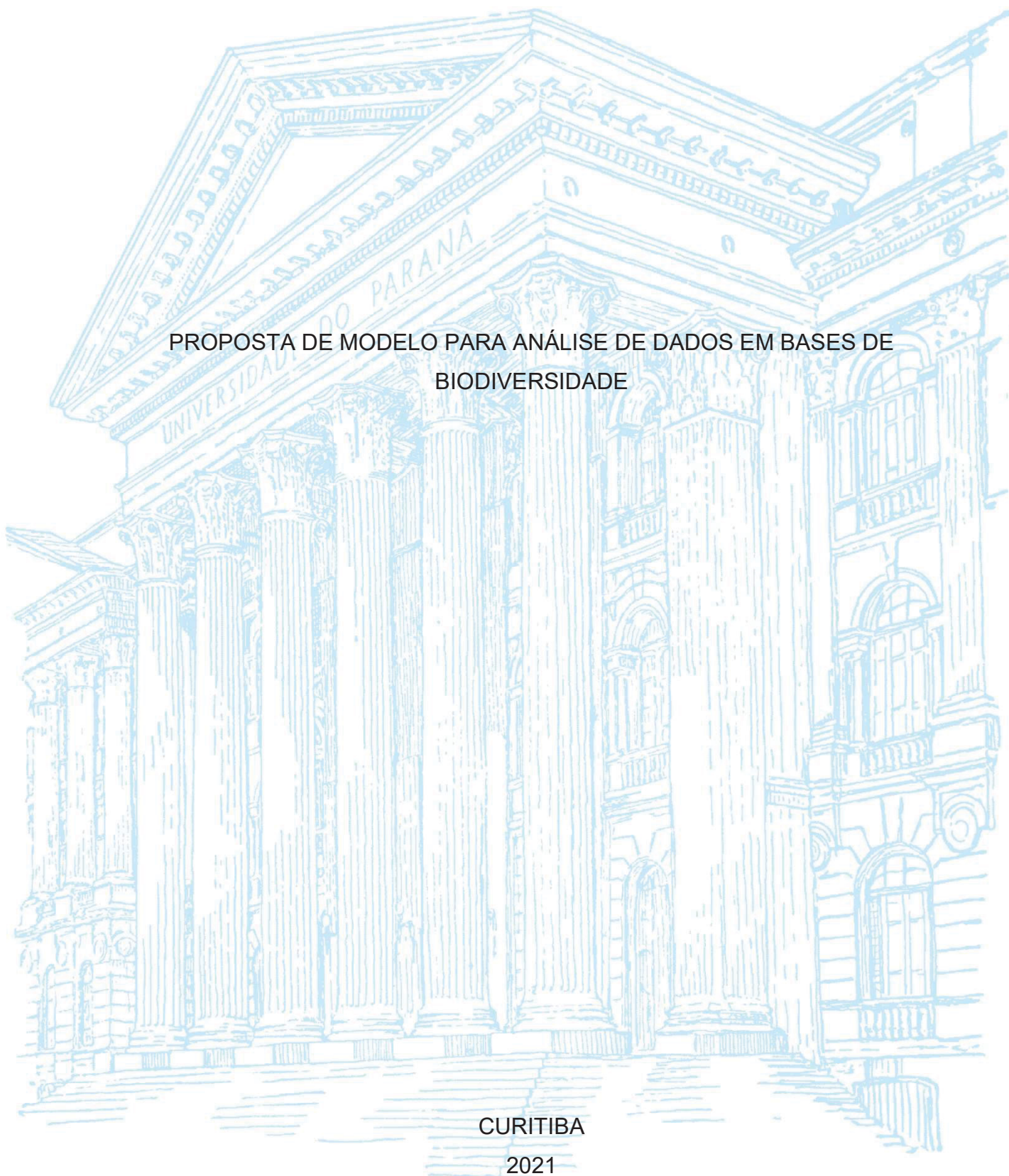
UNIVERSIDADE FEDERAL DO PARANÁ

VICTOR HUGO PANCERA TEDESCHI

PROPOSTA DE MODELO PARA ANÁLISE DE DADOS EM BASES DE
BIODIVERSIDADE

CURITIBA

2021



VICTOR HUGO PANCERA TEDESCHI

PROPOSTA DE MODELO PARA ANÁLISE DE DADOS EM BASES DE
BIODIVERSIDADE

Dissertação apresentada ao curso de Pós-Graduação em Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Gestão da Informação.

Orientadora: Profa. Dra. Denise Fukumi Tsunoda

CURITIBA

2021

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DE CIÊNCIAS SOCIAIS
APLICADAS – SIBI/UFPR COM DADOS FORNECIDOS PELO(A) AUTOR(A)
Bibliotecário: Maria Lidianne Herculano Graciosa – CRB 9/2018

Tedeschi, Victor Hugo Pancera

Proposta de modelo para análise de dados em bases de
biodiversidade / Victor Hugo Pancera Tedeschi. – 2021.
196 p.

Dissertação (Mestrado) - Universidade Federal do Paraná. Programa
de Pós-Graduação em Gestão da Informação, do Setor de Ciências
Sociais Aplicadas.

Orientadora: Denise Fukumi Tsunoda.

Defesa: Curitiba, 2021.

1. Gestão da Informação. 2. Mineração de dados. 3. Metadados.
4. Biodiversidade. I. Universidade Federal do Paraná. Setor de Ciências
Sociais Aplicadas. Programa de Pós-Graduação em Gestão da
Informação. II. Tsunoda, Denise Fukumi. III. Título.

CDD 658.4038



MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS SOCIAIS E APLICADAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO GESTÃO DA
INFORMAÇÃO - 40001016058P1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO DA INFORMAÇÃO da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **VICTOR HUGO PANCERA TEDESCHI** intitulada: **PROPOSTA DE MODELO PARA ANÁLISE DE DADOS EM BASES DE BIODIVERSIDADE**, sob orientação da Profa. Dra. **DENISE FUKUMI TSUNODA**, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua **APROVAÇÃO** no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 26 de Julho de 2021.

Assinatura Eletrônica

29/07/2021 08:01:48.0

DENISE FUKUMI TSUNODA

Presidente da Banca Examinadora

Assinatura Eletrônica

29/07/2021 10:12:07.0

HELENA DE FÁTIMA NUNES SILVA

Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica

29/07/2021 14:25:37.0

EDUARDO CARNEIRO DOS SANTOS

Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Avenida Prefeito Lothário Meissner, 632 - CURITIBA - Paraná - Brasil
CEP 80210-170 - Tel: (41) 3360-4191 - E-mail: ppggi@ufpr.br

Documento assinado eletronicamente de acordo com o disposto na legislação federal Decreto 8539 de 08 de outubro de 2015.

Gerado e autenticado pelo SIGA-UFPR, com a seguinte identificação única: 103328

Para autenticar este documento/assinatura, acesse <https://www.prpgg.ufpr.br/siga/visitante/autenticacaoassinaturas.jsp>
e insira o código 103328

AGRADECIMENTOS

Minha gratidão a todos que contribuíram para que eu conseguisse concretizar esta dissertação.

À minha orientadora e amiga, Prof.^a Denise, por ter acreditado em mim e no meu trabalho. Não consigo representar em palavras toda a admiração que tenho por você, como professora e principalmente pelo ser humano incrível que és.

Aos professores do PPGGI, que contribuíram com suas aulas para ampliar meus conhecimentos acerca da Gestão da Informação.

À secretária do PPGGI, Simone, por sua agilidade, sempre a nos ajudar nas questões acadêmicas.

Aos colegas, tanto do mestrado, quanto do doutorado, pelas conversas nos momentos difíceis e aquelas conversas descontraídas, pela ajuda nas atividades das disciplinas, nas dicas para construir esta dissertação.

À Fundação Araucária pela concessão da bolsa de estudos no programa de mestrado em Gestão da Informação.

“Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes”. (Marthin Luther King)

RESUMO

A diversidade existente no planeta é muito alta, sendo que em algumas projeções esse número poderia alcançar 10 a 30 milhões de espécies. O Brasil abriga, aproximadamente, 13% da biota mundial, conferindo ao país o título de país megadiverso. Para cada espécie que compõe esta biodiversidade, conforme são estudadas, inúmeras informações são geradas, como arquivos de texto, georreferenciamento, recursos multimídia, sequenciamento de DNA dentre outras. Todos estes dados são depositados em repositórios ou bancos de dados e disponibilizados em acesso aberto à comunidade científica e demais interessados. Com o tempo estas bases têm armazenado dados de grande relevância, ponto que se encontra o objetivo geral deste trabalho, “propor um modelo de análise de dados de bases de biodiversidade”. Para alcançar este objetivo foram realizadas algumas etapas, dentre as quais destacam-se o levantamento bibliográfico sobre as bases de dados de biodiversidade e os principais padrões de metadados presentes, e o levantamento das análises de dados em biodiversidade. Para ambos os levantamentos foram aplicadas técnicas de análise bibliométrica com o pacote Bibliometrix do *software* R. A segunda etapa consistiu na construção de um modelo de análises que pudesse ser aplicado às bases. O modelo contempla a aplicação de estatísticas descritivas e técnicas de mineração de dados, com detalhamento das etapas de pré-processamento para preparação dos dados para as análises propostas. Para a validação do modelo proposto foi utilizado um conjunto de dados do portal GBIF, de onde foram filtrados os dados relacionados à fauna brasileira. Os dados analisados correspondem a todos os registros fornecidos pelo Museu Nacional do Rio de Janeiro e foram submetidos às etapas de pré-processamento, e posteriormente à mineração de dados, no *software* WEKA com tarefas de classificação (PART, J48, SVM e Apriori), de associação (Apriori) e de agrupamento (K-Means). Foram recuperados 225.404 registros com 249 colunas cada, referentes ao Museu Nacional do Rio de Janeiro. Na etapa de pré-processamento foi realizada a remoção de dados não relevantes para as análises. No algoritmo Part foram encontradas 58, 33 e 140 regras e o método J48 atingiu taxas de acerto na ordem de 54,27%, 99,67% e 46,69%. O algoritmo Apriori foi utilizado para descoberta de regras de associação e também como classificador. Nas análises com o método K-Means a classe “ano codificado” mostrou-se mais relevante para o agrupamento. Considera-se que a maior contribuição desta pesquisa foi apresentar um modelo que pode ser utilizado em outras pesquisas relacionadas à biodiversidade, auxiliando os pesquisadores na descoberta de informações relevantes a respeito dos organismos presentes no planeta.

Palavras-chave: mineração de dados, análise descritiva, GBIF, Darwin Core, WEKA

ABSTRACT

The diversity existing on the planet is very high, and in some projections this number could reach 10 to 30 million species. Brazil is home to approximately 13% of the world's biota, giving the country the title of megadiverse country. For each species that make up this biodiversity, as they are studied, countless pieces of information are generated, such as text files, georeferencing, multimedia resources, DNA sequencing, among others. All these data are deposited in repositories or databases and made available in open access to the scientific community and other interested parties. Over time these databases have stored data of great relevance, which is where the general objective of this work lies, "to propose a model of biodiversity database data analysis". To achieve this goal, some steps were taken, among which we highlight the bibliographic survey on biodiversity databases and the main metadata standards present, and the survey of biodiversity data analysis. For both surveys, bibliometric analysis techniques were applied using the Bibliometrix package from the R software. The second step consisted in building an analysis model that could be applied to the databases. The model contemplates the application of descriptive statistics and data mining techniques, with details of the pre-processing steps for preparing the data for the proposed analyses. For the validation of the proposed model a dataset from the GBIF portal was used, from which data related to the Brazilian fauna were filtered. The data analyzed correspond to all the records provided by the National Museum of Rio de Janeiro and were submitted to the pre-processing steps, and later to data mining, in the WEKA software with classification (PART, J48, SVM and Apriori), association (Apriori) and clustering (K-Means) tasks. A total of 225,404 records with 249 columns each, referring to the Museu Nacional do Rio de Janeiro, were retrieved. In the pre-processing step, the removal of data not relevant for the analysis was performed. The Part algorithm found 58, 33 and 140 rules, and the J48 method achieved hit rates of 54.27%, 99.67% and 46.69%. The Apriori algorithm was used for association rule discovery and also as a classifier. In the analyses with the K-Means method the class "coded year" proved to be more relevant for clustering. It is considered that the greatest contribution of this research was to present a model that can be used in other research related to biodiversity, helping researchers to discover relevant information about the organisms present on the planet.

Keywords: Data mining, descriptive analysis, GBIF, Darwin Core, WEKA

LISTA DE FIGURAS

FIGURA 1 – O CICLO DE VIDA DO DADO	22
FIGURA 2 – DISTRIBUIÇÃO DOS BIOMAS BRASILEIROS AO LONGO DO TERRITÓRIO NACIONAL E OS LIMITES DA FAIXA DA COSTA MARÍTIMA DO BRASIL	27
FIGURA 3 – DESCRIÇÃO DOS PRINCIPAIS CONJUNTOS DE CARACTERÍSTICAS E EXTENSÕES PRESENTES NO PADRÃO DWC, OS QUAIS SÃO APLICADOS EM DIVERSOS DOMÍNIOS DA ÁREA BIOLÓGICA.	31
FIGURA 4 – REPRESENTAÇÕES GRÁFICAS DOS MODELOS DE CLASSIFICAÇÃO, A) BASEADO EM CONHECIMENTO, B) ÁRVORES, C) CONEXIONISTAS, D) DISTÂNCIA, E) FUNÇÃO E F) PROBABILÍSTICA.....	40
FIGURA 5 FLUXOGRAMA DE DESENVOLVIMENTO DA PESQUISA.....	45
FIGURA 6 – AS 50 PALAVRAS-CHAVES DO AUTOR COM MAIOR OCORRÊNCIA	63
FIGURA 7 – DISTRIBUIÇÃO DOS REGISTROS DA FAUNA PRESENTES NO GBIF NO TERRITÓRIO BRASILEIRO	64
FIGURA 8– ESTRUTURA PRINCIPAL DO MODELO DE ANÁLISES EM BASES DE DADOS DE BIODIVERSIDADE.....	70

LISTA DE GRÁFICOS

GRÁFICO 1 – PRODUÇÃO CIENTÍFICA POR ANO.....	56
GRÁFICO 2 – CONTRIBUIÇÃO DE PRODUÇÃO CIENTÍFICA NO MUNDO.....	56
GRÁFICO 3 – FONTES COM MAIORES NÚMEROS DE PRODUÇÕES.....	57
GRÁFICO 4 – DINÂMICA DAS PALAVRAS-CHAVES DO AUTOR AO LONGO DO TEMPO.....	63
GRÁFICO 5 – NÚMERO DE OCORRÊNCIAS POR TIPO DE REGISTRO PRESENTES NOS DADOS RECUPERADOS SOBRE A FAUNA BRASILEIRA NO GBIF	66
GRÁFICO 6 – DISTRIBUIÇÃO DOS REGISTROS DE OCORRÊNCIA DA FAUNA BRASILEIRA POR SÉRIE DE DADOS.....	67
GRÁFICO 7 – DISTRIBUIÇÃO DAS OCORRÊNCIAS DA FAUNA BRASILEIRA QUANTO A SUA CLASSIFICAÇÃO TAXONÔMICA, DISPONÍVEL NA GBIF	68
GRÁFICO 8 – DISTRIBUIÇÃO DAS VARIÁVEIS PRESENTES NA INSTÂNCIA “ISSUE”	79
GRÁFICO 9 - DISTRIBUIÇÃO DOS DADOS DO “ANO CODIFICADO” EM RELAÇÃO A CLASSE “STATEPROVINCE”, COM VISUALIZAÇÃO DAS VARIÁVEIS DA CLASSE “PHYLUM” SENDO AS VARIÁVEIS VISUALIZÁVEIS: CHORDATA VERDE, ARTHROPODA VERMELHA, MOLUSCA VERDE-ÁGUA E BRYOZOA AZUL ESCURO.....	81
GRÁFICO 10 – RELAÇÃO DAS CLASSES “ <i>typeStatus</i> ” E “ <i>stateProvince</i> ”, COM VISUALIZAÇÃO DAS VARIÁVEIS DA CLASSE “PHYLUM” SENDO AS VARIÁVEIS VISUALIZÁVEIS: CHORDATA PRETO, ARTHROPODA VERMELHA E BRYOZOA AZUL.....	82

LISTA DE QUADROS

QUADRO 1 - REGRAS DE USO DO PADRÃO DARWIN CORE	32
QUADRO 2 – PASSOS PARA REALIZAÇÃO DA PESQUISA	44
QUADRO 3 – PROTOCOLO DE REVISÃO SISTEMA DE LITERATURA APLICADO COM FOCO EM ANÁLISE DE DADOS DA BIODIVERSIDADE.....	49
QUADRO 4 – PROTOCOLO DE REVISÃO SISTEMA DE LITERATURA APLICADO COM FOCO EM BASE DE DADOS DE BIODIVERSIDADE.....	51
QUADRO 6 – AUTORES MAIS PRODUTIVOS	57
QUADRO 6 – AUTORES MAIS CITADOS.....	57
QUADRO 7 – ARQUIVOS RECUPERADOS NA PLATAFORMA GBIF	72
QUADRO 8 – ETAPAS DE PRÉ-PROCESSAMENTO E OS TERMOS / ATRIBUTOS TRABALHADOS EM CADA ETAP.....	74
QUADRO 9 – LOCALIDADES ENCONTRADAS NO ARQUIVO, QUE FORAM REMOVIDOS NO PROCESSO DE LIMPEZA.....	78
QUADRO 10 - RESULTADOS DAS ANÁLISES DO ALGORITMO J48	87

LISTA DE TABELAS

TABELA 1 - TERMOS UTILIZADOS NAS BUSCAS REALIZADAS NA SCOPUS E WEB OF SCIENCE	17
TABELA 2 - DOCUMENTOS COM O MAIOR NÚMERO DE CITAÇÕES NAS REFERÊNCIAS DOS ARTIGOS RECUPERADOS	60
TABELA 3 - ARTIGOS MAIS CITADOS NAS REFERENCIAS	62
TABELA 5 – RESULTADOS DAS ANÁLISES DO ALGORITMO PART.....	83
TABELA 6 – INFORMAÇÕES GERAIS DOS CLUSTERS GERADOS	88

LISTA DE ABREVIATURAS OU SIGLAS

ABCD	– Access to Biological Collections Data schema
ARFF	– <i>AttributeRelation File Format</i>
CSV	– comma-separated-values
DC	– Dublin Core
DNA	– Deoxyribonucleic acid
DwC	– Darwin Core
EML	– Ecological Metadata Language
GBIF	– The Global Biodiversity Information Facility
GEF	– Global Environment Fund
GNU	– <i>General Public License</i>
GEF	– <i>Global Environment Fund</i>
ICMBio	– Instituto Chico Mendes de Conservação da Biodiversidade
IB	– Informática para Biodiversidade
INBio	– National Institute of Biodiversity
KNIME	– Konstanz Information Miner
MCTIC	– Ministério da Ciência, Tecnologia, Inovação e Comunicações
NCEAS	– National Center for Ecological Analysis and Synthesis
PNUMA	– Programa das Nações Unidas para Meio Ambiente
PPGGI	– Pós-Graduação em Gestão da Informação
SiBBR	– Sistema de Informação sobre a Biodiversidade Brasileira
TDWG	– Taxonomic Database Working Group
UNEP	– <i>United Nation Environment Programme</i>
WEKA	– Waikato Environment for Knowledge Analysis
WGS84	– <i>World Geodetic System 1984</i>
XML	– eXtensible Markup Language

SUMÁRIO

1 INTRODUÇÃO	14
1.1 PROBLEMA DE PESQUISA	15
1.2 OBJETIVOS	16
1.3 JUSTIFICATIVA	16
1.3.1 Justificativa para a Ciência.....	16
1.3.2 Justificativa para a área de biodiversidade.....	18
1.3.3 Justificativa para o PPGGI e para a Gestão da Informação	19
1.3.4 Justificativa para os pesquisadores.....	19
1.4 DELIMITAÇÃO DA PESQUISA.....	21
1.5 ESTRUTURA DO DOCUMENTO.....	21
2 REFERENCIAL TEÓRICO.....	22
2.1 DADO, INFORMAÇÃO E CONHECIMENTO	22
2.2 METADADOS.....	24
2.3 BIODIVERSIDADE	25
2.4 PADRÕES ENCONTRADOS NAS BASES DE BIODIVERSIDADE	28
2.4.1 Access to Biological Collections Data schema (ABCD).....	29
2.4.2 Ecological Metadata Language (EML)	29
2.4.3 Plinian Core.....	30
2.4.4 Darwin Core	30
2.5 ANÁLISE DE DADOS.....	35
2.5.1 Limpeza de dados	36
2.5.2 Mineração de dados	38
3 ENCAMINHAMENTOS METODOLÓGICOS	43
3.1 CARACTERIZAÇÃO DA PESQUISA	43
3.2 MATERIAIS E MÉTODOS.....	44
3.2.1 VISÃO GERAL DA PESQUISA	44
3.2.2 WEKA.....	45
3.2.3 <i>Software</i> R e Biblioshiny.....	46
3.2.4 Microsoft Excel	47
3.2.5 Base GBIF	48
3.2.6 Revisão sistemática de literatura.....	49

3.2.7 Escolha da base de dados sobre biodiversidade	50
3.2.8 Métodos de análise	52
4 RESULTADOS E DISCUSSÃO	55
4.1 VISÃO GERAL DE ANÁLISE DE DADOS E BIODIVERSIDADE.....	55
4.2 ANÁLISE DE DADOS NA PLATAFORMA GBIF	65
4.3 MODELO DE ANÁLISE.....	69
4.4 ANÁLISE DESCRITIVA.....	72
4.5 VISUALIZAÇÃO DOS DADOS WEKA	80
4.6 RESULTADOS PART.....	83
4.7 RESULTADOS APRIORI	85
4.8 RESULTADOS SVM	87
4.9 RESULTADOS J48	87
4.10 RESULTADOS K-MEANS.....	88
5 CONSIDERAÇÕES FINAIS E RECOMENDAÇÕES	91
REFERÊNCIAS.....	94

1 INTRODUÇÃO

A diversidade biológica no planeta é muito alta pois, segundo Brusca e Brusca (2007), havia no mundo aproximadamente 1,3 milhões de espécies, sendo que em algumas projeções esse número poderia alcançar 10 a 30 milhões. Segundo o Ministério do Meio Ambiente (MMA)¹, estimam-se mais de 100 milhões de espécies distintas no mundo, apesar de que menos de dois milhões destas sejam conhecidas. Ainda no portal do MMA, a biodiversidade abrange variedades de “espécies de flora, fauna e micro-organismos; as funções ecológicas desempenhadas por estes organismos nos ecossistemas; e as comunidades, habitats e ecossistemas formados por eles”.

O Brasil, segundo estimativas conservadoras, abriga 13% da biota mundial, uma vez que possui o maior sistema fluvial do mundo, além de cinco biomas, rendendo ao país o título de “país megadiverso” (INSTITUTO CHICO MENDES DE CONSERVAÇÃO DA BIODIVERSIDADE, 2018). Segundo o livro vermelho de espécies ameaçadas de extinção, publicado pelo Instituto Chico Mendes em 2018, no Brasil existem, aproximadamente, mais de 117 mil espécies de animais conhecidas e, diariamente, novas descobertas são realizadas.

A partir da descoberta de uma espécie, são geradas informações de maneira contínua e agregadora sobre as características (morfologia, nomenclatura, filogenia, etc.), hábitos (alimentação, comportamento, etc.), distribuição geográfica (avistamentos, registros de coletas, etc.), genética (filogenética, sequenciamento de DNA, etc.) entre outros. Estas informações por sua vez são de interesse de diversos grupos e entidades, os quais necessitam que as informações estejam dispostas de forma compreensível, atualizada e de fácil acesso.

Neste cenário é que se encontra o tema que norteia o presente estudo, uma vez que o grande volume de dados gerado e depositado com atualização em tempo real das bases de dados oportuniza a utilização de métodos e ferramentas de análise destes dados para: descoberta de padrões nas relações entre os dados e o seu propósito de uso, relações entre região e interesse dos dados, tipos de dados mais

¹ BRASIL. Ministério do Meio Ambiente (MMA). Disponível em: <https://www.mma.gov.br/biodiversidade/biodiversidade-brasileira.html#:~:text=Estima-se%20que%20at%C3%A9%20100,nos%20ecossistemas%3B%20e%20as%20comunidades%2C>. Acesso em: 17 nov. 2020.

recorrentes para cada localização do país, se existe alguma relação entre a finalidade do dado com a região do dado de onde ele foi coletado ou com os fornecedores dos dados, quais os principais erros, falhas e/ou incongruências encontrados nos dados presentes. Assim como averiguar a distribuição dos dados nos grupos taxonômicos, quais os tipos de dados mais frequentes, quem são os maiores depositantes de dados na base, origem dos dados, entre outras características pertinentes que podem ser identificadas nas análises exploratórias.

1.1 PROBLEMA DE PESQUISA

A digitalização das coleções é de vital importância para a preservação e compartilhamento da informação das espécies, mas esta abordagem traz novos desafios, como a confiabilidade dos dados digitalizados. Partindo deste ponto, Ruas (2017) destaca a importância da metainformação na curadoria do conteúdo gerado pelos processos de digitalização que garante a fidelidade e autenticidade da informação apresentada, assim como a contextualização na qual a informação foi capturada.

Nestas circunstâncias, o problema identificado que norteará os estudos da presente pesquisa está na premissa de que, a cada dia que passa, o volume de informações acerca da fauna cresce de maneira exponencial, devido a este crescimento vem à tona dúvidas quanto a esses dados, os quais vem sendo gerados e depositados nas mais diversas bases de conhecimento sobre a biodiversidade, não ficando claro se os dados possuem algum tipo de discriminação, ou se possuem algum viés, ou relação direta com o depositante, e se os depositantes são qualificados ou as informações fornecidas são verossímeis, se tais incongruências/ausências possuem algum padrão que possa ser identificado com o qual possa elucidar mais sobre os dados da biodiversidade brasileira.

Assim, a questão de pesquisa é: “Quais padrões podem ser encontrados por meio de análise de dados da fauna brasileira em uma base de biodiversidade?”.

1.2 OBJETIVOS

Com o intuito de responder à questão da pesquisa definiu-se o objetivo geral **propor um modelo de análise de dados de bases de biodiversidade**, e, para o cumprimento deste, foram definidos os seguintes objetivos específicos:

- a) buscar repositórios de biodiversidade e estudar seus formatos de metadados;
- b) identificar os principais métodos e ferramentas de análise de dados, passíveis de serem utilizados nas bases de dados de biodiversidade;
- c) alinhar o principal formato encontrado nos repositórios de bases de biodiversidade aos métodos e ferramentas de análises de dados;
- d) validar o modelo proposto pela aplicação do mesmo em uma base escolhida em um dos repositórios de biodiversidade.

1.3 JUSTIFICATIVA

Para a fundamentação desta pesquisa as justificativas foram divididas em quatro (4) abordagens conforme apresentadas nas próximas subseções.

1.3.1 Justificativa para a Ciência

Com o objetivo de identificar a produção científica e formação do embasamento teórico deste estudo foram realizadas pesquisas sobre o tema, nos acervos das bases científicas nos campos título, resumo e palavras-chave. Os filtros aplicados foram documentos do tipo artigo ou revisão; de acesso aberto; publicados entre 2015 e 2019; e idioma sendo português ou inglês, com os termos de busca: "DATA ANALYSIS" AND BIODIVERSITY, "DATA ANALYSIS" AND "BIODIVERSITY DATABASE", "DATA ANALYSIS" AND BIODIVERSITY AND DATABASE, "DATA ANALYSIS" AND BIODIVERSITY AND FAUNA, "DATA ANALYSIS" AND "BIODIVERSITY INFORMATICS" e "DATA ANALYSIS" AND "PRIMARY BIODIVERSITY DATA".

Em 15 de outubro de 2019 realizou-se uma pesquisa na base Scopus com os descritivos e resultados demonstrados na Tabela 1.

TABELA 1 - TERMOS DE BUSCA UTILIZADOS NAS BUSCAS REALIZADAS NA SCOPUS E WEB OF SCIENCE

Temas de busca	Scopus	WoS
"DATA ANALYSIS" AND BIODIVERSITY	996	438
"DATA ANALYSIS" AND "BIODIVERSITY DATABASE"	3	2
"DATA ANALYSIS" AND BIODIVERSITY AND DATABASE	104	41
"DATA ANALYSIS" AND BIODIVERSITY AND FAUNA	41	15
"DATA ANALYSIS" AND "BIODIVERSITY INFORMATICS"	8	6
"DATA ANALYSIS" AND "PRIMARY BIODIVERSITY DATA"	2	1

Fonte: Autor, 2020

O total de documentos recuperados foi 1.657, dos quais após a leitura dos títulos e resumos, foram selecionados 114 documentos para posterior análise, foram excluídos os documentos que não possuíam relação direta com a análise de biodiversidade ou que possuem um viés relacionado à genética ou métodos de análise genica. Após a leitura completa de todos os documentos não foram encontrados trabalhos com a mesma abordagem do presente projeto. Como exemplo Hortal et al (2007) apresentaram em seu trabalho as limitações nos dados presentes em uma base de dados de sementes, que diferem do presente estudo por trazer outra metodologia de análise dos dados e por tratar de espécies botânicas o presente estudo pertence trazer espécies zoológicas. Os Yesson et al (2007), realizaram análises na mesma base a qual será abordada no presente estudo, contudo os pesquisadores utilizaram apenas uma família de leguminosas para os testes e a metodologia de análise diferente da proposta. O artigo apresentado por Fondes et al (2015), testaram a possibilidade de utilização de um *software* para mineração de dados em uma base de dados sobre a biodiversidade, no qual por meio de seus testes demonstraram o uso potencial destas técnicas como suporte para tomada de decisões no contexto da biodiversidade e na avaliação dos dados de biodiversidade que serão

utilizados nas análises. Estes artigos demonstram o interesse do meio científico no tema proposto pelo presente estudo.

1.3.2 Justificativa para a área de biodiversidade

Os dados sobre a biodiversidade são relevantes para a tomada de decisão e definição de políticas públicas, uma vez que a sustentabilidade tem sido objeto de discussões internacionais. Desta forma, o acesso rápido aos dados atualizados e confiáveis assume importância até então não conhecida. A análise de dados de biodiversidade pode prover benefícios não somente para a preservação da biota, como também econômicos, pela redução do desperdício de recursos ambientais e redução de impactos refletindo na melhoria na qualidade de vida das populações.

Em meio ao cenário no qual o meio ambiente se encontra, os dados sobre a biodiversidade apresentam relevância para prever, estimar e avaliar os impactos causados em crises ambientais naturais ou antrópicas, assim trazendo impactos relevantes na conservação de áreas-chaves, ou auxiliando nas etapas de recuperação das áreas impactadas.

Outros usos destes dados e no monitoramento de espécies, onde pode-se identificar dois usos importantes para a conservação, sendo o primeiro o monitoramento de espécies nativas, no qual são verificadas a abrangência da distribuição da espécie, com enfoque em espécies de interesse ou ameaçadas, entende-se espécies de interesse aquelas as quais são consideradas espécies bandeira ou guarda-chuva, pois com a conservação destas indiretamente muitas outras ou, até mesmo o ecossistema onde elas se encontram são conservados. O segundo uso do bio monitoramento é voltado às espécies invasoras, as quais são introduzidas de maneira acidental ou não, em ambientes onde originalmente estas não ocupavam, nestes ambientes os invasores não encontram predadores que controlem a sua população, assim crescendo descontroladamente e causando diversos impactos ao ecossistema.

As unidades de Conservação (UCs) são definidas pela Lei Nº 9.985, de 2000, como um espaço territorial e seus recursos ambientais, incluindo águas jurisdicionais, com características naturais relevantes, legalmente instituído pelo poder público, com objetivos de administração, ao qual se aplicam garantias adequadas de proteção. O

uso de dados sobre a biodiversidade em unidades é utilizado desde o início da escolha de áreas críticas para a implementação, assim identificando e auxiliando nas melhores estratégias para a implementação de unidades de conservação e para o monitoramento de gerenciamento da unidade, a fim de identificar os impactos decorrentes da sua implementação. Com isso a busca por novas formas de interpretação destes dados se faz necessária, para que os esforços aplicados à conservação de biota sejam mais efetivos e precisos.

1.3.3 Justificativa para o PPGGI e para a Gestão da Informação

Com o intuito de identificar se o tema já foi trabalhado no Programa de Pós-Graduação em Gestão da Informação, da UFPR no dia 11 de outubro de 2020 foram realizadas pesquisas no Repositório Digital da UFPR na seção de Teses & Dissertações, do programa de Pós-graduação em Ciência, Gestão e Tecnologia da Informação, utilizando o descritivo “*biodiversidade*” nos campos título, resumo e palavras-chave (quando disponíveis). Apenas um documento foi encontrado, o qual não possui o mesmo viés que a presente pesquisa, pois trata de dados da digitalização do acervo botânico do museu municipal de Curitiba, já o presente estudo focará na análise dos dados pertencentes a fauna brasileira.

A proposta do presente estudo tem como abordagem a Gestão da Informação no contexto de desenvolver a análise e descoberta de conhecimento (KDD) em grandes volumes de dados, abordando o tratamento de dados e informações, gestão e tecnologia, em consonância à Linha 2 do programa de Pós-Graduação em Gestão da Informação (PPGGI) da Universidade Federal do Paraná. A presente dissertação trará como contribuição a interação de técnicas de análise de dados utilizadas na GI, em um conjunto de dados sobre biodiversidade os quais não foram ainda explorados no programa.

1.3.4 Justificativa para os pesquisadores

Os dados sobre a biodiversidade têm se tornado cada vez mais relevantes, para as mais diversas áreas do conhecimento, podendo-se tomar como base para esta relevância, o aumento recorrente da busca da sustentabilidade, nas mais

diversas atividades humanas. Estes dados são utilizados como base para a tomada de decisão quando da implementação de diversos empreendimentos, tanto públicos como privados, tendo em vista a redução de impactos negativos dos empreendimentos à biota local.

Dados estruturados de forma correta, atualizados e de confiabilidade, disponíveis abertamente para consulta e utilização, podem trazer inúmeros benefícios ao agilizar o processo de levantamento de espécies de um local ou direcionar os esforços das coletas de campo, repercutindo em benefícios econômicos e ambientais.

Dados históricos da biodiversidade são de extrema valia para dimensionar os danos causados por desastres em ambientes naturais, toma-se como exemplo a catástrofe causada pelas queimadas nos territórios ocupados pela floresta amazônica e pantaneira que ocorreram em 2020. Por meio destes dados é possível avaliar os reais danos afligidos à estas regiões, uma vez que a floresta, qualquer que seja, não se limita a um conjunto de árvores e pássaros, mais sim à uma complexa e delicada rede de interações entre os mais diferentes níveis, que se alterados podem trazer problema em escalas nunca antes vistas, as quais podem nunca mais serem restauradas.

Com os dados da biodiversidade é possível observar, com relativa facilidade, as lacunas nos conhecimentos presentes sobre a fauna, auxiliando no direcionamento de pesquisas, como indicado por Rocha, Boeger (2009) que filós inteíros não foram registrados do Brasil, necessitando de pesquisas direcionadas para averiguação desta falha. Tal uso torna-se de grande valia para pesquisadores da área biológica, que podem utilizar estes dados como um contato inicial com o grupo ou com a área que será foco de seus estudos. Dentre a utilização dada por estes profissionais/pesquisadores, estão os estudos da dispersão de espécies no território, relação entre espécies, monitoramento da dispersão de espécies invasoras ou de espécies em risco.

Contudo, se tais dados não forem provenientes de fontes confiáveis, mesmo sendo fornecidos de maneira correta ou redigidos com todas as informações corretas, podem tornar as pesquisas irrelevantes ou com uma acuidade muito baixa. Com isso, realizar uma análise dos dados sobre a fauna brasileira presentes nesta base se torna relevante.

1.4 DELIMITAÇÃO DA PESQUISA

Tendo em vista a validação do modelo proposto, esta pesquisa utiliza apenas um repositório, delimitando os dados para apenas espécies de fauna de ocorrência no território brasileiro.

1.5 ESTRUTURA DO DOCUMENTO

Este documento é organizado com a apresentação do referencial teórico no Capítulo 2 apoiado nos temas Dado, Informação e Conhecimento; Biodiversidade; Visão geral da análise de dados da biodiversidade; Padrões encontrados em bases de biodiversidade; análise de dados.

A metodologia de pesquisa é apresentada no Capítulo 3 por meio de sua caracterização, definição do ambiente e procedimentos metodológicos adotados. O resultado da análise dos dados da base de biodiversidade é apresentado no Capítulo 4 e, finalmente, no Capítulo 5 são realizadas as considerações finais da pesquisa, apresentando a confrontação entre os objetivos e os resultados, as principais contribuições do trabalho e as sugestões para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo são apresentados os principais temas relacionados à pesquisa em pauta: Dado, Informação e Conhecimento; Biodiversidade e repositórios de biodiversidade; Padrões utilizados em bases de biodiversidade e Análise de dados.

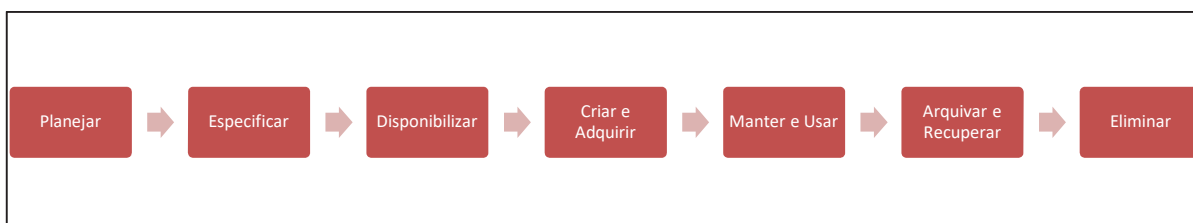
2.1 DADO, INFORMAÇÃO E CONHECIMENTO

Os ativos (recursos com valor reconhecido) variam conforme a atividade econômica das organizações (governamentais ou não governamentais), porém há ativos comuns a todas as organizações como recursos financeiros e humanos, dado, informação e conhecimento. Apenas recentemente estes três últimos foram considerados ativos de uma organização, isto porque a correta utilização dos mesmos possibilita os profissionais a tomarem decisões e realizarem ações mais assertivas e conscientes (DAMA, 2012).

Entender o que é dado, informação e conhecimento é importante para compreender o porquê de terem se tornados ativos. Davenport (1998, p. 19) destaca que “não é fácil distinguir, na prática, dados, informação e conhecimento” e define “dados como ‘observações sobre o estado do mundo’”. Para Dama, (2012, p. 10) “dados são a representação de fatos em forma de: texto, números, gráficos, imagens, som ou vídeo. Tecnicamente, ‘dados’ é o plural da palavra em latim ‘datum’, que significa ‘um fato’. No entanto, as pessoas comumente usam o termo como uma coisa singular”.

Assim como qualquer ativo o dado possui um ciclo de vida, demonstrado na Figura 1.

FIGURA 1 – O CICLO DE VIDA DO DADO



FONTE: Adaptado de Dama, (2012, p. 12).

Esse ativo é criado a partir de uma necessidade informacional específica ou demanda prevista, que desdobra-se em um plano de ação e análise de viabilidade (planejar) identificando o escopo da necessidade, resultado esperado, tratamentos necessários (especificar), origem e fontes de coleta, formatos (disponibilizar), formas de aquisição (criar e adquirir), onde armazenar e como serão utilizados (manter, usar, arquivar e recuperar) e se será necessário a eliminação (eliminar) (SANT'ANA, 2016).

Os dados possuem duas formas: estruturada e não estruturada. Um dado para ser estruturado precisa ser rotulado e armazenado em um banco de dados, arquivos planos (*flat files*) e documentos eletrônicos (ex. planilhas e relatórios). Um dado não estruturado não possui rótulo e pode ser armazenado em *e-mails*, arquivos de imagem, vídeo, gráficos e áudio (DAVENPORT, 1998; DAMA., 2012).

Para Dama, (2012, p. 12 e 10) o dado flui, entrando e saindo de repositórios (seja por consultas à sistemas, planilhas eletrônicas, *e-mails*, mensageiro eletrônico ou relatórios), “e é empacotado para ser entregue em produtos denominados de informação”, e “informação são dados em contexto”. O contexto pode ser a definição das regras de negócio, o prazo, o formato e a relevância do dado, que auxilia na interpretação gerando informações significativas (agregação de valor ao negócio).

Yau (2013) complementa que o dado sem contexto é inútil assim como qualquer ação que for tomada sobre ele. Saber o quem, como, quando, onde, por que e como faz com que o dado se torne informação clara para os leitores e os direcionam corretamente. As decisões baseadas em dados sem contexto custam bilhões de dólares devido a produtos encalhados, uma reengenharia de processos não funcionais e aquisição de equipamentos que não produzem (DAVENPORT, 1998).

Mesmo que um computador possa através do processamento de dados produzir um gráfico, uma tabela ou relatório, é necessário que a figura humana escolha como representar esse produto. Pessoas transformam dados em informações atribuindo a eles relevância e propósito. A informação é o elo de ligação entre o dado e o conhecimento que se pode obter por meio de análise (DAVENPORT, 1998; TARAPANOFF, 2006).

O termo informação não possui necessariamente em seu uso clareza ou exatidão, fazendo com que poucos saibam ao que estão se referindo. Por isso para esta pesquisa a informação são dados dotados de relevância e propósito (CHOO, 2003; DAVENPORT, 1998; DAMA, 2012; TARAPANOFF, 2006; YAU, 2013).

A relevância é um indicador para uso da informação, pois ela objetiva responder uma pergunta ou resolver um problema. Quando uma pessoa seleciona e processa uma informação baseada na sua relevância ocorre um processo de mudança no seu estado de conhecimento e capacidade de agir. Devido ao envolvimento do ser humano no processo de uso da informação se torna confuso, desordenado e sujeito aos caprichos do indivíduo (CHOO, 2003).

Para Davenport (1998, p. 113) assim como o dado “é o *uso* da informação, não sua simples existência, que permite aos gestores tomar decisões melhores sobre produtos e processos, aprender com os clientes e com a concorrência, monitorar os resultados de seus atos”, a informação se torna atrativa ao uso quando é oportuna, rica em detalhes, possui sequência e casualidade e, principalmente contexto ou significado.

A informação contribui para o conhecimento. O conhecimento é a compreensão, a consciência, conhecimento, e o reconhecimento de uma situação e familiaridade com a sua complexidade. Conhecimento é a informação em perspectiva, integrado em um ponto de vista com base no reconhecimento e interpretação de padrões, tais como tendências, formadas com outras informações e experiências. Pode também incluir hipóteses e teorias sobre causas. O conhecimento pode ser explícito – o que uma organização aceita como verdadeiro – ou tácito, que está dentro da cabeça dos indivíduos. Nós ganhamos em conhecimento quando compreendemos o significado da informação (DAMA, 2012, p. 11).

Conhecimento é informação processada e interpretada por indivíduos tornando-o individual, por isso, se torna um ativo difícil de gerenciar. Ele é individual porque o conjunto particular de experiências; a percepção, participação e reação do próprio indivíduo em uma determinada situação se tornam base para dar à informação um contexto, um significado e uma interpretação. (CHOO, 2003; DAVENPORT, 1998; TARAPANOFF, 2006).

2.2 METADADOS

Segundo Carlos Barbieri (2020) a definição de metadados é que são dados sobre dados, os quais colocam o dado em um contexto.

Os metadados devem estar associados a um dado, atribuindo a ele contexto e significado, sob diferentes óticas, como características (nome, peso, tipo, comprimento, etc.), relacionamentos do dado algo (localizado em mantido por, etc.),

formas de tratamento (fórmulas cálculos, protocolos, etc.), regras que definem a obrigatoriedade daquele dado (valores, regras de qualidade, etc.), informações históricas sobre o dado (inventado por, descrito por, etc.), entre outros. Os dados e os metadados se transmutam, dependendo dos contextos a qual são submetidos, assim um determinado metadado assume a função de dados, e o dados tona-se metadado em seu lugar.

O autor Barbieri (2020, p.220) define:

Quando os metadados são falhos ou mal gerenciados, potencializam-se os problemas de qualidade de dados. Muito simples: um “profiling” de dados, buscando erros de conteúdo ou de estrutura poderá, numa base de dados, ser influenciado pela baixa qualidade dos metadados (ou seja, de suas definições, regras, etc.). Um módulo que esteja fazendo integração de bases de dados, baseada em campos comuns a ambas as bases, produzirá erros, caso os metadados dos campos de integração não estejam consistentes e coerentes. (BARBIERI, 2020, p. 220)

A fim de solucionar estes problemas são utilizados padrões de metadados que facilitam a troca de informações entre entidades. Por meio destes padrões são definidos os locais, tipos dos dados e se são dados estruturados, não estruturados ou semiestruturados.

2.3 BIODIVERSIDADE

Antes de falar sobre a biodiversidade, faz-se necessário trazer a sua origem, uma vez que a palavra biodiversidade provém da contração do termo “diversidade biológica”. O termo “diversidade biológica”, segundo Magurran (2013), localizou como referência mais antiga um documento datado de 1955, que o utilizou no contexto de variação intraespecífica no comportamento e história de vida. A autora traz que possivelmente existem exemplos de utilização do termo até mesmo anteriores a este ano. O termo se tornou mais utilizado nos anos 80, sendo nesta época que realizaram a dissecação do termo em três: diversidade genética (diversidade intraespecífica, ou seja, diversidade dentro de uma única espécie), diversidade de espécies (número de espécies) e diversidade ecológica (diversidade de comunidades ecológicas).

O termo biodiversidade tem sua origem mais recente, podendo ser rastreada até um evento específico, no qual foi proposto em 1985 durante o planejamento do

Fórum Nacional de BioDiversidade (*National Forum on BioDiversity*) que ocorreu em 1986, em Washington DC, EUA. Como resultado deste fórum, em 1988 foi publicado por Edward O. Wilson, a coletânea de resultados sob o título “*Biodiversity*” (Biodiversidade). Logo após o seu surgimento, o termo tornou-se muito popular, ultrapassando o termo “diversidade Biológica” em publicações.

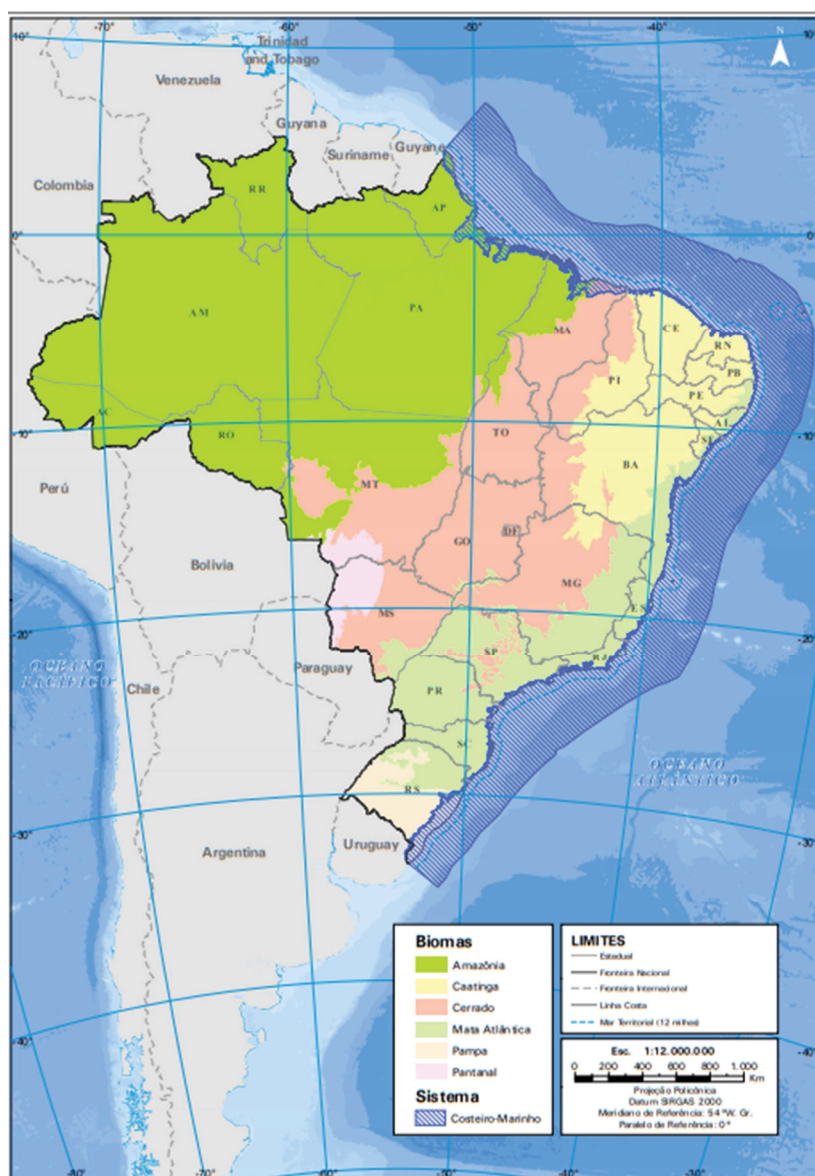
Tanto “diversidade biológica” quanto biodiversidade, possuem o mesmo significado podendo ser utilizadas de forma intercambiável.

No Brasil em 1992, ocorreu a Conferência das Nações Unidas sobre o Meio Ambiente e Desenvolvimento, realizada no Rio de Janeiro, a qual ficou conhecida como ECO-92, como resultado criou-se a *Convention on Biological Diversity* (Convenção da Diversidade Biológica - CBD), que em seu documento, apresentou a definição oficial de “diversidade biológica” ou biodiversidade:

Diversidade biológica significa a variabilidade entre organismos vivos de todas as partes incluindo, entre outros, terrestres, marinhos ou outros sistemas aquáticos e os complexos ecológicos de que eles fazem parte; compreendendo ainda a diversidade dentro de espécie, entre espécies e de ecossistema. (HEYWOOD, WATSON, 1995, p. 8)

O Brasil detém mais de 13% da biota do planeta, possuindo assim uma das maiores riquezas de espécies do planeta, característica essas que inspirou o conceito de país megadiverso (INSTITUTO CHICO MENDES DE CONSERVAÇÃO DA BIODIVERSIDADE, 2018). Esta enorme diversidade de espécies presentes no Brasil se dá por conta de suas dimensões continentais e a presença de seis biomas, sendo eles: Amazônia, Caatinga, Cerrado, Mata Atlântica, Pampa e Pantanal, com destaque para o Cerrado e a Mata Atlântica, os quais são considerados *hotspots*, ou seja, áreas de grande riqueza de espécies e de endemismo (entende-se endemismo como a ocorrência exclusiva de uma espécie em uma determinada região geográfica), fazendo destas áreas prioridade para a conservação em um nível mundial.

FIGURA 2 – DISTRIBUIÇÃO DOS BIOMAS BRASILEIROS AO LONGO DO TERRITÓRIO NACIONAL E OS LIMITES DA FAIXA DA COSTA MARÍTIMA DO BRASIL



FONTE: IBGE, 2019

Além dos biomas supracitados, o Brasil apresenta um dos maiores sistemas fluviais do mundo, abrigando cerca de 15% de toda a água doce existente no planeta. Esse sistema é distribuído em quatro bacias principais e três secundárias, sendo as bacias principais a Amazônica, a Tocantins-Araguaia, a Platina e a São Francisco. É válido destacar que a bacia Amazônica é a maior do mundo, e ocupa mais da metade do território brasileiro. Em questão de biodiversidade, cada bacia torna-se relevante, pois a diversidade de espécies é única para cada uma delas.

A zona costeira e marinha do Brasil ocupa aproximadamente 3,5 milhões de quilômetros quadrados, sendo uma das maiores faixas costeiras do mundo com 7.400 km de extensão que abrangem ambientes extraordinariamente diversos. No litoral

brasileiro é possível encontrar uma enorme variedade de ecossistemas que incluem manguezais, recifes de corais, dunas, restingas, praias arenosas, costões rochosos, lagoas e estuários. Esta diversidade de ambientes tem como suporte a composição de águas presentes no litoral brasileiro, águas frias nas regiões sul e sudeste e águas quentes nas costas nordeste e norte. Estas regiões apresentam inúmeras espécies da flora e da fauna endêmicas, por conta das peculiaridades presentes em cada ecossistema.

O Livro Vermelho da Fauna Brasileira Ameaçada de Extinção, do Instituto Chico Mendes de Conservação da Biodiversidade (2018), apresenta que são reconhecidas 117.096 espécies de animais, com estimativas de que estes números ultrapassem 137 mil. São aproximadamente 9.000 espécies de vertebrados e 94.000 de artrópodes, estimativas recentes mostram que entre os vertebrados são cerca de 4.545 espécies de peixes, 1.080 de anfíbios, 773 de répteis, 1.919 de aves e 701 de mamíferos, dos artrópodes a maior parte é de insetos, com cerca de 83.000, destaca-se também os aracnídeos com 6.200 espécies e moluscos, com aproximadamente 3.100 espécies. É válido ressaltar que os números citados já podem apresentar alterações, uma vez que, frequentemente, estão sendo realizadas descobertas de novas espécies e os grupos taxonômicos são constantemente revisados.

2.4 PADRÕES ENCONTRADOS NAS BASES DE BIODIVERSIDADE

É possível encontrar diversos padrões para o compartilhamento de dados de biodiversidade que foram desenvolvidos para facilitar tanto a disseminação quanto a integração dos dados por diversos sistemas. Nas ciências biológicas é comum o uso de metadados em suas disciplinas, a exemplo da zoologia sistemática, pois toda a descrição de uma nova espécie deve estar de acordo com o código internacional de nomenclatura zoológica, no qual apresenta registros que devem ser representados de acordo com um conjunto de metadados.

O primeiro padrão de metadados foi o Dublin Core (DC), criado em 1995, tendo sua utilização voltada para a descrição de recursos web. Em seu núcleo básico, o DC possui 15 elementos que permitem, de forma simples, descrever a definição dos elementos de forma a ser facilmente entendido pelas pessoas, e ser eficiente, sendo supostamente adequado para a representação de diversos tipos de recursos digitais e analógicos. Contudo, os 15 elementos presentes no DC são muito genéricos e

muitas vezes não possibilitam que os conteúdos dos recursos informacionais sejam descritos até a exaustão.

Já no domínio da biodiversidade, foram criados padrões de metadados que buscassem uma melhor organização de coleções biológicas, informações sobre espécies, dados de pesquisas ecológicas, entre outras formas de obtenção de dados sobre a biodiversidade. Os padrões que são mundialmente utilizados são: *Access to Biological Collections Data schema* (ABCD), *Ecological Metadata Language* (EML), *Plinian Core* e *Darwin Core* (DwC).

2.4.1 Access to Biological Collections Data schema (ABCD)

O seu desenvolvimento teve início em 2001, em 2005 foi ratificado como um dos padrões do TDWG (Taxonomic Database Working Group) (TDWG, 2007). É um padrão abrangente para o acesso e troca de dados sobre espécimes e observações. O desenvolvimento deste padrão foi voltado para aplicação em coleções de espécimes em museus de história natural, contudo também pode ser utilizado para casos gerais de ocorrência de espécies (TDWG, 2007).

O ABCD tem como base XML, o que significa que a estrutura dos dados pode ser facilmente processada pelo *software* (HOLETSCHEK ET AL, 2012). Este padrão teve duas versões as quais foram mais adotadas, sendo a versão 1.2 e a versão 2.06 (TDWG, 2007; HOLETSCHEK et al, 2012), em 2019 foi publicado a ABCD 3.0 (FICHTMULLER, 2019)

2.4.2 Ecological Metadata Language (EML)

O *Ecological Metadata Language* é um projeto *open source*, que possui objetivo de prover metadados de alta qualidade para descrever dados relevantes para as mais diversas disciplinas o qual envolvam pesquisas nas áreas da ecologia e meio ambiente. O EML foi originalmente desenvolvido por Matthew Jones no ano de 1997 no *National Center for Ecological Analysis and Synthesis* (NCEAS), baseado no relatório *ESA Committee on the Future of Long-Term Ecological Data* e no artigo sob o título “*Nongeospatial metadata for the ecological sciences*”, escrito por William K. Michener e colaboradores, publicado no mesmo ano (JONES et al, 2019). As

primeiras versões do EML foram utilizadas internamente na NCEAS, apenas na versão 2 o EML se tornou aberto e mantido pela comunidade (JONES et al, 2019).

O documento EML pode ser feito com um simples editor de texto, por meio linguagens como o R e *Python*, para propósitos gerais pode-se utilizar ferramentas XML e pode ser desenvolvido o documento por meio de ferramentas editor de metadados baseado em web (JONES et al., 2019).

2.4.3 Plinian Core

Este padrão foi desenvolvido por meio de uma parceria entre *National Institute of Biodiversity* (INBio), da Costa Rica, com o braço espanhol do GBIF. O desenvolvimento iniciou-se em 2004, com o intuito de desenvolver um sistema que funcione de forma a agregar, gestar e publicar informações no nível de espécies (TDWG, 2020).

O Plinian Core é um conjunto de termos usados para descrever os tipos de propriedades relacionados com classificação taxonômica. Este padrão foi desenhado para que os requerimentos para inclusão de dados sejam de forma facilitada, independente, abertos para integração de dados de suporte de múltiplas bases de dados e com a capacidade de ser implementado em diferentes níveis de “granularidade”. Atualmente este padrão encontra-se na versão 3.2.2 (TDWG, 2020)

2.4.4 Darwin Core

A primeira vez em que o padrão surgiu foi por volta de 1999, mas o mesmo só foi ratificado como padrão em 2009 (WIECZOREK et al, 2012). Este padrão foi desenvolvido pelo *Darwin Core Task Group*, subdivisão do *Biodiversity Information Standards* (TDWG), o qual utilizou como base para a criação o Padrão Dublin Core. Este padrão foi o escolhido pelo GBIF, por conta disso e tem sido aceito e aplicado por mais de 340 instituições, em 43 países, incluindo o Brasil.

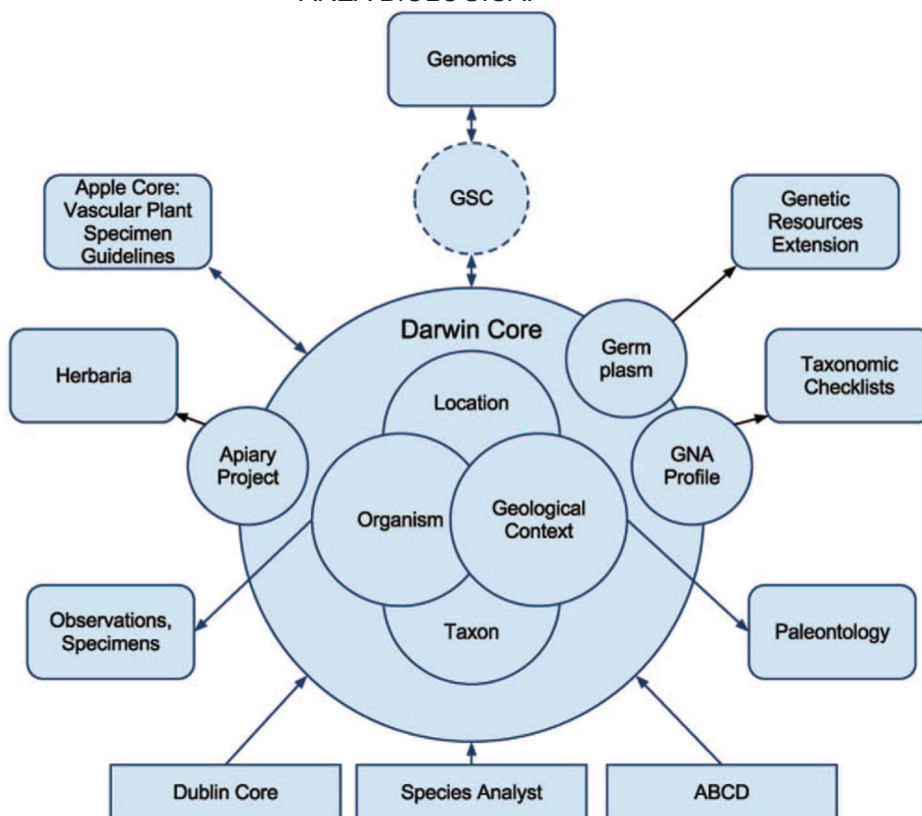
Fundamentalmente, o padrão Darwin Core possui uma série de termos bem definidos, para que os mesmos possam ser interpretados por pessoas ou máquinas (WIECZOREK et al, 2012). Estes termos dos metadados estão divididos em duas categorias, *Record-level Terms* e *Auxiliary Terms*, com isso a primeira agrupa

metadados administrativos do registro, e a segunda possui campos para descrição do espécime, os quais abrangem dados taxonômicos, de geolocalização e outros, permitindo a inclusão de informações da sequência genética associado com o registro. (SOARES, 2019).

Conforme apresentado por Soares (2019), sobre a divisão das classes dos metadados:

O DwC engloba diversos núcleos, que são classes de metadados que podem ser combinadas nos registros. São eles: *Record-level*, *Occurrence*, *Organism*, *MaterialSample*, *Event*, *Location*, *GeologicalContext*, *Identification*, *Taxon*, *MeasurementOrFact*, *ResourceRelationship*, *UseWithIRI*, *LivingSpecimen*, *PreservedSpecimen*, *FossilSpecimen*, *HumanObservation* e *MachineObservation*. (SOARES, 2019, p. 88-89).

FIGURA 3 – DESCRIÇÃO DOS PRINCIPAIS CONJUNTOS DE CARACTERÍSTICAS E EXTENSÕES PRESENTES NO PADRÃO DWC, OS QUAIS SÃO APLICADOS EM DIVERSOS DOMÍNIOS DA ÁREA BIOLÓGICA.



FONTE: Wieczorek et al, 2012

A Figura 3 apresenta no centro os conjuntos de características básicas presentes no DwC, os quais consistem em Organismo, Classificação taxonômica, Localização e contexto geológico. O padrão tem influências dos padrões Dublin Core (Utilizado como base para criação do DwC) e o padrão ABCD os quais compartilham muitas características em comum, sendo considerados como equivalentes. Ao redor

são representadas as relações do núcleo principal do DwC, com as extensões específicas, para cada subárea da biologia, como genética, paleontologia, botânica etc.

As extensões de metadados são conjuntos de elementos que não compõem o núcleo básico do DwC, os quais são desenvolvidos para aplicações específicas, como por exemplo representar informações que não poderiam ser completamente representadas pelos conjuntos de termos contidos no núcleo básico (SOARES, 2019). A criação de extensões de metadados se dá mais comumente por três objetivos, sendo o de definir novos elementos, refinar elementos e para a utilização de esquemas de codificação e vocábulos controlados.

Para a criação de um registro em DwC, um conjunto de regras devem ser seguidas, as quais estão representadas no Quadro 1. Quanto à criação dos registros, o padrão possui uma versão simples, o *Simple DwC*, o qual consiste em um grupo de elementos pré-definidos (DARWIN CORE TASK GROUP, 2009).

QUADRO 1 - REGRAS DE USO DO PADRÃO DARWIN CORE

REGRA	DEFINIÇÃO
1	Qualquer nome de termo do Darwin Core pode ser usado como um nome de campo.
2	Nenhum nome de campo pode ser repetido em um registro.
3	Não utilizar uma classe (<i>Occurrence</i> , <i>Organism</i> , <i>MaterialSample</i> , <i>LivingSpecimen</i> , <i>PreservedSpecimen</i> , <i>FossilSpecimen</i> , <i>Event</i> , <i>HumanObservation</i> , <i>MachineObservation</i> , <i>Location</i> , <i>GeologicalContext</i> , <i>Identification</i> , <i>Taxon</i>) como campo.
4	Forneça dados em quantos campos você puder.
5	Use o campo <i>dcterms: type</i> para fornecer o nome da classe do tipo Dublin Core (<i>PhysicalObject</i> , <i>StillImage</i> , <i>MovingImage</i> , <i>Sound</i> , <i>Text</i>) que o registro representa.
6	Use o campo <i>basisOfRecord</i> para fornecer o nome da classe mais específica do Darwin Core (<i>LivingSpecimen</i> , <i>PreservedSpecimen</i> , <i>FossilSpecimen</i> , <i>MaterialSample</i> , <i>HumanObservation</i> , <i>MachineObservation</i> , <i>Event</i> , <i>Occurrence</i> , <i>Taxon</i> , <i>Identification</i> , <i>Organism</i> , <i>Location</i> , <i>GeologicalContext</i> , <i>MeasurementOrFact</i> , <i>ResourceRelationship</i>) que o registro representa.
7	Preencha os campos com dados que correspondam à definição do campo.
8	Use o vocabulário controlado para os valores dos campos que os recomendam.
9	Se os dados forem considerados sensíveis, use <i>informationWithheld</i> para dizer isso.
10	Se os dados forem compartilhados com qualidade inferior a original, use <i>dataGeneralizations</i> para isso.

FONTE: Soares, 2019

As regras apresentadas no Quadro 1 orientam quanto às melhores práticas para elaboração do registro de dados.

O Núcleo básico do padrão de metadados DwC, apresenta 12 classes de termos, que são: *Record-level*; *Occurrence*, *Organism*; *MaterialSample*; *Event*; *Location*; *GeologicalContext*; *Identification*; *Taxon*; *MeasurementOrFact* e *ResourceRelationship*.

Cada classe citada possui conjunto de termos ou elementos possíveis de serem utilizados, os quais serão apresentados nos apêndices 1 a 12, trazendo o seu nome, termo ou elemento, sua definição e exemplos

A classe *Record-level*, tratado no apêndice 1, tem como função descrever o tipo do documento que está sendo apresentado, descrevendo os dados administrativos dos recursos. Entre os elementos contidos nesta classe temos o “type”, que é a forma do registro, podendo ser: *LivingSpecimen*, *PreservedSpecimen*, *fossilSpecimen*, *humanObservation* e *machineObservation*; esta classe contém as informações referente as licenças e direitos cedidos pelas pessoas ou instituições que fornecerão os dados, os quais devem ser sempre observados ao utilizar os dados. Os principais elementos contidos nessa classe são os “ID” e “Code”, os quais são identificadores únicos para aquele conjunto de dados, instituição, coleção, etc., sendo úteis para a recuperação de conjuntos únicos específicos de informações, caso seja de conhecimento os códigos.

O Apêndice 2 apresenta os elementos presentes na classe *Occurrence*, que tem por função apresentar os elementos referentes a evidência da espécie na natureza, quem foi o observador, seu comportamento, se possui alguma medida associada ao registro e as referências ligadas a aquele registro

O Apêndice 3 representa os elementos encontrado na classe *Organism* (Organismo), nesta classe são apresentados os elementos referentes ao organismo, facilitando a ligação entre as classes “*Identification*” e “*Occurrence*”.

Na classe *MaterialSample* são apresentados os elementos descritos no apêndice 4, esta classe contém informações pertencentes à uma amostra. Em coleções biológicas, o material de uma amostragem é tipicamente coletado e preservado ou submetido a processos que possam destruí-lo parcial ou totalmente (como extração de DNA), portando os dados apresentados neste elemento por conter informações sobre um organismo inteiro (espécimes coletados) ou por partes

(conteúdo estomacal, sangue, etc.), ou mesmo materiais não biológicos (amostra de solo).

A classe *Event*, apêndice 5, apresenta os elementos referentes à condição temporal na qual ocorreram os registros, sendo apresentados a data, horário do registro, e também os protocolos de amostragem e métodos aplicados para a obtenção do registro.

A classe *Location*, tratada no apêndice 6, apresenta dados referentes à localização geográfica do registro, apresentando as informações referentes à descrição do local do mais amplo (continente) ao mais específico (cidade, bairro, etc.), as coordenadas de latitude e longitude referentes aos registros, e elementos para a identificação de organismos que se encontram em ambientes aquáticos, os quais podem ser identificados nos elementos: *waterBody*, *island* e *islandGroup*.

A classe *GeologicalContex* (apêndice 7), apresenta os contextos geológicos do registro, trazendo o tempo geológico, *biostratigraphy*, *lithostratigraphy* e *crono-stratigraphy*, informações muito importantes para os registros fósseis.

O apêndice 8 apresenta a classe *Identification*, os elementos sobre a identificação da espécie incluem as informações dos responsáveis pela identificação dos organismos apresentados no registro, a data da identificação e as referências utilizadas para a identificação.

No apêndice 9 são apresentados os elementos pertencentes a classe *Taxon*, na qual são dispostas as informações referentes à classificação taxonômica dos organismos do registro. As informações contidas nesta classe devem seguir os Códigos Internacionais de Nomenclatura, os quais padronizam os nomes científicos utilizados.

O Apêndice 10 apresenta os elementos pertencentes à classe *MeasurementOrFact*, onde são descritas informações associadas a outros elementos onde são apresentadas as medidas ou fatos sobre medidas de algum recurso, sendo apresentado o tipo de medição, o valor referente à essa medição, a unidade e o método de medição e quem realizou essas medições.

No Apêndice 11 são apresentados os elementos pertencentes à classe *ResourceRelationship*, utilizada para apresentar as relações entre recursos no registro de metadados, podendo relacionar dois ou mais elementos que apresentam afinidade entre si, especificando quem, quando e qual a natureza desta afinidade.

A Classe *UseWithIRI* está representada no apêndice 12, os elementos presentes nesta classe fazem referência a todos os elementos pertencentes ao DwC que devem ter um esquema de codificação com conjunto universal de caracteres.

2.5 ANÁLISE DE DADOS

A análise de dados segundo Silva (2016), consiste em conhecer o contexto no qual o dado está inserido e como eles ocorrem. Pinheiro et al (2009) descrevem a análise de dados como um método de identificar comportamentos médios e discrepantes, independência entre variáveis, revelar tendências, entre outros, a partir de um grupo de dados, e com auxílio de recursos computacionais diferenciar o que é estrutural do que é ruído para, por fim, resumir as informações contidas em grupos de dados para o uso de forma consciente em tomadas de decisão.

Em um primeiro contato com a base pode ser realizada a análise exploratória, que consiste em uma técnica de tratamento de dados, para obter um primeiro contato com os dados disponíveis, realizando uma sondagem, sem a utilização de uma fundamentação matemática mais rigorosa (PINHEIRO et al, 2009). Conforme aumenta a quantidade dos dados a serem utilizados na análise exploratória, faz-se necessário o uso de ferramentas, sendo que tal análise extrapola a capacidade humana em uma investigação manual ou demanda tempo desnecessário sem a utilização de uma ferramenta adequada. Com o uso de tais ferramentas, é possível apresentar os diferentes aspectos presentes nos dados, com o propósito de encontrar “pistas” sobre suas interações (SILVA, 2016).

A estatística descritiva descreve e sintetiza as principais características passíveis de serem observadas em um conjunto de dados por meio de tabelas, gráficos e medidas-resumo, permitindo ao pesquisador melhor compreensão do comportamento dos dados.

A análise é baseada no conjunto de dados em estudo (amostra), sem tirar quaisquer conclusões ou inferências acerca da população. A estatística descritiva é utilizada por pesquisadores para estudar uma única variável (estatística descritiva univariada), duas variáveis (estatística descritiva bivariada) ou mais de duas variáveis (estatística descritiva multivariada) (FÁVERO; BELFIORE, 2017).

Na estatística descritiva univariada, ou seja, com apenas uma variável, podem ser avaliados a frequência de ocorrência de um conjunto de observações, a

representação da distribuição de uma variável e medidas representativas de uma série de dados, como medidas de posição ou localização, medidas de dispersão ou variabilidade e medidas de forma (assimetria e curtose) (FÁVERO; BELFIORE, 2017).

Ainda nas análises descritivas temos análise descritiva bivariada, a qual tem por objetivo estudar as relações entre duas variáveis, podendo ser uma associação para variáveis qualitativas e correlação para variáveis quantitativas (FÁVERO e BELFIORE, 2017).

Em determinados casos, faz-se necessário estudar a inter-relação com múltiplas variáveis, a fim de avaliar a existência de eventuais associações, correlações ou identificação de grupos, para estas situações utiliza-se as técnicas multivariadas exploratórias. Tais técnicas, também conhecidas como técnicas de interdependência, tem o potencial de serem utilizadas em todas as áreas do conhecimento em que o objetivo seja o de estudar relação entre variáveis, sem que haja a intenção de criar modelos confirmatórios. Segundo Fávero e Belfiore (2017, p. 305)

Não existe, portanto, a definição de uma variável preditora em modelos exploratórios, e neste sentido, seus principais objetivos referem-se à redução ou simplificação estrutural dos dados, a classificação ou agrupamento de observação e variáveis, a investigação da existência de correlação entre variáveis métricas ou associação entre variáveis categóricas e entre suas categorias, a elaboração de rankings de desempenho de observação a partir de várias e a construção de mapas perceptuais. As técnicas exploratórias são consideradas extremamente relevantes para que se desenvolvam diagnósticos acerca do comportamento dos dados em anais, e neste sentido, seus mais diversos procedimentos são comumente adotados de forma preliminar, ou até mesmo simultânea, a aplicação de determinado modelo confirmatório. (FÁVERO e BELFIORE, 2017, p 305)

2.5.1 Limpeza de dados

Tais análises, anteriormente citadas, nos auxiliam no processo de reconhecimento dos dados presentes no conjunto de dados, realizando a função de pré-processamento dos dados. Tal etapa se faz necessária pois não raramente os dados apresentam algumas falhas ou estão organizados de forma inadequada, para serem submetidos a alguns algoritmos, gerando assim dificuldades para a descoberta de conhecimento nos dados. Problemas típicos associados à baixa qualidade de dados são ausência de valores, dados ruidosos, valores inconsistentes, atributos de naturezas distintas e redundância de dados

Para solucionar estes problemas, os dados são submetidos a um processo de limpeza ou *data cleaner*, o qual tem por objetivo amenizar problemas advindos de valores ausente ou valores ruidosos.

Segundo Silva (2016, p. 47), as soluções mais comuns para resolver o problema de valores ausentes são:

- Remoção do exemplar em que ocorre a falta do valor: trata-se de uma solução bastante simples, porém não é efetiva, principalmente quando a quantidade total de exemplares no conjunto de dados é pequena ou quando a ausência dos valores ocorre em um grande número de exemplares.

- Preenchimento manual dos valores: se houver a possibilidade de fazê-lo, ou seja, se o processo de aquisição de dados puder ser executado novamente ou se houver um especialista no contexto, é possível encontrar o valor correto para determinado atributo de determinado exemplar; entretanto, também se torna inviável quando a quantidade de valores ausentes é grande, e corre-se o risco de impor um viés aos dados, caso o preenchimento seja feito a partir de um especialista com conhecimento parcial do contexto dos dados.

- Preenchimento automático dos valores: o preenchimento automático pode ser feito a partir do estabelecimento de um valor constante (o mais frequente no conjunto de dados para o atributo em questão), um valor médio ou mediano (considerando os demais valores presentes no atributo em questão) ou com base no resultado de um método de análise preditiva (um modelo de regressão, por exemplo). Contudo, há de sempre se tomar o cuidado de não associar um viés aos dados. (SILVA, 2016, p. 47)

Ao referir-se a valores ruídos, trata-se de valores cuja referências originais foram modificadas, portanto, consiste em um erro de medida ou valores, consideravelmente, diferentes da maioria dos outros valores do conjunto de dados, também conhecidos como *outliers*.

Ainda nesse sentido, Silva (2016, p.47), aponta duas abordagens para o tratamento de dados ruidosos:

- Inspeção e correção manual: a análise exploratória pode dar condições ao analista de dados de perceber a existência de ruídos nos dados (por exemplo, com a análise de gráficos de caixas ou com a observação da forma da distribuição de frequências). Esses valores ruidosos, se em pouca quantidade, podem ser manualmente corrigidos. Se em quantidades maiores, sua influência na análise de dados pode ser amenizada com procedimentos de suavização que usam a criação de faixas de valores guiadas por medidas de posição, como média ou mediana e medidas separatrizes.

- Identificação e limpeza automática: existem algoritmos que podem ser usados para suavizar, ou anular, ruídos em conjuntos de dados. Algoritmos que resolvem problemas de predição e de agrupamento podem ser modelados para identificar e separar exemplares ruidosos de um conjunto de dados (um exemplo de algoritmo que pode ser usado para este fim é o DBSCAN, veja Capítulo 4), ou para mapear um valor ruidoso para uma classe ou grupo de dados ou para uma faixa de valores que o torne compatível com os valores esperados. A depender do tipo de dado sob

análise, técnicas de tratamento de sinais podem ser aplicadas para separar valores ruidosos de não ruidosos. (SILVA, 2016, p. 47)

2.5.2 Mineração de dados

Apesar da presente popularidade do termo nos dias atuais, estudos para sistematizar a descoberta de padrões a partir de bases de dados não é algo recente. Estatísticos da área de análise de dados, desde os anos 1960, esforçavam-se por meio de métodos indutivos para aprimorar os conhecimentos nesta área. Com os anos, esta metodologia foi chamada de “*Data Fishing*” ou “*Data Dredging*”, até os anos 1990 quando surgiu o termo “*Data Mining*” ou Mineração de Dados (SILVA, 2016). O termo mineração de dados foi cunhado como alusão aos processos de mineração, no qual consiste na extração de minerais valiosos com o uso de ferramentas de uma mina, com relação a alusão podemos comparar a base de dados com a mina, na qual serão aplicados os algoritmos (ferramentas), para a obtenção do conhecimento o qual seria o mineral valioso (CASTRO, FERRARI, 2016)

A mineração de Dados, é definida como:

...um processo automático ou semiautomático de explorar analiticamente grandes bases de dados, com a finalidade de descobrir padrões relevantes que ocorrem nos dados e que sejam importantes para embasar a assimilação de informação importante, suportando a geração de conhecimento... (SILVA, 2016, p. 7)

Em outras palavras, a mineração consiste na implementação e técnica, por meio de algoritmos computacionais, que são capazes de receber um conjunto de dados na entrada e desenvolver como saída, apresentação de associações, agrupamentos, classificações, regressões, padrões frequentes, correlações, análise de grupos, detecção de desvios, entre outros (SILVA, 2016).

Segundo Castro e Ferrari, (2016, p. 7), a mineração de dados é interdisciplinar e multidisciplinar, envolvendo o conhecimento das áreas de banco de dados, estatística, aprendizagem de máquina, computação de alto desempenho, reconhecimento de padrões, computação natural, visualização de dados, recuperação da informação, processamento de imagens e de sinais, análise espacial de dados, inteligência artificial, entre outras.

A funcionalidade da mineração de dados pode ser classificada segundo Castro e Ferrari, (2016), em duas características: as descritivas e as preditivas. As descritivas procuram realizar a caracterização geral dos dados analisados, já as preditivas procuram fazer inferência a partir dos dados objetivando previsões quando ao seu comportamento.

2.5.2.1 Classificação

Castro e Ferrari (2016), classificar significa atribuir um rótulo, o qual pode ser chamado de *classe*, de acordo com a categoria a qual ele pertence.

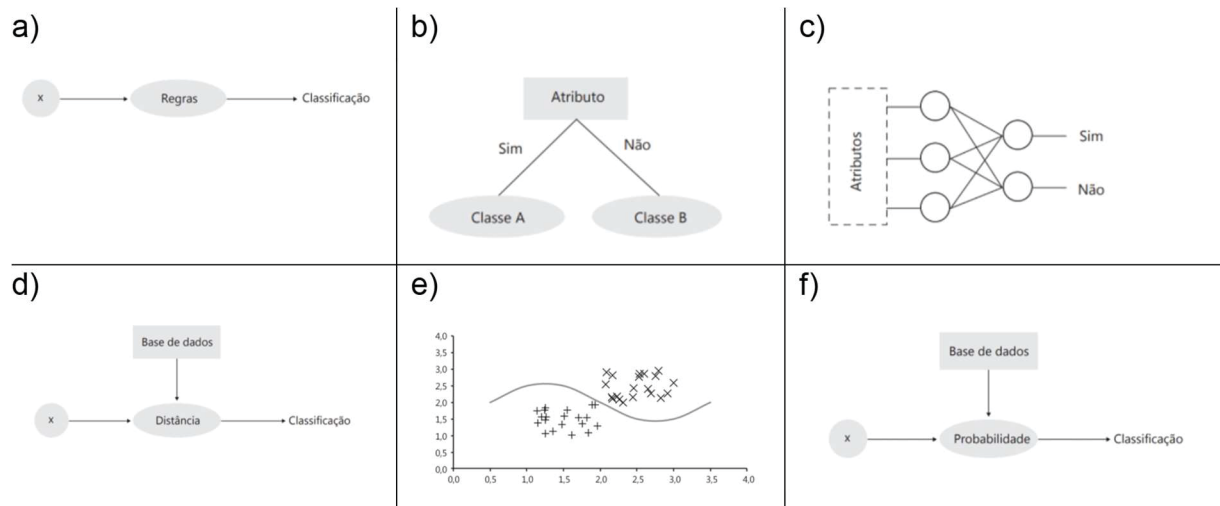
Segundo Silva, (2016, p.75), denomina-se classificação como, “... o processo pelo qual se determina um mapeamento capaz de indicar a qual classe pertence qualquer exemplar de um domínio sob análise, com base em um conjunto de dados já classificado.”

A Classificação pode ser dividida em pelo menos dois tipos, a classificação binária onde os dados a serem classificados são o máximo de 2 variáveis, e os classificadores multiclases, utilizado para conjuntos de dados com mais de 2 variáveis.

De acordo com Castro e Ferrari (2016), os algoritmos de classificação podem ser separados de acordo com a sua estrutura e podem ser encontrados: a) baseados em conhecimentos, o qual opera por meio de um conjunto de regras utilizadas para atribuir a classe a um objeto, caso sejam satisfeitas condições preestabelecidas; b) baseado em árvores - apresenta um objeto em uma estrutura de árvores para agrupamento de dados, onde nesta árvore os nós representam testes sobre o atributo, os ramos representam os resultados desses testes, e nós e folhas os rótulos das classes, sendo válido destacar que em uma árvore de classificação todos os níveis possuem um papel importante no processo classificatório; c) Conexionistas baseados em redes de unidades interconectados, os mais comuns são as Redes Neurais Artificiais; d) baseados em distância, onde se calcula a distância de um objeto cujo se deseja conhecer a classe em relação a um ou mais objetos rotulados, ao fim a classe do objeto passa-se a ser do grupo que estiver a menor distância; e) Baseado em função são modelos paramétricos em funções predefinidas, onde os parâmetros são ajustados durante o processo de treinamento do algoritmo, após o treinamento um novo objeto sem classe conhecida é apresentado, e seu valor calculado e que

representa, de alguma forma, a classe deste objeto; f) Probabilísticos no qual permite atribuir uma probabilidade de um objeto pertencer a uma ou mais classes possíveis.

FIGURA 4 - REPRESENTAÇÕES GRÁFICAS DOS MODELOS DE CLASSIFICAÇÃO, A) BASEADO EM CONHECIMENTO, B) ÁRVORES, C) CONEXIONISTAS, D) DISTÂNCIA, E) FUNÇÃO E F) PROBABILÍSTICA.



FONTE: Castro e Ferrari (2016)

2.5.2.2 Associação

A mineração de regras de associação é uma técnica, que por meio da formação de regras, constrói as relações entre itens de uma base de dados transacional. Diferente dos algoritmos de agrupamento, que buscam as relações de similaridade entre objetos, já as regras de associação buscam relações entre os itens que compõem a base (CASTRO, 2016).

A aplicação da mineração de associação tem como objetivo encontrar regras fortes de acordo com alguma medida do grau de interesse da própria regra. As tarefas de associação e correlação diferenciam-se entre si, pois, enquanto a associação busca relações de co-ocorrência entre dois itens, o que significa quando um ocorre o outro também ocorre, a correlação avalia a dependência linear entre os itens (CASTRO, 2016).

As regras de associação não são diferentes das regras de classificação, exceto pelo fato de que elas podem ser usadas, não apenas a classe, mas para prever qualquer atributo. Esta característica confere-lhe a liberdade também de prever combinações de atributos (CASTRO, 2016). Além disso, as regras de associação não

são planejadas para serem utilizadas em conjunto, como em regras de classificação. Diferentes regras de associação expressam diferentes regularidades da base de dados e sendo geralmente utilizadas para estimar relações distintas entre os itens. Mesmo que pequena, pode derivar uma grande quantidade de regras de associação, normalmente se objetiva a derivação de regras que suportem um grande número de transações, o qual possuam uma confiança razoável.

Os autores Castro (2016, p. 236) mencionam que, os requisitos para determinar a força de uma regra, estão associados a dois conceitos centrais:

Suporte: o suporte, ou cobertura, de uma regra de associação é o número de transações para as quais ela faz a predição correta. Também pode ser entendida como a utilidade de uma dada regra.

Confiança: a confiança, ou acurácia, de uma regra é o número de transações que ela prediz corretamente proporcionalmente às transações para as quais ela se aplica. Também pode ser entendida como a certeza de uma dada regra. (CASTRO, 2016, p. 236).

O algoritmo Apriori descrito por Castro, (2016, p.247):

O algoritmo Apriori é o método mais conhecido para a mineração de regras de associação e emprega busca em profundidade 10 e gera conjuntos de itens candidatos de k elementos a partir de conjuntos de itens com $k - 1$ elementos. Os itens candidatos não frequentes são eliminados, e toda a base de dados é rastreada e os conjuntos de itens frequentes, obtidos a partir dos conjuntos de itens candidatos. (CASTRO, 2016, p.247)

Uma das principais aplicações do Apriori é na descoberta de padrões de compras. Por este motivo, também é conhecido como método de *Market Basket Analysis*.

2.5.2.3 Agrupamento

A análise de agrupamento tem como objetivo verificar a existência de comportamentos semelhantes entre observações em relação a determinadas variáveis, a fim de descobrir grupos ou *clusters*, no qual prevaleça uma homogeneidade interna dos grupos ainda que heterogêneos entre si, facilitando a visualização dos grupos.

Para realizar uma análise de agrupamento, se faz necessário a definição das medidas de distância ou de semelhança, as quais servem de base para a observações. Um determinado esquema de aglomeração deverá ser definido entre os métodos hierárquicos e não hierárquicos.

Segundo Fávero e Belfiore, (2017, p. 310) diferenciam os esquemas hierárquicos e não hierárquicos como:

Enquanto os esquemas hierárquicos permitem a identificação do ordenamento e da alocação das observações, oferecendo possibilidades para que o pesquisador estude, avalie e decida sobre a quantidade de agrupamentos formados, nos esquemas não hierárquicos, parte-se de uma quantidade conhecida de *clusters* e, a partir de então, é elaborada a alocação das observações nesses *clusters*, com posterior avaliação da representatividade de cada variável para a formação deles. (BELFIORE, 2017, p. 310)

Entre os métodos de aglomeração não hierárquicos, o mais popular é o *K-means* (K- Medias), Silva, (2016, p.154), descreve o algoritmo:

Esse algoritmo tem como objetivo encontrar partições no conjunto de dados de forma que k grupos disjuntos de exemplares sejam descobertos, sendo que k é um parâmetro de entrada para o algoritmo. A busca pela descoberta do conjunto de k partições é interativa e classicamente iniciada a partir da escolha aleatória de k vetores distintos que têm o papel de representar centroides (ou protótipos) para grupos. Como esses vetores devem representar grupos de exemplares, eles devem estar localizados no mesmo espaço, ou seja, os vetores k possuem d coordenadas, em que d é o número de atributos descritivos do conjunto de dados. A partir da escolha inicial para os centroides, o algoritmo procede verificando quais exemplares são mais similares a quais centroides, ajustando os centroides aos exemplares mais similares a eles e, dessa forma, estabelecendo as partições. (SILVA, 2016, p.154)

Esta seção apresentou os principais conceitos que norteiam a pesquisa em pauta, trazendo as principais características dos metadados presentes nas bases de dados sobre biodiversidade, bem como as principais formas de análise que serão aplicadas na pesquisa em pauta. A próxima seção apresenta o percurso metodológico adotado para atingimento dos objetivos.

3 ENCAMINHAMENTOS METODOLÓGICOS

Neste capítulo são apresentados os procedimentos e características adotadas ao longo da pesquisa. Sendo apresentados os capítulos 3.1 CARACTERIZAÇÃO DA PESQUISA; 3.2 MATERIAIS E MÉTODOS; 3.2.1 VISÃO GERAL DA PESQUISA; 3.2.2 WEKA; 3.2.3 *Software R e Biblioshiny*; 3.2.4 Microsoft Excel; 3.2.5 Base GBIF; 3.2.6 Revisão sistemática de literatura; 3.2.7 Escolha da base de dados sobre biodiversidade e 3.2.8 Métodos de análise.

3.1 CARACTERIZAÇÃO DA PESQUISA

O propósito da pesquisa pode ser classificado como descritiva, uma vez que procura descrever e detalhar as características presentes nas bases existentes, a fim e determinar as características compartilhadas e as possíveis exclusividades.

Quanto à abordagem a pesquisa se caracteriza como mista, sendo qualitativa por meio da seleção de elementos de caráter classificatório para as informações a serem utilizadas no modelo proposto, os quais não podem ser apresentados de forma numérica. E de caráter quantitativo quanto à análise dos dados relacionados a recorrência de registros, número de variáveis, e afins, informações as quais são melhor representadas por meios numéricos.

Quanto ao delineamento da pesquisa ela se caracteriza como metodológica (VERGARA, 1997), tendo em vista que se refere a construção de uma proposta de análise dos dados da biodiversidade, por meio de métodos originalmente empregados em outro tipo de dados

A natureza da pesquisa é aplicada, uma vez que objetiva delinear uma metodologia

QUADRO 2 – PASSOS PARA REALIZAÇÃO DA PESQUISA

Id	Descrição	Motivador	Resultado
1	Pesquisa na base de periódicos	Identificação de trabalhos já desenvolvidos sobre o tema	Trabalhos encontrados não possuem método semelhante ao trabalho
2	Identificar as bases de dados de Biodiversidade	Delimitação do objeto de estudo	Serão utilizadas as bases de dados sobre a biodiversidade presente no GBIF
3	Escolha das bases a serem utilizadas	Escolha da amostra	Dados da Fauna pertencente ao território brasileiro
4	Desenvolvimento do modelo	Adquirir um instrumento de avaliação	Modelo apresentado na seção 4.3
5	Aplicação do modelo	Validar a efetividade do modelo	Resultados apresentados nas seções 4.4 a 4.10
6	Conclusão	Conclusão da pesquisa e alinhamento dos objetivos.	Considerações apresentadas na seção 5

FONTE: O autor (2021).

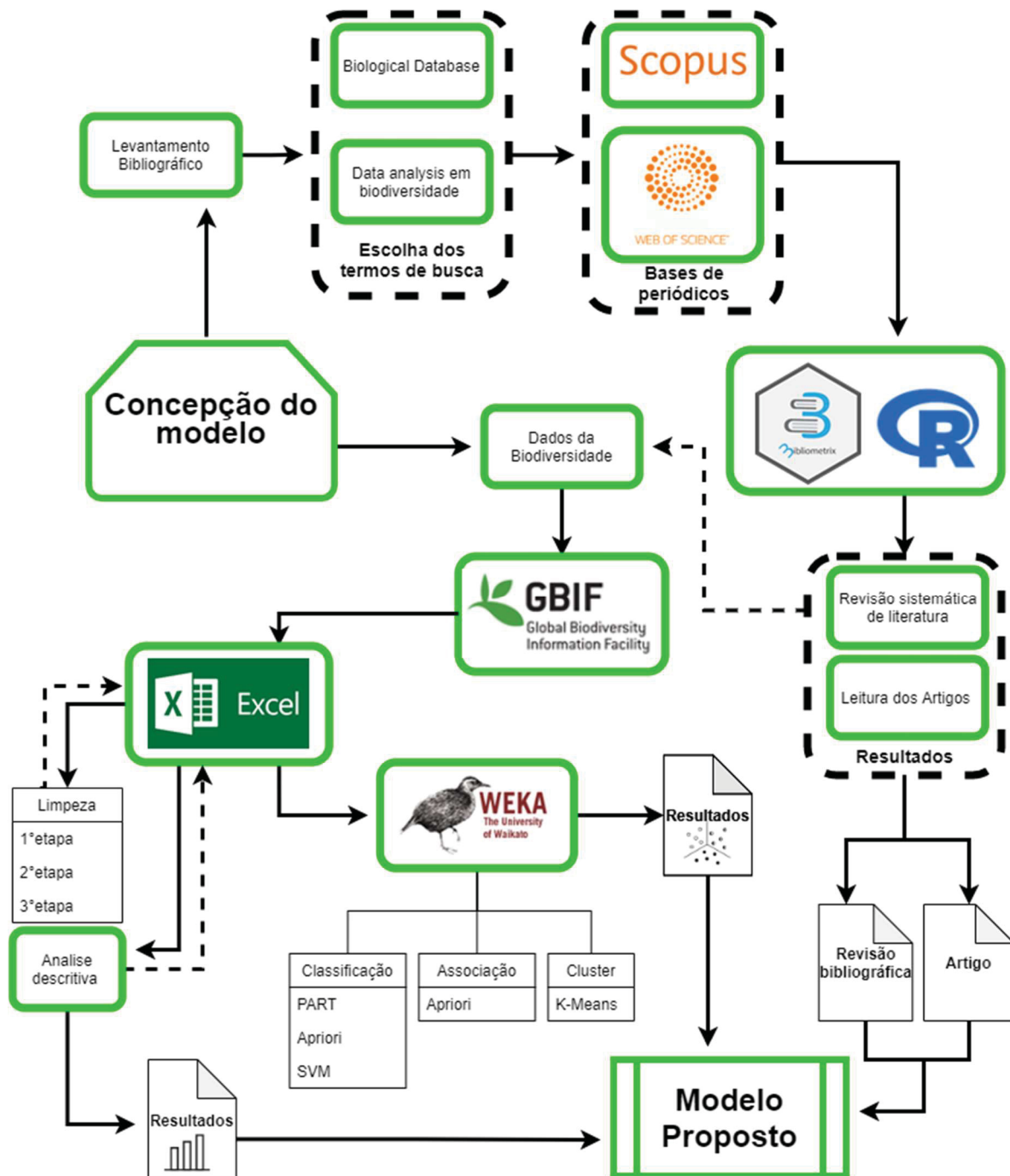
3.2 MATERIAIS E MÉTODOS

Neste capítulo serão apresentados os *softwares* utilizados na pesquisa e apresentando a base onde serão obtidos os dados utilizados nas análises. Além do Excel para análises de estatística descritiva tais como: média, desvio padrão, mínimos e máximos, as próximas seções apresentam as ferramentas complementares.

3.2.1 VISÃO GERAL DA PESQUISA

A FIGURA 5 apresenta o fluxograma de desenvolvimento da pesquisa. As etapas e *softwares* são descritos nos subcapítulos 3.2.2 ao 3.2.8.

FIGURA 5 -FLUXOFRAMA DE DESENVOLVIMENTO DA PESQUISA



FONTE: O autor (2021)

3.2.2 WEKA

Para a realização das análises será utilizado o *software* WEKA (*Waikato Environment for Knowledge Analysis*), desenvolvido pela Universidade de Waikato na Nova Zelândia, o qual possui uma coleção de algoritmos de aprendizado de máquina e de processamentos de dados (FRANK, HALL, WITTEN, 2016). O programa possui uma ampla variedade de algoritmos que podem ser aplicados de maneira simples em

conjuntos de dados, com foco no presente trabalho, a função que possui maior relevância são os métodos de aprendizado com os quais podemos aprender mais sobre o conjunto de dados.

O *software* apresenta entre os algoritmos, os algoritmos de aprendizagem divididos entre classificadores Bayesianos, árvores de decisões (*Trees*), regras (*Rules*), funções (*Funtions*), entre outros. A também os algoritmos de cluster e de Associação, os qual buscam regras de associação, com destaque para o algoritmo de cluster *CobWeb* e *SimplesKMeans*. Tem também o *Attribute selection*, o que possibilita acesso a diversos métodos para a seleção de atributos, provendo avaliadores de atributos e método de busca (FRANK, HALL, WITTEN, 2016).

O WEKA tem como formato nativo para o seu funcionamento e ARFF (*AttributeRelation File Format*), que consiste em uma lista de instâncias, e o valores de atributos para data instância separadas por virgula (FRANK, HALL, WITTEN, 2016). O programa também aceita em arquivos em CSV (*Comma Separated Values*), formato que pode ser gerado pelo Microsoft® Excel. O presente estudo utiliza o Microsoft Excel 2016.

Para as análises por mineração de dados, é utilizado o *software* WEKA versão 3.8.3 de 2018, para Windows, programa disponível no site oficial do WEKA (<https://www.cs.waikato.ac.nz/ml/weka/>).

3.2.3 *Software* R e Biblioshiny

O R é, ao mesmo tempo, uma linguagem de programação e um ambiente computacional para estatística e foi desenvolvido pelos estatísticos Ross Ihaka e Robert Gerntleman, no departamento de Estatística da Universidade de Auckland, da Nova Zelândia, na década de 90.

O *software* R está disponível como *software* livre, sob os termos da *Free Software Foundation's GNU (General Public License)* no formato de código fonte. Este código pode ser compilado e executado em uma variedade de plataformas e sistemas como o Windows, Linux e MacOs (R Projet, 2020).

O R possui uma variedade de recursos para análises estatística, como modelos lineares e não lineares, testes estatísticos clássicos, análise de serie temporal, classificação, agrupamentos, entre outros. O programa e altamente

expansível, por meio de ferramentas (*Tool's*), desenvolvidas pela comunidade de usuários (R Projet, 2020).

O Bibliometrix é um pacote de ferramentas livres para o *software* R, utilizados em pesquisas cientométricas e bibliométricas, incluindo todos os principais métodos de análise bibliométrica. No pacote encontram-se rotinas para a importação de dados bibliográficos das bases de periódicos *SCOPUS*, *Clarivate Analytics' Web of Science*, *PubMed*, *Digital Science Dimentions* e *Cochrane databases* (ARIA, CUCCURULLO, 2016).

A ferramenta realiza análises bibliométricas, matrizes de cocitação, ligações, análise de colaboração científica e análises de correlação de palavras. Ainda a ferramenta é flexível podendo ser rapidamente integrada aos demais pacotes do R (ARIA, CUCCURULLO, 2016).

Na versão 2.0 foi introduzido o *Biblioshiny*, uma interface gráfica que facilita a utilização dos recursos disponibilizados pelo Bibliometrix, sem a necessidade de desenvolver habilidade de programação na linguagem R (ARIA, CUCCURULLO, 2016).

3.2.4 Microsoft Excel

O Excel é um editor de planilhas, desenvolvido pela empresa Microsoft, para utilização em computadores com sistema operacional Microsoft Windows, estando disponível também para o sistema macOS e sistemas operacionais para dispositivos móveis tais como Android ou iOS.

Programas de planilha como o Excel possuem características que propiciam formas de interação únicas com os dados, dentre as características é válido destacar: a) criação de fórmulas facilitada com a interface *point and click*, b) células relativas e absolutas uma vez que as variáveis estejam nomeadas, c) interação e cálculos múltiplos e d) recálculo automático perante mudança de valores nas células (BAIER, NEUWIRTH, 2007).

Neste trabalho a ferramenta é utilizada para limpeza e visualização dos dados e aplicação de métodos de estatística descritiva.

3.2.5 Base GBIF

O GBIF é uma rede internacional e uma infraestrutura de dados, com o princípio de prover para qualquer um em qualquer lugar, acesso livre aos dados sobre todos os tipos de vida na terra.

Os dados contidos neste repositório, são fornecidos por instituições ao redor do planeta, as quais utilizam o mesmo padrão de dados e ferramentas *opensource*, facilitando o compartilhamento das informações sobre qualquer espécie que tenha sido encontrada. As fontes da informação provêm de diversos recursos, desde registros de espécimes preservado em museus, que podem datar de séculos atrás, até fotos com dados de georreferenciamento compartilhados por naturalistas amadores.

Todas as fontes são reunidas pelo uso do padrão Darwin Core, no qual são reunidos pela rede do GBIF que atualmente possui mais de 1,6 bilhão de registros de ocorrência de espécies. Aqueles que publicam os dados fornecem acesso aberto aos seus conjuntos de dados usando definições de licença *Creative Commons*.

O surgimento do GBIF se deu por uma recomendação do subgrupo de informática para a biodiversidade do fórum de Megasciência da Organização e Desenvolvimento Econômico em 1999, no relatório gerado por este grupo concluiu que “ é necessário um mecanismo internacional para os dados e informações sobre biodiversidade acessíveis em todo o mundo”, relatando que este mecanismo poderia trazer muitos benefícios no futuro.

O Brasil também está presente nesta rede, por meio da plataforma online SiBBr, a qual foi criada pelo Ministério da Ciência, Tecnologia e Inovação (MCTI) com suporte técnico do *United Nation Environment Programme* (UNEP) e suporte financeiro do *Global Environment Fund* (GEF). O SiBBr tem como propósito realizar a gestão e uso da informação da biodiversidade, com o intuito do Brasil expandir a sua capacidade de conservação e uso da biodiversidade. Esta plataforma busca encorajar e facilitar a publicação integração, acesso e uso da informação da biodiversidade.

O GBIF e a base de dados que será utilizada para testar o modelo de análise proposto. Nesta base estão depositados dados coletados de registros de ocorrências presentes no território brasileiro. Para a coleta dos dados, o Portal GBIF dispõe de ferramentas de filtros e exportação.

Para fins desta pesquisa, foram aplicados os filtros por país ou área (*Country or área*) onde foi selecionada a opção do Brasil, e no campo de nomes científicos foi selecionado o reino Animalia, para assim obter-se apenas os resultados referente à fauna. Os resultados decorrentes desta pesquisa, são disponibilizados pela plataforma em dois formatos o de registos completos no formato DwC e registos pré-processados pelo sistema do GBIF. Foram coletados os registos completos, os quais foram submetidos a etapas de limpeza e padronização para utilização no ambiente do WEKA. Ainda no portal são fornecidas estatísticas básicas referentes aos registos recuperados, conforme apresentado na seção 4.2.

3.2.6 Revisão sistemática de literatura

A fim de compreender como tema do presente estudo está sendo trabalhado, buscar referência para compor o referencial teórico e métodos para compor a proposta apresentada, foi realizada uma revisão sistemática de literatura, a qual seguiu o protocolo descrito no Quadro 3.

QUADRO 3 – PROTOCOLO DE REVISÃO SISTEMA DE LITERATURA APLICADO COM FOCO EM ANÁLISE DE DADOS DA BIODIVERSIDADE

Questão de pesquisa	Quais as abordagens utilizadas para a realização de análises em dados oriundos de repositórios de biodiversidade?
Objetivo (s)	Estudar as abordagens utilizadas nas análises nos repositórios de periódicos;
População/amostra/foco	Análise de dados da biodiversidade
Bases de dados	Web of Science e Scopus
Justificativas bases de dados	Bases como maior abrangência

Estratégia de busca	"DATA ANALYSIS" AND BIODIVERSITY, "DATA ANALYSIS" AND "BIODIVERSITY DATABASE", "DATA ANALYSIS" AND BIODIVERSITY AND DATABASE, "DATA ANALYSIS" AND BIODIVERSITY AND FAUNA, "DATA ANALYSIS" AND "BIODIVERSITY INFORMATICS" e "DATA ANALYSIS" AND "PRIMARY BIODIVERSITY DATA".
Crítérios de inclusão/seleção	Fauna, Animais, análise de dados. Métodos de análise
Crítérios de exclusão	Botânica, plantas, proteínas, DNA, RNA, genética
Recorte temporal	Todos os artigos até 2020

FONTE: O autor (2020)

Os artigos resultantes deste processo foram submetidos a análises utilizando o *software* R, com a biblioteca Bibliometrix na interface do *Biblioshiny*, no qual foram realizadas as análises da distribuição dos documentos recuperados em tipos (Artigos, anais de congresso e Capítulo de livros), número de autores, palavras chaves dos autores e estendidas (geradas automaticamente pela base da Scopus), produção científica por ano, distribuição dos documentos por país, revistas com o número de publicações entre os documentos, pesquisadores mais citados, autores que mais produziram, os 5 artigos mais referenciados, os 5 artigos mais citados nas referências, palavras chaves mais utilizadas, dinâmica das palavras chaves ao longo do tempo e análise de coocorrência das palavras chaves.

Os 5 artigos mais referenciados e os 5 artigos mais citados nas referências, foram lidos na íntegra para compor como base a análise. Os resultados referentes a estas análises realizadas são apresentados no capítulo 4, seção 4.1.

3.2.7 Escolha da base de dados sobre biodiversidade

A fim de identificar as bases passíveis de serem utilizadas como fonte de dados para o modelo que será proposto, utilizou-se de uma revisão sistematizada, com o seguinte protocolo de pesquisa.

QUADRO 4 – PROTOCOLO DE REVISÃO SISTEMA DE LITERATURA APLICADO COM FOCO EM BASE DE DADOS DE BIODIVERSIDADE.

Questão de pesquisa	Quais são os repositórios existentes que contém dados sobre a biodiversidade da fauna brasileira?
Objetivo (s)	Buscar repositórios nacionais de biodiversidade e estudar seus formatos de metadados;
População/amostra/foco	Bases de dados de biodiversidade
Bases de dados	Web of Science e Scopus
Justificativas bases de dados	As bases Web of Science e Scopus pela grande abrangência
Estratégia de busca	"BIODIVERSITY DATABASE", "FAUNA DATABASE" e "TAXONOMIC DATABASE"
Critérios de inclusão/seleção	Fauna, Animais, bases abrangentes
Critérios de exclusão	Bases limitadas a um grupo taxonômico, bases de proteínas, bases genéticas, bases botânicas
Recorte temporal	Todos os artigos até 2020

FONTE: O autor (2020)

Os documentos recuperados foram ao primeiro momento, submetidos a análises utilizando o *software* R, com a biblioteca Bibliometrix na interface do *Biblioshiny*, onde foram realizadas análise da distribuição dos documentos recuperados em tipos (Artigos, anais de congresso e Capítulo de livros), número de autores, palavras chaves dos autores e estendidas (geradas automaticamente pela base da Scopus), produção científica por ano, distribuição dos documentos por país, revistas com o número de publicações entre os documentos, pesquisadores mais citados, autores que mais produziram, os 5 artigos mais referenciados, os 5 artigos mais citados nas referências, palavras chaves mais utilizadas, dinâmica das palavras chaves ao longo do tempo e análise de coocorrência das palavras chaves.

Os resultados desta análise foram publicados em um artigo sob o título “BANCO DE DADOS SOBRE BIODIVERSIDADE: UMA ANÁLISE TEMPORAL DA PRODUÇÃO CIENTÍFICA”. A fim de complementar a pesquisa, foram realizadas buscas no Google Acadêmico e na base *BioOne*, onde os documentos oriundos de todas as buscas, para a identificação da sua relevância para o tema, por meio da leitura dos títulos e resumos, aqueles identificados com informação relevantes, foram realizadas a leitura completa do artigo, e extraídas as informações referentes as bases.

As bases que mais se destacaram, foram submetidas a uma segunda etapa foram realizadas referente a sua fonte, se possuíam site, portal ou outro meio digital de acesso. Na escolha da base, o principal critério foi a abrangência, volume de dados, “popularidade” (entende-se está popularidade como uma alta utilização nos artigos encontrados).

Os resultados referentes a essas análises foram publicados no artigo acima citados, e compuseram as informações apresentadas na seção 2.4 das referências e na base escolhida e detalhada na seção 3.2.5 dos encaminhamentos metodológicos.

3.2.8 Métodos de análise

Buscou-se nestes dados as relações entre os dados e o seu propósito de uso, quais as principais extensões presentes nos dados, relação região e interesse dos dados, quais os tipos de dados mais recorrentes para cada localização do país, se existe alguma relação entre a finalidade do dado com a região dos dados de onde ele foi coletado ou com os fornecedores dos dados, quais os principais erros, falhas e/ou incongruências encontrados nos dados presentes.

A fim de buscar respostas, os dados foram, primeiramente, submetidos a análises exploratórias e descritivas a fim de buscar maiores esclarecimentos. Cabe destacar, que algumas análises foram realizadas na própria plataforma, conforme apresentado na seção de resultados 4.2, para as demais foi utilizado o Microsoft Excel, versão 2016.

Decorrente das dificuldades técnicas, não foi possível trabalhar com os dados em sua totalidade. Contudo, optou-se por um recorte (amostra) representativo do conjunto de dados para a realização das análises.

Para a seleção deste conjunto de dados, retornou-se à plataforma GBIF, e foi refeita a pesquisa. Em seguida, foi utilizada a área “Publisher” e recuperou-se dados no intervalo entre 100 e 500 mil registros de ocorrência, com distribuição de registros por todo o território brasileiro. Esta distribuição foi observada com o auxílio da ferramenta “Map” disponível na plataforma do GBIF. Optou-se pela utilização do filtro “Publisher” ao invés do “Dataset”, uma vez que uma mesma instituição pode disponibilizar mais de uma série de dados, sendo observada uma prática comum entre os fornecedores das informações a de dividir os seus conjuntos de dados por grupo taxonômico de interesse. Assim, realizando-se a busca pelo depositante recuperam-se todas as séries de dados fornecidas. Foi escolhido o conjunto de dados referente ao Museu Nacional do Rio de Janeiro

Realizada a etapa de descrição, e com a identificação prévia dos dados que se encontram no material recuperado, realizou-se um processo de limpeza dos dados (*Data Cleaning*), onde foi realizada em 3 etapas. A primeira etapa foi realizada a remoção das colunas / atributos as quais não foram preenchidas (portanto não continham dados). Na segunda etapa foram removidas as colunas / atributos que não possuísssem variação nos dados ou que não eram passíveis de serem utilizadas nas análises

Na terceira etapa foi realizada a adequação dos dados para as análises, onde foram realizadas correções dos caracteres, codificação de termos e exclusão de linhas com dados incorretos. Foram identificados registros erroneamente marcados como pertencentes ao território brasileiro, os quais foram excluídos, sem sequencia foram realizadas as codificações dos atributos. Foi realizado um processo de remoção ou substituição de caracteres especiais, correção de dados desconfigurados e padronização dos dados que haviam sido escritos de diferentes formas.

Ao final da limpeza o arquivo dos dados foi convertido para o formato aceito pelo programa WEKA, afim se serem submetidos a etapa de mineração de dados.

Os dados foram submetidos a mineração de dados por meio da ferramenta WEKA aplicadas regras de classificação e agrupamento (clusterização) pelo algoritmo K-Means, métodos de árvore de decisão e regras de classificação por meio do J48 e PART uma vez que ambos possuem o C4.5 como base, nas apresentações de resultados distintas, e a regra de associação por meio do Apriori, o qual também foi utilizado como classificador. Após a leitura dos artigos, também foi incluído o método *Support Vector Machines* pelos resultados positivos apresentados pelo pesquisador

Miguel A. Acevedo e demais colaboradores. Este trabalho está detalhado na próxima seção de “resultados e discussão”.

4 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados os resultados referentes aos métodos empregados na bibliografia e nos dados obtidos na análise dos dados biodiversidade, e estes resultados serão confrontados com as informações apresentadas por outros autores.

4.1 VISÃO GERAL DE ANÁLISE DE DADOS E BIODIVERSIDADE

Para compor a visão geral sobre a análise de dados em biodiversidade foi utilizada a metodologia de análise sistemática de literatura com o auxílio de *software* de bibliometria. As pesquisas nas duas bases foram realizadas no dia 10 de outubro de 2020 na base da Scopus e Web of Science, sendo utilizado, para a busca, os termos apresentados na Tabela 1, apresentada anteriormente na justificativa (Páginas 17), escolha foi pelos campos título, resumo e palavra-chave.

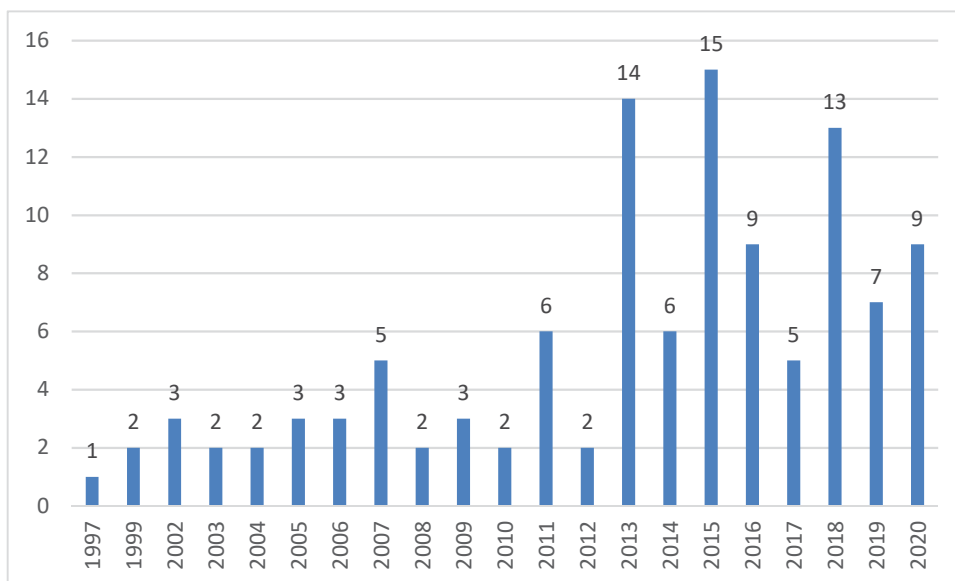
Após as buscas iniciais, foi realizada a seleção dos documentos que possuem alguma relação com o trabalho. Do total de 1.657 documentos recuperados, após seleção dos mais relevantes e remoção de duplicados, restaram 114 documentos que foram submetidos a análise bibliométrica.

Destes 72 são artigos publicados, 20 anais de conferências e 4 capítulos de livros. A primeira publicação encontrada se refere ao ano 1997. Os 114 documentos recuperados foram escritos por 603 autores. Foram identificadas 1275 palavras-chave estendidas (palavras geradas automaticamente pela base) e 603 palavras-chave dos autores.

Conforme o Gráfico 1, é possível observar um baixo volume de produções sobre o assunto no período de 1997 a 2012, variando a produção entre 0 a 6, já no ano de 2013 observa-se um grande aumento das produções, sendo 14 artigos publicados 14 documentos somente neste ano, e 2 grandes picos de produção nos anos de 2015 com 15 documentos e 2018 com 13.

Já o ano de 2020 apresenta 9 documentos mesmo que no período em que foi realizado o levantamento o ano não havia terminado, o que já ultrapassa a produção científica de 2019. E observado o crescimento de 11,61% por ano na produção científica sobre o tema.

GRÁFICO 1 - PRODUÇÃO CIENTÍFICA POR ANO

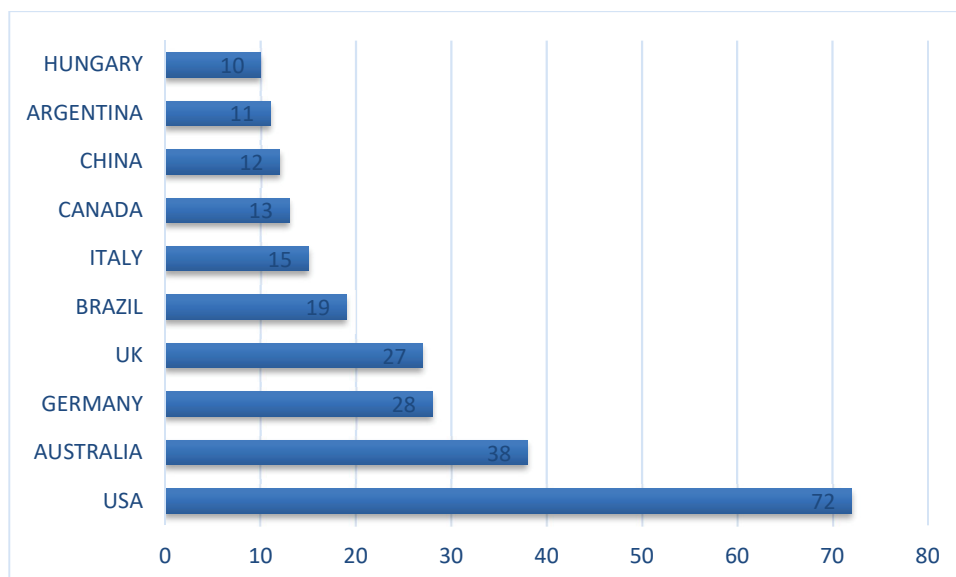


FONTE: Autor, 2021

A distribuição dos documentos por países mostra os Estados Unidos da América (72) com maior número de publicações nesta área, seguido pelo Austrália (38), isso mostra uma predominância de documentos na língua inglesa. Tal resultado pode ter sofrido influência devido a utilização de termos de busca na língua inglesa.

Por outro lado, diversos periódicos em diversos idiomas utilizam o *abstract* como um dos elementos obrigatórios. O Brasil (19) aparece em quinto lugar em número de publicações, conforme demonstra o Gráfico 2.

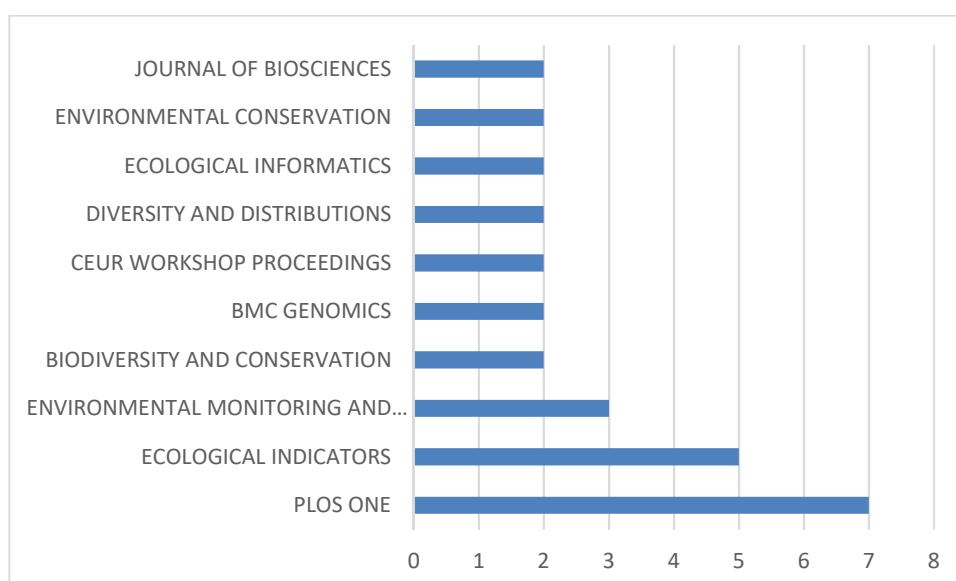
GRÁFICO 2 – Contribuição de produção científica no mundo



Fonte: Autor, 2021

Dentre os 114 documentos recuperados na consulta, a PLOS ONE é a que apresenta o maior número de artigos publicados, no total de 7, seguido pela *Ecological Indicators* com 5. O Gráfico 3 apresenta as 10 fontes com maior número de publicações.

GRÁFICO 3 - FONTES COM MAIORES NÚMEROS DE PRODUÇÕES



Fonte: Autor, 2021

Os cinco (5) pesquisadores mais produtivos estão apresentados no Quadro 3 e os cinco (5) mais citados no Quadro 4. O autor mais produtivo (4) o doutor Paul Roe, professor pesquisador da *Queensland University of Technology*. O autor com o maior número de citações (48) é James D. Nichols pesquisador sênior da *U. S. Geological Survey no Patuxent Wildlife Research Center*.

QUADRO 5 – AUTORES MAIS PRODUTIVOS

Autores	Artigos
ROE P	4
TOWSEY M	3
WIMMER J	3

QUADRO 6 – AUTORES MAIS CITADOS

Autores	citações
NICHOLS J D	48
ALBUQUERQUE U P	30
GASTON K J	29

AGGARWAL A	2
ALLARD A	2

Fonte: Autor, 2021

MARGULES C R	27
PRESSEY R L	27

Fonte: Autor, 2021

Os artigos mais referenciados estão listados na Tabela 2. O artigo mais citado foi escrito por Daniel H. Huson do *Center for Bioinformatics da Tübingen University*, sob o título “*MEGAN analysis of metagenomic data*”, e apresenta o *software* MEGAN por meio da comparação dos resultados deste *software* em relação a outras ferramentas utilizadas por pesquisadores desta área. O *software* é utilizado no estudo de genética, com enfoque no artigo para biodiversidade de microrganismos, utilizando séries de dados presentes em bases de microrganismos.

O artigo sob o título “*Current databases on biological variation: pros, cons and progress*” foi o segundo trabalho mais citado com 818, foi escrito por C. Ricos e colaboradores, e trata a variação da informação biológica em bases de dados de laboratórios médicos. Durante a análise realizada identificaram dados discrepantes, que necessitaram ser removidos, e em seu artigo recomendam que os profissionais que trabalham nestes laboratórios sejam encorajados a verificar estes resultados discrepantes. Apesar deste artigo não possuir uma ligação direta com a proposta da presente dissertação, a recomendação pode ser replicada para os dados da biodiversidade, para evitar que dados discrepantes sejam disseminados, evitando-se assim análises imprecisas ou incorretas.

O terceiro artigo mais citado com 277, com o título de “*Eukaryotic genome size databases*”, escrito por T. Ryan Gregory e colaboradores em 2007, no artigo são tratados do tamanho da informação que tem sido enviada para três bases de dados independentes que hospedam o genoma de organismos Eucariotas (ou seja, seres vivos com células dotadas de núcleo), mais especificamente de uma base para plantas, outra para fungos e outra para animais. De maneira geral, este conjunto de bases apresentam acesso livre para os dados do genoma de mais 10.000 espécies. E tratado no artigo da atualização na qual as bases apresentaram até período de 2005, mas não foi apresentado nenhum padrão para os dados das bases, sendo que cada base possui uma organização própria.

O artigo sob o título de “*How global is the global biodiversity information facility?*”, escrito por Chris Yesson e colaboradores, publicado em 2007, apresenta-

se em quarto lugar na listagem dos artigos mais citados, com 180 citações. O artigo apresenta a uma indagação de quão global são as informações sobre a biodiversidade presentes no *Global Biodiversity Information Facility* (GBIF). O GBIF é uma organização internacional que possui como missão de disponibilizar gratuitamente e livremente os dados sobre a biodiversidade do mundo. Até o ano da publicação do artigo, o GBIF fornecia mais 120 milhões de registros em suas bases, os autores utilizaram para o estudo os registros pertencentes a plantas da família Leguminosa e, observando a cobertura e acurácia dos dados presentes na base. Foram encontrados mais de meio milhão de registros desta família, cobrindo segundo os autores, 31% das espécies da família, e constataram que em suma maioria os registros estão corretos, sendo que 84% passaram nos critérios analisados, sendo entre os critérios a precisão do georreferenciamento dos registros. Os autores concluem que o GBIF ainda não possui os dados necessários para a visualização da biodiversidade, e que os próximos desafios sejam a negociação para a aquisição de dados que ainda não estejam disponíveis para o público e o encorajamento da digitalização e incorporação de registros a seus acervos.

O quinto artigo mais referenciado, com 169 citações, foi escrito por Joaquin Hortal em colaboração com Jorge M. Lobo e Alberto Jimenez-Valverde, sob o título de "*Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife*", publicado em 2007. No artigo os autores analisaram três limitações em uma base de dados de biodiversidade, com dados da ilha Tenerife do arquipélago espanhol das ilhas Canárias, base BIOTA-Canárias. Os itens que foram avaliados são, primeiro a inadequação dos dados para descrever os padrões de riqueza devido a um viés de amostragem; segundo avaliaram a falta de exaustividade na compilação dos dados dos levantamentos, e terceiro a falta de cobertura nas variações ambientais que podem afetar a distribuição dos organismos. Na análise de riqueza foi escolhido um grupo de sementes de plantas bem conhecidas, para assim facilitar a comparação dos resultados. Nos resultados encontraram que os dados presentes nas bases estão incompletos e enviesados, sendo necessário a inclusão de protocolos afim de controlar a qualidade dos dados inseridos nestas e outras bases, os autores reforçam que esta e outras bases são muito importantes para o estudo da biodiversidade. A Tabela 2 apresenta as referências mais citadas e mais recorrentes nas referências dos documentos recuperados.

TABELA 2 - DOCUMENTOS COM O MAIOR NÚMERO DE CITAÇÕES NAS REFERÊNCIAS DOS ARTIGOS RECUPERADOS

Artigo	Total citações
HUSON, Daniel H. et al. MEGAN analysis of metagenomic data. Genome research , v. 17, n. 3, p. 377-386, 2007.	1.686
RICÓS, C1 et al. Current databases on biological variation: pros, cons and progress. Scandinavian journal of clinical and laboratory investigation , v. 59, n. 7, p. 491-500, 1999.	818
GREGORY, T. Ryan et al. Eukaryotic genome size databases. Nucleic acids research , v. 35, n. suppl_1, p. D332-D338, 2007.	277
YESSON, Chris et al. How global is the global biodiversity information facility?. PloS one , v. 2, n. 11, p. e1124, 2007.	180
HORTAL, Joaquín; LOBO, Jorge M.; JIMÉNEZ-VALVERDE, ALBERTO. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. Conservation Biology , v. 21, n. 3, p. 853-863, 2007.	169

FONTE: AUTOR, 2021

A Tabela 3 apresenta as referências mais citadas e recorrentes nas referências dos documentos recuperados.

O artigo mais encontrado nas referências sobre o título, “*The quiet revolution: biodiversity informatics and the internet*”, escrito por Frank A. Bisby em 2000, o artigo trata do surgimento de iniciativas globais relacionadas a distribuição e armazenamento de informações relacionadas a biodiversidade na internet. Em seu artigo são tratados de duas iniciativas no desenvolvimento de systems que permitam a interação e síntese do conhecimento em amplas matrizes de sistemas locais, a iniciativa a qual é mais tratado no artigo *Species* 2000.

O segundo e terceiro artigos mais citados nas referências foram escritos pelo mesmo autor Miguel A. Acevedo e colaboradores, o primeiro sobre o título “*Automated classification of bird and amphibian calls using machine learning: a comparison of methods*” publicado em 2009, em seu artigo realizaram a comparação entre três algoritmos de aprendizado de máquina aplicados para a classificação dos sons de 12 animais (nove sapos e três pássaros). Os algoritmos utilizados na pesquisa: análise discriminante linear, árvore de decisão e *Support Vector Machines* (SVM). Foram utilizados no total de 10.061 de sons isolados dos animais, para treino e teste dos algoritmos, estes sons foram retirados dos sons gravados por um sistema automatizado em quatorze locais de Porto rico, onde foram coletados um total de

2.132 minutos de gravação. No artigo, ainda foram verificadas as classificações corretas de cada método sendo que o *Support Vector Machines* apresentou 94,95% de acertos com 0,94% de falso positivo, árvore de decisão apresentou 89,20% com 1,25% de falso positivo e análise discriminante linear 71,45% de acertos com 1,98% de falso positivos. Segundo os resultados o *Support Vector Machines* mostra-se promissor para aplicações em conjunto com sistemas digitais de gravação, podendo contribuir em muito com a cobertura espacial e temporal de conjuntos de dados biológicos.

O segundo artigo do mesmo autor, sobre o título “*Using Automated Digital Recording Systems as effective tools for the monitoring of birds and Amphibiens*”, publicado em 2006, aborda a concepção e teste de um sistema digital de gravação, com o intuito de promover uma maior quantidade e qualidade de dados sobre a biodiversidade de aves e anfíbios. No artigo foram realizados testes comparando o esse novo método automatizado com o tradicional, os testes foram realizados em 10 locais, no município de Toa Baja, próximo à costa norte de Porto Rico. Os resultados obtidos foram de 17 espécies de pássaros pelo método automático em relação a 16 espécies pelo tradicional, e de 5 espécies de anfíbios pelo método automático em relação a 4 espécies encontradas pelo método tradicional. O método apresentado produziu boa quantidade e qualidade de dados, sendo apontado pelo autor que o método apresenta três vantagens adicionais que são: gravações permanentes de censos, coleta de dados 24 horas por dia e a possibilidade de automação na identificação. Os dois artigos anteriores têm relação com a presente dissertação, ao ponto que o primeiro artigo trata de uma possibilidade de utilização dos dados biológicos de qualidade, e o segundo artigo de uma das possíveis fontes dos dados que podem estar disponíveis nas bases de dados sobre biodiversidade.

O quarto artigo mais citado nas referências (3), sobre o título “*Search and clustering orders of magnitude faster than BLAST*”, escrito por Robert C. Edgar, publicado em 2010. Em seu artigo são tratados dois novos algoritmos (UCLAST e USERCH), utilizados em pesquisas em grandes conjuntos de dados de sequências de DNA. A motivação para o seu trabalho se dá por conta que as sequências de dados biológicos têm se acumulado de maneira mais rápida que a velocidade de crescimentos de processamento da computação, assim motivando a busca por métodos com maior rendimento nas análises. Estes algoritmos apresentaram um desempenho melhor que o algoritmo BLAST (explicado no próximo parágrafo), tendo

maior velocidade, menor uso de memória, com maior sensibilidade e classificação de series de dados muito maiores.

O Artigo “*Basic Local Alignment Search Tool (BLAST)*”, escrito por Stephen F. Altschul e colaboradores, publicado em 1990, está em quinto artigo mais citado nas referências. O artigo trata sobre uma nova ferramenta para a época, de comparação de sequência rápida, que tem o nome de *Basic Local Alignment Search Tool (BLAST)*. No artigo são apresentadas as vantagens da ferramenta em comparação aos recursos da época.

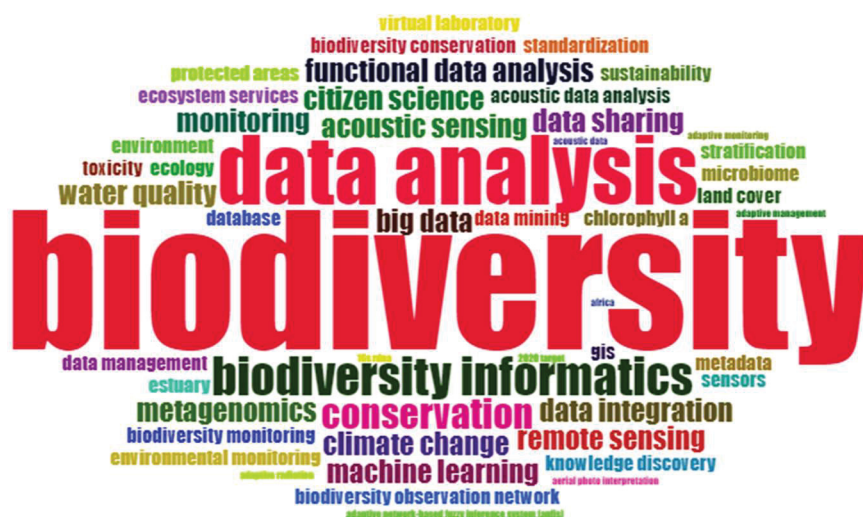
TABELA 3 - ARTIGOS MAIS CITADOS NAS REFERENCIAS

Artigos	Citações
BISBY, Frank A. The quiet revolution: biodiversity informatics and the internet. Science , v. 289, n. 5488, p. 2309-2312, 2000.	4
ACEVEDO, Miguel A. et al. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. Ecological Informatics , v. 4, n. 4, p. 206-214, 2009.	3
ACEVEDO, Miguel A.; VILLANUEVA-RIVERA, LUIS J. From the field: Using automated digital recording systems as effective tools for the monitoring of birds and amphibians. Wildlife Society Bulletin , v. 34, n. 1, p. 211-214, 2006.	3
EDGAR, Robert C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics , v. 26, n. 19, p. 2460-2461, 2010.	3
ALTSCHUL, Stephen F. et al. Basic local alignment search tool. Journal of molecular biology , v. 215, n. 3, p. 403-410, 1990.	2

Fonte: Autor, 2021

A palavras-chaves mais utilizadas pelos autores *Biodiversity* (18), seguido por *data analysis* (9), *biodiversity informatics* (5), *conservation* (4) e *acoustic sensing* (3), estas frequências das palavras provavelmente foi afetada pela metodologia de seleção dos documentos realizada nesta pesquisa, os dois termos mais relevantes foram utilizados como termos de busca na recuperação dos documentos. A Figura 6 apresenta as 50 palavras-chave com maior ocorrência das 1033.

FIGURA 6 – AS 50 PALAVRAS-CHAVES DO AUTOR COM MAIOR OCORRÊNCIA



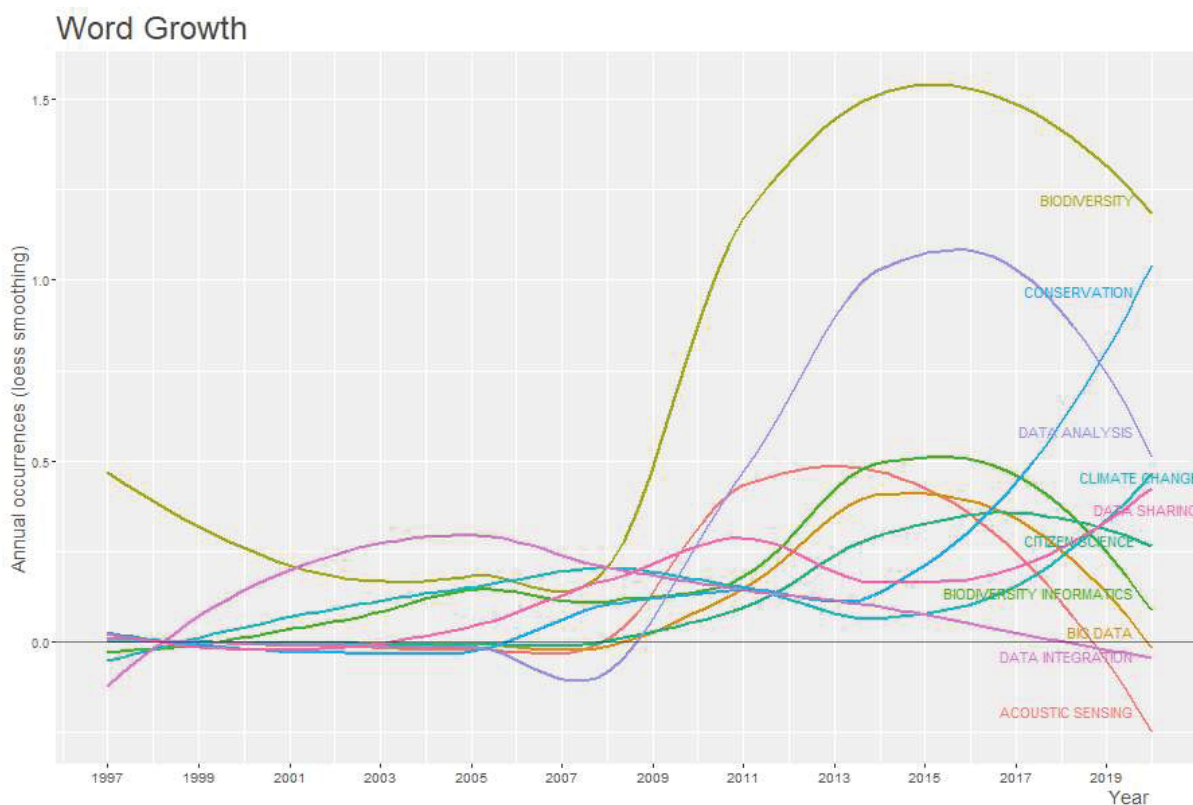
Fonte: Autor, 2021

Com relação às palavras-chave estendidas, a palavra mais utilizada é Biodiversity (137), seguido por *data analyses* (43), essas palavras foram utilizadas nos termos de busca, por este motivo torna-se claro as suas posições. Na sequência os termos *Article* (34), *environmental monitoring* (27), *nonhuman* (22), *animals* (20) e *phylogeny* (20). E nos resumos as palavras mais frequentes *data* (440), *biodiversity* (185), *species* (183), *analysis* (167), *monitoring* (66). As palavras mais frequentes nos títulos dos documentos são *data* (73), *biodiversity* (39), *analysis* (36), *environmental* (11) e *study* (11).

As 10 palavras-chaves dos autores de maiores frequências estão representadas de acordo o número de ocorrências acumuladas ao longo do tempo no Gráfico 4. No gráfico é possível observar ainda o crescimento das palavras *conservation*, *climate change* e *data sharing*, o que pode indicar o interesse pela utilização da análise de dados biológicos para a conservação das espécies, e para estudos dos efeitos ecológicos da mudança climática, sendo possível graças ao aumento do compartilhamento de dados. É possível observar uma queda na utilização

das palavras *Biodiversity* e *Data analysis*, mesmo elas estando em evidência graças ao método aplicado na busca dos artigos.

GRÁFICO 4 - DINÂMICA DAS PALAVRAS-CHAVES DO AUTOR AO LONGO DO TEMPO



Fonte: Autor, 2021

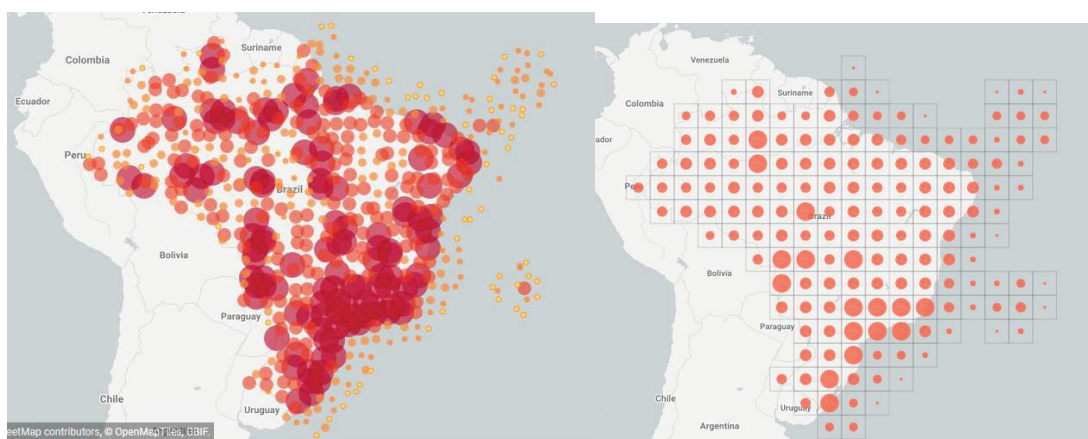
Na análise de co-ocorrência das palavras-chaves, percebe-se algumas relações: a) *biodiversity* – *data analysis* – *climate change*, *gis*, *toxicity*, *water quality estuary*; b) *data sharing* – *protected áreas*; c) *acoustic data analysis* – *acoustic sensing*. É possível observar no grupo “a”, uma relação das palavras biodiversidade e análise de dados com aspectos de qualidade ambiental, como mudança climática, toxicidade e qualidade de águas, o que indica uma forte relação da utilização dos dados sobre a biodiversidade na observação de variações no ambiente. A sigla GIS presente no grupo “a” significa Sistema de Informação Geográfica (Geographic Information System), dados muito presentes e utilizados na seleção de áreas de estudo ou nos registros de espécies.

4.2 ANÁLISE DE DADOS NA PLATAFORMA GBIF

No Portal do GBIF, no dia 16 de novembro de 2020, para a obtenção dos dados, utilizou-se a área de pesquisa com o campo de país ou área (*Country or area*), o Brasil e foram retornados 16.915.050 resultados. Em seguida recorreu-se ao campo de nome científico (*Scientific name*), e selecionado o reino Animalia. Nesta etapa foram recuperados 8.725.923 registros.

A própria plataforma GBIF fornece algumas análises estatísticas dos dados recuperados como, por exemplo, a distribuição nacional na Figura 7.

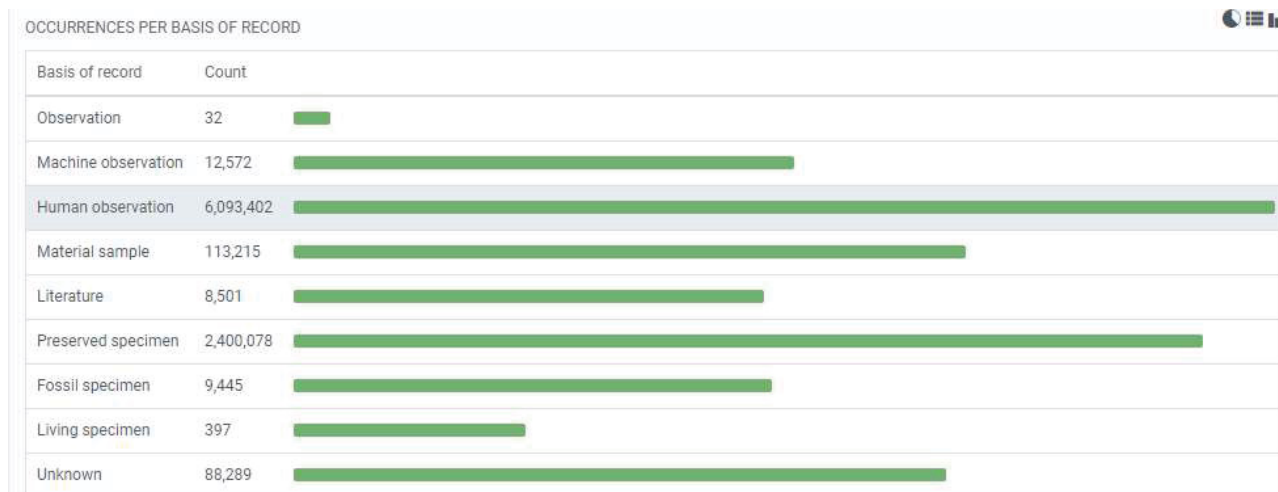
FIGURA 7 – DISTRIBUIÇÃO DOS REGISTROS DA FAUNA PRESENTES NO GBIF NO TERRITÓRIO BRASILEIRO



Fonte: Autor, 2021

Na Figura 7 é possível verificar a distribuição dos registros de ocorrência dos animais no território marítimo e terrestre brasileiro. Observa-se uma maior concentração de ocorrências na região sudeste e sul do país, além da maior quantidade de registros concentrados ao longo da costa, no rio Amazonas, em regiões pantaneiras e região central do Brasil. Chama atenção ainda a heterogeneidade dos registros no país, indicando que muitas áreas estão carentes de levantamentos de registros, advertindo possivelmente a necessidade de mais trabalhos de levantamentos nestas regiões, ou a base apresenta uma baixa adesão dos pesquisadores brasileiros.

GRÁFICO 5 - NÚMERO DE OCORRÊNCIAS POR TIPO DE REGISTRO PRESENTES NOS DADOS RECUPERADOS SOBRE A FAUNA BRASILEIRA NO GBIF



Fonte: Autor, 2021

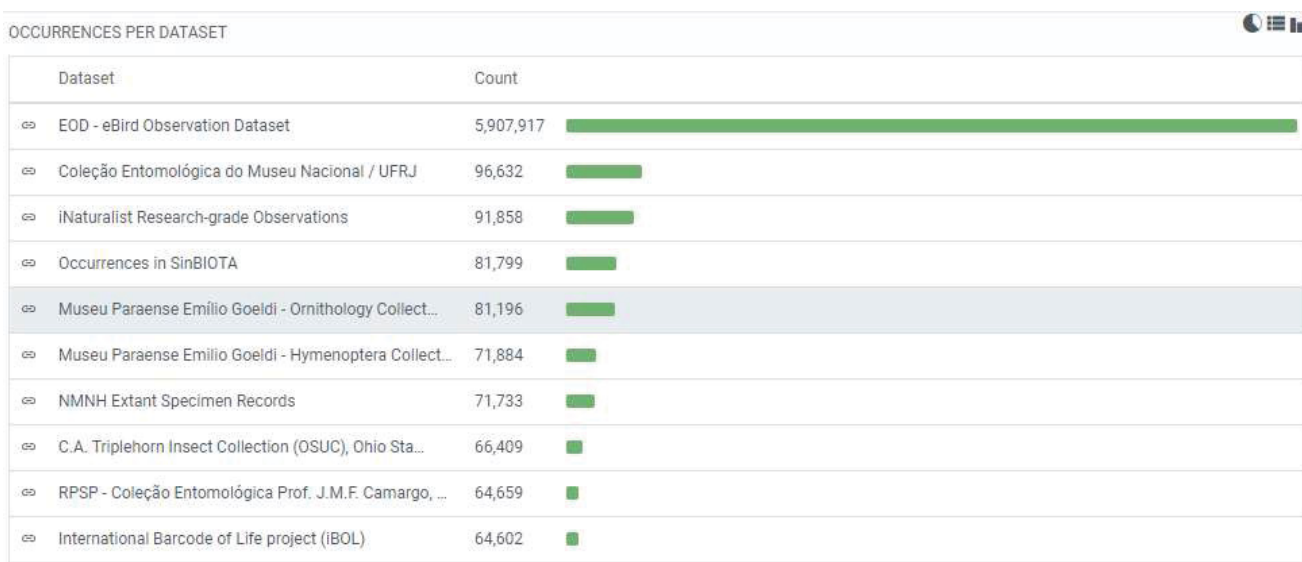
O Gráfico 5 sobre os tipos de registros de ocorrências: observação (*observation*), observação por máquina (*Machine Observation*), Observação humana (*Human observation*), Amostra de material (*Material Sample*), literatura (*Literature*), espécimes preservados (*Preserved Specimen*), espécimes fósseis (*Fossil Specimen*), espécimes vivos (*Living specimen*) e desconhecido (*Unknown*).

Nota-se que o maior número de registros se deu por observação humana (*Human observation*), este tipo de registro na base consiste de um animal o que foi apenas observado em ambiente natural, sendo o registro apenas por foto ou por “anotação” do pesquisador, em muitos casos este tipo de registro pode ser considerado sensível, pois a correta identificação do espécime pode estar comprometida, e dependendo do registro o mesmo não pode ser corrigido ou revisado, ao longo da dissertação podemos averiguar mais sobre esses dados.

O segundo maior número de registro é de espécimes preservados (*Preserved Specimen*), estes registros consistem dos registros dos animais que foram coletados e estão depositados e preservados em coleções biológicas. Eles apresentam uma maior confiabilidade, pois os materiais podem ser revisados e conferidos ao longo do tempo, por conta da relevância deste tipo de registro e devido ao número de instituições que apresentam coleções biológicas, esperava-se que o número destes registros fosse o maior, o que pode indicar uma ausência destas instituições na base, algo que poderá ser averiguado posteriormente.

Aqui observa-se o grande número dos registros desconhecidos (*Unknow*), este fator deve ser averiguado, para verificar se possuem alguma relação com outros fatores, como termos erroneamente preenchidos ou utilização de termos incomuns.

GRÁFICO 6 - DISTRIBUIÇÃO DOS REGISTROS DE OCORRÊNCIA DA FAUNA BRASILEIRA POR SERIA DE DADOS



Fonte: Autor, 2021

No Gráfico 6 são apresentados os 10 maiores *Datasets* presentes nos conjuntos de dados obtidos. Na pré-análise fornecida não foi possível visualizar quantos *Datasets*, no todo, foram recuperados, dado que será buscado em análises futuras.

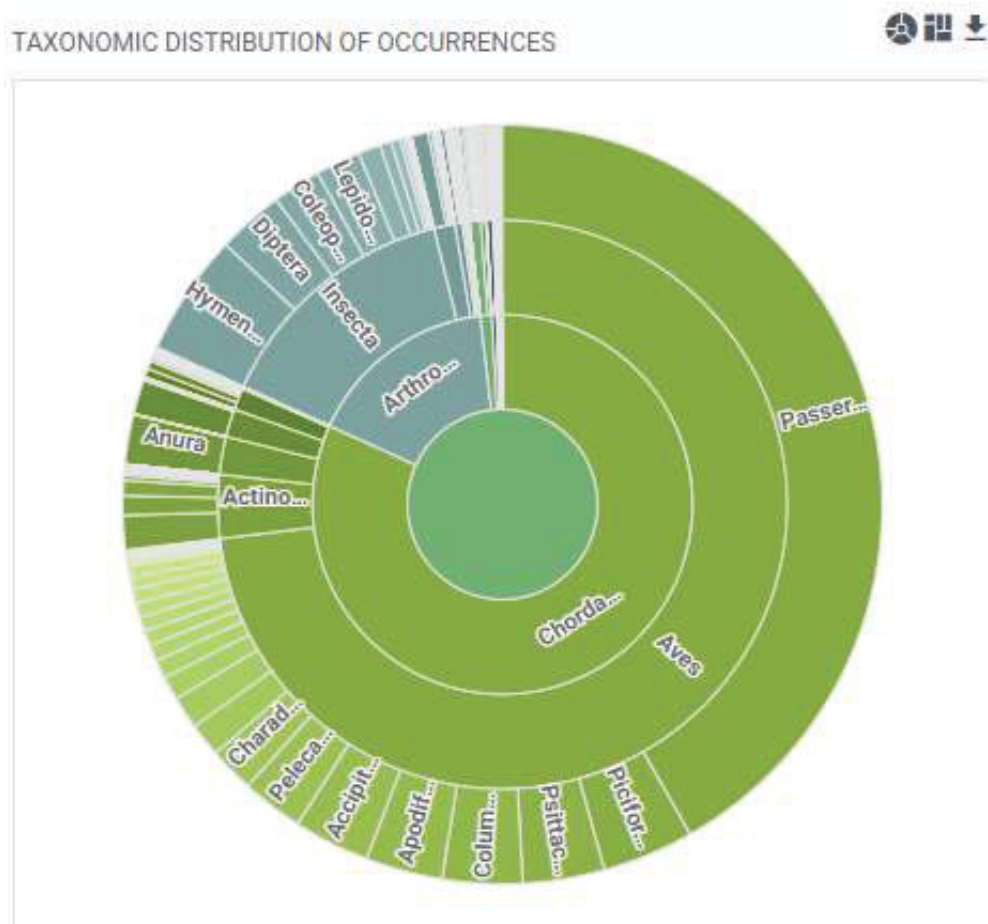
O que mais chama atenção é o *Dataset* apresentado em primeiro lugar, o *EOD* – *eBird Observation Dataset*, apresenta 5.907.917 milhões de registros, o qual equivale a 67,59% referentes aos registros da fauna brasileira na base do GBIF.

O *eBird* é uma iniciativa de cunho colaborativo se aproximando da ciência cidadão, no qual qualquer observador de aves, amador ou pesquisador, podem compartilhar seus conhecimentos e lista de aves observadas (eBird, 2020). O principal objetivo desta iniciativa é aumentar a quantidade de dados sobre a biodiversidade das aves, para isso recrutando e aumentando o engajamento em nível global, como resultados o *eBird* se tornou uma das maiores fontes de dados sobre a biodiversidade das aves, contribuindo muito para o conhecimento das dinâmicas populacionais, e tendo impacto direto na conservação das aves e de seus habitats. Os dados

fornechos ao *eBird* são disponibilizados livremente para o uso. Com o seu engajamento global, e o acesso facilitado por meio do aplicativo mobile, torna-se plausível que apresente o maior volume de dados em um *Dataset* individual.

Os outros *Datasets* apresentam volume de dados em comparação com o *eBird* muito pequeno, contudo é válido destacar a presença de instituições nacionais e internacionais de coleções biológicas. As coleções nacionais entre os 10 maiores *Datasets* são: Coleção Entomológica do Museu Nacional –UFRJ com 96.632, Museu Paranaense Emilio Goeldi – coleções de Ornitologia (81.196) e Hymenoptera (71.884) e RPSP - Coleção Entomológica Prof. J.M.F. Camargo com 64.659, e uma coleção internacional C.A. *Triplehorn Insect Collection* (OSUC), *Ohio State University* com 66.409.

GRÁFICO 7 - DISTRIBUIÇÃO DAS OCORRÊNCIAS DA FAUNA DE BRASILEIRA QUANTO A SUA CLASSIFICAÇÃO TAXONÔMICA, DISPONÍVEL NA GBIF



Fonte: Autor, 2021

O Gráfico 7 mostra a distribuição dos registros de ocorrência em relação aos grupos taxonômicos. Observa-se que a esmagadora maioria dos registros presentes na base são de aves, isso se deve ao fato de o maior *Dataset* presente nos dados recuperados serem pertencentes a *eBird*, como evidenciado anteriormente. Tal distribuição afasta-se e muito do que se esperava encontrar, se a base tivesse registros mais consonantes com a diversidade taxonômica de Metazoa, era esperado uma predominância do grupo Insecta (Insetos), uma vez que são os animais mais numerosos do planeta, tal fato pode ser evidenciado quando analisamos as informações apresentadas no Catálogo Taxonômico da Fauna do Brasil (acesso em 28 de fevereiro de 2021), consta que 119.575 espécies reconhecidas.

Deste, cerca de 94.000 são espécies de artrópodes e apenas cerca de 3.000 espécies de aves, e uma diferença de quase 32 vezes, de um grupo para o outro, além do que para se realizar estudos dessas espécies, tanto de morfologia quanto de área de ocorrência, diversos exemplares são registrados, portanto era de se esperar que o grupo dos insetos fosse os mais frequentes.

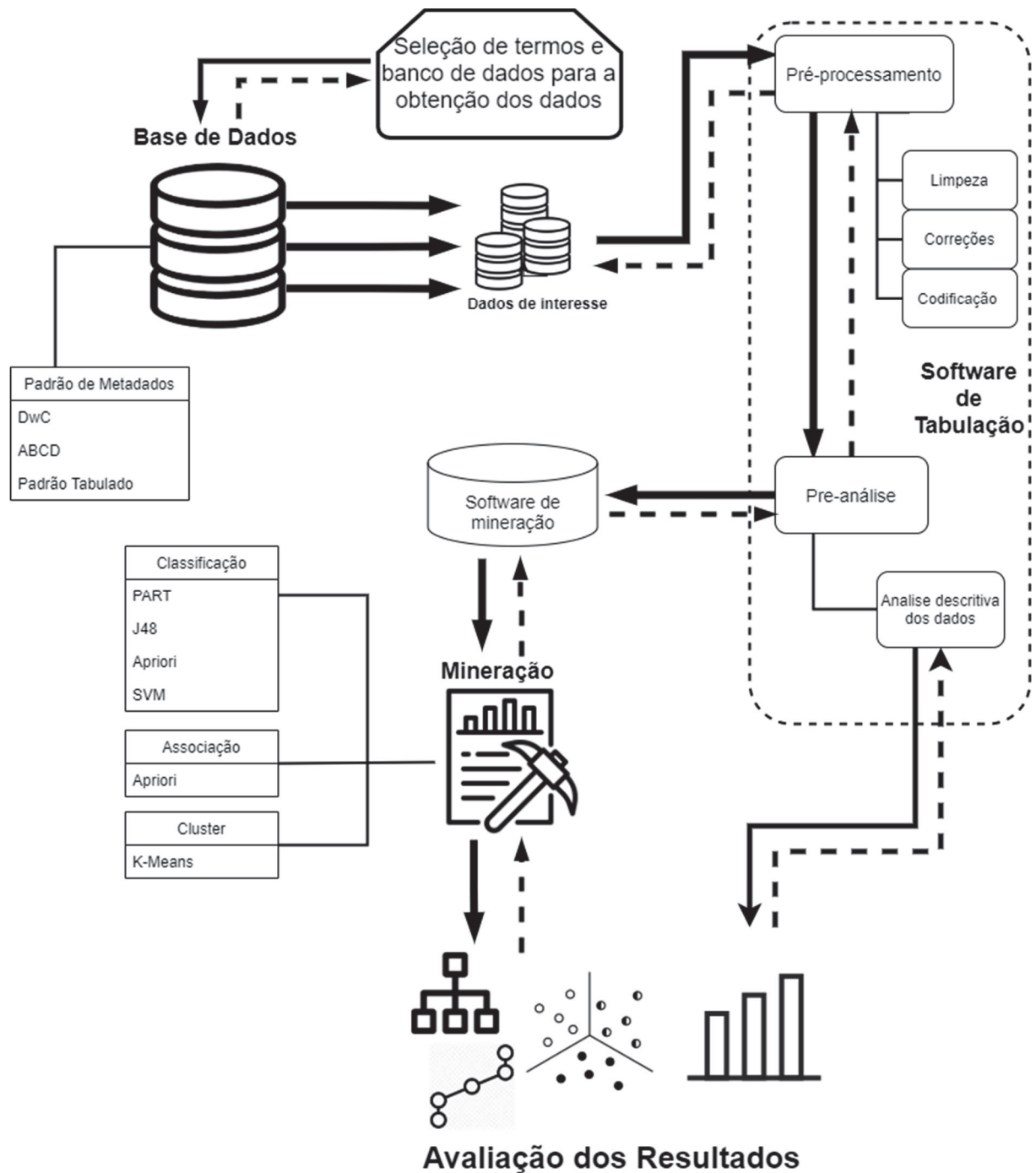
As 10 espécies com maior número de registros, são pássaros, evidenciando novamente a interferência na distribuição dos registros pela *eBird*, sendo necessário em análises futuras, a realização de dois tipos de análise: um com o conjunto de dados do *eBird* e ou sem.

A espécie com o maior número de registros é o *Pitangus sulphuratus* (Linnaeus, 1766), mais popularmente conhecido no território brasileiro como Bem-te-vi, apresenta mais de 101mil registros, tal volume de registros se deve a interligação com os hábitos e distribuição da espécie, assim como por ser muito registrado por observadores de aves, vinculados ao *eBird*.

4.3 MODELO DE ANÁLISE

O modelo definido neste trabalho busca a obtenção de padrões e descoberta de conhecimento em bases de biodiversidade, buscando a classificação e agrupamento das informações as quais podem indicar a necessidade ou não de estudos mais intensos em determinados grupos, ou a identificação de padrões inesperados.

FIGURA 8 – ESTRUTURA PRINCIPAL DO MODELO DE ANÁLISES EM BASES DE DADOS DE BIODIVERSIDADE



FONTE: O autor 2021

A etapa inicial compreende a escolha dos termos de busca, ou seja, as informações que se pretende obter (Grupo taxonômico específico, Área territorial, etc.) e a base de onde pretende-se extrair os dados para serem submetidos a análise.

A base em questão precisa apresentar os dados em um padrão de metadados (sendo os padrões Darwin Core (DwC) e Access to Biological Collections Data (ABCD) comumente encontrados em bases de biodiversidade) ou os dados precisam estar tabulados ou serem passíveis de tabulação, para assim poderem ser aplicadas as demais etapas.

Os dados de interesse então passam para a etapa de pré-processamento, onde são “limpos”. Esta limpeza consiste na remoção de dados ruidosos, campos vazios e termos ou atributos que não possam ser analisados.

A próxima etapa consiste na eventual correção de dados para a análise decorrentes de problemas de escrita, digitação ou desconfiguração dos dados.

Na sequência existe a codificação ou padronização dos termos, os quais apresentam muita variação, realizando o agrupamento dos dados em grupos, como por exemplo uma instância “Ano” onde apresenta os anos de maneira linear, sendo codificado, por exemplo, por décadas.

A pré-análise apresenta visão geral da base de dados. Nesta etapa é realizada a análise descritiva dos dados. Por meio da análise descritiva e da estatística básica, busca-se aprofundar nos dados, procurando seus significados, sua disposição, distribuição, seus desvios, dentre outros. É comum nesta etapa a identificação de dados “problemáticos”, que escaparam do pré-processamento, assim necessitando do retorno para o pré-processamento para tratamentos.

O pré-processamento e pré-análise ocorrem com o auxílio de um *software* de tabulação (exemplo: Excel), sendo indicado na Figura 8 por uma caixa pontilhada. Tal destaque se faz necessário pois, por ocorrerem no mesmo *software*, estas etapas apresentam uma interação dinâmica, muitas vezes ocorrendo em conjunto.

Os dados que passaram na etapa de pré-análise, são carregados em um *software* de mineração, onde serão submetidos aos algoritmos de Classificação (PART, J48, Apriori e SVM), Associação (Apriori) e Clusterização (K-Means), gerando ao final diversos resultados que serão avaliados.

A avaliação dos resultados se dá pela interpretação dos resultados estatísticos (taxa de acerto, taxa de erro, matriz de confusão e outros), em conjunto com as análises de um especialista na área de conhecimento dos dados, uma vez que apenas um especialista tem o arcabouço necessário para interpretar os resultados encontrados no processo de mineração e avaliar se os padrões encontrados possuem interesse e são úteis para a área de aplicação. A etapa de avaliação de resultados

incluída no modelo contempla tanto as avaliações estatísticas de desempenho do método quanto a relevância na área para conferir a validação dos resultados encontrados.

4.4 ANÁLISE DESCRITIVA

O conjunto de dados recuperados da plataforma do GBIF, em formato compactado *WinRar*, continha 8.725.923 registros até 16 de novembro de 2020. Após descompactação do arquivo, foram encontrados 5 arquivos de texto no formato *.txt*: *Citations*, multimídia, *Occurrence*, *Rights* e *verbatim*; 2 arquivos referentes ao formato de metadados em formato *.xml*: *meta* e *metadata*, e uma pasta de arquivo contendo um total de 2.451 arquivos em formato *.xml* referentes aos *Datasets*, no qual os nomes dos arquivos são referentes à identificação do pacote de dados ao qual estão vinculados, exemplo “00adaef8-991b-47aa-9c33-12e2d46a5258”.

QUADRO 7 – ARQUIVOS RECUPERADOS NA PLATAFORMA GBIF

Nome do arquivo	Formato	Tamanho
<i>Citations</i>	.txt	668 KB
Meta	.xml	36 KB
Metadata	.xml	278 KB
Multimídia	.txt	94.589 KB
<i>Occurrence</i>	.txt	8.890.130 KB
<i>Rights</i>	.txt	445 KB
verbatim	.txt	5.004.842 KB
“Pasta” Dataset	.xml	18.245 KB total

Fonte: Autor, 2021

Os arquivos no formato *.xml* foram abertos com no Microsoft Excel, sendo possível a análise dos conteúdos dos arquivos. O arquivo denominado “Meta”, apresenta dados referentes ao set de dados os quais foram obtidos na plataforma GBIF, contendo informações referentes ao GBIF como: URL da plataforma; serviço de download; contato do suporte; informações sobre os termos utilizados para a

seleção deste conjunto de dados; data da aquisição dos dados; idioma e quais outros dados estão inclusos nestes registros.

O arquivo “Metadata” apresenta 13 termos, nos quais são apresentados: codificação; terminação de linhas e campos; index; padrão de informações; tipos das informações; termos presentes no conjunto de dados (termos do núcleo base do DwC, termos dos sistemas do GBIF e termos das extensões), neste termo estão presentes 484 links URL para glossários dos termos presentes nos registros.

A Pasta de nome “*Dataset*” apresenta em seus arquivos dados sobre os pesquisadores ou instituições que forneceram os dados para o sistema do GBIF. Foram selecionados para essa avaliação 10 arquivos de maneira aleatória, os quais foram abertos no Excel a fim de descrever as suas informações. Nestes arquivos selecionados observou-se diferença no número de atributos fornecidas pelos depositantes, variando de 29 a 75. Os atributos em comum observadas são: *packgeldID*; *system*; *scope*; *ns3:lang*; *alternateldenfier*; *title*; *givenName*; *surName*; *surName2*; *organizationName*; *organizationName3*; *deliveryPoint*; *city*; *country*; *electronicMailAddress*; *onlineUrl*; *role*; *pubDate*; *language*; *para*; *url*; *citetitle*; *scope4*; *url5*; *function*; *givenName6*; *surName7*; *electronicMailAddress8*; *onlineUrl9*; *dateStamp*; *citation*.

O arquivo *Citations*, contém uma lista textual de referências dos dados coletados nas bases, com por exemplo “Levatich T, Ligocki S (2020). EOD - *eBird Observation Dataset*. Cornell Lab. of Ornithology. Occurrence Dataset <https://doi.org/10.15468/aomfnb> accessed via GBIF.org on 2020-11-16.”

O Arquivo identificado como “multimídia” contém 15 termos, onde apresentam as características referentes aos arquivos de mídia associados aos registros de ocorrência, apresentando: *gbifID*; *type*; *format*; *identifier*; *references*; *title*; *description*; *source*; *audience*; *created*; *creator*; *contributor*; *Publisher*; *license*; *RightsHolder*.

Os “*Rights*” apresentam uma lista dos *Dataset* e os direitos de acessos disponibilizados pelos depositantes. Por exemplo, o *Dataset*: “*Dataset: Neuroptera collection of National Museum of Nature and Science*” e os direitos referentes a este *Dataset*: “*Rights as supplied*: <http://creativecommons.org/licenses/by/4.0/legalcode>”.

Os arquivos “*Verbatim*” e “*Occurrence*”, não foram possíveis de serem abertos no Microsoft Excel, devido ao grande número de linhas e devido ao tamanho dos arquivos, os quais ultrapassavam a capacidade do computador. Contudo foi possível realizar a pré-visualização por meio da função “consulta dos dados” no Microsoft

Excel, e com isso foi possível averiguar os termos contidos em cada análise. Dos arquivos recuperados no GBIF, optou-se pela utilização apenas do arquivo “*Occurrence*”, uma vez que ele apresenta os dados relevantes para a pesquisa. O “Meta” apresenta as mesmas informações já trazidas na metodologia. Os arquivos “*Citations*”, “*Rights*”, “*Verbatim*” e os arquivos da pasta “*Dataset*”, apresentam os mesmos dados presentes no “*Occurrence*”. Os dados presentes no “*Metadata*”, serão utilizados apenas para consultar os termos presentes na análise exploratória dos dados.

Foi escolhido o conjunto de dados referente ao Museu Nacional do Rio de Janeiro, o qual apresenta 225.404 registros de ocorrência, distribuídos por todo o território do Brasil. Foram encontrados no arquivo 249 atributos.

Estes dados foram submetidos ao pré-processamento dos dados, na qual a primeira etapa foi a remoção das colunas / atributos as quais não foram preenchidas (portanto não continham dados). Em seguida foram removidas as colunas / atributos que não possuísem variação nos dados ou que não eram passíveis de serem utilizadas na análise como os “ID”. A terceira etapa consistiu na correção de dados desconfigurados, codificação de parâmetros e exclusão de dados incorretos.

Na primeira etapa foram excluídas 127 colunas / atributos. Nesta primeira etapa foi excluído aproximadamente 51% da base por dados faltantes. Dentre os dados excluídos nesta primeira etapa, é válido ressaltar os termos relacionados ao tempo geológico (apresentados no Apêndice 7), muito associados a registros fósseis, tornando-os importantes dados da catalogação dos espécimes, sendo sabido que no Museu Nacional se abriga uma coleção considerável de fósseis.

Na segunda etapa foram excluídos 84 atributos, sendo aproximadamente 33,7% da base original. Um dos principais critérios para a exclusão destes atributos foi por possuírem dados dispersos com valores que não poderiam ser codificados para serem utilizados nas análises. A maioria dos atributos excluídas apresentou valores únicos para cada registro, como no caso dos campos intensificados como “ID” (*institutionID*; *OccurrenceID*; *eventID*; *acceptedNameUsageID*) ou campos com codificações (*institutionCode*; *collectionCode*; *ownerInstitutionCode*; *catalogNumber*; *otherCatalogNumbers*; *fieldNumber*; *DatasetKey*; *taxonKey*; *acceptedTaxonKey*; *kingdomKey*; *phylumKey*; *classKey*; *orderKey*; *familyKey*; *genusKey*; *subgenusKey*; *speciesKey*). Ao analisar cada um dos termos, foram identificadas alguns atributos que não contão nas listas de termos disponibilizadas pelo TDWG (*level0Gid*;

level0Name; level1Gid; level1Name; level2Gid; level2Name; level3Gid; level3Name). Por conta disto, e de conterem informações iguais a outros termos e codificações que não se enquadrariam na análise, estes termos foram excluídos. Os termos que apresentaram informações repetidas também foram removidos, reduzindo o volume de dados. Os termos referentes à instituição que forneceu os dados foram removidos nesta etapa por conta de todos os dados pertencerem à mesma instituição.

QUADRO 8 – ETAPAS DE PRÉ-PROCESSAMENTO E OS TERMOS / ATRIBUTOS TRABALHADOS EM CADA ETAPA

Etapas	Termos / atributos excluídas, modificadas e analisadas
Primeira etapa	<i>Abstract; accrualMethod; accrualPeriodicity; accrualPolicy; alternative; audience; available; bibliographicCitation; conformsTo; contributor; coverage; created; creator; date; dateAccepted; dateCopyrighted; dateSubmitted; description; educationLevel; extent; format; hasFormat; hasPart; hasVersion; instructionalMethod; isFormatOf; isPartOf; isReferencedBy; isReplacedBy; isRequiredBy; isVersionOf; issued; mediator; médium; provenance; references; relation; replaces; requires; Rights; source; spatial; subject; tableOfContents; temporal; title; valid; collectionID; DatasetID; dataGeneralizations; recordNumber; organismQuantity; organismQuantityType; reproductiveCondition; behavior; establishmentMeans; associatedSequences; organismID; organismName; organismScope; associatedOrganisms; organismRemarks; materialSampleID; parentEventID; startDayOfYear; endDayOfYear; verbatimEventDate; samplingEffort; sampleSizeValue; sampleSizeUnit; locationID; higherGeographyID; minimumDistanceAboveSurfaceInMeters; maximumDistanceAboveSurfaceInMeters; locationAccordingTo; coordinatePrecision; pointRadiusSpatialFit; verbatimSRS; footprintWKT; footprintSRS; footprintSpatialFit; georeferenceVerificationStatus; georeferenceRemarks; geologicalContextID; earliestEonOrLowestEonothem;</i>

	<i>latestEonOrHighestEonothem; earliestEraOrLowestErathem;</i> <i>latestEraOrHighestErathem; earliestPeriodOrLowestSystem;</i> <i>latestPeriodOrHighestSystem; earliestEpochOrLowestSeries;</i> <i>latestEpochOrHighestSeries; earliestAgeOrLowestStage;</i> <i>latestAgeOrHighestStage; lowestBiostratigraphicZone;</i> <i>highestBiostratigraphicZone; lithostratigraphicTerms; group;</i> <i>formation; member; bed; identificationID;</i> <i>identificationReferences; identificationVerificationStatus;</i> <i>taxonID; scientificNameID; parentNameUsageID;</i> <i>originalNameUsageID; nameAccordingToID;</i> <i>namePublishedInID; taxonConceptID; acceptedNameUsage;</i> <i>parentNameUsage; nameAccordingTo; namePublishedInYear;</i> <i>higherClassification; subgenus; verbatimTaxonRank;</i> <i>vernacularName; nomenclaturalCode; nomenclaturalStatus;</i> <i>distanceAboveSurface; distanceAboveSurfaceAccuracy;</i> <i>typifiedName; relativeOrganismQuantity; recordedByID;</i> <i>identifiedByID;</i>
Segunda etapa	<i>Language; license; modified; Publisher; RightsHolder;</i> <i>institutionID; institutionCode; collectionCode; DatasetName;</i> <i>ownerInstitutionCode; basisOfRecord; informationWithheld;</i> <i>OccurrenceID; catalogNumber; recordedBy; lifeStage;</i> <i>OccurrenceStatus; otherCatalogNumbers; OccurrenceRemarks;</i> <i>associatedOccurrences; eventID; fieldNumber; eventDate;</i> <i>eventTime; habitat; fieldNotes; eventRemarks; continente;</i> <i>waterBody; islandGroup; countryCode; county; municipality;</i> <i>locality; verbatimLocality; locationRemarks; decimalLatitude;</i> <i>decimalLongitude; coordinateUncertaintyInMeters;</i> <i>verbatimCoordinateSystem; georeferencedBy;</i> <i>georeferencedDate; identificationQualifier; identifiedBy;</i> <i>dateIdentified; identificationRemarks; acceptedNameUsageID;</i> <i>namePublishedIn; kingdom; taxonRemarks; DatasetKey;</i> <i>publishingCountry; lastInterpreted; elevation; elevationAccuracy;</i> <i>depth; depthAccuracy; mediaType; taxonKey;</i>

	<i>acceptedTaxonKey; kingdomKey; phylumKey; classKey; orderKey; familyKey; genusKey; subgenusKey; speciesKey; species; genericName; acceptedScientificName; verbatimScientificName; protocol; lastParsed; lastCrawled; repatriated; level0Gid; level0Name; level1Gid; level1Name; level2Gid; level2Name; level3Gid; level3Name</i>
Terceira etapa	<i>Type; dynamicProperties; individualCount; sex; preparations; disposition; associatedReferences; associatedTaxa; previousIdentifications; year; month; day; samplingProtocol; island.; stateProvince; verbatimElevation; verbatimDepth; georeferenceProtocol; georeferenceSources; typeStatus; scientificName; originalNameUsage; phylum; class; order; Family; genus; specificEpithet; infraspecificEpithet; taxonRank; taxonomicStatus; issue; hasCoordinate; hasGeospatialIssues;</i>

Fonte: Autor, 2021

Na terceira etapa foi realizada a adequação dos dados para as análises, onde foram realizadas correções dos caracteres, codificação de termos e exclusão de linhas com dados incorretos. Foram identificados registros erroneamente marcados como pertencentes ao território brasileiro, contudo, ao analisar o termo “*stateProvince*” foi possível identificar estas localidades, às quais estão apresentados na tabela 4, as localidades foram verificadas por meio de buscas na internet e as linhas com registros em localidades incorretas foram excluídas, sendo retirados da base o total de 402 registros. Ainda neste termo, foi realizada a padronização de siglas para os estados (exemplo: Rio de Janeiro para RJ).

QUADRO 9 – LOCALIDADES ENCONTRADAS NO ARQUIVO, QUE FORAM REMOVIDOS NO PROCESSO DE LIMPEZA

Localidades erradas deletadas

Bogota; caiena; caqueta; bolivia; caracas; cartago; cauca; chamchamayo; Chimborazo; Chiriquí; Coroico; Cuernavaca; cundinamarca; Cusco; DelSara; Drophae; ElPorvenir; ElTopo; Guantanamo; Guerrero; Huancabamba;Huila; Iadiple; Ichilo; Inambari; Iquitos; Jalaba; Jimenez; Juanqui; JP;LaMerced; LaPaz; LaRioja; loc?; macas; Meta; Misiones; Mocoa; Morelo; Moyobamba; Muzo; Orizaba; Oxapampa; Panama; Pebas; Pozuzo; Pucallpa; Putumayo; Quito; Riobamba; SanLorenzo; SanPedroSula; SanRamÃ³n; SantaCruzdeLaSierra; SantigodeCuba; Senegal; Tabasco; Tacuba; Tacuta; Tampico; TarapacÃ¡; Tarapoto; TingoMaria; Umbria; VillavicÃªncio; Yurimagua; Zamora; ChocÃ³

Fonte: Autor, 2021

No termo “*scientificName*”, foram identificadas 588 espécies com o total de 12.273 registros, nos quais ao invés de conter os nomes completo das espécies (gênero e epíteto específico), encontrava-se o símbolo de interrogação (“?”) no lugar dos gêneros. Para a correção, foram utilizadas as informações contidas na instância “genus”. Para os registros que não foram encontrados nesta instância, foram utilizados o epíteto, o nome do descritor e o ano nas buscas na internet, a fim de completar a os dados da espécie. Observou-se que os termos que necessitaram destas buscas eram, em sua maioria, registros de espécies extintas ou fósseis, o que torna a situação destes registros preocupantes, uma vez que muitos dos termos utilizados para identificar características importantes de fósseis, como por exemplo, o registro geológico, foram excluídos na primeira etapa por não conterem nenhum dado.

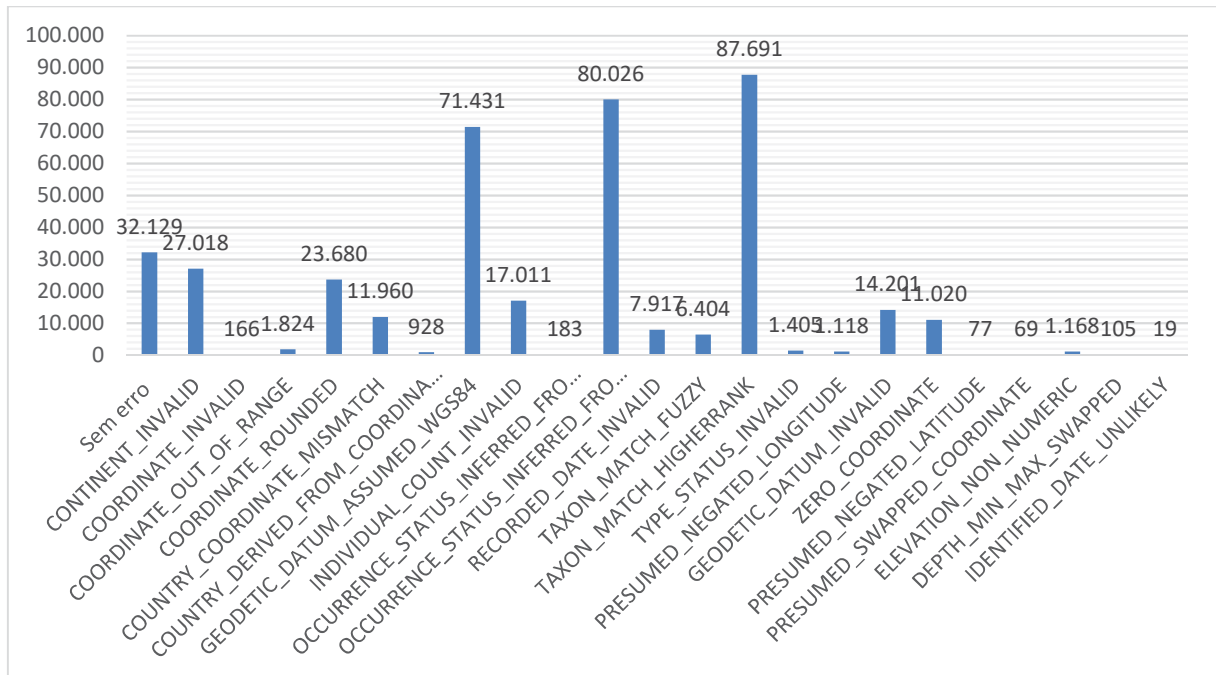
Foi realizada a codificação das instâncias “year” (transformada em décadas), “Elevation” (intervalos de altitude) e “Depth” (convertido a intervalos de profundidade), a fim de padronizar e reduzir as informações contidas, possibilitando as análises pelos algoritmos.

As demais instâncias passaram por um processo de remoção ou substituição de caracteres especiais (substituição das vírgulas “,” por sublinhado “_” ou remoção dos pontos “.”), correção de dados desconfigurados (exemplo: “*Madaglymbus alutaceus* (RÃ©gimbart. 1900)” para “*Madaglymbus alutaceus Regimbart 1900*”) e

padronização dos dados que haviam sido escritos de diferentes formas, como por exemplo: álcool 70%, Alcool70% e Alcool 70%.

O atributo “*issue*” apresenta informações sobre os erros encontrados no registro durante a etapa de processamento e interpretação do sistema do GBIF. Observou-se que um mesmo registro pode apresentar mais de um erro identificado pelo sistema, sendo que 32.129 registros não apresentam nenhum erro. Para a interpretação das informações contidas neste elemento, foi consultado o *GITHUB* do GBIF. Dentre os erros, os cinco mais presentes são: “*TAXON_MATCH_HIGHERRANK*” (87.691) o qual apresenta os registros onde a ligação do conjunto taxonômico somente é possível com níveis mais altos de classificação, impossibilitando de se fazer com o nome científico; “*OCCURRENCE_STATUS_INFERRED_FROM_INDIVIDUAL_COUNT*” (80.026), no qual o status da ocorrência foi inferido a partir do valor de contagem individual; “*GEODETIC_DATUM_ASSUMED_WGS84*” (71.431) já que os dados não foram indicados ou interpretados pelo sistema, assume-se que as coordenadas interpretadas estão baseadas no WGS84; “*CONTINENT_INVALID*” (27.018) os valores referentes ao continente não são interpretáveis; “*COORDINATE_ROUNDED*” (23.680) coordenada original modificada por arredondamento para 5 decimais.

GRÁFICO 8 – DISTRIBUIÇÃO DAS INSTÂNCIAS PRESENTES NO ATRIBUTO “ISSUE”.



Fonte: Autor, 2021

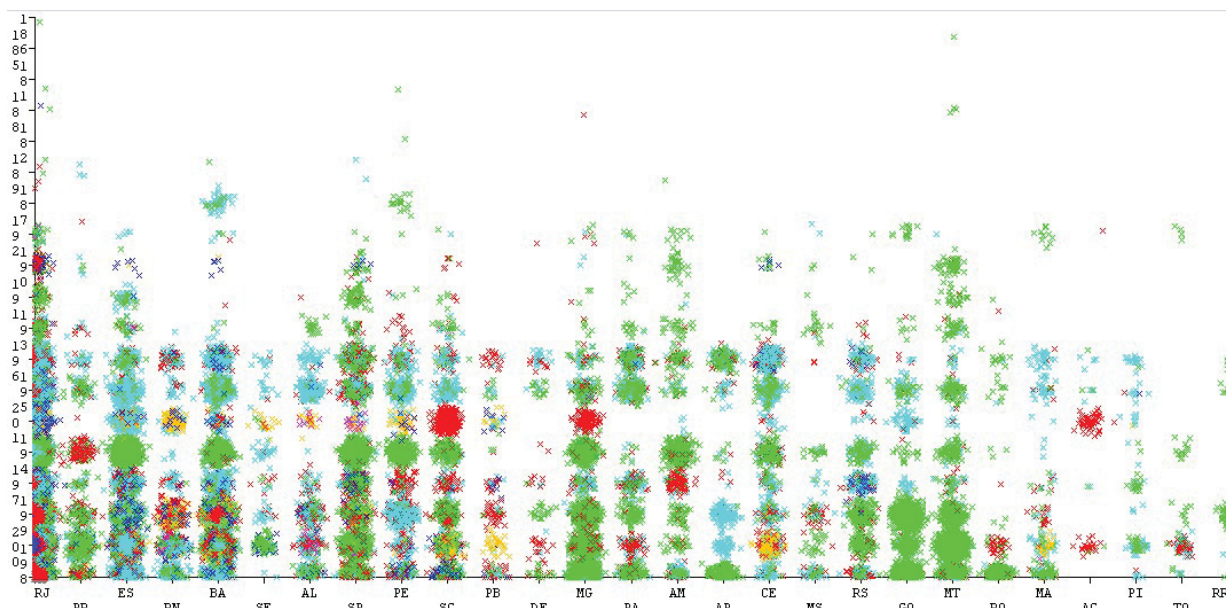
4.5 VISUALIZAÇÃO DOS DADOS WEKA

O *software* WEKA apresenta uma ferramenta de visualização dos dados, onde são gerados gráficos que correlacionam os atributos. Nos dados analisados, foram gerados o total de 1.024 gráficos, sendo que o *software* utiliza um sistema de cores para representar diferentes atributos em cada gráfico, possibilitando a visualização geral dos dados e a pré-análise de suas interações. Foram analisados os gráficos correspondentes às classes “ano codificado”, “*stateProvince*” e à classe “*phylum*”, como parâmetro de visualização. Não foram realizadas análises nos níveis taxonômicos mais específicos (família, gênero, espécie, epíteto), por conta do alto número de elementos que dificultam a visualização. Esta limitação também foi observada nas outras análises.

O Gráfico 9 apresenta a relação entre “ano codificado” e “*stateProvince*”, onde é possível observar a evolução e a transição dos grupos de animais registrados em cada estado brasileiro. Os estados AC, PI, TO e RR apresentam registros concentrados em poucos anos, AC tem concentrado os dados nos anos 2001-2010 e 2011-2017, PI mais concentrados em 1971-1980 e 2001-2010, TO mais concentrados em 2001-2010 e RR mais concentrados nos anos 1991-2000. As maiores concentrações de registros encontrados são referentes aos estados do Rio de Janeiro (RJ), São Paulo (SP), Espírito Santo (ES) e Bahia (BA), sendo observadas maiores

concentrações destes registos nos anos 1901-1910 a 2011-2017. Por meio das cores é possível observar quais Filos estão mais representados em cada estado. Representado pela cor verde está o filo Chordata, sendo o filo mais representativo nos estados SP, MG, GO, MT, RO, PI e RR. Representado pela cor vermelha temos o filo Arthropoda. Os seus dados apresentam-se distribuídos ao longo de todo o gráfico, podendo observar conjuntos de dados deste filo nos estados RJ, PR, BA, SC, MG, AM, RO, e AC. Na cor verde-água está representado o filo Molusca. Observa-se nos anos de 1951-1960 a 1961- 1970 a presença de registro deste filo ao longo de quase todos os estados. Também é possível observar concentração de registos nos estados RJ, ES, BA, Al, PE, AP e CE. Em azul escuro o Filo Bryozoa apresenta registos esparsos pelos estados, com concentrações significativas nos estados RJ, ES, RN, BA, SC e RS. Os demais filis não foram possíveis de serem visualizados. O estado que aparentemente apresenta o maior número de filis é o RJ, o que já era esperado, dado que o museu que disponibilizou estes dados reside neste estado.

GRÁFICO 9 - DISTRIBUIÇÃO DOS DADOS DO “ANO CODIFICADO” EM RELAÇÃO A CLASSE “STATEPROVINCE”, COM VISUALIZAÇÃO DAS VARIVEIS DA CLASSE “*PHYLUM*” SENDO AS VARIÁVEIS VISUALIZÁVEIS: CHORDATA VERDE, ARTHROPODA VERMELHA, MOLUSCA VERDE-ÁGUA E BRYOZOA AZUL ESCURO.

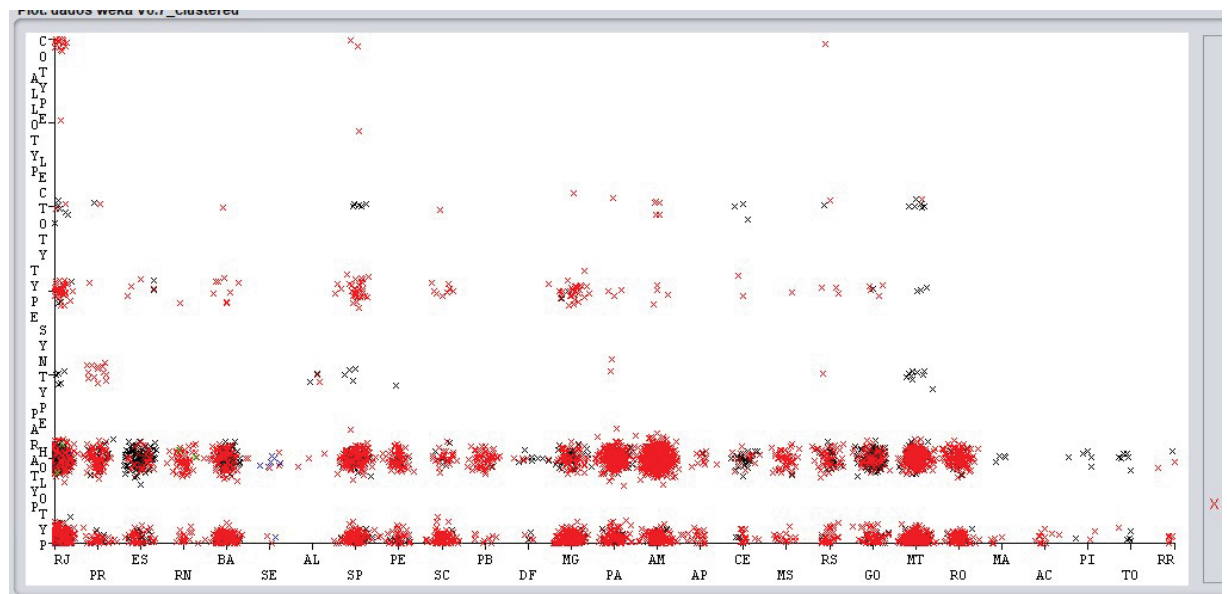


Fonte: Autor, 2021

O Gráfico 10 apresenta a distribuição em cada estado das chamadas “espécimes tipo”, utilizadas para a descrição das espécies em estudos taxonômicos. Essa correlação permite a identificação de possíveis grupos focais de estudo realizados pela instituição, a origem dos materiais coletados e as regiões de maior interesse e esforço amostral. No gráfico estão representados por cores os filos Artropoda em vermelho, Chordata em preto e Bryozoa em Azul. Os Artrópodes são os que aparecem em maior quantidade, destacando as concentrações nos estados RJ, BA, MG, PA, AM, MT e RO. Os Chordatos apresentam-se em maior número como Paratypus (espécimes que também foram utilizados na descrição da espécie), do que como Holotypus (espécime escolhido pelo descritor como a que melhor representa a espécie entre os espécimes utilizados na descrição da espécie). Este filo está mais presente nos estados ES, BA, CE e GO. O grupo Bryozoa apresenta-se apenas no estado do SE, sendo também o com maior número de ocorrências neste estado. No entanto, pode ocorrer que os filos Arthropoda e Chordata, por possuírem mais registros, estejam mascarando no gráfico a presença do filo Bryozoa nos demais estados.

GRÁFICO 10 – RELAÇÃO DAS CLASSES “*typeStatus*” E “*stateProvince*”, COM VISUALIZAÇÃO DAS VARIÁVEIS DA CLASSE “*PHYLUM*” SENDO AS

VARIÁVEIS VISUALIZÁVEIS: CHORDATA PRETO, ARTHROPODA VERMELHA E BRYOZOA AZUL



Fonte: Autor, 2021

4.6 RESULTADOS PART

Foram realizadas as análises em três atributos metas, sendo “ano codificado”, “*phylum*”, “*stateProvince*”. Os valores de precisão da classificação, erro absoluto e número de regras encontrados estão apresentados na tabela 5.

TABELA 5 – RESULTADOS DAS ANÁLISES DO ALGORITMO PART

Atributo meta	Ano Codificado	<i>Phylum</i>	<i>stateProvince</i>
Classificações corretas	43,7074%	99,3586%	34,9979%
Classificações incorretas	56,2926%	0,6414%	65,0021%
Kappa statistic	0,2734	0,9885	0,1273
Mean absolute error	0,0734	0,0021	0,0598
Root mean squared error	0,1909	0,0276	0,1728
Relative absolute error	85,3911%	2,4566%	93,3928%
Relative absolute squared error	92,0739%	13,3218%	96,5842%
Total Number of Instances	73.555	224.969	211.964
Ignored Class Unknown instances	151.446	32	13.037

Nº de regras	58	33	140
--------------	----	----	-----

Fonte: Autor, 2021

O atributo “Ano Codificado” apresentou uma taxa baixa de acerto de 43,71%, ao analisar a matriz de confusão (apêndice 15), é possível observar que a maior parte dos erros ocorreram nos anos 1981-90 (a), 1991-00 (c) e 2001-10 (b), nos anos 1931-40, 1911-20, 1901-10, 1921-30, 1871-80, 1891-1900, 1821-30, 1881-90, 1811-20, 1851-60 e 1861-70. Não ocorreu nenhuma marcação correta, isso ocorre devido a maior parte dos registros estarem concentrados nos anos 1981-90, 1991-00 e 2001-10, criando-se uma tendência de classificar os registros entre esses anos.

Foram encontradas 58 regras (Apêndice 15), nas quais é possível observar a relação de alguns grupos taxonômicos com anos específicos, o que pode indicar que nestes determinados anos estes grupos tiveram de alguma forma um enfoque especial na instituição, como os indicados pelas regras 1 a 7, relacionando registros no nível taxonômico de Gênero com anos específicos. As regras que exibem níveis taxonômicos mais altos, apresentam em sua maioria outra variável da regra, destaca-se o “*hasGeospatialIssues*”, que indica se os registros têm algum erro relacionado aos dados de georeferência, apontando assim a necessidade de correção ou revisão dos registros daqueles anos específicos, como o apresentado nas regras 11, 12, 17 e 30, o que indica que possuem erros nos dados de georeferenciamento. As regras 14, 19, 20, 21, 24, 31, 33, 35, 36, 38, 39, 40 e 41, mostram uma relação de determinados grupos taxonômicos com estados em determinados períodos, indicando um maior interesse nesses grupos em um período específico de tempo. As regras 49, 51 e 52 correspondem ao mesmo período de 1951-60, onde são relacionados determinados grupos taxonômicos com a dinâmica de doação do material.

O atributo *Phylum* apresentou alta taxa de acertos (99,36%). Ao analisar a matriz de confusão (Apêndice 16), observou-se que os erros se concentraram em dois atributos, sendo o Arthropoda (b) e o Chordata (c), os quais são as variáveis com maior número de registros. As instâncias Tardigrada, Platyhelminthes, Annelida, Sipuncula, Nemertea, Nematomorpha, Echinodermata, Brachiopoda, Nematoda, não tiveram nenhuma classificação correta.

Foram encontradas 33 regras (Apêndice 16), em sua maioria as regras encontradas relacionam os filos com os táxons mais específicos pertencentes àquele filo. Nas regras apresentou-se uma relação entre alguns táxons com o nível

taxonômico do registro, uma vez que nem todos os espécimes são identificados a níveis mais específicos (gênero, espécie ou subespécie). Estas regras apontam a necessidade de apuração destes dados a fim de torná-los mais específicos, ou indicam grupos que necessitam de estudos. Estas inferências aqui apresentadas são mais aplicáveis quando grupos menores são identificados. Nestes resultados, apresenta-se uma visão macro dos grupos taxonômicos, com o intuito de direcionar para grupos mais específicos, podendo ser aplicada a mesma proposta deste trabalho para grupos taxonômicos específicos, como a classe Insecta.

O atributo “*stateProvince*” apresentou baixa taxa de acerto (35%). Ao analisar a matriz de confusão (apêndice 17), os erros se concentraram nos estados RJ, ES, BA, MG, PA, AM, GO e MT. as instâncias PR, SE, AL, SP, PE, SC, PB, DF, MG, PA, AM, AP, CE, MS, RS, GO, MT, RO, MA, AC, PI, TO, RR não tiveram nenhuma classificação correta.

Foram encontradas 140 regras (apêndice 17), as quais relacionam grupos taxonômicos de diversos níveis hierárquicos com estados específicos, contudo, devido à baixa taxa de acerto, as regras encontradas não condizem com a realidade, como exemplo a regra 23, que vincula uma espécie de peixe muito comum no território brasileiro apenas ao estado do RJ. Tal resultado seria interessante se identificasse uma espécie endêmica (espécie que ocorre exclusivamente em uma determinada região geográfica). Das regras encontradas, 101 estão relacionadas ao RJ. Isso se deve ao excesso de classificações neste estado.

4.7 RESULTADOS APRIORI

Foi aplicado o algoritmo Apriori como associação e como classificador. Como classificador, foi aplicado nos atributos meta ano codificado, *stateProvince* e *phylum*. Para cada atributo, foram realizadas análises com taxa mínima de confiabilidade 50% e número máximo de regras 10.000 (dez mil). Este valor de regras foi escolhido com o intuito de se obter o máximo de regras de classificação possível.

Como Associação, foram encontradas 647 regras, as quais estão apresentadas no Apêndice 13. Destas regras, foram separadas 17 mais relevantes para a análise.

As regras 12 e 106 apresentam a relação entre a variável Álcool 70 do atributo “*preparation*”, com a variável Mollusca do atributo “*phylum*”, o que indica que este grupo está conservado em via líquida na coleção. No caso específico do filo Mollusca,

demonstra que a maioria dos exemplares presentes nesta coleção possuem preservados os tecidos moles destes animais e/ou que a maior parte dos exemplares preservados nesta coleção são de animais sem concha. Segundo Brusca e Brusca (2013), são conhecidas aproximadamente 93.000 espécies viventes e 70.000 fósseis, os quais contemplam os animais conhecidos como caracóis, caramujos, mexilhões, ostras (que possuem concha), lesmas, lulas e polvos (que não possuem conchas). Sendo necessário análises mais profundas neste filo para verificar os reais motivos destas regras.

As regras 163, 164, 168, 288, 326 e 351, apresentam a relação entre o *taxonRank=FAMILY* com o *phylum=Arthropoda*, o que indica que este filo apresenta muitos registros com os espécimes identificados apenas no nível taxonômico de família, indicando um grupo de interesse para trabalhos de revisão taxonômica. Contudo, este filo deve passar novamente por análises a fim de se identificar quais grupos dentro dele são os que apresentam esta característica. É válido ressaltar que os resultados apresentados são exclusivamente sobre a coleção do Museu Nacional, outras instituições podem apresentar resultados diferentes.

Nas regras 258, 363 e 406, são regras relacionadas os *phylum=Chordata* com *taxonRank=SPECIES*, o que demonstra que os espécimes do filo Chordata presentes na coleção estão identificados ao nível taxonômico de espécie. Diferente do apresentado nas regras do filo Arthropoda, os registros do filo Chordata podem não representar um grupo de interesse para revisões taxonômicas, contudo, análises mais específicas são necessárias.

As regras anteriormente apresentadas com relação aos filos Arthropoda e Chordata, apresentam o que é observado na rotina dos que trabalham com a fauna. Os grupos taxonômicos pertencentes ao filo Arthropoda, devido à sua elevada diversidade, apresenta uma carência de especialistas, tornando comum a identificação dos espécimes de determinados grupos apenas ao nível taxonômico de família, dado a ausência de chaves de identificação dicotômicas para auxiliar na identificação para níveis mais específicos. Por outro lado, o filo Chordata apresenta seus grupos mais bem estudados e mais conhecidos. Uma possível causa é a atração dos pesquisadores por grupos mais “carismáticos”, como por exemplo os felinos (onça-pintada, gato-do-mato, etc.) e quelônios marinhos (tartarugas marinhas), onde por interesses pessoais, mais profissionais se dedicam a estudarem estes grupos.

O único estado que esteve presente nas regras foi o Rio de Janeiro (*stateProvince*=RJ), mostrando a influência do local onde a instituição se encontra. As regras 382 e 493 relacionam que os registros do RJ possuem suas identificações aceitas (*taxonomicStatus*=*ACCEPTED*), então estes registros estão intensificados com nomes taxonomicamente corretos, o que aumenta a credibilidade dos mesmos. As regras 391 e 493 indicam que os registros do RJ não possuem coordenadas (*hasCoordinate*=*false*), o que indica que estes registros não podem ser utilizados para se averiguar a dispersão das espécies ao longo do estado.

Como classificador para os atributos “ano codificado” e “*stateProvince*” não foram encontradas regras de classificação com o algoritmo Apriori. Para o atributo “*phylum*”, foram encontradas 32 regras (apêndice 14)., as quais são semelhantes as regras encontradas como associação.

4.8 RESULTADOS SVM

Foram aplicados dois conjuntos de algoritmos do SVM disponíveis no *software* WEKA. Os algoritmos SMO e o LibSVM (externo) não obtiveram resultados. Isso se dá pelo fato de os dados utilizados nesta pesquisa não apresentarem delimitações claras para que o algoritmo trace vetores entre os atributos analisados. Este resultado pode ter sido causado devido à qualidade dos dados que, mesmo após o processo de limpeza, ainda não apresentam elementos claros para a utilização deste algoritmo. Contudo, o mesmo algoritmo deve ser utilizado em outras bases, para assim averiguar se o problema foi causado exclusivamente nesta base ou se o algoritmo não é apropriado para este tipo de análise.

4.9 RESULTADOS J48

Foram realizadas as análises em três atributos meta, sendo “ano codificado”, “*phylum*”, “*stateProvince*”. Os valores referentes a número de floras, tamanho da árvore, classificação correta e incorreta, são apresentadas no Quadro 7.

QUADRO 10 - RESULTADOS DAS ANÁLISES DO ALGORITMO J48

	“ano codificado”	“ <i>phylum</i> ”	“ <i>stateProvince</i> ”
<i>Number of Leaves</i>	13.960	322	13.961
<i>Size of the tree</i>	13.961	332	13.963
<i>Correctly Classified</i>	54.2737 %	99.6715%	46.6924 %
<i>Incorrectly Classified</i>	45.7263 %	0.3285 %	53.3076 %

Fonte: Autor, 2021

O atributo “Ano Codificado” apresentou uma taxa de acerto de 54,27%, ao analisar a matriz de confusão (Apêndice 18), é possível observar que a maior parte dos erros ocorreram nos anos 1941-50 (e), 1951-60 (g), 1971-80 (d), 1961-70 (h), 1981-90 (a), 1991-00 (c), 2001-10 (b) e 2011-17 (f), nos anos, 1871-80, 1891-1900, 1821-30, 1881-90, 1811-20, 1851-60 e 1861-70, não ocorreu nenhuma marcação correta, isso acontece devido a maior parte dos registros estarem concentrados nos anos 1981-90, 1991-00 e 2001-10, criando-se uma tendência de classificar os registros entre esses anos.

O atributo *Phylum* apresentou alta taxa de acertos (99,67%). Ao analisar a matriz de confusão (Apêndice 18), observou-se que os erros se concentraram em dois atributos, sendo o Arthropoda (b) e o Chordata (c), os quais são as variáveis com maior número de registros. As instâncias Tardigrada, Platyhelminthes e Nemertea, não tiveram nenhuma classificação correta.

O atributo “*stateProvince*” apresentou taxa de acerto de 46.69%. Ao analisar a matriz de confusão (Apêndice 18), os erros se concentraram nos estados RJ, ES, BA, MG, PA, AM, GO e MT. Todas as instâncias apresentam alguma classificação correta.

As árvores geradas nos três atributos meta, apresentaram número de folhas (13.960) e o tamanho em largura (13.961) em valores muito elevados, o que inviabilizou a representação dos resultados de forma gráfica para visualização e apresentação neste documento.

4.10 RESULTADOS K-MEANS

Foi realizado análise utilizando a $k = 13$, buscando gerar número de clusters igual a instância *phylum*, os clusters encontrados são apresentados na Tabela 6. Os

cluster 0,4 e 10 são os que apresentam o maior número de registros, juntos englobam 47% de todos os registros.

TABELA 6 – INFORMAÇÕES GERAIS DOS CLUSTERS GERADOS

Cluster	Nº de registro	Porcentagem
0	54.758	24%
1	15.959	-7%
2	20.667	-9%
3	6.534	-3%
4	29.489	-13%
5	6.683	-3%
6	16.469	-7%
7	7.025	-3%
8	7.349	-3%
9	7.543	-3%
10	21.832	-10%
11	21.291	-9%
12	9.402	-4%

Fonte: Autor, 2021

Ao analisar os centroides dos grupos observa-se que as classes *Type*, *dynamicProperties*, *sex*, *lifeStage*, *Preparations*, *disposition*, *associatedReferences*, *associatedTaxa*, *previousIdentifications*, *samplingProtocol*, *island.*, *Elevation*, *Depth*, *georeferenceProtocol*, *georeferenceSources*, *typeStatus*, *originalNameUsage*, *specificEpithet*, *infraspecificEpithet*, *taxonomicStatus*, apresentam a mesmas variáveis. Nas classes phylum, a qual foi o motivo de escolha do valor de k, apresentou apenas como variáveis nos centróides apenas três filos Arthropoda (7), Chordata (4) e Mollusca (1).

Foi aplicada a avaliação dos grupos por meio das classes “phylum” e “ano codificado”. Para a classe “phylum” as instâncias incorretamente cluterizadas foram de 66.161% dos registros. Para os cluters 1, 7, 8, 9, 10 e 12, não foram identificados na avaliação atribuídos a classe phylum, para os demais clusters foram atribuídos os seguinte atributos: Cluster 0 - Arthropoda; Cluster 2 - Annelida; Cluster 3 -

Echinodermata; Cluster 4- Chordata; Cluster 5- Nematomorpha; Cluster 6 - Bryozoa; Cluster 11- Mollusca.

Para a classe “ano codificado”, as instâncias incorretamente cluterizadas foram de 24, 6381%, apenas no cluster 5 não foi identificado nenhum atribuído, para os demais clusters foram atribuídos os seguintes atributos: Cluster 0 - 1901-10; Cluster 1 - 1981-90; Cluster 2 - 1931-40; Cluster 3 - 1961-70; Cluster 4 - 1991-2000; Cluster 6 - 2011-17; Cluster 7 - 1951-60; Cluster 8 - 1921-30; Cluster 9 - 1941-50; Cluster 10 - 1911-20; Cluster 11 - 2001-10; Cluster 12 - 1971-80.

5 CONSIDERAÇÕES FINAIS E RECOMENDAÇÕES

Este trabalho apresentou uma proposta de modelo de análise de dados para dados de biodiversidade, utilizando métodos de análise descritiva e técnicas de mineração de dados para interpretar, classificar e identificar interações entre os dados analisados. Utilizou-se como fonte de estudos os dados da fauna brasileira presentes na plataforma do GBIF.

Ao longo desta pesquisa, a plataforma do GBIF se mostrou relevante devido ao volume e à estruturação dos dados que registram a biodiversidade brasileira. Contudo, ainda há muito o que ser inserido para que a plataforma nos forneça uma visão mais completa da biodiversidade presente no país, dado a grandiosidade dela. O desbalanço nos registros presentes, deve-se principalmente aos grupos animais, os quais geram maior interesse por parte dos pesquisadores ou da população, como evidenciado nas análises com relação ao *eBird*, que se trata de um projeto de ciência cidadã (capítulo 4.2, p. 67).

Observou-se que há poucos registros relacionados a instituições de pesquisa, o que pode indicar uma não adesão dessas instituições, tornando necessárias averiguações futuras. Apesar destes problemas, a plataforma ainda é um terreno fértil para análises da biodiversidade.

Quanto aos objetivos específicos desta pesquisa, considera-se que o primeiro, que tratou de ‘buscar repositórios de biodiversidade e estudar seus formatos de metadados’, foi atendido. Estes resultados são apresentados no Capítulo 2, seção 2.4, onde foram descritos os principais formatos de metadados presentes nas bases de biodiversidade, e no Capítulo 3, subseção 3.2.5, onde foi apresentado o repositório GBIF, de onde foram obtidos os dados utilizados nesta pesquisa. Decorrente deste objetivo específico, também foi produzido um artigo publicado referente às bases de dados biológicas.

O segundo objetivo específico, que foi ‘identificar os principais métodos e ferramentas de análise de dados, passíveis de serem utilizados nas bases de dados de biodiversidade’ foi atendido, conforme apresentado nos capítulos 2, seção 2.5 e Capítulo 4, seção 4.1, em relação aos métodos de análise, e no Capítulo 3, subseções 3.2.2 e 3.2.4 referentes às ferramentas. A escolha destas se deu pela leitura dos métodos e ferramentas recuperados nas leituras do primeiro objetivo específico e pela

familiaridade do autor, destacando que outros *softwares* de tabulação e de mineração podem ser utilizados.

O terceiro objetivo específico, que foi 'alinhar o principal formato encontrado nos repositórios de bases de biodiversidade aos métodos e ferramentas de análises de dados' foi atendido, conforme apresentado no Capítulo 4, seção 4.3, onde este objetivo ocorreu em conjunto com o desenvolvimento do modelo de análise proposto nesta pesquisa, utilizando o conhecimento adquirido nos objetivos anteriores para a elaboração do objetivo principal.

O quarto objetivo específico, que foi 'validar o modelo proposto pela aplicação do mesmo em uma base escolhida em um dos repositórios de biodiversidade', foi atendido, conforme apresentado no Capítulo 4 seções 4.4 a 4.10, demonstrando a aplicabilidade do modelo proposto e destacando as condições não ideais apresentadas no conjunto de dados utilizados nesta pesquisa.

Mediante os resultados obtidos no desenvolvimento dos objetivos específicos, considera-se atendido o objetivo geral da pesquisa, que foi 'propor um modelo de análise de dados de bases de biodiversidade'. Espera-se que o modelo proposto neste trabalho (Capítulo 4, seção 4.3, Figura 8), possa ser utilizado em outras pesquisas relacionadas à biodiversidade, auxiliando os pesquisadores na descoberta de informações relevantes para suas necessidades.

A questão norteadora proposta 'quais padrões podem ser encontrados por meio de análise de dados da fauna brasileira em uma base de biodiversidade?' foi respondida por meio da aplicação do modelo proposto, uma vez que, conforme apresentado nas seções 4.4 a 4.10 e nos Apêndices 13 a 17, é possível visualizar os padrões presentes no conjunto de dados do Museu Nacional do Rio de Janeiro disponível no portal do GBIF.

Alguns padrões, por exemplo os descobertos pelo APRIORI, apresentaram a relação entre o *taxonRank*=FAMILY com o *phylum*=Arthropoda, indicando que este filo apresenta muitos registros com os espécimes identificados apenas no nível taxonômico de família, indicando um grupo de interesse para trabalhos de revisão taxonômica. Outras regras descobertas relacionam o *phylum*=Chordata com *taxonRank*=SPECIES, o que demonstra que os espécimes do filo Chordata presentes na coleção estão identificados ao nível taxonômico de espécie. Diferente do apresentado nas regras do filo Arthropoda, os registros do filo Chordata podem não representar um grupo de interesse para revisões taxonômicas. Ainda assim, todas os

resultados, para fins de validação, devem passar pela análise de especialistas da área, conforme previsto no modelo proposto.

Considera-se que a maior contribuição desta pesquisa foi apresentar um modelo que pode ser utilizado em outras pesquisas relacionadas à biodiversidade, auxiliando os pesquisadores na descoberta de informações relevantes a respeito dos organismos presentes no planeta, por meio de uma abordagem diferenciada de análise dos dados.

Recomenda-se a aplicação do modelo proposto a outros conjuntos de dados, como os fornecidos por projetos de ciência cidadã, ou conjuntos de dados que apresentem registros mais completos do que os utilizados nesta pesquisa, ou ainda dados obtidos em outras bases. Ressalta-se que, neste trabalho, o modelo de análise foi utilizado em níveis taxonômicos superiores, tornando necessária a aplicação do mesmo em níveis mais específicos.

REFERÊNCIAS

- ACEVEDO, M. A. et al. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. **Ecological Informatics**, v. 4, n. 4, p. 206-214, 2009.
- ACEVEDO, M. A.; VILLANUEVA-RIVERA, L. J. From the field: using automated digital recording systems as effective tools for the monitoring of birds and amphibians. **Wildlife Society Bulletin**, v. 34, n. 1, p. 211-214, 2006
- ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403-410, 1990
- ARIA, M.; CUCCURULLO, C. **Bibliometrix**. 2016 Disponível em: <https://www.bibliometrix.org/index.html>. Acesso em: 6 jan. 2020.
- BAIER, T.; NEUWIRTH, E. Excel: COM:: R. **Computational Statistics**, v. 22, n. 1, p. 91-108, 2007.
- BISBY, F. A. The quiet revolution: biodiversity informatics and the internet. **Science**, v. 289, n. 5488, p. 2309-2312, 2000.
- BOEGER, W. A. et al. **Catálogo Taxonômico da Fauna do Brasil**. 2016. Disponível em: <http://fauna.jbrj.gov.br/fauna/listaBrasil/PrincipalUC/PrincipalUC.do?lingua=pt>. Acesso em: 28 de fev. de 2021.
- BRASIL. Lei nº 9985, institui o Sistema Nacional de Unidades de Conservação da Natureza. **Diário Oficial Da União**, Brasília, 2000.
- BRUSCA, R. C.; BRUSCA, G. J. **Invertebrados**. Rio de Janeiro: Guanabara Koogan, 2007.
- CASTRO, L. N. de; FERRARI, D. G. **Introdução à mineração de dados**: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.
- BARBIERI, C. **Governança de dados**: práticas, conceitos e novos caminhos. Rio de Janeiro: Alta Books, 2020.
- CHOO, C. W. **A organização do conhecimento**: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. São Paulo: Senac, 2003.
- CONVENTION ON BIOLOGICAL DIVERSITY. [Report]. Rio de Janeiro: United Nations, 1992. Disponível em <https://www.cbd.int/doc/legal/cbd-en.pdf>. Acesso em: 16 set. 2020.
- DAMA, O. G. The DAMA guide to the data management body of knowledge. **DAMA International**, 2012.

DARWIN CORE MAINTENANCE GROUP (DwCMG). List of Darwin Core terms. Biodiversity Information Standards (TDWG). **Zenodo**, 2020. Disponível em: <https://dwc.tdwg.org/terms/>. Acesso em: 15 mar. 2021.

DAVENPORT, T. H. **Ecologia da informação**: por que só a tecnologia não basta para o sucesso na era da informação. São Paulo : Futura, 1998

JONES, M. B.; O'BRIAN, M.; MECUM, B.; BOTTIGER, C.; SCHILDHAUER, M.; MAIER, M. WHITEAKER, T., STEVAN, E.; CHONG, S. **Ecological Metadata Language version 2.2.0**. KNB Data Repository. 2019. Disponível em: <https://eml.ecoinformatics.org/>. Acesso em: 14 set. 2020.

EDGAR, R. C. Search and clustering orders of magnitude faster than BLAST. **Bioinformatics**, v. 26, n. 19, p. 2460-2461, 2010.

FÁVERO, L. P.; BELFIORE, P. **Manual de análise de dados**: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Elsevier Brasil, 2017.

FICHTMUELLER, D. et al. ABCD 3.0 ready to use. **Biodiversity Information Science and Standards**, v. 3, p. e37214, 2019.

FRANK, E.; HALL, M. A.; WITTEN, I. H. **The WEKA workbench**. Morgan Kaufmann, 2016.

FONTES, S. G.; STANZANI, S. L.; CORREA, P. L. P.. A data mining framework for primary biodiversity data analysis. In: **New Contributions in Information Systems and Technologies**. Springer, Cham, 2015. p. 813-821.

GITHUB. Enum *Occurrence*Issue. Disponível em: <https://gbif.github.io/gbif-api/apidocs/org/gbif/api/vocabulary/OccurrenceIssue.html>. Acesso em: 26 de maio de 2021.

GREGORY, T. R. et al. Eukaryotic genome size databases. **Nucleic acids research**, v. 35, n. suppl_1, p. D332-D338, 2007.

HEYWOOD, V. H.; WATSON, R. T. **Global biodiversity assessment**. Cambridge: Cambridge university press, 1995.

HOLETSCHEK J.; DRÖGE G.; GÜNTSCH A.; BERENDSOHN W.G. The ABCD of primary biodiversity data access. **Plant Biosystems** - An International Journal Dealing with all Aspects of Plant Biology. doi: 10.1080/11263504.2012.740085, 146:771–779, 2012

HORTAL, J.; LOBO, J. M.; JIMÉNEZ-VALVERDE, A.. Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. **Conservation Biology**, v. 21, n. 3, p. 853-863, 2007.

HUSON, D. H. et al. MEGAN analysis of metagenomic data. **Genome research**, v. 17, n. 3, p. 377-386, 2007.

IBGE (INSTITUTO BRAZILEIRO DE GEOGRAFIA E ESTATISTICA). Biomas e sistema costeiro-marinho do Brasil: compatível com a escala 1: 250 000. **Série Relatórios Metodológicos. v. 45**, Rio de Janeiro: IBGE, 2019. Disponível em: <https://www.ibge.gov.br/geociencias/informacoes-ambientais/15842-biomas.html?edicao=25799&t=acesso-ao-produto> , Acesso em: 5 jan. 2020.

ICMBIO (INSTITUTO CHICO MENDES DE CONSERVAÇÃO DA BIODIVERSIDADE). **Livro vermelho da fauna Brasileira ameaçada de extinção**. 2018.

LEVATICH. T; LIGOCKI, S.; GERBRACHT, J. **EOD - eBird Observation Dataset**. Disponível em: <https://www.gbif.org/pt/Dataset/4fa7b334-ce0d-4e88-aaae-2e0c138d049e#description>. Acesso em: 16 dez. 2020.

MAGURRAN, A. E. **Medindo a diversidade biológica**. Curitiba: Editora da UFPR, v. 261, 2011.

PINHEIRO, J. I. D.; DA CUNHA, S. B.; CARVAJAL, S. R. **Estatística Básica-a Arte de Trabalhar com Dados**. Elsevier Brasil, 2009.

R Project. **What is R?**. Disponível em: <https://www.r-project.org/about.html>. Acesso em: 6 jan. 2020.

ROCHA, R. M.; BOEGER, W. A. **Estado da arte e perspectivas para a Zoologia no Brasil**. Sociedade Brasileira de Zoologia. Editora UFPR. Curitiba, 2009.

RICÓS, C1 et al. Current databases on biological variation: pros, cons and progress. **Scandinavian Journal of Clinical and Laboratory Investigation**, v. 59, n. 7, p. 491-500, 1999.

RUA, J. DIGITALIZAÇÃO, PRESERVAÇÃO E ACESSO: contributos para o projeto Museu Digital da U. PORTO. **Páginas a&b: arquivos e bibliotecas**, 199-229, 2017

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação & Informação**, v. 21, n. 2, p. 116-142, 2016.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017.

SOARES, F. M.. **Princípios para a criação de uma extensão de metadados sobre interações ecológicas na agrobiodiversidade para o padrão Darwin Core**. 2019. Dissertação (Mestrado de Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, 2019.

TARAPANOFF, K. **Inteligência, informação e conhecimento em corporações**. Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), 2006.

TDWG (TAXONOMIC DATABASES WORKING GROUP). **ABCD Shema – Task Group on Access to Biological Collection Data**. 2007 Disponível em: <https://www.bgbm.org/tdwg/CODATA/Schema/default.htm>. Acesso em: 10 out. 2020.

TDWG (TAXONOMIC DATABASES WORKING GROUP). **PlinianCore**. Disponível em: <https://github.com/tdwg/PlinianCore/wiki/About>. Acesso em: 10 out. 2020

TEDESCHI, V. H. P.; TSUNODA, D. F. Banco de dados sobre biodiversidade: uma análise temporal da produção científica. **Revista Científica Multidisciplinar Núcleo do Conhecimento**, v. 6, n.9, , p. 68-81. 2020.

THE CORNELLLAB OF ORNITHOLOGY. **EBird**. Disponível em: <https://ebird.org/home>. Acesso em: 16 dez. 2020

WIECZOREK, J. et al. Darwin Core: an evolving community-developed biodiversity data standard. **PloS one**, v. 7, n. 1, p. e29715, 2012.

YAU, N. **Data points**: visualization that means something. John Wiley & Sons, 2013.

YESSON, C. et al. How global is the global biodiversity information facility?. **PloS One**, v. 2, n. 11, p. e1124, 2007.

APENDICE 1 – TABELA DA CLASSE RECORD-LEVEL

Elemento	Definição	Exemplos
Type	Natureza ou classe do recurso	•
Modified	Última data na qual o recurso foi alterado. E recomendado a utilização de um esquema de codificação, como a ISO 8601:2004	• Som • Imagem • Imagem em movimento • Objeto Físico • Texto
Language	Idioma do recurso. E recomendado a utilização de vocábulo controlado, como a ISO 639	• En (par Inglês) • Es (para espanhol) • Pt (para português)
License	Documento legal que permite oficialmente para a utilização do recurso	• http://creativecommons.org/publicdomain/zero/1.0/legalcode
RightsHolder	Organização ou pessoa que detém os direitos sobre o recurso	• The Regents of the University of California
accessRights	Informações de quem pode acessar os recursos disponibilizados ou indicação do status de segurança. Direitos de acesso podendo incluir informações sobre acesso ou restrições de privacidade, segurança ou políticas	• not-for-profit use only • https://www.fieldmuseum.org/field-museum-natural-history-conditions-and-suggested-norms-use-collections-data-and-images

bibliographicCitation	Uma referência vinculada com o registro como uma declaração indicando como este registro deverá ser citado quando for utilizado. A recomendação e de se utilizar detalhes suficientes para identificar o registro da maneira menos ambígua possível	• Oliver P. Pearson. 1985. Los tuco-tucos (genera Ctenomys) de los Parques Nacionales Lanin y Nahuel Huapi, Argentina Historia Natural, 5(37):337-343.
References	recurso relacionado que e referenciado, citado ou apontado de outra forma pelo recurso descrito	• http://arctos.database.museum/guid/MVZ:Mamm:165861 • http://www.catalogueoflife.org/annual-checklist/show_species_details.php?record_id=6197868
institutionID	Identificador utilizado pela instituição que possui a custódia do(s) Objeto(s) ou das informações mencionadas no registro	• http://grbio.org/cool/km06-gtbn • http://biocol.org/urn:lsid:biocol.org:col:34777 • http://grbio.org/cool/km06-gtbn
collectionID	Identificador da coleção ou conjunto de dados, partindo do qual o registro e derivado. Em amostras físicas, e recomendado a utilização do identificador de cadastro da coleção.	• http://biocol.org/urn:lsid:biocol.org:col:1001 • http://grbio.org/cool/p5fp-c036
DatasetID	Identificador para o conjunto de dados, podendo ser um identificador global ou um identificador específico para uma coleção ou instituição	b15d4952-7d20-46f1-8a3e-556a512b04c5

		•
InstitutionCode	Nome ou sigla, utilizada pela instituição que possui custódia do objeto ou das informações mencionadas no registro	• MVZ • FMNH • CLO • UCMP
CollectionCode	Nome, acrônimo, coden ou inicialismo que identifica a coleção ou conjunto de dados do qual o registro foi derivado	• Mammals • Hildebrandt • EIRD • VP
DatasetName	Identificação do conjunto de dados do qual o registro foi derivado	• Grinnell Resurvey Mammals • Lacey Ctenomys Recaptures
ownerinstitutionCode	Nome ou sigla em uso pela instituição que possui a propriedade dos objetos ou das informações mencionadas no registro	• NPS • APN • InBio
basisOfRecord	Termo obrigatório para registros de ocorrências, sendo natureza específica do registro de dados. E recomendado a utilização do vocábulo controlado como o indicado pelo Darwin Core Type Vocabulary	• PreservedSpecimen • FossilSpecimen • LivingSpecimen • MaterialSample • Event • HumanObservation • MachineObservation • Taxon

		• Occurrence
InformationWithheld	Informações adicionais que existem, mas que não foram compartilhados no registro	• location information not given for endangered species • collector identities withheld ask about tissue samples
dataGeneralizations	Ações feitas com o intuito de tornar as informações menos específicas ouu completos, do que em sua forma original, os dados fornecidos. Sugere-se que os dados alternativos de maior completitude possam estar disponíveis	• Coordinates generalized from original GPS coordinates to the nearest half degree grid cell
dynamicProperties	Uma lista continua e com valores separados por “; ”, de medições, fatos, características ou afirmações adicionais sobre o registro. Pretende-se fornecer um mecanismo para estruturar contedos.	• {"heightInMeters":1.5} • {"iucnStatus":"vulnerable", "taxonDistribution":"Neuquén, Argentina"} • {"relativeHumidity":28, "airTemperatureInCelsius":22, "sampleSizeInKilograms":10}

FONTE: DwCMG (2020), Soares (2019) e Wieczorek et al. (2012)

APENDICE 2 - TABELA DA CLASSE OCCURRENCE

Elemento	Descrição	Exemplos
OccurrenceID	Termo obrigatório para os registros de ocorrência, sendo um identificador para a ocorrência. Na ausência de um identificador único, global e persistente, pode se gerar um a partir da combinação de identificadores no registro que podem tornar-se o ID único Global.	• http://arctos.database.museum/guid/MSB:Mamm:233627 • 000866d2-c177-4648-a200-ead4007051b9 • urn:catalog:UWBM:Bird:89776
catalogNumber	Identificador, preferencialmente único, para o registro dentro do conjunto de dados ou coleções	• 145732 • 2008.1334 • R-4313
recordNumber	Identificador para o registro dentro do conjunto de dados ou coleções, o qual deve ser preferencialmente único	• OPP 7101
recordedBy	Lista de nomes de pessoas, grupos ou organizações responsáveis pelo registro da ocorrência original, a lista deve ser continua com os valores separados por “;”. o coletor ou observador primeiro, especialmente aquele que dá um identificador pessoal, deve ser o primeiro da lista	• José E. Crespo • Oliver P. Pearson Anita K. Pearson
individualCount	Número de indivíduos representados e presentes no momento da ocorrência	• 0 • 1 • 25

OrganismQuantity	Número ou valor de enumeração para a quantidade de organismos	• 27 • 12,5 • r
organismQuantityType	Sistema utilizado para quantificar a quantidade de organismos correspondentes ao registro	• Individuals • %biomass • BraunBlanquetScale
Sex	Sexo do(s) indivíduo(s) biológico(s) representado(s) na ocorrência, recomendado a utilização de vocabulário controlado	• Female • Male • hermaphrodite
lifeStage	Idade ou fase de vida dos indivíduos quando a ocorrência foi registrada, recomendado o uso de vocábulo controlado	• egg • eft • juvenile • adult
ReproductiveCondition	Condição reprodutiva dos indivíduos biológicos representados na ocorrência. Recomendado a utilizações de vocábulo controlado	• non-reproductive • pregnant • in bloom • fruit-bearing
Behavior	Descrição do comportamento mostrado pelo espécime quando a ocorrência foi registrada, recomendado a utilização de vocábulo controlado	• roosting • foraging • running
establishmentMeans	Processos pelos qual os indivíduos representados na ocorrência ficaram estabelecidos no local, recomendado a utilização de vocábulo controlado	• native • nativeReintroduced

	- introduced - introducedAssistedColonisation - vagrant - uncertain
degreeOfEstablishment	O grau em que um organismo sobrevive, se reproduz e expande seu alcance em um determinado lugar e tempo, sendo recomendado a utilização de vocábulo controlado - native - captive - cultivated - released - failing - casual - reproducing - established - colonising - invasive - widespreadInvasive
Pathway	O processo pelo qual um Organismo chega a um determinado lugar em um determinado momento, sendo recomendado a utilização de vocábulo controlado - releasedForUse - otherEscape - transportContaminant - transportStowaway - corridor - unaided

OccurrenceStatus	Definição sobre a presença ou ausência de um táxon em um local, recomendado a utilização de vocábulo controlado	• presente • absent
Preparations	Lista de métodos e preparações de preservação dos espécimes, a lista deve ser contínua e com valores separados por “ ”	• fósfil • cast • photograph • DNA extract • skin skull skeleton • whole animal (ETOH) tissue (EDTA)
Disposition	Estado atual de um exemplar em respeito à coleção identificada em collectionCode ou collectionID, recomendado a utilização de vocábulo controlado	• in collection • missing • voucher elsewhere • duplicates elsewhere
AssociatedMedia	Lista de identificadores das mídias associada com a ocorrência, a lista deve ser contínua e com valores separados por “ ”	• https://arctos.databases.e.museum/media/10520962 https://arctos.databases.e.museum/media/10520964
associatedReferences	Lista de identificadores referentes a publicações, referência bibliográfica, da literatura associada a ocorrência	• http://www.sciencemag.org/cgi/content/abstract/322/5899/261 • Christopher J. Conroy, Jennifer L. Neuwald.

2008. Phylogeographic study of the California vole, Microtus californicus Journal of Mammalogy, 89(3):755-767.	
associatedSequence	Lista de identificadores referente a publicações e referências bibliográficas de informações da sequência genética associada com a ocorrência
• http://www.ncbi.nlm.nih.gov/nucleotide/U34853.1 • http://www.ncbi.nlm.nih.gov/nucleotide/GU328060 http://www.ncbi.nlm.nih.gov/nucleotide/AF326093	
associatedTaxa	Lista de identificadores ou nomes de táxons e suas associações com a ocorrência
• "host": "Quercus alba" • "parasitoid of": "Cyclocephala signaticollis" "predator of": "Apis mellifera"	
otherCatalogNumbers	Lista de números de catálogo completos anteriores, alternativos ou outros identificadores usados pelo homem para a mesma ocorrência, seja no conjunto de dados ou na coleção atual ou em qualquer outro conjunto de dados. Recomendado a separação dos valores utilizando barra vertical " "
• FMNH:Mammal:1234 • NPS YELL06778 MBG 33424	

Occurrences	Remarks	Comments or notes on the occurrence	
			• found dead on road

FONTE: DwCMG (2020), Soares (2019) e Wierzchorek et al. (2012)

292063?seid=3177392 http://arctos.database.museum/guid/MSB:Mamm:292063?seid=3609139	
associatedOrganisms	•
	• "sibling
	of": "DWNS:Mamm http://arctos.database.museum/guid/MSB:Mamm:292063?seid=3609139
	of": "MSB:Mamm http://arctos.database.museum/guid/MSB:Mamm:292063?seid=3609139
	"parent of": "MSB:Mamm http://arctos.database.museum/guid/MSB:Mamm:292063?seid=3609139
PreviousIdentifications	• of": "MSB:Mamm http://arctos.database.museum/guid/MSB:Mamm:292063?seid=3609139
	• Chalepidae
	• Pinus abies
	• Anthus sp., field ID by G. Iglesias
	Cicero 2009-02-12 based on morphology
organismRemarks	One of a litter of six
	•

FONTE: DwCMG (2020), Soares (2019) e Wieczorek et al. (2012)

APENDICE 4 - TABELA DA CLASSE MATERIALSAMPLE

Elementos	descrição	exemplos
materialSampleId	Identificador para a classe materialSample, na ausência de um identificador único, global e persistente, construir um a partir de uma combinações de identificadores no registro que pode torna-lo único globalmente	06809dc5-f143-459a-be1a-6f03e63fc083

FONTE: DwCMG (2020), Soares (2019) e Wieczorek et al. (2012)

APENDICE 5 - TABELA DA CLASSE EVENT

Elementos	Descrição	Exemplos
eventID	Identificador para o conjunto de informações associadas a um evento, recomenda-se um identificador único e global ou um identificador específico para o conjunto de dados	• INBO:VIS:Ev:00009375
parentEventID	Identificador par o evento mais amplo que agrupa e potencialmente outros eventos	• A1
fieldNumber	Identificador dado ao evento no campo, servindo muitas vezes como link entre as notas de campo e o evento	• RV Sol 87-03-08
eventDate	A data-hora ou intervalo durante o qual um evento ocorreu. Para ocorrências, é a data e hora em que o evento foi registrado.	• 1963-03-08T14:07-0600 • 2009-02-20T08:40Z • 2018-08-29T15:19 • 1809-02-12 • 1906-06 • 1971 • 2007-03- • 01T13:00:00Z/2008-05-11T15:30:00Z • 1900/1909 • 2007-11-13/15
eventTime	Hora ou intervalo ao qual ocorreu o evento, recomendado a utilização de codificação coma ISO 8601:2004	• 14:07-0600 • 08:40:21Z • 13:00:00Z/15:30:00Z

startDayOfYear	Primeiro dia ordinal do ano em que ocorreu o evento, 1 para 1 de janeiro e 365 para 31 de dezembro, exceto em um ano bissexto, que no caso é 366	• 1 • 366 • 365
endDayOfYear	Ultimo dia ordinal do ano em que ocorreu o evento, 1 para 1 de janeiro e 365 para 31 de dezembro, exceto em um ano bissexto, que no caso é 366	• 365 • 32 • 1
Year	Ano de quatro dígitos em que ocorreu o evento, de acordo com o calendário da Era Comum	• 1160 • 2008
Month	Mês, em número inteiro, em que ocorreu o evento	• 10 • 1 • 12
Day	Dia, em número inteiro do mês em que ocorreu o evento	• 9 • 28 • 31
Verbatim EventDate	Representação textual e original da data e as informações de tempo para o evento	• spring 1910 • Marzo 2002 • 1999-03-XX • 17IV1934
Habitat	Categoria ou descrição do habitat em que ocorreu o evento	• oak savana • pre-cordillera steppe •

samplingProtocol	Nome, referência ou descrição do método ou protocolo usado durante um Evento	• UV light trap • bottom trawl • point count • Penguins from space: faecal stains reveal the location of emperor penguin colonies, https://doi.org/10.1111/j.1466-8238.2009.00467.x • Takats et al. 2001. Guidelines for Nocturnal Owl Monitoring in North America. Beaverhill Bird Observatory and Bird Studies Canada, Edmonton, Alberta. 32 pp. • http://www.bsc-eoc.org/download/Owl.pdf
sampleSizeValue	Valor numérico para uma medição do tamanho de uma amostra em um evento de amostragem	• 5 metros

sampleSizeUnit	Unidade de medida do tamanho de uma amostra em um evento de amostragem	• Minute • Hour • day • metre • square metre • cubic metre • 40 trap-nights • 10 observer-hours • 10 km by foot • 30 km by car • Notes available in the Grinnell-Miller Library. • After the recent rains the river is nearly at flood stage
samplingEffort	Quantidade de esforço gasto durante um evento	
fieldNotes	Pode ser um indicador da existência de, ou uma referência a ou o texto das anotações feitas no campo sobre o Evento	
eventRemarks	Comentários ou notas sobre o evento	

FONTE: DwCMG (2020), Soares (2019) e Wieczorek et al. (2012)

APENDICE 6 - TABELA DA CLASSE LOCATION

Elementos	Descrição	Exemplos
loc	Identificador para o conjunto de informações do local. Podendo ser um	• https://opencontext.org/su
ati	identificador único e global ou um específico dentro do conjunto de dados	• bjects/768A875F-E205-4D0B-DE55-BAB7598D0FD1
onl		
D		
higherGeographyID	Identificador da região geográfica dentro da qual ocorreu o evento. Recomendado a utilização de vocábulo controlado	• http://vocab.getty.edu/tgn/1002002
higherGeography	Uma lista de nomes geográficos menos específicas do que as informações contidas no termo locality	• North Atlantic Ocean • South America Argentina Patagonia Parque Nacional Nahuel Huapi Neuquén Los Lagos • Neuquén • Los Lagos
continent	Nome do continente em que a localização ocorre, recomendado a utilização de vocábulo controlado	• Africa • Antartctica • Asia • Europe • North America • Oceania • South America

waterBody	Nome do corpo de agua onde ocorreu o registro, recomentado a utilização de vocábulo controlado	- Indian Ocean - Baltic Sea - Hudson River - Lago Nahuel Huapi
islandGroup	Nome do grupo de ilhas em que ocorreu o registro, recomentado a utilização de vocábulo controlado	- Alexander Archipelago - Archipiélago Diego Ramírez - Seychelles
island	Nome da ilha do local do registro, podendo ser recitado também o nome da ilha nos casos da ocorrência ter ocorrido nas proximidade com a ilha, recomentado a utilização de vocábulo controlado	- Nosy Be - Bikini Atoll - Vancouver - Viti Levu - Zanzibar
Country	Nome do País ou maior unidade administrativa em que o registro ocorreu, recomentado a utilização de vocábulo controlado	- Denmark - Colombia - España
countryCode	Código padrão para o país em que o registro ocorre, recomentado a utilização de vocábulo controlado	- AR - SV
stateProvince	Nome da região administrativa menor que país, sendo estado ou condado, em que o registro ocorreu, recomendado a utilização de vocábulo controlado	- Montana, - Minas Gerais - Córdoba

County	Nome completo de uma região administrativa menor que <i>stateProvince</i> , neste caso município, de ocorrendo do registro, não deve-se utilizar abreviação neste campo	• Missoula • Los Lagos • Mataró
Municipality	Nome completo da região administrativa menor do que county, sendo distritos ou cidade, no qual ocorreu o registro, não deve-se utilizar abreviação neste campo	• Holzminden • Araçatuba • Ga-Segonyana
Locality	Descrição específica do local, contendo informações geográficas de menor especificidade, podendo ser fornecido informações de outros elementos geográficos, podendo conter informações para corrigir imprecisões nos termos anteriores ou padronizar a descrição	• Bariloche, 25 km NNE via Ruta Nacional 40 (=Ruta 237)
verbatimLocality	A descrição textual e original do lugar	• 25 km NNE Bariloche por R. Nac. 237
minimumElevation InMeters	Limite inferior do intervalo de elevação em metros	• -100 • 802
maximumElevation InMeters	Limite superior do intervalo de elevação em metros	• -205 • 1236
verbatimElevation	Descrição original da elevação do local	• 100-200 m
minimumDepthInMeters	Menor profundidade de um intervalo de profundidade abaixo da superfície, em metros	• 0 • 100
maximumDepthInMeters	Maior profundidade de um intervalo de profundidade abaixo da superfície, em metros	• 0 • 200
verbatimDepth	Descrição original da profundidade abaixo da superfície.	• 100-200 m

minimumDistance- Above SurcelnMeters	Menor distancia de um intervalo de distância a partir e uma superfície de referência em direção vertical, em metros. Sendo utilizado valores positivos para registros acima da superfície e valores negativos para registros abaixo da superfície do local	• -1.5 • 4.2 • 300m • 300 • 20m • 0 • -1.5
maximumDistance AboveSurface InMeters	Maior distancia de um intervalo de distância a partir e uma superfície de referência em direção vertical, em metros. Sendo utilizado valores positivos para registros acima da superfície e valores negativos para registros abaixo da superfície do local	• -1.5 • 4.2 • 300m • 300 • 20m • 0 • -1.5
Location AccordingTo	Informações sobre a fonte de informação do local específico de registro, podendo ser uma publicação, instituição ou grupo de indivíduos	• Getty Thesaurus of Geographic Names • GADM
locationRemarks	Comentários ou notas sobre o local	• under water since 2005
decimalLatitude	Latitude geográfica do centro geográfico do local, valor dever ser apresentado em graus decimais, utilizando o sistema de referência espacial dado em geodeticDatum.	• -41.0983423

decimalLongitude	Longitude geográfica do centro geográfico do local, valor dever ser apresentado em graus decimais, utilizando o sistema de referência espacial dado em geodeticDatum.	-121.1761111
geodeticDatum	O elipsoide, datum geodésico, ou sistema de referência espacial (SRS) em que as coordenadas geográficas dadas decimalLatitude e decimalLongitude estão baseadas. A prática recomendada é usar o código EPSG como um vocabulário controlado para fornecer um SRS, se conhecido. Caso contrário, use um vocabulário controlado para o nome ou código do datum geodésico, se conhecido. Caso contrário, use um vocabulário controlado para o nome ou código do elipsoide, se conhecido. Se nenhum desses é conhecido, use o valor "desconhecido"	- EPSG:4326 - WGS84 - NAD27 - Campo Inchauspe - European 1950 - Clarke 1866 - Unknown
Coordinate	Distancia horizontal a partir de uma latitude e longitude decimais dadas	30
UncertaintyInMetes	descrevendo o círculo menor que contem a totalidade do local, recomentado deixar o valor em branco caso a incerteza seja desconhecida, não pode ser estimada, não se aplica, sendo que o valor zero não e valido para esse termo	71
coordinatePrecision	Representação decimal da precisão das coordenadas apresentadas nos termos anteriores	0.00001 0.000278 0.01667 1.0
pointRadius	A relação entre a área do ponto-raio para a área da representação espacial	0
SpatialIFit	verdadeira original ou especifica do local. Os valores permitidos neste termo ou menor do que ou igual a 1, sendo 1 correspondente a 100% de sobreposição.	1 1.5708

Verbatim	Coordenadas espaciais originais e textuais do local	• 41 05 54S 121 05 34W
Coordinates		• 17T 630000 4833400
verbatimLatitude	latitude textual e original do Local	• 41 05 54.03S
verbatimLongitude	A longitude textual e original do Local	• 121d 10' 34" W
Verbatim	sistema de coordenadas espaciais utilizadas no registro, recomendado a	• decimal degrees
CoordinateSystem	utilização de vocábulo controlado	• degrees decimal minutes • degrees minutes seconds • UTM
verbatimSRS	O elipsoide, datum geodésico, ou SRS para as coordenadas dadas em verbatimLatitude e verbatimLongitude, ou verbatimCoordinates no qual estão baseadas. A prática recomendada é usar o código EPSG como um vocabulário controlado para fornecer um SRS, se conhecido. Caso contrário, use um vocabulário controlado para o nome ou código do datum geodésico, se conhecido. Caso contrário, use um vocabulário controlado para o nome ou código do elipsoide, se conhecido. Se nenhum deles é conhecido, use o valor "desconhecido"	• Unknown • EPSG:4326 • WGS84 • NAD27 • Campo Inchauspe • European 1950 • Clarke 1866
footprintWKT	Representação dos dados espaciais em formato de texto, que define o local.	• POLYGON ((10 20, 11 20, 11 21, 10 21, 10 20))
footprintSRS	Representação dos dados espaciais em formato de texto, que define o local. Não use este termo para descrever o SRS	• GEOGCS["GCS_WGS_1984", DATUM["D_WGS_1984"]

• SPHEROID["WGS_1984",6378137,298.257223563]], PRIMEM["Greenwich",0] • UNIT["Degree",0.0174532925199433]]	
footprintSpatialFit	Relação entre a área do footprint e a área da representação espacial verdadeira do local. Os valores permitidos neste termo 0 ou menor do que ou igual a 1, sendo 1 correspondente a 100% de sobreposição.
georeferencedBy	Lista de nomes de pessoas, grupos ou organizações que determinaram o georreferenciamento para o registro
• Brad Millen (ROM) • Kristina Yamamoto Janet Fang	
georeferencedDate	Data em que o local foi georeferenciado, recomendado esquema de codificação, como a ISO 8601:2004
• 1963-03-08T14:07-0600 • 2009-02-20T08:40Z • 2018-08-29T15:19 • 1809-02-12 • 1906-06 • 1971 • 2007-03-01T13:00:00Z/2008-05-11T15:30:00Z • 1900/1909 • 2007-11-13/15	
georeferenceProtocol	Descrição ou referência aos métodos usados para determinar a descrição espacial, coordenadas e incertezas
I	Georeferencing Quick Reference Guide (Zermoglio

et al. 2020, https://doi.org/10.35035/e09p-h128	
georeferenceSource	Uma lista de mapas, dicionários geográficos ou outros recursos usados para o georreferenciar o local, descrito com a suficiente especificidade para permitir que qualquer pessoa, no futuro, possa usar os mesmos recursos - USGS 1:24000 Florence Montana Quad 1967 Terrametrics 2008 on Google Earth - GeoLocate https://www.geonames.org/
Georeference VerificationStatus	Descrição categórica da medida em que o georreferenciamento foi verificado para representar da melhor maneira possível a descrição espacial. recomendado a utilização de vocábulo controlado - requires verification - verified by collector - verified by curator
georeferenceRemarks	Notas ou comentários sobre a determinação da descrição espacial, explicando suposições feitas em adição ou oposição daqueles formalizados Assumed distance by road (Hwy. 101)

FONTE: DwCMG (2020), Soares (2019) e Wieczorek et al. (2012)

APENDICE 7 - TABELA DA CLASSE GEOLOGICALCONTEXT

Elementos	Descrição	Exemplos
geologicalContextID	Identificador para o conjunto de informações associadas ao contexto geológico, podendo ser um identificador único com um identificador específico dentro do conjunto de dados	• https://opendatacommons.org/licenses/by/4.0/
earliestEonOrLowestEonOr	Nome completo do eon geológico mais antigo possível ou eonothem cronoestratigráfico inferior ou nome informal atribuível ao horizonte estratigráfico do qual o item catalogado foi coletado	• Phanerozoic • Proterozoic
LatesEonOrHighestEonOr	Nome completo do eon geológico mais recente possível ou eonothem cronoestratigráfico inferior ou nome informal atribuível ao horizonte estratigráfico do qual o item catalogado foi coletado	• Phanerozoic • Proterozoic
earliestEraOrLowestEraOr	Nome completo da era geocronológica mais antiga possível ou eratema cronoestratigráfico inferior atribuível ao horizonte estratigráfico do qual o item catalogado foi coletado	• Cenozoic • Mesozoic
latesEraOrHighestEraOr	Nome completo da era geocronológica mais recente possível ou eratema cronoestratigráfico inferior atribuível ao horizonte estratigráfico do qual o item catalogado foi coletado	• Cenozoic • Mesozoic
earliestPeriodOrLowestPeriodOr	Nome completo do período geocronológico mais antigo possível ou sistema cronoestratigráfico inferior atribuível ao horizonte estratigráfico do qual o item catalogado foi coletado	• Neogene • Tertiary • Quaternary

latestPeriodOr HighestSystem	Nome completo do período geocronológico mais recente possível ou sistema cronoestratigráfico inferior atribuível ao horizonte estratigráfico do qual o item catalogado foi coletado	• Neogene • Tertiary • Quaternary
earliestEpochOr LowestSeries	Nome completo da época geocronológica mais antiga possível ou séries cronoestratigráfico inferior atribuível ao horizonte estratigráfico do qual o item catalogado foi coletado.	• Holocene • Pleistocene • Ibexian Series
latesEpochOr HighestSeries	Nome completo da época geocronológica mais recente possível ou séries cronoestratigráfico superior atribuível ao horizonte estratigráfico do qual o item catalogado foi coletado	• Holocene • Pleistocene • Ibexian Series
earliestAgeOr LowesStage	nome completo da idade geocronológica mais antiga possível ou andar cronoestratigráfico inferior atribuível ao horizonte estratigráfico do qual o item catalogado foi coletado	• Atlantic • Boreal • Skullrockian
latestAgeOr HighestStage	Nome completo da idade geocronológica mais recente possível ou andar cronoestratigráfico superior atribuível ao horizonte estratigráfico do qual o item catalogado foi coletado	• Atlantic • Boreal • Skullrockian
Lowest Biostratigraphic Zone	nome completo da zona bioestratigráfica inferior do horizonte estratigráfico do qual o item catalogado foi coletado.	• Maastrichtian
Highest Biostratigraphic Zone	nome completo da zona bioestratigráfica superior do horizonte estratigráfico do qual o item catalogado foi coletado	• Blancan

Lithostratigraphic Terms	Combinação de todos os nomes litoestratigraficos para a rocha da qual i item catalogado foi coletado	• Pleistocene-Weichselien
Group	nome completo do grupo litoestratigráfica do qual o item catalogado foi coletado	• Bathurst • Lower Wealden
Formation	Nome completo da formação litoestratigráfica do qual o item catalogado foi coletado	• Notch Peak Formation • House Limestone • Fillmore Formation
Member	Nome completo do membro litoestratigráfica do qual o item catalogado foi coletado	• Lava Dam Member • Hellnmaria Member
bed	Nome completo da Camada litoestratigráfica da qual o item catalogado foi coletado	• Harlem coal

FONTE: DwCMG (2020), Soares (2019) e Wieczorek et al. (2012)

APENDICE 8 - TABELA DA CLASSE IDENTIFICATION

Elementos	Descrição	Exemplos
identificationID	Identificador utilizado para identificar o bloco de informações que estão associadas com a atribuição de um nome científico. Podendo ser um identificador único global ou um identificador específico dentro do conjunto de dados	9992
identificationQualifier	Breve frase ou um termo padrão para expressar as dúvidas do determinador sobre a identificação	- aff. agrifolia var. oxyadenia - cf. var. oxyadenia - Quercus agrifolia cf. var. oxyadenia
typeStatus	Lista de tipos nomenclaturas aplicadas ao espécime (status do tipo, nome científico tipificado, publicação), a lista deve ser continua com valores separados por “ ”	- holotype of Ctenomys sociabilis. Pearson O. P., and M. I. Christie. 1985. Historia Natural, 5(37):388 - holotype of Pinus abies holotype of Picea abies
identifiedBy	Lista de nomes de pessoas, grupos ou organizações que atribuíram o táxon ao espécime, a lista deve ser continua com valores separados por “,”	- James L. Patton - Theodore Pappenfuss Robert Macey

		tiger salamanders (Ambystoma tigrinum) (Amphibia: Caudata). Herpetologica, 53(1), 30-49
identificationStatus	Indicador categórico da média em que a identificação taxonômica foi verificada como correta, recomendado a utilização de vocábulo controlado.	• 0
identificationRemarks	Comentários ou anotação sobre a identificação	• Distinguished between Anthus correndera and Anthus hellmayri based on the comparative lengths of the uñas.

FONTE: DwCMG (2020), Soares (2019) e Wieczorek et al. (2012)

APENDICE 9 - TABELA DA CLASSE TAXON

Elementos	Descrição	Exemplos
taxonID	Termo obrigatório para alista de espécies, sendo um identificador para o conjunto de informações do táxon, podendo ser um identificador único e global ou um identificador específico dentro do conjunto de dados	• 8fa58e08-08de-4ac1-b69c-1235340b7001 • 32567 • https://www.gbif.org/species/212
scientificNameID	Identificador para os detalhes nomenclaturas do nome científico, não sendo detalhes taxonômicos	• urn:lsid:ipni.org:names:37829-1:1.3
acceptedName UsageID	Identificador para o nome em uso do nome atual valido (zoologia) ou aceite (botânica) do taxo	• 8fa58e08-08de-4ac1-b69c-1235340b7001
parentNameUsage ID	Identificador para o nome em uso do táxon parental diretamente mais próximo de maior nível hierárquico do elemento mais específico do scientificName	• https://www.gbif.org/species/2684876
originalNameUsage eID	identificador para o uso do nome (significado documentado do nome de acordo com uma fonte) no qual o elemento terminal de ScientificName foi originalmente estabelecido sob as regras do nomenclaturalCode associado.	• https://www.gbif.org/species/2685484
nameAccording ToID	Identificador para a fonte em que está definido ou implícito a circunscrição do conceito do táxon específico.	• https://doi.org/10.1016/j.sbspro.2015.12.001

namePublished	Identificador para a publicação em que o scientificName foi originalmente publicado sob as regras do código de nomenclatura	
lnID		
taxonConceptID	Identificador para o conceito taxonômico a que o registro se refere, este campo não e para ser utilizado para detalhes de nomenclatura de um táxon	8fa58e08-08de-4ac1-b69c-1235340b7001
scientificName	Nome científico completo, com autoria e informações de data se conhecidas, sendo parte de uma identificação, o nome deve ser em nível mais baixo da classificação taxonômica possível. Neste termo não se devem conter qualificações de identificação, que devem antes ser fornecidos no termo identificationQualifier.	- Coleoptera - Vespertilionidae - Manis - Ctenomys - sociabilis - Ambystoma tigrinum - diaboli - Roptrocerus - typographi (Györfi, 1952) - Quercus agrifolia - var. oxyadenia (Torr.) J.T. Howell - Tamias - minimum(valid name for Eutamias minimum).
acceptedName	Nome completo, com autoria e informações de data se conhecida, do nome atual valido (zoologia) ou aceito (botânico) do táxon	
Usage		

parentNameUsage	nome completo, com informações de autoria e data, se conhecido, do táxon pai de classificação superior direto e mais próximo (em uma classificação) do elemento mais específico do Nome científico.	- Rubiaceae - Gruiformes - Testudinae
originalName	nome do táxon, com autoria e data, se conhecida, da maneira que apareceu originalmente	Pinus abies
Usage	quando foi publicado pela primeira vez sob as regras do código de nomenclatura associado ao registro	Gasterosteus saltatrix Linnaeus 1768
nemAccordingTo	Referência para a fonte em que está definida ou implícita a circunscrição do conceito do táxon específico. Tradicionalmente representado por a expressão do Latim "sensu" ou "Sec". Para táxons que resultam de identificações, devem ser dadas as referências para as chaves, monografias, especialistas e outras fontes relacionadas com o registro	McCranie, J. R., D. B. Wake, and L. D. Wilson. 1996. The taxonomic status of Bolitoglossa schmidti, with comments on the biology of the Mesoamerican salamander Bolitoglossa dofleini (Caudata: Plethodontidae). Carib. J. Sci. 32:395-398.

	- Werner Greuter 2008 - Lilljeborg 1861, Upsala Univ. Arsskrift, Math. Naturvet., pp. 4, 5
namePublishedIn	referência para a publicação em que o scientificName foi originalmente publicado sob as regras do código de nomenclatura associado - Pearson O. P., and M. I. Christie. 1985. Historia Natural, 5(37):388 - Forel, Auguste, Diagnosies provisoires de quelques espèces nouvelles de fourmis de Madagascar, récoltées par M. Grandidier., Annales de la Societe Entomologique de Belgique, Comptes-

		rendus des Seances 30, 1886
namePublished InYear	Lista de nomes de táxons terminando no nível imediatamente superior ao nome do táxon referenciado no registro do táxon, a lista deve ser contínua e com os valores separados por “,”. E recomendado iniciar a lista com o nível mais alto	• 1915 • 2008
Higher Classification	Lista de nomes de táxons terminando no nível imediatamente superior ao nome do táxon referenciado no registro do táxon, a lista deve ser contínua e com os valores separados por “ ”. E recomendado iniciar a lista com o nível mais alto	Plantae Tracheophyta Magnoliopsida Ranunculales Ranunculaceae Ranunculus • Animalia • Animalia Chordata Vertebrata Mammalia Theria Eutheria Rodentia Hystriognatha Hystriognathi Ctenomyidae Ctenomyini Ctenomys
Kingdom	nome científico completo do reino em que o táxon é classificado	• Animalia

- Archaea - Bacteria - Chromista - Fungi - Plantae - Protozoa - Viruses	
Phylum	Nome científico completo do filo em que o táxon é classificado
class	Nome científico completo da classe em que o táxon é classificado
Order	Nome científico completo da ordem em que o táxon é classificado
family	nome científico completo da família em que o táxon é classificado
Genus	nome científico completo do gênero em que o táxon é classificado
Subgenus	Nome científico completo do subgênero em que o táxon é classificado. Os valores devem incluir o gênero para evitar confusão com o homônimo
specificEpithet	nome do epíteto específico do scientificName
- Concolor - gottschei	

infraspecificEpith et	Nome do epíteto infra específico de menor categoria ou hierarquia do scientificName, excluindo qualquer designação de categoria taxonômica	• concolor • oxyadenia • sayi
taxonRank	nível taxonômico do nome mais específico em scientificName, sendo recomendado a utilização de vocábulo controlado	• subspecies • varietas • forma • species • genus
Verbatim TaxonRank	nível taxonômico do nome mais específico em scientificName como aparece no registro original	• Agamospecies • sub-lesus • prole • apomict • nothogrex • sp. • subsp. • var.
scientificName Authorship	Informação sobre a autoria do scientificName, formado de acordo com o as convenções do código de nomenclatura	• (Torr.) J.T. Howell • (Martinovský) Tzvelev • (Györfi, 1952)
vernacularName	Nome comum ou vernáculo	• Andean Condor

		• Condor Andino • American Eagle • Gänsegeier
nomencIaturalcod	Código de nomenclatura sob o qual o scientificName e construído. Sendo recomendado a utilização de um vocábulo controlado	•
Taxonomic Status	Condição de uso do scientificName com um rotulo para um táxon, requerido opinião taxonômica para definir o escopo de um táxon, regras de prioridade são utilizadas para definir a condição taxonômica da nomenclatura contida neste âmbito, combinados com a opinião de especialistas, devendo estar associada a uma referência taxonômica específica que define o conceito. Sendo recomendado a pratica de vocábulo controlado	• ICN • ICZN • BC • ICNCP • BioCode
Nomenclatural Status	Estado do nome e da sua conformidade com as normas pertinentes da nomenclatura em relação com a publicação original. Baseia-se essencialmente em um algoritmo de acordo com as regras de negócio do código. Ele não exige nenhuma opinião taxonômica	• nom. ambig. • nom. illeg. • nom. subnud.
taxonRemarks	Comentários ou notas sobre o táxon ou nome	• this name is a misspelling in common use

FONTE: DwCMG (2020), Soares (2019) e Wieczorek et al. (2012)

APENDICE 10 - TABELA DA CLASSE MEASUREMENTORFACT

Elementos	Descrição	Exemplos
measurementID	Identificador para as medições ou fatores, informações relativas as medidas, fatos características, ou afirmações relacionadas ao registro. Podendo ser um identificador único global ou um específico dentro do conjunto de dados	9c752d22-b09a-11e8-96f8-529269fb1459
measurementType	Natureza da medição, fato, característica ou afirmação, e recomendado o uso de vocábulo controlado	- tail length - temperature - trap line length - survey área - trap type
measurementValue	valor da medição, fato, característica ou afirmação	45 20 1 14.5 - UV-light
Measurement Accuracy	descrição do erro potencial associado com o measurementValue	0.01 - normal distribution with variation of 2 m
measurementUnit	unidades associadas com o measurementValue, recomendado a utilização das unidades do SI	- mm - C - Km

		ha
Measurement	Lista de nomes de pessoas, grupos ou organizações que determinaram o valor de	Rob Guralnick
DeterminedBy	medições ou fatos, esta lista deve ser continua com os valores separados por " "	Peter Desmet Stijn Van Hoey
Measurement	Data em que a Medições ou fatos foi feita. Recomendado a utilizações de esquema	1963-03-08T14:07-0600
DeterminedDate	de codificação, como a ISO 8601:2004	2009-02-20T08:40Z 2018-08-29T15:19 1809-02-12 1906-06 2007-03- 01T13:00:00Z/2008-05- 11T15:30:00Z 1900/1909 2007-11-13/15
Measurement	Descrição ou referência para o método ou protocolo utilizado para determinar a	minimum convex polygon
Method	medida, fato, característica ou afirmação.	- around burrow entrances - barometric altimeter
Measurement	Comentários ou notas que acompanham as medições ou fatos	tip of tail missing
Remarks		

FONTE: DwCMG (2020), Soares (2019) e Wieczorek et al. (2012)

APENDICE 11 - TABELA DA CLASSE RESOURCERELATIONSHIP

Elementos	Descrição	Exemplos
Resource RelationshipID	Identificador de uma instancia de relação entre um recurso e outro	• 04b16710-b09c-11e8-96f8-529269fb1459
resourceID	Identificador para o recurso que e o tema da relação	• f809b9e0-b09b-11e8-96f8-529269fb1459
relatedResource ID	Identificador para um recurso relacionado	• dc609808-b09b-11e8-96f8-529269fb1459
Relationship OfResource	Relação do recurso identificado em relatedResourceID ao sujeito, sendo recomendado a utilização de vocábulo controlado	• sameAs • duplicate of • mother of • endoparasite of • host to • sibling of • valid synonym of • located within
Relationship AccordingTo	Fonte que estabeleceu a relacionamento entre os dois recursos, esta fonte pode ser uma pessoa, organização ou referência publicada	• Julie Woodruff
Relationship EstablishedDate	Data e hora na qual foi estabelecida a relação entre os dois recursos, sendo recomendado a utilização de um esquema de codificação, como a ISO 8601:2004	• 1963-03-08T14:07-0600 • 2009-02-20T08:40Z • 2018-08-29T15:19

		• 1809-02-12 • 1906-06 • 2007-03-01T13:00:00Z/2008-05-11T15:30:00Z • 1900/1909 • 2007-11-13/15
Relationship	Comentários ou anotações sobre a relação entre os dois recursos	
Remarks		• mother and offspring collected from the same nest • pollinator captured in the act

FONTE: DwCMG (2020), Soares (2019) e Wieczorek et al. (2012)

APENDICE 12 - TABELA DA CLASSE USEWITHIRI

Elementos	Descrição	Exemplos
Language		•
inDescribedPlace	Usado para vincular uma instancia de	•
identifiedBy	Pessoa grupo ou organização que atribuiu o Taxon ao sujeito	•
recordedBy	Pessoa, grupo ou organização responsável por registrar a ocorrência original.	•
toTaxon	Use para vincular uma classe “Identification” sujeita a uma entidade taxonômica, como um táxon, conceito de táxon ou uso de nome de táxon. Uma "propriedade de conveniência" que substitui os termos de valor literal do Darwin Core relacionados a entidades taxonômicas.	•
inCollection	Use para vincular qualquer recurso de assunto que faça parte de uma coleção à coleção que contém o recurso. A melhor prática recomendada é usar um IRI de um registro controlado. Uma "propriedade de conveniência" que substitui termos de valor literal relacionados a coleções e instituições	•
georeferencedBy	pessoa, grupo ou organização que determinou a georreferência	•
Behavior	Descrição do comportamento mostrado pelo sujeito quando a ocorrência foi registrada, recomendado o uso de vocábulo controlado	•
dataGeneralizations	Ações tomadas para tornar os dados compartilhados menos específicos ou completos do que em sua forma original. Sugere que dados alternativos de qualidade superior possam estar disponíveis mediante solicitação	•
degreeOf Establishment		•

Disposition	Estado atual de uma amostra em relação à coleção identificada em collectionCode ou collectionID, sendo recomendado a utilização de vocábulo controlado.	•
Earliest	Use para vincular uma instância “GeologicalContext” a períodos cronoestratigráficos no nível	•
Geochronologica	mais baixo possível em uma hierarquia padronizada. Use essa propriedade para apontar para o período geológico possível mais antigo a partir do qual o item catalogado foi coletado	
IEra		
Establishment	processo pelo qual o (s) indivíduo (s) biológico (s) representado (s) na Ocorrência foram estabelecidos no local, e recomendado o uso de vocábulo controlado	•
Means		
fieldNotes	Texto das anotações feitas em campo sobre o evento do registro, podendo ser também um indicador de existência ou referencia	•
fielsNumber	Identificador fornecido ao evento no campo, frequentemente serve como um link entre as anotações de campo e o Evento.	•
footprintSRS	Uma representação de WKT do SRS para o footprintWKT do local. Não use esse termo para descrever o SRS de decimalLatitude e decimalLongitude, mesmo que seja o mesmo do footprintWKT - use o geodeticDatum. Os termos no namespace dwciri devem ser usados no RDF com objetos não literais	•
footprintWKT	Uma representação WKT da forma (pegada, geometria) que define o local. Um local pode ter uma representação de raio de ponto (consulte decimalLatitude) e uma representação de pegada, e elas podem diferir uma da outra. Os termos no namespace dwciri devem ser usados no RDF com objetos não literais	•

Fron	Use para vincular uma instância “GeologicalContext” a uma unidade litoestratigráfica	•
Lithostratigraphic	identificada pelo IRI no nível mais	
Unit	baixo possível em uma hierarquia, e recomendado o uso de vocábulo controlado	
GeodeticDatum	O elipsoide, dado geodésico ou SRS no qual as coordenadas geográficas são fornecidas em decimalLatitude e decimalLongitude como base. A melhor prática recomendada é usar um IRI para o código EPSG do SRS, se conhecido. Caso contrário, use um IRI ou vocabulário controlado para o nome ou código do dado geodésico, se conhecido. Caso contrário, use um IRI ou vocabulário controlado para o nome ou código do elipsoide, se conhecido. Se nada disso for conhecido, use o valor unknown Uma descrição ou referência aos métodos usados para determinar a pegada espacial, coordenadas e incertezas. Os termos no namespace dwciri devem ser usados no RDF com objetos não literais.	•
Georeference		
Protocol		
Georeference	Mapa, gazeta ou outro recurso usado para georreferenciar o local. Os termos no namespace	•
Sources	dwciri devem ser usados no RDF com objetos não literais.	
Georeferenceverification	Descrição categórica da extensão em que a georreferência foi verificada para representar a	•
Status	melhor descrição espacial	
	possível, recomendado a utilização de vocábulo controlado	
Habitat	categoria ou descrição do habitat em que o Evento ocorreu	•
Identification	Valor controlado para expressar as dúvidas do determinante sobre a identificação	•
Qualifier		

Identification	Indicador categórico da extensão em que a identificação	•
VerificationStatus	taxonômica foi verificada como correta, recomendado a utilização do vocábulo controlado	
inDataset	Use para vincular um registro de conjunto de dados de assunto ao conjunto de dados que o contém. Um nome literal de sequência de caracteres do conjunto de dados pode ser fornecido usando o termo “DatasetName”.	•
Information	Informações adicionais que existem, mas que não foram	•
Withheld	compartilhadas no registro fornecido.	
Latest	Use para vincular uma instância “GeologicalContext” a períodos	•
GeochronologicalEra	cronoestratigráficos no nível mais baixo possível em uma hierarquia padronizada, recomendado a utilização do vocábulo controlado	
lifeStage	classe etária ou estágio de vida do(s) indivíduo(s) biológico(s) quando a Ocorrência foi registrada, recomendado a utilização do vocábulo controlado	•
Location	Informações sobre a fonte dessas informações de localização.	•
AccordingTo	Pode ser uma publicação, instituição ou equipe de indivíduos	
Measurement	Pessoa, grupo ou organização que determinou o valor do	•
DeterminedBy	MeasurementOrFact.	
Measurement	Método ou protocolo usado para determinar a medida, fato,	•
Method	característica ou afirmação	
Measurement	natureza da medida, fato, característica ou afirmação, recomendado a utilização do	•
Type	vocábulo controlado	
measurementUnit	As unidades associadas com o measurementValue	•

OccurrenceStatus	declaração sobre a presença ou ausência de um táxon em um local, recomendado a utilização do vocábulo controlado	•
Organismo	tipo de sistema de quantificação usado para a quantidade de organismos	•
QuantityType		
pathway		•
preparations	método de preparação ou preservação de uma amostra.	•
recordNumber	Identificador fornecido à Ocorrência quando foi gravado. Frequentemente serve como um link entre as anotações de campo e um registro de Ocorrência, como o número de um coletor de amostras	•
Reproductive Condition	condição reprodutiva do(s) indivíduo(s) biológico(s) representado(s) na Ocorrência , recomendado a utilização do vocábulo controlado	•
Sample	unidade de medida do tamanho (duração, duração, área ou volume) de uma amostra em um evento de amostragem, recomendado a utilização do vocábulo controlado	•
SizeUnit		
Sampling Protocol	método ou protocolo usado durante um evento	•
Sex	sexo do(s) indivíduo(s) biológico(s) representado(s) na Ocorrência, recomendado a utilização do vocábulo controlado	•
typeStatus	Tipo nomenclatural (status do tipo, nome científico tipificado, publicação) aplicado ao assunto	•
Verbatim Coordinate System	sistema de coordenadas espaciais para verbatimLatitude e verbatimLongitude ou verbatimCoordinates do local, recomendado a utilização do vocábulo controlado	•

verbatimSRS	Elipsoide, o dado geodésico ou o SRS no qual as coordenadas fornecidas em <code>verbatimLatitude</code> e <code>verbatimLongitude</code>, ou <code>verbatimCoordinates</code> são baseadas. A melhor prática recomendada é usar um IRI para o código • EPSG do SRS, se conhecido. Caso contrário, use um IRI ou vocabulário controlado para o nome ou código do dado geodésico, se conhecido. Caso contrário, use um IRI ou vocabulário controlado para o nome ou código do elipsoide, se conhecido. Se nada disso for conhecido, use o valor <code>unknown</code>.
--------------------	--

FONTE:DwCMG (2020), Soares (2019) e Wieczorek et al. (2012)

APENDICE 13 – RESULTADOS DA ANÁLISE DE ASSOCIAÇÃO PELO APRIORI

Regras encontradas na análise de associação Apriori

1. disposition=Na Colecao *phylum*=Mollusca 23998 ==> type=Collection 23998 <conf:(1)> lift:(4.26) lev:(0.08) [18363] conv:(18363.19)
2. disposition=Na Colecao *phylum*=Mollusca *hasGeospatialIssues*=false 23789 ==> type=Collection 23789 <conf:(1)> lift:(4.26) lev:(0.08) [18203] conv:(18203.26)
3. preparations=Álcool 70 *phylum*=Mollusca 23764 ==> type=Collection 23764 <conf:(1)> lift:(4.26) lev:(0.08) [18184] conv:(18184.13)
4. preparations=Alcool 70 disposition=Na Colecao *phylum*=Mollusca 23629 ==> type=Collection 23629 <conf:(1)> lift:(4.26) lev:(0.08) [18080] conv:(18080.83)
5. preparations=Alcool 70 *phylum*=Mollusca *hasGeospatialIssues*=false 23555 ==> type=Collection 23555 <conf:(1)> lift:(4.26) lev:(0.08) [18024] conv:(18024.21)
6. preparations=Alcool 70 disposition=Na Colecao *phylum*=Mollusca *hasGeospatialIssues*=false 23420 ==> type=Collection 23420 <conf:(1)> lift:(4.26) lev:(0.08) [17920] conv:(17920.9)
7. *phylum*=Mollusca 24139 ==> type=Collection 24137 <conf:(1)> lift:(4.26) lev:(0.08) [18469] conv:(6157.03)
8. *phylum*=Mollusca *hasGeospatialIssues*=false 23930 ==> type=Collection 23928 <conf:(1)> lift:(4.26) lev:(0.08) [18309] conv:(6103.72)
9. type=Collection disposition=Na Colecao 24005 ==> *phylum*=Mollusca 23998 <conf:(1)> lift:(9.32) lev:(0.1) [21422] conv:(2678.71)
10. type=Collection disposition=Na Colecao *hasGeospatialIssues*=false 23796 ==> *phylum*=Mollusca 23789 <conf:(1)> lift:(9.32) lev:(0.09) [21236] conv:(2655.38)
11. type=Collection preparations=Alcool 70 23771 ==> *phylum*=Mollusca 23764 <conf:(1)> lift:(9.32) lev:(0.09) [21213] conv:(2652.59)
12. type=Collection preparations=Alcool 70 disposition=Na Colecao 23636 ==> *phylum*=Mollusca 23629 <conf:(1)> lift:(9.32) lev:(0.09) [21093] conv:(2637.53)
13. type=Collection preparations=Alcool 70 *hasGeospatialIssues*=false 23562 ==> *phylum*=Mollusca 23555 <conf:(1)> lift:(9.32) lev:(0.09) [21027] conv:(2629.27)
14. *phylum*=Chordata *hasCoordinate*=false 30150 ==> *hasGeospatialIssues*=false 30141 <conf:(1)> lift:(1.07) lev:(0.01) [1860] conv:(186.92)
15. type=Collection preparations=Alcool 70 disposition=Na Colecao *hasGeospatialIssues*=false 23427 ==> *phylum*=Mollusca 23420 <conf:(1)> lift:(9.32) lev:(0.09) [20906] conv:(2614.21)
16. *phylum*=Chordata *taxonomicStatus*=ACCEPTED *hasCoordinate*=false 27521 ==> *hasGeospatialIssues*=false 27512 <conf:(1)> lift:(1.07) lev:(0.01) [1697] conv:(170.62)

17. *phylum*=Chordata *taxonRank*=SPECIES *hasCoordinate*=false 23631 ==>
hasGeospatialIssues=false 23622 <conf:(1)> lift:(1.07) lev:(0.01) [1456] conv:(146.5)
18. *phylum*=Arthropoda *taxonRank*=FAMILY *hasCoordinate*=false 67438 ==>
taxonomicStatus=ACCEPTED 67357 <conf:(1)> lift:(1.12) lev:(0.03) [7175] conv:(88.5)
19. *phylum*=Arthropoda *taxonRank*=FAMILY *hasCoordinate*=false *hasGeospatialIssues*=false 67218
==> *taxonomicStatus*=ACCEPTED 67137 <conf:(1)> lift:(1.12) lev:(0.03) [7152] conv:(88.21)
20. *taxonRank*=FAMILY *hasCoordinate*=false 69935 ==> *taxonomicStatus*=ACCEPTED 69846
<conf:(1)> lift:(1.12) lev:(0.03) [7436] conv:(83.62)
21. *taxonRank*=FAMILY *hasCoordinate*=false *hasGeospatialIssues*=false 69713 ==>
taxonomicStatus=ACCEPTED 69624 <conf:(1)> lift:(1.12) lev:(0.03) [7412] conv:(83.35)
22. *phylum*=Arthropoda *taxonRank*=FAMILY *hasGeospatialIssues*=false 68217 ==>
taxonomicStatus=ACCEPTED 68128 <conf:(1)> lift:(1.12) lev:(0.03) [7251] conv:(81.56)
23. *type*=Collection *phylum*=Chordata 28639 ==> *hasGeospatialIssues*=false 28600 <conf:(1)>
lift:(1.06) lev:(0.01) [1736] conv:(44.39)
24. *taxonRank*=FAMILY *hasGeospatialIssues*=false 71808 ==> *taxonomicStatus*=ACCEPTED 71709
<conf:(1)> lift:(1.12) lev:(0.03) [7628] conv:(77.27)
25. *type*=Collection *phylum*=Chordata *taxonomicStatus*=ACCEPTED 26545 ==>
hasGeospatialIssues=false 26507 <conf:(1)> lift:(1.06) lev:(0.01) [1607] conv:(42.2)
26. *taxonRank*=FAMILY 75549 ==> *taxonomicStatus*=ACCEPTED 75331 <conf:(1)> lift:(1.12)
lev:(0.04) [7911] conv:(37.12)
27. *phylum*=Arthropoda *taxonRank*=FAMILY 71924 ==> *taxonomicStatus*=ACCEPTED 71716
<conf:(1)> lift:(1.12) lev:(0.03) [7531] conv:(37.03)
28. *taxonRank*=FAMILY *hasCoordinate*=false 69935 ==> *hasGeospatialIssues*=false 69713
<conf:(1)> lift:(1.06) lev:(0.02) [4113] conv:(19.44)
29. *taxonRank*=FAMILY *taxonomicStatus*=ACCEPTED *hasCoordinate*=false 69846 ==>
hasGeospatialIssues=false 69624 <conf:(1)> lift:(1.06) lev:(0.02) [4108] conv:(19.42)
30. *phylum*=Arthropoda *taxonRank*=FAMILY *hasCoordinate*=false 67438 ==>
hasGeospatialIssues=false 67218 <conf:(1)> lift:(1.06) lev:(0.02) [3960] conv:(18.92)
31. *phylum*=Arthropoda *taxonRank*=FAMILY *taxonomicStatus*=ACCEPTED *hasCoordinate*=false
67357 ==> *hasGeospatialIssues*=false 67137 <conf:(1)> lift:(1.06) lev:(0.02) [3955] conv:(18.9)
32. *taxonRank*=FAMILY *hasCoordinate*=false 69935 ==> *taxonomicStatus*=ACCEPTED
hasGeospatialIssues=false 69624 <conf:(1)> lift:(1.19) lev:(0.05) [11280] conv:(37.15)
33. *phylum*=Arthropoda *taxonRank*=FAMILY *hasCoordinate*=false 67438 ==>
taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 67137 <conf:(1)> lift:(1.19) lev:(0.05)
[10876] conv:(37.01)
34. *type*=Collection *taxonRank*=SPECIES *taxonomicStatus*=ACCEPTED 36026 ==>
hasGeospatialIssues=false 35865 <conf:(1)> lift:(1.06) lev:(0.01) [2072] conv:(13.79)

35. type=Collection taxonRank=SPECIES 39891 ==> *hasGeospatialIssues*=false 39710 <conf:(1)> lift:(1.06) lev:(0.01) [2292] conv:(13.59)
36. type=Collection taxonomicStatus=ACCEPTED 48580 ==> *hasGeospatialIssues*=false 48355 <conf:(1)> lift:(1.06) lev:(0.01) [2786] conv:(13.33)
37. type=Collection 52831 ==> *hasGeospatialIssues*=false 52583 <conf:(1)> lift:(1.06) lev:(0.01) [3027] conv:(13.15)
38. type=Collection preparations=Alcool 70 23771 ==> disposition=Na Colecao 23636 <conf:(0.99)> lift:(8.31) lev:(0.09) [20791] conv:(153.87)
39. preparations=Alcool 70 *phylum*=Mollusca 23764 ==> disposition=Na Colecao 23629 <conf:(0.99)> lift:(8.31) lev:(0.09) [20785] conv:(153.82)
40. type=Collection preparations=Alcool 70 *phylum*=Mollusca 23764 ==> disposition=Na Colecao 23629 <conf:(0.99)> lift:(8.31) lev:(0.09) [20785] conv:(153.82)
41. preparations=Alcool 70 *phylum*=Mollusca 23764 ==> type=Collection disposition=Na Colecao 23629 <conf:(0.99)> lift:(9.32) lev:(0.09) [21093] conv:(156.09)
42. type=Collection preparations=Alcool 70 *hasGeospatialIssues*=false 23562 ==> disposition=Na Colecao 23427 <conf:(0.99)> lift:(8.31) lev:(0.09) [20607] conv:(152.52)
43. preparations=Alcool 70 *phylum*=Mollusca *hasGeospatialIssues*=false 23555 ==> disposition=Na Colecao 23420 <conf:(0.99)> lift:(8.31) lev:(0.09) [20601] conv:(152.47)
44. type=Collection preparations=Alcool 70 *phylum*=Mollusca *hasGeospatialIssues*=false 23555 ==> disposition=Na Colecao 23420 <conf:(0.99)> lift:(8.31) lev:(0.09) [20601] conv:(152.47)
45. preparations=Alcool 70 *phylum*=Mollusca *hasGeospatialIssues*=false 23555 ==> type=Collection disposition=Na Colecao 23420 <conf:(0.99)> lift:(9.32) lev:(0.09) [20906] conv:(154.72)
46. type=Collection *phylum*=Mollusca 24137 ==> disposition=Na Colecao 23998 <conf:(0.99)> lift:(8.31) lev:(0.09) [21109] conv:(151.77)
47. type=Collection *phylum*=Mollusca *hasGeospatialIssues*=false 23928 ==> disposition=Na Colecao 23789 <conf:(0.99)> lift:(8.31) lev:(0.09) [20925] conv:(150.46)
48. *phylum*=Mollusca 24139 ==> disposition=Na Colecao 23998 <conf:(0.99)> lift:(8.31) lev:(0.09) [21109] conv:(149.65)
49. *phylum*=Mollusca 24139 ==> type=Collection disposition=Na Colecao 23998 <conf:(0.99)> lift:(9.32) lev:(0.1) [21422] conv:(151.86)
50. *phylum*=Mollusca *hasGeospatialIssues*=false 23930 ==> disposition=Na Colecao 23789 <conf:(0.99)> lift:(8.31) lev:(0.09) [20925] conv:(148.35)
51. *phylum*=Mollusca *hasGeospatialIssues*=false 23930 ==> type=Collection disposition=Na Colecao 23789 <conf:(0.99)> lift:(9.32) lev:(0.09) [21235] conv:(150.54)
52. type=Collection preparations=Alcool 70 23771 ==> disposition=Na Colecao *phylum*=Mollusca 23629 <conf:(0.99)> lift:(9.32) lev:(0.09) [21093] conv:(148.5)
53. type=Collection preparations=Alcool 70 *hasGeospatialIssues*=false 23562 ==> disposition=Na Colecao *phylum*=Mollusca 23420 <conf:(0.99)> lift:(9.32) lev:(0.09) [20906] conv:(147.2)

54. type=Collection taxonRank=SPECIES hasCoordinate=true 24127 ==> *hasGeospatialIssues*=false 23951 <conf:(0.99)> lift:(1.06) lev:(0.01) [1319] conv:(8.45)
55. type=Collection taxonomicStatus=ACCEPTED hasCoordinate=true 28841 ==> *hasGeospatialIssues*=false 28621 <conf:(0.99)> lift:(1.06) lev:(0.01) [1568] conv:(8.09)
56. type=Collection hasCoordinate=true 30908 ==> *hasGeospatialIssues*=false 30667 <conf:(0.99)> lift:(1.06) lev:(0.01) [1675] conv:(7.92)
57. *phylum*=Mollusca 24139 ==> *hasGeospatialIssues*=false 23930 <conf:(0.99)> lift:(1.06) lev:(0.01) [1287] conv:(7.13)
58. type=Collection *phylum*=Mollusca 24137 ==> *hasGeospatialIssues*=false 23928 <conf:(0.99)> lift:(1.06) lev:(0.01) [1287] conv:(7.13)
59. type=Collection disposition=Na Colecao 24005 ==> *hasGeospatialIssues*=false 23796 <conf:(0.99)> lift:(1.06) lev:(0.01) [1279] conv:(7.09)
60. disposition=Na Colecao *phylum*=Mollusca 23998 ==> *hasGeospatialIssues*=false 23789 <conf:(0.99)> lift:(1.06) lev:(0.01) [1278] conv:(7.08)
61. type=Collection disposition=Na Colecao *phylum*=Mollusca 23998 ==> *hasGeospatialIssues*=false 23789 <conf:(0.99)> lift:(1.06) lev:(0.01) [1278] conv:(7.08)
62. disposition=Na Colecao *phylum*=Mollusca 23998 ==> type=Collection *hasGeospatialIssues*=false 23789 <conf:(0.99)> lift:(4.24) lev:(0.08) [18180] conv:(87.57)
63. *phylum*=Mollusca 24139 ==> type=Collection *hasGeospatialIssues*=false 23928 <conf:(0.99)> lift:(4.24) lev:(0.08) [18286] conv:(87.25)
64. type=Collection preparations=Alcool 70 23771 ==> *hasGeospatialIssues*=false 23562 <conf:(0.99)> lift:(1.06) lev:(0.01) [1264] conv:(7.02)
65. preparations=Alcool 70 *phylum*=Mollusca 23764 ==> *hasGeospatialIssues*=false 23555 <conf:(0.99)> lift:(1.06) lev:(0.01) [1264] conv:(7.02)
66. type=Collection preparations=Alcool 70 *phylum*=Mollusca 23764 ==> *hasGeospatialIssues*=false 23555 <conf:(0.99)> lift:(1.06) lev:(0.01) [1264] conv:(7.02)
67. preparations=Alcool 70 *phylum*=Mollusca 23764 ==> type=Collection *hasGeospatialIssues*=false 23555 <conf:(0.99)> lift:(4.24) lev:(0.08) [18001] conv:(86.72)
68. type=Collection preparations=Alcool 70 disposition=Na Colecao 23636 ==> *hasGeospatialIssues*=false 23427 <conf:(0.99)> lift:(1.06) lev:(0.01) [1256] conv:(6.98)
69. preparations=Alcool 70 disposition=Na Colecao *phylum*=Mollusca 23629 ==> *hasGeospatialIssues*=false 23420 <conf:(0.99)> lift:(1.06) lev:(0.01) [1255] conv:(6.98)
70. type=Collection preparations=Alcool 70 disposition=Na Colecao *phylum*=Mollusca 23629 ==> *hasGeospatialIssues*=false 23420 <conf:(0.99)> lift:(1.06) lev:(0.01) [1255] conv:(6.98)
71. preparations=Alcool 70 disposition=Na Colecao *phylum*=Mollusca 23629 ==> type=Collection *hasGeospatialIssues*=false 23420 <conf:(0.99)> lift:(4.24) lev:(0.08) [17897] conv:(86.22)
72. type=Collection disposition=Na Colecao 24005 ==> *phylum*=Mollusca *hasGeospatialIssues*=false 23789 <conf:(0.99)> lift:(9.32) lev:(0.09) [21235] conv:(98.86)

73. type=Collection preparations=Alcool 70 23771 ==> *phylum*=Mollusca *hasGeospatialIssues*=false 23555 <conf:(0.99)> lift:(9.32) lev:(0.09) [21026] conv:(97.89)
74. type=Collection preparations=Alcool 70 disposition=Na Colecao 23636 ==> *phylum*=Mollusca *hasGeospatialIssues*=false 23420 <conf:(0.99)> lift:(9.32) lev:(0.09) [20906] conv:(97.34)
75. disposition=Na Colecao taxonomicStatus=ACCEPTED 24616 ==> *hasGeospatialIssues*=false 24384 <conf:(0.99)> lift:(1.06) lev:(0.01) [1294] conv:(6.55)
76. disposition=Na Colecao 26927 ==> *hasGeospatialIssues*=false 26673 <conf:(0.99)> lift:(1.06) lev:(0.01) [1415] conv:(6.55)
77. preparations=Alcool 70 disposition=Na Colecao taxonomicStatus=ACCEPTED 24071 ==> *hasGeospatialIssues*=false 23841 <conf:(0.99)> lift:(1.06) lev:(0.01) [1262] conv:(6.46)
78. preparations=Alcool 70 disposition=Na Colecao 26339 ==> *hasGeospatialIssues*=false 26087 <conf:(0.99)> lift:(1.06) lev:(0.01) [1380] conv:(6.45)
79. *phylum*=Chordata 57290 ==> *hasGeospatialIssues*=false 56652 <conf:(0.99)> lift:(1.05) lev:(0.01) [2913] conv:(5.56)
80. *phylum*=Chordata taxonomicStatus=ACCEPTED 53504 ==> *hasGeospatialIssues*=false 52902 <conf:(0.99)> lift:(1.05) lev:(0.01) [2714] conv:(5.5)
81. *stateProvince*=RJ *hasCoordinate*=false 41004 ==> *hasGeospatialIssues*=false 40486 <conf:(0.99)> lift:(1.05) lev:(0.01) [2024] conv:(4.9)
82. *phylum*=Chordata taxonRank=SPECIES 45798 ==> *hasGeospatialIssues*=false 45214 <conf:(0.99)> lift:(1.05) lev:(0.01) [2255] conv:(4.85)
83. *phylum*=Chordata taxonRank=SPECIES taxonomicStatus=ACCEPTED 42358 ==> *hasGeospatialIssues*=false 41810 <conf:(0.99)> lift:(1.05) lev:(0.01) [2077] conv:(4.78)
84. *hasCoordinate*=false 153570 ==> *hasGeospatialIssues*=false 151581 <conf:(0.99)> lift:(1.05) lev:(0.03) [7531] conv:(4.78)
85. taxonomicStatus=ACCEPTED *hasCoordinate*=false 133466 ==> *hasGeospatialIssues*=false 131640 <conf:(0.99)> lift:(1.05) lev:(0.03) [6448] conv:(4.53)
86. *stateProvince*=RJ taxonomicStatus=ACCEPTED *hasCoordinate*=false 35268 ==> *hasGeospatialIssues*=false 34773 <conf:(0.99)> lift:(1.05) lev:(0.01) [1691] conv:(4.41)
87. type=Collection *phylum*=Mollusca 24137 ==> disposition=Na Colecao *hasGeospatialIssues*=false 23789 <conf:(0.99)> lift:(8.31) lev:(0.09) [20927] conv:(60.96)
88. type=Collection preparations=Alcool 70 23771 ==> disposition=Na Colecao *hasGeospatialIssues*=false 23427 <conf:(0.99)> lift:(8.31) lev:(0.09) [20609] conv:(60.73)
89. preparations=Alcool 70 *phylum*=Mollusca 23764 ==> disposition=Na Colecao *hasGeospatialIssues*=false 23420 <conf:(0.99)> lift:(8.31) lev:(0.09) [20602] conv:(60.72)
90. type=Collection preparations=Alcool 70 *phylum*=Mollusca 23764 ==> disposition=Na Colecao *hasGeospatialIssues*=false 23420 <conf:(0.99)> lift:(8.31) lev:(0.09) [20602] conv:(60.72)
91. preparations=Alcool 70 *phylum*=Mollusca 23764 ==> type=Collection disposition=Na Colecao *hasGeospatialIssues*=false 23420 <conf:(0.99)> lift:(9.32) lev:(0.09) [20906] conv:(61.6)

92. *phylum*=Mollusca 24139 ==> disposition=Na Colecao *hasGeospatialIssues*=false 23789
<conf:(0.99)> lift:(8.31) lev:(0.09) [20927] conv:(60.62)
93. *phylum*=Mollusca 24139 ==> type=Collection disposition=Na Colecao *hasGeospatialIssues*=false 23789 <conf:(0.99)> lift:(9.32) lev:(0.09) [21236] conv:(61.5)
94. *phylum*=Arthropoda taxonRank=FAMILY taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 68128 ==> hasCoordinate=false 67137 <conf:(0.99)> lift:(1.44) lev:(0.09) [20637] conv:(21.8)
95. *phylum*=Arthropoda taxonRank=FAMILY *hasGeospatialIssues*=false 68217 ==> hasCoordinate=false 67218 <conf:(0.99)> lift:(1.44) lev:(0.09) [20657] conv:(21.66)
96. type=Collection preparations=Alcool 70 23771 ==> disposition=Na Colecao *phylum*=Mollusca *hasGeospatialIssues*=false 23420 <conf:(0.99)> lift:(9.32) lev:(0.09) [20906] conv:(60.39)
97. type=Collection disposition=Na Colecao 24005 ==> preparations=Alcool 70 23636
<conf:(0.98)> lift:(4.23) lev:(0.08) [18048] conv:(49.78)
98. disposition=Na Colecao *phylum*=Mollusca 23998 ==> preparations=Alcool 70 23629
<conf:(0.98)> lift:(4.23) lev:(0.08) [18042] conv:(49.76)
99. type=Collection disposition=Na Colecao *phylum*=Mollusca 23998 ==> preparations=Alcool 70 23629 <conf:(0.98)> lift:(4.23) lev:(0.08) [18042] conv:(49.76)
100. disposition=Na Colecao *phylum*=Mollusca 23998 ==> type=Collection preparations=Alcool 70 23629 <conf:(0.98)> lift:(9.32) lev:(0.09) [21093] conv:(58.01)
101. type=Collection *phylum*=Mollusca 24137 ==> preparations=Alcool 70 23764 <conf:(0.98)> lift:(4.23) lev:(0.08) [18145] conv:(49.51)
102. type=Collection disposition=Na Colecao *hasGeospatialIssues*=false 23796 ==> preparations=Alcool 70 23427 <conf:(0.98)> lift:(4.23) lev:(0.08) [17887] conv:(49.34)
103. disposition=Na Colecao *phylum*=Mollusca *hasGeospatialIssues*=false 23789 ==> preparations=Alcool 70 23420 <conf:(0.98)> lift:(4.23) lev:(0.08) [17882] conv:(49.33)
104. type=Collection disposition=Na Colecao *phylum*=Mollusca *hasGeospatialIssues*=false 23789 ==> preparations=Alcool 70 23420 <conf:(0.98)> lift:(4.23) lev:(0.08) [17882] conv:(49.33)
105. disposition=Na Colecao *phylum*=Mollusca *hasGeospatialIssues*=false 23789 ==> type=Collection preparations=Alcool 70 23420 <conf:(0.98)> lift:(9.32) lev:(0.09) [20906] conv:(57.5)
106. *phylum*=Mollusca 24139 ==> preparations=Alcool 70 23764 <conf:(0.98)> lift:(4.23) lev:(0.08) [18145] conv:(49.26)
107. *phylum*=Mollusca 24139 ==> type=Collection preparations=Alcool 70 23764 <conf:(0.98)> lift:(9.32) lev:(0.09) [21213] conv:(57.42)
108. type=Collection *phylum*=Mollusca *hasGeospatialIssues*=false 23928 ==> preparations=Alcool 70 23555 <conf:(0.98)> lift:(4.23) lev:(0.08) [17985] conv:(49.09)
109. type=Collection disposition=Na Colecao 24005 ==> preparations=Alcool 70 *phylum*=Mollusca 23629 <conf:(0.98)> lift:(9.32) lev:(0.09) [21093] conv:(56.95)
110. *phylum*=Mollusca *hasGeospatialIssues*=false 23930 ==> preparations=Alcool 70 23555
<conf:(0.98)> lift:(4.23) lev:(0.08) [17984] conv:(48.83)

111. *phylum*=Mollusca *hasGeospatialIssues*=false 23930 ==> type=Collection preparations=Alcool 70 23555 <conf:(0.98)> lift:(9.32) lev:(0.09) [21026] conv:(56.92)
112. type=Collection disposition=Na Colecao *hasGeospatialIssues*=false 23796 ==> preparations=Alcool 70 *phylum*=Mollusca 23420 <conf:(0.98)> lift:(9.32) lev:(0.09) [20906] conv:(56.45)
113. *phylum*=Arthropoda taxonRank=FAMILY *hasGeospatialIssues*=false 68217 ==> taxonomicStatus=ACCEPTED hasCoordinate=false 67137 <conf:(0.98)> lift:(1.66) lev:(0.12) [26672] conv:(25.67)
114. *stateProvince*=RJ *phylum*=Arthropoda hasCoordinate=false 29520 ==> *hasGeospatialIssues*=false 29012 <conf:(0.98)> lift:(1.05) lev:(0.01) [1322] conv:(3.6)
115. *phylum*=Arthropoda hasCoordinate=false 113097 ==> *hasGeospatialIssues*=false 111138 <conf:(0.98)> lift:(1.05) lev:(0.02) [5052] conv:(3.58)
116. *phylum*=Arthropoda taxonomicStatus=ACCEPTED hasCoordinate=false 96382 ==> *hasGeospatialIssues*=false 94584 <conf:(0.98)> lift:(1.05) lev:(0.02) [4177] conv:(3.32)
117. *stateProvince*=RJ *phylum*=Arthropoda taxonomicStatus=ACCEPTED hasCoordinate=false 24835 ==> *hasGeospatialIssues*=false 24350 <conf:(0.98)> lift:(1.05) lev:(0) [1054] conv:(3.17)
118. type=PhysicalObject *phylum*=Chordata 28651 ==> *hasGeospatialIssues*=false 28052 <conf:(0.98)> lift:(1.04) lev:(0.01) [1177] conv:(2.96)
119. type=PhysicalObject *phylum*=Chordata taxonomicStatus=ACCEPTED 26959 ==> *hasGeospatialIssues*=false 26395 <conf:(0.98)> lift:(1.04) lev:(0) [1107] conv:(2.96)
120. type=Collection *phylum*=Mollusca 24137 ==> preparations=Alcool 70 disposition=Na Colecao 23629 <conf:(0.98)> lift:(8.36) lev:(0.09) [20803] conv:(41.87)
121. *phylum*=Mollusca 24139 ==> preparations=Alcool 70 disposition=Na Colecao 23629 <conf:(0.98)> lift:(8.36) lev:(0.09) [20803] conv:(41.71)
122. *phylum*=Mollusca 24139 ==> type=Collection preparations=Alcool 70 disposition=Na Colecao 23629 <conf:(0.98)> lift:(9.32) lev:(0.09) [21093] conv:(42.28)
123. type=Collection *phylum*=Mollusca *hasGeospatialIssues*=false 23928 ==> preparations=Alcool 70 disposition=Na Colecao 23420 <conf:(0.98)> lift:(8.36) lev:(0.09) [20618] conv:(41.51)
124. *phylum*=Mollusca *hasGeospatialIssues*=false 23930 ==> preparations=Alcool 70 disposition=Na Colecao 23420 <conf:(0.98)> lift:(8.36) lev:(0.09) [20618] conv:(41.35)
125. *phylum*=Mollusca *hasGeospatialIssues*=false 23930 ==> type=Collection preparations=Alcool 70 disposition=Na Colecao 23420 <conf:(0.98)> lift:(9.32) lev:(0.09) [20906] conv:(41.91)
126. type=PhysicalObject *phylum*=Chordata taxonRank=SPECIES taxonomicStatus=ACCEPTED 24440 ==> *hasGeospatialIssues*=false 23908 <conf:(0.98)> lift:(1.04) lev:(0) [983] conv:(2.84)
127. type=PhysicalObject *phylum*=Chordata taxonRank=SPECIES 25967 ==> *hasGeospatialIssues*=false 25400 <conf:(0.98)> lift:(1.04) lev:(0) [1042] conv:(2.83)
128. disposition=Na Colecao 26927 ==> preparations=Alcool 70 26339 <conf:(0.98)> lift:(4.2) lev:(0.09) [20071] conv:(35.07)

129. disposition=Na Colecao *hasGeospatialIssues*=false 26673 ==> preparations=Alcool 70 26087
<conf:(0.98)> lift:(4.2) lev:(0.09) [19878] conv:(34.86)
130. disposition=Na Colecao taxonomicStatus=ACCEPTED 24616 ==> preparations=Alcool 70 24071
<conf:(0.98)> lift:(4.2) lev:(0.08) [18341] conv:(34.59)
131. disposition=Na Colecao taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 24384 ==> preparations=Alcool 70 23841 <conf:(0.98)> lift:(4.2) lev:(0.08) [18165] conv:(34.39)
132. *phylum*=Chordata taxonomicStatus=ACCEPTED hasCoordinate=true 25983 ==> *hasGeospatialIssues*=false 25390 <conf:(0.98)> lift:(1.04) lev:(0) [1017] conv:(2.71)
133. *phylum*=Chordata hasCoordinate=true 27140 ==> *hasGeospatialIssues*=false 26511
<conf:(0.98)> lift:(1.04) lev:(0) [1053] conv:(2.67)
134. type=Collection disposition=Na Colecao 24005 ==> preparations=Alcool 70
hasGeospatialIssues=false 23427 <conf:(0.98)> lift:(5.35) lev:(0.08) [19044] conv:(33.89)
135. disposition=Na Colecao *phylum*=Mollusca 23998 ==> preparations=Alcool 70
hasGeospatialIssues=false 23420 <conf:(0.98)> lift:(5.35) lev:(0.08) [19038] conv:(33.88)
136. type=Collection disposition=Na Colecao *phylum*=Mollusca 23998 ==> preparations=Alcool 70
hasGeospatialIssues=false 23420 <conf:(0.98)> lift:(5.35) lev:(0.08) [19038] conv:(33.88)
137. disposition=Na Colecao *phylum*=Mollusca 23998 ==> type=Collection preparations=Alcool 70
hasGeospatialIssues=false 23420 <conf:(0.98)> lift:(9.32) lev:(0.09) [20906] conv:(37.11)
138. type=Collection *phylum*=Mollusca 24137 ==> preparations=Alcool 70 *hasGeospatialIssues*=false 23555
<conf:(0.98)> lift:(5.35) lev:(0.09) [19148] conv:(33.84)
139. *phylum*=Mollusca 24139 ==> preparations=Alcool 70 *hasGeospatialIssues*=false 23555
<conf:(0.98)> lift:(5.34) lev:(0.09) [19147] conv:(33.73)
140. *phylum*=Mollusca 24139 ==> type=Collection preparations=Alcool 70 *hasGeospatialIssues*=false 23555
<conf:(0.98)> lift:(9.32) lev:(0.09) [21027] conv:(36.94)
141. type=Collection disposition=Na Colecao 24005 ==> preparations=Alcool 70 *phylum*=Mollusca
hasGeospatialIssues=false 23420 <conf:(0.98)> lift:(9.32) lev:(0.09) [20906] conv:(36.68)
142. taxonRank=SPECIES hasCoordinate=false 43168 ==> *hasGeospatialIssues*=false 42032
<conf:(0.97)> lift:(1.04) lev:(0.01) [1540] conv:(2.35)
143. taxonRank=SPECIES taxonomicStatus=ACCEPTED hasCoordinate=false 38326 ==> *hasGeospatialIssues*=false 37259
<conf:(0.97)> lift:(1.04) lev:(0.01) [1309] conv:(2.22)
144. taxonRank=FAMILY taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 71709 ==> hasCoordinate=false 69624
<conf:(0.97)> lift:(1.42) lev:(0.09) [20680] conv:(10.91)
145. taxonRank=FAMILY *hasGeospatialIssues*=false 71808 ==> hasCoordinate=false 69713
<conf:(0.97)> lift:(1.42) lev:(0.09) [20701] conv:(10.88)
146. type=Collection *phylum*=Mollusca 24137 ==> preparations=Alcool 70 disposition=Na Colecao
hasGeospatialIssues=false 23420 <conf:(0.97)> lift:(8.37) lev:(0.09) [20621] conv:(29.72)
147. *phylum*=Mollusca 24139 ==> preparations=Alcool 70 disposition=Na Colecao
hasGeospatialIssues=false 23420 <conf:(0.97)> lift:(8.37) lev:(0.09) [20621] conv:(29.64)

148. *phylum*=Mollusca 24139 ==> *type*=Collection *preparations*=Alcool 70 *disposition*=Na Colecao *hasGeospatialIssues*=false 23420 <conf:(0.97)> lift:(9.32) lev:(0.09) [20906] conv:(30.04)
149. *taxonRank*=FAMILY *hasGeospatialIssues*=false 71808 ==> *taxonomicStatus*=ACCEPTED *hasCoordinate*=false 69624 <conf:(0.97)> lift:(1.63) lev:(0.12) [27028] conv:(13.37)
150. *disposition*=Na Colecao 26927 ==> *preparations*=Alcool 70 *hasGeospatialIssues*=false 26087 <conf:(0.97)> lift:(5.31) lev:(0.09) [21170] conv:(26.17)
151. *disposition*=Na Colecao *taxonomicStatus*=ACCEPTED 24616 ==> *preparations*=Alcool 70 *hasGeospatialIssues*=false 23841 <conf:(0.97)> lift:(5.3) lev:(0.09) [19346] conv:(25.93)
152. *taxonRank*=FAMILY *taxonomicStatus*=ACCEPTED *hasCoordinate*=false 69846 ==> *phylum*=Arthropoda 67357 <conf:(0.96)> lift:(1.6) lev:(0.11) [25244] conv:(11.14)
153. *taxonRank*=FAMILY *hasCoordinate*=false 69935 ==> *phylum*=Arthropoda 67438 <conf:(0.96)> lift:(1.6) lev:(0.11) [25272] conv:(11.12)
154. *taxonRank*=FAMILY *taxonomicStatus*=ACCEPTED *hasCoordinate*=false *hasGeospatialIssues*=false 69624 ==> *phylum*=Arthropoda 67137 <conf:(0.96)> lift:(1.6) lev:(0.11) [25158] conv:(11.11)
155. *taxonRank*=FAMILY *hasCoordinate*=false *hasGeospatialIssues*=false 69713 ==> *phylum*=Arthropoda 67218 <conf:(0.96)> lift:(1.6) lev:(0.11) [25185] conv:(11.09)
156. *taxonRank*=FAMILY *hasCoordinate*=false 69935 ==> *phylum*=Arthropoda *taxonomicStatus*=ACCEPTED 67357 <conf:(0.96)> lift:(1.84) lev:(0.14) [30776] conv:(12.93)
157. *taxonRank*=FAMILY *hasCoordinate*=false *hasGeospatialIssues*=false 69713 ==> *phylum*=Arthropoda *taxonomicStatus*=ACCEPTED 67137 <conf:(0.96)> lift:(1.84) lev:(0.14) [30672] conv:(12.9)
158. *taxonRank*=FAMILY *taxonomicStatus*=ACCEPTED *hasCoordinate*=false 69846 ==> *phylum*=Arthropoda *hasGeospatialIssues*=false 67137 <conf:(0.96)> lift:(1.76) lev:(0.13) [29054] conv:(11.72)
159. *taxonRank*=FAMILY *hasCoordinate*=false 69935 ==> *phylum*=Arthropoda *hasGeospatialIssues*=false 67218 <conf:(0.96)> lift:(1.76) lev:(0.13) [29086] conv:(11.7)
160. *taxonRank*=FAMILY *hasCoordinate*=false 69935 ==> *phylum*=Arthropoda *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 67137 <conf:(0.96)> lift:(2.05) lev:(0.15) [34342] conv:(13.27)
161. *phylum*=Chordata *hasCoordinate*=true *hasGeospatialIssues*=false 26511 ==> *taxonomicStatus*=ACCEPTED 25390 <conf:(0.96)> lift:(1.07) lev:(0.01) [1731] conv:(2.54)
162. *phylum*=Chordata *hasCoordinate*=true 27140 ==> *taxonomicStatus*=ACCEPTED 25983 <conf:(0.96)> lift:(1.07) lev:(0.01) [1763] conv:(2.52)
163. *taxonRank*=FAMILY 75549 ==> *phylum*=Arthropoda 71924 <conf:(0.95)> lift:(1.58) lev:(0.12) [26373] conv:(8.27)
164. *taxonRank*=FAMILY *taxonomicStatus*=ACCEPTED 75331 ==> *phylum*=Arthropoda 71716 <conf:(0.95)> lift:(1.58) lev:(0.12) [26296] conv:(8.27)

165. taxonRank=FAMILY taxonomicStatus=ACCEPTED 75331 ==> *hasGeospatialIssues*=false 71709
<conf:(0.95)> lift:(1.01) lev:(0) [1048] conv:(1.29)
166. type=PhysicalObject hasCoordinate=true *hasGeospatialIssues*=false 26478 ==>
taxonomicStatus=ACCEPTED 25171 <conf:(0.95)> lift:(1.07) lev:(0.01) [1542] conv:(2.18)
167. taxonRank=FAMILY 75549 ==> *hasGeospatialIssues*=false 71808 <conf:(0.95)> lift:(1.01)
lev:(0) [942] conv:(1.25)
168. taxonRank=FAMILY taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 71709 ==>
phylum=Arthropoda 68128 <conf:(0.95)> lift:(1.58) lev:(0.11) [24892] conv:(7.95)
169. taxonRank=FAMILY *hasGeospatialIssues*=false 71808 ==> *phylum*=Arthropoda 68217
<conf:(0.95)> lift:(1.58) lev:(0.11) [24921] conv:(7.94)
170. *phylum*=Arthropoda taxonRank=FAMILY taxonomicStatus=ACCEPTED 71716 ==>
hasGeospatialIssues=false 68128 <conf:(0.95)> lift:(1.01) lev:(0) [858] conv:(1.24)
171. taxonRank=FAMILY 75549 ==> *phylum*=Arthropoda taxonomicStatus=ACCEPTED 71716
<conf:(0.95)> lift:(1.81) lev:(0.14) [32199] conv:(9.4)
172. taxonRank=FAMILY 75549 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 71709
<conf:(0.95)> lift:(1.14) lev:(0.04) [8682] conv:(3.26)
173. taxonRank=FAMILY *hasGeospatialIssues*=false 71808 ==> *phylum*=Arthropoda
taxonomicStatus=ACCEPTED 68128 <conf:(0.95)> lift:(1.81) lev:(0.14) [30567] conv:(9.3)
174. *phylum*=Arthropoda taxonRank=FAMILY 71924 ==> *hasGeospatialIssues*=false 68217
<conf:(0.95)> lift:(1.01) lev:(0) [751] conv:(1.2)
175. type=PhysicalObject hasCoordinate=true 38124 ==> taxonomicStatus=ACCEPTED 36148
<conf:(0.95)> lift:(1.06) lev:(0.01) [2126] conv:(2.08)
176. *phylum*=Arthropoda taxonRank=FAMILY 71924 ==> taxonomicStatus=ACCEPTED
hasGeospatialIssues=false 68128 <conf:(0.95)> lift:(1.14) lev:(0.04) [8125] conv:(3.14)
177. type=PhysicalObject preparations=Alcool 70 25994 ==> taxonomicStatus=ACCEPTED 24517
<conf:(0.94)> lift:(1.06) lev:(0.01) [1320] conv:(1.89)
178. hasCoordinate=true *hasGeospatialIssues*=false 59471 ==> taxonomicStatus=ACCEPTED 56067
<conf:(0.94)> lift:(1.06) lev:(0.01) [2995] conv:(1.88)
179. type=PhysicalObject *hasGeospatialIssues*=false 43924 ==> taxonomicStatus=ACCEPTED 41405
<conf:(0.94)> lift:(1.06) lev:(0.01) [2207] conv:(1.88)
180. type=PhysicalObject 55597 ==> taxonomicStatus=ACCEPTED 52408 <conf:(0.94)> lift:(1.06)
lev:(0.01) [2793] conv:(1.88)
181. hasCoordinate=true 71431 ==> taxonomicStatus=ACCEPTED 67323 <conf:(0.94)> lift:(1.06)
lev:(0.02) [3578] conv:(1.87)
182. type=PhysicalObject preparations=Alcool 70 hasCoordinate=true 24275 ==>
taxonomicStatus=ACCEPTED 22872 <conf:(0.94)> lift:(1.06) lev:(0.01) [1209] conv:(1.86)
183. type=PhysicalObject *phylum*=Chordata taxonRank=SPECIES *hasGeospatialIssues*=false 25400
==> taxonomicStatus=ACCEPTED 23908 <conf:(0.94)> lift:(1.05) lev:(0.01) [1241] conv:(1.83)

184. type=PhysicalObject *phylum*=Chordata taxonRank=SPECIES 25967 ==> taxonomicStatus=ACCEPTED 24440 <conf:(0.94)> lift:(1.05) lev:(0.01) [1267] conv:(1.83)
185. type=PhysicalObject *phylum*=Chordata 28651 ==> taxonomicStatus=ACCEPTED 26959 <conf:(0.94)> lift:(1.05) lev:(0.01) [1391] conv:(1.82)
186. type=PhysicalObject *phylum*=Chordata *hasGeospatialIssues*=false 28052 ==> taxonomicStatus=ACCEPTED 26395 <conf:(0.94)> lift:(1.05) lev:(0.01) [1361] conv:(1.82)
187. *stateProvince*=RJ taxonRank=SPECIES taxonomicStatus=ACCEPTED 28029 ==> *hasGeospatialIssues*=false 26329 <conf:(0.94)> lift:(1) lev:(0) [37] conv:(1.02)
188. *phylum*=Arthropoda taxonRank=FAMILY taxonomicStatus=ACCEPTED 71716 ==> hasCoordinate=false 67357 <conf:(0.94)> lift:(1.38) lev:(0.08) [18408] conv:(5.22)
189. *stateProvince*=RJ taxonRank=SPECIES 31263 ==> *hasGeospatialIssues*=false 29348 <conf:(0.94)> lift:(1) lev:(0) [23] conv:(1.01)
190. taxonRank=SPECIES taxonomicStatus=ACCEPTED 85664 ==> *hasGeospatialIssues*=false 80405 <conf:(0.94)> lift:(1) lev:(0) [51] conv:(1.01)
191. taxonRank=SPECIES 94136 ==> *hasGeospatialIssues*=false 88292 <conf:(0.94)> lift:(1) lev:(-0) [-8] conv:(1)
192. *phylum*=Arthropoda taxonRank=FAMILY 71924 ==> hasCoordinate=false 67438 <conf:(0.94)> lift:(1.37) lev:(0.08) [18347] conv:(5.09)
193. *phylum*=Arthropoda taxonRank=FAMILY 71924 ==> taxonomicStatus=ACCEPTED hasCoordinate=false 67357 <conf:(0.94)> lift:(1.58) lev:(0.11) [24693] conv:(6.41)
194. taxonRank=FAMILY taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 71709 ==> *phylum*=Arthropoda hasCoordinate=false 67137 <conf:(0.94)> lift:(1.86) lev:(0.14) [31092] conv:(7.8)
195. *phylum*=Arthropoda taxonRank=FAMILY taxonomicStatus=ACCEPTED 71716 ==> hasCoordinate=false *hasGeospatialIssues*=false 67137 <conf:(0.94)> lift:(1.39) lev:(0.08) [18822] conv:(5.11)
196. taxonRank=FAMILY *hasGeospatialIssues*=false 71808 ==> *phylum*=Arthropoda hasCoordinate=false 67218 <conf:(0.94)> lift:(1.86) lev:(0.14) [31123] conv:(7.78)
197. type=PhysicalObject taxonRank=SPECIES *hasGeospatialIssues*=false 34773 ==> taxonomicStatus=ACCEPTED 32544 <conf:(0.94)> lift:(1.05) lev:(0.01) [1512] conv:(1.68)
198. *phylum*=Chordata hasCoordinate=true 27140 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 25390 <conf:(0.94)> lift:(1.12) lev:(0.01) [2748] conv:(2.57)
199. taxonRank=FAMILY *hasGeospatialIssues*=false 71808 ==> *phylum*=Arthropoda taxonomicStatus=ACCEPTED hasCoordinate=false 67137 <conf:(0.93)> lift:(2.18) lev:(0.16) [36377] conv:(8.79)
200. taxonomicStatus=ACCEPTED 200789 ==> *hasGeospatialIssues*=false 187707 <conf:(0.93)> lift:(1) lev:(-0) [-634] conv:(0.95)
201. *phylum*=Arthropoda taxonRank=FAMILY 71924 ==> hasCoordinate=false *hasGeospatialIssues*=false 67218 <conf:(0.93)> lift:(1.39) lev:(0.08) [18763] conv:(4.99)

202. *phylum*=Chordata 57290 ==> taxonomicStatus=ACCEPTED 53504 <conf:(0.93)> lift:(1.05) lev:(0.01) [2378] conv:(1.63)
203. type=PhysicalObject preparations=Alcool 70 25994 ==> hasCoordinate=true 24275 <conf:(0.93)> lift:(2.94) lev:(0.07) [16022] conv:(10.31)
204. *phylum*=Chordata *hasGeospatialIssues*=false 56652 ==> taxonomicStatus=ACCEPTED 52902 <conf:(0.93)> lift:(1.05) lev:(0.01) [2346] conv:(1.63)
205. *phylum*=Arthropoda taxonRank=FAMILY 71924 ==> taxonomicStatus=ACCEPTED hasCoordinate=false *hasGeospatialIssues*=false 67137 <conf:(0.93)> lift:(1.6) lev:(0.11) [25056] conv:(6.23)
206. type=Collection hasCoordinate=true *hasGeospatialIssues*=false 30667 ==> taxonomicStatus=ACCEPTED 28621 <conf:(0.93)> lift:(1.05) lev:(0.01) [1254] conv:(1.61)
207. type=Collection hasCoordinate=true 30908 ==> taxonomicStatus=ACCEPTED 28841 <conf:(0.93)> lift:(1.05) lev:(0.01) [1258] conv:(1.61)
208. type=PhysicalObject preparations=Alcool 70 taxonomicStatus=ACCEPTED 24517 ==> hasCoordinate=true 22872 <conf:(0.93)> lift:(2.94) lev:(0.07) [15088] conv:(10.17)
209. type=PhysicalObject taxonRank=SPECIES hasCoordinate=true 24684 ==> taxonomicStatus=ACCEPTED 23027 <conf:(0.93)> lift:(1.05) lev:(0) [999] conv:(1.6)
210. taxonRank=SPECIES hasCoordinate=true *hasGeospatialIssues*=false 46260 ==> taxonomicStatus=ACCEPTED 43146 <conf:(0.93)> lift:(1.05) lev:(0.01) [1863] conv:(1.6)
211. preparations=Alcool 70 hasCoordinate=true 42661 ==> taxonomicStatus=ACCEPTED 39781 <conf:(0.93)> lift:(1.04) lev:(0.01) [1710] conv:(1.59)
212. *stateProvince*=RJ hasCoordinate=true 25705 ==> taxonomicStatus=ACCEPTED 23943 <conf:(0.93)> lift:(1.04) lev:(0) [1004] conv:(1.57)
213. preparations=Alcool 70 52374 ==> taxonomicStatus=ACCEPTED 48751 <conf:(0.93)> lift:(1.04) lev:(0.01) [2012] conv:(1.56)
214. type=PhysicalObject taxonRank=SPECIES 39269 ==> taxonomicStatus=ACCEPTED 36547 <conf:(0.93)> lift:(1.04) lev:(0.01) [1503] conv:(1.55)
215. preparations=Alcool 70 hasCoordinate=true *hasGeospatialIssues*=false 31394 ==> taxonomicStatus=ACCEPTED 29164 <conf:(0.93)> lift:(1.04) lev:(0.01) [1148] conv:(1.51)
216. taxonRank=SPECIES hasCoordinate=true 50968 ==> taxonomicStatus=ACCEPTED 47338 <conf:(0.93)> lift:(1.04) lev:(0.01) [1854] conv:(1.51)
217. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> taxonomicStatus=ACCEPTED 38109 <conf:(0.93)> lift:(1.04) lev:(0.01) [1450] conv:(1.49)
218. taxonRank=FAMILY taxonomicStatus=ACCEPTED 75331 ==> hasCoordinate=false 69846 <conf:(0.93)> lift:(1.36) lev:(0.08) [18430] conv:(4.36)
219. type=Collection *phylum*=Chordata 28639 ==> taxonomicStatus=ACCEPTED 26545 <conf:(0.93)> lift:(1.04) lev:(0) [987] conv:(1.47)

220. type=Collection *phylum*=Chordata *hasGeospatialIssues*=false 28600 ==> taxonomicStatus=ACCEPTED 26507 <conf:(0.93)> lift:(1.04) lev:(0) [984] conv:(1.47)
221. type=Collection hasCoordinate=true 30908 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 28621 <conf:(0.93)> lift:(1.11) lev:(0.01) [2836] conv:(2.24)
222. taxonRank=FAMILY 75549 ==> hasCoordinate=false 69935 <conf:(0.93)> lift:(1.36) lev:(0.08) [18370] conv:(4.27)
223. type=Collection *phylum*=Chordata 28639 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 26507 <conf:(0.93)> lift:(1.11) lev:(0.01) [2614] conv:(2.23)
224. taxonRank=GENUS 32471 ==> *hasGeospatialIssues*=false 30046 <conf:(0.93)> lift:(0.99) lev:(-0) [-411] conv:(0.83)
225. *phylum*=Chordata taxonRank=SPECIES 45798 ==> taxonomicStatus=ACCEPTED 42358 <conf:(0.92)> lift:(1.04) lev:(0.01) [1488] conv:(1.43)
226. *phylum*=Chordata taxonRank=SPECIES *hasGeospatialIssues*=false 45214 ==> taxonomicStatus=ACCEPTED 41810 <conf:(0.92)> lift:(1.04) lev:(0.01) [1461] conv:(1.43)
227. taxonRank=FAMILY 75549 ==> taxonomicStatus=ACCEPTED hasCoordinate=false 69846 <conf:(0.92)> lift:(1.56) lev:(0.11) [25031] conv:(5.39)
228. taxonRank=FAMILY taxonomicStatus=ACCEPTED 75331 ==> hasCoordinate=false *hasGeospatialIssues*=false 69624 <conf:(0.92)> lift:(1.37) lev:(0.08) [18874] conv:(4.31)
229. taxonRank=GENUS taxonomicStatus=ACCEPTED 29947 ==> *hasGeospatialIssues*=false 27678 <conf:(0.92)> lift:(0.99) lev:(-0) [-412] conv:(0.82)
230. *phylum*=Chordata 57290 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 52902 <conf:(0.92)> lift:(1.11) lev:(0.02) [5107] conv:(2.16)
231. taxonRank=FAMILY 75549 ==> hasCoordinate=false *hasGeospatialIssues*=false 69713 <conf:(0.92)> lift:(1.37) lev:(0.08) [18816] conv:(4.22)
232. taxonRank=GENUS 32471 ==> taxonomicStatus=ACCEPTED 29947 <conf:(0.92)> lift:(1.03) lev:(0) [970] conv:(1.38)
233. *stateProvince*=RJ 66709 ==> *hasGeospatialIssues*=false 61516 <conf:(0.92)> lift:(0.98) lev:(-0) [-1057] conv:(0.8)
234. taxonRank=FAMILY 75549 ==> taxonomicStatus=ACCEPTED hasCoordinate=false *hasGeospatialIssues*=false 69624 <conf:(0.92)> lift:(1.58) lev:(0.11) [25422] conv:(5.29)
235. type=PhysicalObject *phylum*=Chordata 28651 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 26395 <conf:(0.92)> lift:(1.1) lev:(0.01) [2492] conv:(2.1)
236. taxonRank=GENUS *hasGeospatialIssues*=false 30046 ==> taxonomicStatus=ACCEPTED 27678 <conf:(0.92)> lift:(1.03) lev:(0) [865] conv:(1.36)
237. type=PhysicalObject *phylum*=Chordata taxonRank=SPECIES 25967 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 23908 <conf:(0.92)> lift:(1.1) lev:(0.01) [2245] conv:(2.09)

238. type=Collection *hasGeospatialIssues*=false 52583 ==> taxonomicStatus=ACCEPTED 48355
<conf:(0.92)> lift:(1.03) lev:(0.01) [1430] conv:(1.34)

239. type=Collection 52831 ==> taxonomicStatus=ACCEPTED 48580 <conf:(0.92)> lift:(1.03)
lev:(0.01) [1434] conv:(1.34)

240. stateProvince=RJ taxonomicStatus=ACCEPTED 59211 ==> *hasGeospatialIssues*=false 54335
<conf:(0.92)> lift:(0.98) lev:(-0.01) [-1205] conv:(0.75)

241. type=Collection 52831 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 48355
<conf:(0.92)> lift:(1.1) lev:(0.02) [4280] conv:(1.96)

242. disposition=Na Colecao *hasGeospatialIssues*=false 26673 ==> taxonomicStatus=ACCEPTED
24384 <conf:(0.91)> lift:(1.02) lev:(0) [581] conv:(1.25)

243. disposition=Na Colecao 26927 ==> taxonomicStatus=ACCEPTED 24616 <conf:(0.91)>
lift:(1.02) lev:(0) [586] conv:(1.25)

244. preparations=Alcool 70 taxonRank=SPECIES *hasGeospatialIssues*=false 30987 ==>
taxonomicStatus=ACCEPTED 28323 <conf:(0.91)> lift:(1.02) lev:(0) [670] conv:(1.25)

245. preparations=Alcool 70 disposition=Na Colecao *hasGeospatialIssues*=false 26087 ==>
taxonomicStatus=ACCEPTED 23841 <conf:(0.91)> lift:(1.02) lev:(0) [561] conv:(1.25)

246. preparations=Alcool 70 disposition=Na Colecao 26339 ==> taxonomicStatus=ACCEPTED 24071
<conf:(0.91)> lift:(1.02) lev:(0) [566] conv:(1.25)

247. phylum=Chordata taxonRank=SPECIES 45798 ==> taxonomicStatus=ACCEPTED
hasGeospatialIssues=false 41810 <conf:(0.91)> lift:(1.09) lev:(0.02) [3603] conv:(1.9)

248. phylum=Chordata *hasCoordinate*=false 30150 ==> taxonomicStatus=ACCEPTED 27521
<conf:(0.91)> lift:(1.02) lev:(0) [615] conv:(1.23)

249. phylum=Chordata *hasCoordinate*=false *hasGeospatialIssues*=false 30141 ==>
taxonomicStatus=ACCEPTED 27512 <conf:(0.91)> lift:(1.02) lev:(0) [614] conv:(1.23)

250. phylum=Chordata *hasCoordinate*=false 30150 ==> taxonomicStatus=ACCEPTED
hasGeospatialIssues=false 27512 <conf:(0.91)> lift:(1.09) lev:(0.01) [2359] conv:(1.89)

251. taxonRank=SPECIES taxonomicStatus=ACCEPTED *hasCoordinate*=true 47338 ==>
hasGeospatialIssues=false 43146 <conf:(0.91)> lift:(0.97) lev:(-0.01) [-1257] conv:(0.7)

252. taxonRank=SPECIES *hasGeospatialIssues*=false 88292 ==> taxonomicStatus=ACCEPTED 80405
<conf:(0.91)> lift:(1.02) lev:(0.01) [1613] conv:(1.2)

253. preparations=Alcool 70 taxonRank=SPECIES 35096 ==> taxonomicStatus=ACCEPTED 31956
<conf:(0.91)> lift:(1.02) lev:(0) [636] conv:(1.2)

254. preparations=Alcool 70 taxonRank=SPECIES *hasCoordinate*=true 27919 ==>
taxonomicStatus=ACCEPTED 25417 <conf:(0.91)> lift:(1.02) lev:(0) [502] conv:(1.2)

255. taxonRank=SPECIES 94136 ==> taxonomicStatus=ACCEPTED 85664 <conf:(0.91)> lift:(1.02)
lev:(0.01) [1657] conv:(1.2)

256. taxonRank=SPECIES *hasCoordinate*=true 50968 ==> *hasGeospatialIssues*=false 46260
<conf:(0.91)> lift:(0.97) lev:(-0.01) [-1548] conv:(0.67)

257. type=PhysicalObject *phylum*=Chordata taxonomicStatus=ACCEPTED 26959 ==> taxonRank=SPECIES 24440 <conf:(0.91)> lift:(2.17) lev:(0.06) [13160] conv:(6.22)
258. type=PhysicalObject *phylum*=Chordata 28651 ==> taxonRank=SPECIES 25967 <conf:(0.91)> lift:(2.17) lev:(0.06) [13979] conv:(6.21)
259. *phylum*=Arthropoda *hasGeospatialIssues*=false 122680 ==> hasCoordinate=false 111138 <conf:(0.91)> lift:(1.33) lev:(0.12) [27405] conv:(3.37)
260. type=PhysicalObject *phylum*=Chordata taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 26395 ==> taxonRank=SPECIES 23908 <conf:(0.91)> lift:(2.16) lev:(0.06) [12864] conv:(6.17)
261. disposition=Na Colecao 26927 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 24384 <conf:(0.91)> lift:(1.09) lev:(0.01) [1920] conv:(1.75)
262. type=PhysicalObject *phylum*=Chordata *hasGeospatialIssues*=false 28052 ==> taxonRank=SPECIES 25400 <conf:(0.91)> lift:(2.16) lev:(0.06) [13663] conv:(6.15)
263. preparations=Alcool 70 disposition=Na Colecao 26339 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 23841 <conf:(0.91)> lift:(1.08) lev:(0.01) [1867] conv:(1.75)
264. taxonRank=FAMILY taxonomicStatus=ACCEPTED 75331 ==> *phylum*=Arthropoda *hasGeospatialIssues*=false 68128 <conf:(0.9)> lift:(1.66) lev:(0.12) [27054] conv:(4.76)
265. *phylum*=Arthropoda 135660 ==> *hasGeospatialIssues*=false 122680 <conf:(0.9)> lift:(0.96) lev:(-0.02) [-4569] conv:(0.65)
266. type=Collection taxonRank=SPECIES *hasGeospatialIssues*=false 39710 ==> taxonomicStatus=ACCEPTED 35865 <conf:(0.9)> lift:(1.01) lev:(0) [428] conv:(1.11)
267. type=Collection taxonRank=SPECIES 39891 ==> taxonomicStatus=ACCEPTED 36026 <conf:(0.9)> lift:(1.01) lev:(0) [427] conv:(1.11)
268. taxonRank=FAMILY 75549 ==> *phylum*=Arthropoda *hasGeospatialIssues*=false 68217 <conf:(0.9)> lift:(1.66) lev:(0.12) [27024] conv:(4.69)
269. taxonRank=FAMILY 75549 ==> *phylum*=Arthropoda taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 68128 <conf:(0.9)> lift:(1.92) lev:(0.15) [32701] conv:(5.41)
270. type=Collection taxonRank=SPECIES 39891 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 35865 <conf:(0.9)> lift:(1.08) lev:(0.01) [2585] conv:(1.64)
271. preparations=Alcool 70 disposition=Na Colecao *hasGeospatialIssues*=false 26087 ==> type=Collection 23427 <conf:(0.9)> lift:(3.82) lev:(0.08) [17301] conv:(7.5)
272. preparations=Alcool 70 disposition=Na Colecao *hasGeospatialIssues*=false 26087 ==> *phylum*=Mollusca 23420 <conf:(0.9)> lift:(8.37) lev:(0.09) [20621] conv:(8.73)
273. preparations=Alcool 70 disposition=Na Colecao *hasGeospatialIssues*=false 26087 ==> type=Collection *phylum*=Mollusca 23420 <conf:(0.9)> lift:(8.37) lev:(0.09) [20621] conv:(8.73)
274. preparations=Alcool 70 disposition=Na Colecao 26339 ==> type=Collection 23636 <conf:(0.9)> lift:(3.82) lev:(0.08) [17451] conv:(7.45)
275. *stateProvince*=RJ taxonRank=SPECIES *hasGeospatialIssues*=false 29348 ==> taxonomicStatus=ACCEPTED 26329 <conf:(0.9)> lift:(1.01) lev:(0) [139] conv:(1.05)

276. preparations=Alcool 70 disposition=Na Colecao 26339 ==> *phylum*=Mollusca 23629
<conf:(0.9)> lift:(8.36) lev:(0.09) [20803] conv:(8.67)
277. preparations=Alcool 70 disposition=Na Colecao 26339 ==> type=Collection *phylum*=Mollusca 23629
<conf:(0.9)> lift:(8.36) lev:(0.09) [20803] conv:(8.67)
278. *stateProvince*=RJ taxonRank=SPECIES 31263 ==> taxonomicStatus=ACCEPTED 28029
<conf:(0.9)> lift:(1) lev:(0) [130] conv:(1.04)
279. *phylum*=Arthropoda taxonomicStatus=ACCEPTED 117690 ==> *hasGeospatialIssues*=false 105509
<conf:(0.9)> lift:(0.96) lev:(-0.02) [-4884] conv:(0.6)
280. *phylum*=Arthropoda taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 105509 ==>
hasCoordinate=false 94584 <conf:(0.9)> lift:(1.31) lev:(0.1) [22570] conv:(3.07)
281. taxonRank=FAMILY taxonomicStatus=ACCEPTED 75331 ==> *phylum*=Arthropoda
hasCoordinate=false 67357 <conf:(0.89)> lift:(1.78) lev:(0.13) [29491] conv:(4.7)
282. disposition=Na Colecao 26927 ==> preparations=Alcool 70 taxonomicStatus=ACCEPTED 24071
<conf:(0.89)> lift:(4.13) lev:(0.08) [18236] conv:(7.38)
283. disposition=Na Colecao *hasGeospatialIssues*=false 26673 ==> preparations=Alcool 70
taxonomicStatus=ACCEPTED 23841 <conf:(0.89)> lift:(4.13) lev:(0.08) [18061] conv:(7.38)
284. taxonRank=FAMILY 75549 ==> *phylum*=Arthropoda *hasCoordinate*=false 67438 <conf:(0.89)>
lift:(1.78) lev:(0.13) [29463] conv:(4.63)
285. disposition=Na Colecao *hasGeospatialIssues*=false 26673 ==> type=Collection 23796
<conf:(0.89)> lift:(3.8) lev:(0.08) [17533] conv:(7.09)
286. disposition=Na Colecao *hasGeospatialIssues*=false 26673 ==> *phylum*=Mollusca 23789
<conf:(0.89)> lift:(8.31) lev:(0.09) [20927] conv:(8.25)
287. disposition=Na Colecao *hasGeospatialIssues*=false 26673 ==> type=Collection *phylum*=Mollusca 23789
<conf:(0.89)> lift:(8.31) lev:(0.09) [20927] conv:(8.25)
288. taxonRank=FAMILY 75549 ==> *phylum*=Arthropoda taxonomicStatus=ACCEPTED
hasCoordinate=false 67357 <conf:(0.89)> lift:(2.08) lev:(0.16) [34994] conv:(5.27)
289. disposition=Na Colecao 26927 ==> type=Collection 24005 <conf:(0.89)> lift:(3.8) lev:(0.08)
[17682] conv:(7.05)
290. taxonRank=FAMILY taxonomicStatus=ACCEPTED 75331 ==> *phylum*=Arthropoda
hasCoordinate=false *hasGeospatialIssues*=false 67137 <conf:(0.89)> lift:(1.8) lev:(0.13) [29927]
conv:(4.65)
291. disposition=Na Colecao 26927 ==> *phylum*=Mollusca 23998 <conf:(0.89)> lift:(8.31) lev:(0.09)
[21109] conv:(8.2)
292. disposition=Na Colecao 26927 ==> type=Collection *phylum*=Mollusca 23998 <conf:(0.89)>
lift:(8.31) lev:(0.09) [21109] conv:(8.2)
293. type=PhysicalObject taxonRank=SPECIES taxonomicStatus=ACCEPTED 36547 ==>
hasGeospatialIssues=false 32544 <conf:(0.89)> lift:(0.95) lev:(-0.01) [-1737] conv:(0.57)

294. *taxonRank*=FAMILY 75549 ==> *phylum*=Arthropoda *hasCoordinate*=false
hasGeospatialIssues=false 67218 <conf:(0.89)> lift:(1.8) lev:(0.13) [29900] conv:(4.59)
295. *preparations*=Alcool 70 *disposition*=Na Colecao 26339 ==> *type*=Collection
hasGeospatialIssues=false 23427 <conf:(0.89)> lift:(3.81) lev:(0.08) [17271] conv:(6.93)
296. *hasGeospatialIssues*=false 211052 ==> *taxonomicStatus*=ACCEPTED 187707 <conf:(0.89)>
lift:(1) lev:(-0) [-634] conv:(0.97)
297. *preparations*=Alcool 70 *disposition*=Na Colecao 26339 ==> *phylum*=Mollusca
hasGeospatialIssues=false 23420 <conf:(0.89)> lift:(8.36) lev:(0.09) [20618] conv:(8.06)
298. *preparations*=Alcool 70 *disposition*=Na Colecao 26339 ==> *type*=Collection *phylum*=Mollusca
hasGeospatialIssues=false 23420 <conf:(0.89)> lift:(8.36) lev:(0.09) [20618] conv:(8.06)
299. *taxonRank*=FAMILY 75549 ==> *phylum*=Arthropoda *taxonomicStatus*=ACCEPTED
hasCoordinate=false *hasGeospatialIssues*=false 67137 <conf:(0.89)> lift:(2.11) lev:(0.16) [35378]
conv:(5.21)
300. *taxonRank*=SPECIES *hasCoordinate*=false 43168 ==> *taxonomicStatus*=ACCEPTED 38326
<conf:(0.89)> lift:(0.99) lev:(-0) [-196] conv:(0.96)
301. *stateProvince*=RJ 66709 ==> *taxonomicStatus*=ACCEPTED 59211 <conf:(0.89)> lift:(0.99) lev:(-0)
[-319] conv:(0.96)
302. *type*=PhysicalObject *phylum*=Chordata *taxonomicStatus*=ACCEPTED 26959 ==>
taxonRank=SPECIES *hasGeospatialIssues*=false 23908 <conf:(0.89)> lift:(2.26) lev:(0.06) [13329]
conv:(5.37)
303. *type*=PhysicalObject *phylum*=Chordata 28651 ==> *taxonRank*=SPECIES
hasGeospatialIssues=false 25400 <conf:(0.89)> lift:(2.26) lev:(0.06) [14157] conv:(5.35)
304. *taxonRank*=SPECIES *hasCoordinate*=false *hasGeospatialIssues*=false 42032 ==>
taxonomicStatus=ACCEPTED 37259 <conf:(0.89)> lift:(0.99) lev:(-0) [-250] conv:(0.95)
305. *preparations*=Alcool 70 *taxonRank*=SPECIES *taxonomicStatus*=ACCEPTED 31956 ==>
hasGeospatialIssues=false 28323 <conf:(0.89)> lift:(0.94) lev:(-0.01) [-1651] conv:(0.55)
306. *type*=PhysicalObject *taxonRank*=SPECIES 39269 ==> *hasGeospatialIssues*=false 34773
<conf:(0.89)> lift:(0.94) lev:(-0.01) [-2061] conv:(0.54)
307. *disposition*=Na Colecao 26927 ==> *preparations*=Alcool 70 *taxonomicStatus*=ACCEPTED
hasGeospatialIssues=false 23841 <conf:(0.89)> lift:(5.23) lev:(0.09) [19280] conv:(7.25)
308. *disposition*=Na Colecao 26927 ==> *type*=Collection *hasGeospatialIssues*=false 23796
<conf:(0.88)> lift:(3.78) lev:(0.08) [17503] conv:(6.59)
309. *disposition*=Na Colecao 26927 ==> *phylum*=Mollusca *hasGeospatialIssues*=false 23789
<conf:(0.88)> lift:(8.31) lev:(0.09) [20925] conv:(7.67)
310. *disposition*=Na Colecao 26927 ==> *type*=Collection *phylum*=Mollusca *hasGeospatialIssues*=false
23789 <conf:(0.88)> lift:(8.31) lev:(0.09) [20925] conv:(7.67)
311. *stateProvince*=RJ *hasGeospatialIssues*=false 61516 ==> *taxonomicStatus*=ACCEPTED 54335
<conf:(0.88)> lift:(0.99) lev:(-0) [-561] conv:(0.92)

312. preparations=Alcool 70 taxonRank=SPECIES 35096 ==> *hasGeospatialIssues*=false 30987
<conf:(0.88)> lift:(0.94) lev:(-0.01) [-1933] conv:(0.53)
313. type=PhysicalObject preparations=Alcool 70 25994 ==> taxonomicStatus=ACCEPTED
hasCoordinate=true 22872 <conf:(0.88)> lift:(2.94) lev:(0.07) [15094] conv:(5.83)
314. disposition=Na Colecao *hasGeospatialIssues*=false 26673 ==> type=Collection
preparations=Alcool 70 23427 <conf:(0.88)> lift:(8.31) lev:(0.09) [20609] conv:(7.35)
315. disposition=Na Colecao *hasGeospatialIssues*=false 26673 ==> preparations=Alcool 70
phylum=Mollusca 23420 <conf:(0.88)> lift:(8.31) lev:(0.09) [20602] conv:(7.33)
316. disposition=Na Colecao *hasGeospatialIssues*=false 26673 ==> type=Collection
preparations=Alcool 70 *phylum*=Mollusca 23420 <conf:(0.88)> lift:(8.31) lev:(0.09) [20602]
conv:(7.33)
317. disposition=Na Colecao 26927 ==> type=Collection preparations=Alcool 70 23636
<conf:(0.88)> lift:(8.31) lev:(0.09) [20791] conv:(7.32)
318. disposition=Na Colecao 26927 ==> preparations=Alcool 70 *phylum*=Mollusca 23629
<conf:(0.88)> lift:(8.31) lev:(0.09) [20785] conv:(7.3)
319. disposition=Na Colecao 26927 ==> type=Collection preparations=Alcool 70 *phylum*=Mollusca
23629 <conf:(0.88)> lift:(8.31) lev:(0.09) [20785] conv:(7.3)
320. disposition=Na Colecao 26927 ==> type=Collection preparations=Alcool 70
hasGeospatialIssues=false 23427 <conf:(0.87)> lift:(8.31) lev:(0.09) [20607] conv:(6.89)
321. disposition=Na Colecao 26927 ==> preparations=Alcool 70 *phylum*=Mollusca
hasGeospatialIssues=false 23420 <conf:(0.87)> lift:(8.31) lev:(0.09) [20601] conv:(6.87)
322. disposition=Na Colecao 26927 ==> type=Collection preparations=Alcool 70 *phylum*=Mollusca
hasGeospatialIssues=false 23420 <conf:(0.87)> lift:(8.31) lev:(0.09) [20601] conv:(6.87)
323. hasCoordinate=false 153570 ==> taxonomicStatus=ACCEPTED 133466 <conf:(0.87)> lift:(0.97)
lev:(-0.02) [-3578] conv:(0.82)
324. hasCoordinate=false *hasGeospatialIssues*=false 151581 ==> taxonomicStatus=ACCEPTED
131640 <conf:(0.87)> lift:(0.97) lev:(-0.02) [-3629] conv:(0.82)
325. *stateProvince*=RJ *phylum*=Arthropoda 38657 ==> *hasGeospatialIssues*=false 33543
<conf:(0.87)> lift:(0.93) lev:(-0.01) [-2717] conv:(0.47)
326. *phylum*=Arthropoda 135660 ==> taxonomicStatus=ACCEPTED 117690 <conf:(0.87)> lift:(0.97)
lev:(-0.01) [-3371] conv:(0.81)
327. *stateProvince*=RJ *phylum*=Arthropoda *hasGeospatialIssues*=false 33543 ==>
hasCoordinate=false 29012 <conf:(0.86)> lift:(1.27) lev:(0.03) [6117] conv:(2.35)
328. *stateProvince*=RJ *phylum*=Arthropoda 38657 ==> taxonomicStatus=ACCEPTED 33419
<conf:(0.86)> lift:(0.97) lev:(-0) [-1078] conv:(0.79)
329. taxonRank=SPECIES hasCoordinate=false 43168 ==> taxonomicStatus=ACCEPTED
hasGeospatialIssues=false 37259 <conf:(0.86)> lift:(1.03) lev:(0.01) [1246] conv:(1.21)

330. *stateProvince*=RJ *hasCoordinate*=false 41004 ==> *taxonomicStatus*=ACCEPTED 35268
<conf:(0.86)> lift:(0.96) lev:(-0.01) [-1323] conv:(0.77)
331. *phylum*=Arthropoda *hasGeospatialIssues*=false 122680 ==> *taxonomicStatus*=ACCEPTED 105509 <conf:(0.86)> lift:(0.96) lev:(-0.02) [-3969] conv:(0.77)
332. *stateProvince*=RJ *hasCoordinate*=false *hasGeospatialIssues*=false 40486 ==> *taxonomicStatus*=ACCEPTED 34773 <conf:(0.86)> lift:(0.96) lev:(-0.01) [-1356] conv:(0.76)
333. *hasCoordinate*=false 153570 ==> *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 131640 <conf:(0.86)> lift:(1.03) lev:(0.02) [3524] conv:(1.16)
334. *stateProvince*=RJ *phylum*=Arthropoda *taxonomicStatus*=ACCEPTED 33419 ==> *hasGeospatialIssues*=false 28619 <conf:(0.86)> lift:(0.91) lev:(-0.01) [-2728] conv:(0.43)
335. *taxonRank*=SPECIES 94136 ==> *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 80405 <conf:(0.85)> lift:(1.02) lev:(0.01) [1872] conv:(1.14)
336. *preparations*=Alcool 70 *taxonRank*=SPECIES *hasCoordinate*=true 27919 ==> *hasGeospatialIssues*=false 23822 <conf:(0.85)> lift:(0.91) lev:(-0.01) [-2366] conv:(0.42)
337. *stateProvince*=RJ *phylum*=Arthropoda *hasGeospatialIssues*=false 33543 ==> *taxonomicStatus*=ACCEPTED 28619 <conf:(0.85)> lift:(0.96) lev:(-0.01) [-1314] conv:(0.73)
338. *type*=PhysicalObject *phylum*=Chordata 28651 ==> *taxonRank*=SPECIES *taxonomicStatus*=ACCEPTED 24440 <conf:(0.85)> lift:(2.24) lev:(0.06) [13531] conv:(4.21)
339. *taxonRank*=GENUS 32471 ==> *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 27678 <conf:(0.85)> lift:(1.02) lev:(0) [589] conv:(1.12)
340. *type*=PhysicalObject *phylum*=Chordata *hasGeospatialIssues*=false 28052 ==> *taxonRank*=SPECIES *taxonomicStatus*=ACCEPTED 23908 <conf:(0.85)> lift:(2.24) lev:(0.06) [13227] conv:(4.19)
341. *phylum*=Arthropoda *hasCoordinate*=false 113097 ==> *taxonomicStatus*=ACCEPTED 96382 <conf:(0.85)> lift:(0.95) lev:(-0.02) [-4544] conv:(0.73)
342. *phylum*=Arthropoda *hasCoordinate*=false *hasGeospatialIssues*=false 111138 ==> *taxonomicStatus*=ACCEPTED 94584 <conf:(0.85)> lift:(0.95) lev:(-0.02) [-4594] conv:(0.72)
343. *stateProvince*=RJ *phylum*=Arthropoda *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 28619 ==> *hasCoordinate*=false 24350 <conf:(0.85)> lift:(1.25) lev:(0.02) [4816] conv:(2.13)
344. *stateProvince*=RJ *hasCoordinate*=false 41004 ==> *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 34773 <conf:(0.85)> lift:(1.02) lev:(0) [565] conv:(1.09)
345. *taxonRank*=SPECIES *hasCoordinate*=true 50968 ==> *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 43146 <conf:(0.85)> lift:(1.01) lev:(0) [625] conv:(1.08)
346. *stateProvince*=RJ *taxonRank*=SPECIES 31263 ==> *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 26329 <conf:(0.84)> lift:(1.01) lev:(0) [247] conv:(1.05)
347. *stateProvince*=RJ *phylum*=Arthropoda *hasCoordinate*=false 29520 ==> *taxonomicStatus*=ACCEPTED 24835 <conf:(0.84)> lift:(0.94) lev:(-0.01) [-1508] conv:(0.68)

348. *stateProvince*=RJ *phylum*=Arthropoda *hasCoordinate*=false *hasGeospatialIssues*=false 29012
==> *taxonomicStatus*=ACCEPTED 24350 <conf:(0.84)> lift:(0.94) lev:(-0.01) [-1540] conv:(0.67)
349. *phylum*=Arthropoda *hasCoordinate*=false 113097 ==> *taxonomicStatus*=ACCEPTED
hasGeospatialIssues=false 94584 <conf:(0.84)> lift:(1) lev:(0) [232] conv:(1.01)
350. *type*=PhysicalObject *phylum*=Chordata 28651 ==> *taxonRank*=SPECIES
taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 23908 <conf:(0.83)> lift:(2.34) lev:(0.06)
[13669] conv:(3.88)
351. *phylum*=Arthropoda 135660 ==> *hasCoordinate*=false 113097 <conf:(0.83)> lift:(1.22)
lev:(0.09) [20504] conv:(1.91)
352. *taxonomicStatus*=ACCEPTED *hasCoordinate*=true 67323 ==> *hasGeospatialIssues*=false 56067
<conf:(0.83)> lift:(0.89) lev:(-0.03) [-7082] conv:(0.37)
353. *hasCoordinate*=true 71431 ==> *hasGeospatialIssues*=false 59471 <conf:(0.83)> lift:(0.89) lev:(-
0.03) [-7531] conv:(0.37)
354. *type*=PhysicalObject *taxonRank*=SPECIES 39269 ==> *taxonomicStatus*=ACCEPTED
hasGeospatialIssues=false 32544 <conf:(0.83)> lift:(0.99) lev:(-0) [-216] conv:(0.97)
355. *stateProvince*=RJ *phylum*=Arthropoda *hasCoordinate*=false 29520 ==>
taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 24350 <conf:(0.82)> lift:(0.99) lev:(-0) [-277]
conv:(0.95)
356. *phylum*=Arthropoda 135660 ==> *hasCoordinate*=false *hasGeospatialIssues*=false 111138
<conf:(0.82)> lift:(1.22) lev:(0.09) [19745] conv:(1.81)
357. *phylum*=Arthropoda *taxonomicStatus*=ACCEPTED 117690 ==> *hasCoordinate*=false 96382
<conf:(0.82)> lift:(1.2) lev:(0.07) [16055] conv:(1.75)
358. *preparations*=Alcool 70 *taxonomicStatus*=ACCEPTED 48751 ==> *hasCoordinate*=true 39781
<conf:(0.82)> lift:(2.57) lev:(0.11) [24304] conv:(3.71)
359. *preparations*=Alcool 70 52374 ==> *hasCoordinate*=true 42661 <conf:(0.81)> lift:(2.57)
lev:(0.12) [26033] conv:(3.68)
360. *stateProvince*=RJ 66709 ==> *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 54335
<conf:(0.81)> lift:(0.98) lev:(-0.01) [-1316] conv:(0.89)
361. *preparations*=Alcool 70 *taxonRank*=SPECIES 35096 ==> *taxonomicStatus*=ACCEPTED
hasGeospatialIssues=false 28323 <conf:(0.81)> lift:(0.97) lev:(-0) [-955] conv:(0.86)
362. *phylum*=Arthropoda *taxonomicStatus*=ACCEPTED 117690 ==> *hasCoordinate*=false
hasGeospatialIssues=false 94584 <conf:(0.8)> lift:(1.19) lev:(0.07) [15297] conv:(1.66)
363. *phylum*=Chordata 57290 ==> *taxonRank*=SPECIES 45798 <conf:(0.8)> lift:(1.91) lev:(0.1)
[21828] conv:(2.9)
364. *phylum*=Chordata *hasGeospatialIssues*=false 56652 ==> *taxonRank*=SPECIES 45214
<conf:(0.8)> lift:(1.91) lev:(0.1) [21511] conv:(2.88)
365. *preparations*=Alcool 70 *taxonRank*=SPECIES 35096 ==> *hasCoordinate*=true 27919
<conf:(0.8)> lift:(2.51) lev:(0.07) [16777] conv:(3.34)

366. preparations=Alcool 70 taxonRank=SPECIES taxonomicStatus=ACCEPTED 31956 ==> hasCoordinate=true 25417 <conf:(0.8)> lift:(2.51) lev:(0.07) [15271] conv:(3.34)
367. *phylum*=Chordata taxonomicStatus=ACCEPTED 53504 ==> taxonRank=SPECIES 42358 <conf:(0.79)> lift:(1.89) lev:(0.09) [19972] conv:(2.79)
368. type=PhysicalObject *hasGeospatialIssues*=false 43924 ==> taxonRank=SPECIES 34773 <conf:(0.79)> lift:(1.89) lev:(0.07) [16396] conv:(2.79)
369. *phylum*=Chordata taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 52902 ==> taxonRank=SPECIES 41810 <conf:(0.79)> lift:(1.89) lev:(0.09) [19676] conv:(2.77)
370. type=PhysicalObject taxonomicStatus=ACCEPTED 52408 ==> *hasGeospatialIssues*=false 41405 <conf:(0.79)> lift:(0.84) lev:(-0.03) [-7753] conv:(0.3)
371. type=PhysicalObject 55597 ==> *hasGeospatialIssues*=false 43924 <conf:(0.79)> lift:(0.84) lev:(-0.04) [-8226] conv:(0.3)
372. *phylum*=Chordata 57290 ==> taxonRank=SPECIES *hasGeospatialIssues*=false 45214 <conf:(0.79)> lift:(2.01) lev:(0.1) [22732] conv:(2.88)
373. type=PhysicalObject taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 41405 ==> taxonRank=SPECIES 32544 <conf:(0.79)> lift:(1.88) lev:(0.07) [15220] conv:(2.72)
374. hasCoordinate=true 71431 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 56067 <conf:(0.78)> lift:(0.94) lev:(-0.02) [-3524] conv:(0.77)
375. preparations=Alcool 70 52374 ==> *hasGeospatialIssues*=false 41079 <conf:(0.78)> lift:(0.84) lev:(-0.04) [-8048] conv:(0.29)
376. *phylum*=Chordata hasCoordinate=false 30150 ==> taxonRank=SPECIES 23631 <conf:(0.78)> lift:(1.87) lev:(0.05) [11016] conv:(2.69)
377. *phylum*=Chordata hasCoordinate=false *hasGeospatialIssues*=false 30141 ==> taxonRank=SPECIES 23622 <conf:(0.78)> lift:(1.87) lev:(0.05) [11011] conv:(2.69)
378. *phylum*=Chordata hasCoordinate=false 30150 ==> taxonRank=SPECIES *hasGeospatialIssues*=false 23622 <conf:(0.78)> lift:(2) lev:(0.05) [11790] conv:(2.81)
379. preparations=Alcool 70 taxonomicStatus=ACCEPTED 48751 ==> *hasGeospatialIssues*=false 38109 <conf:(0.78)> lift:(0.83) lev:(-0.03) [-7619] conv:(0.28)
380. *phylum*=Chordata taxonomicStatus=ACCEPTED 53504 ==> taxonRank=SPECIES *hasGeospatialIssues*=false 41810 <conf:(0.78)> lift:(1.99) lev:(0.09) [20814] conv:(2.78)
381. type=Collection hasCoordinate=true *hasGeospatialIssues*=false 30667 ==> taxonRank=SPECIES 23951 <conf:(0.78)> lift:(1.87) lev:(0.05) [11120] conv:(2.66)
382. type=Collection hasCoordinate=true 30908 ==> taxonRank=SPECIES 24127 <conf:(0.78)> lift:(1.87) lev:(0.05) [11195] conv:(2.65)
383. hasCoordinate=true *hasGeospatialIssues*=false 59471 ==> taxonRank=SPECIES 46260 <conf:(0.78)> lift:(1.86) lev:(0.1) [21378] conv:(2.62)
384. *phylum*=Arthropoda 135660 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 105509 <conf:(0.78)> lift:(0.93) lev:(-0.03) [-7665] conv:(0.75)

385. type=Collection hasCoordinate=true 30908 ==> taxonRank=SPECIES *hasGeospatialIssues*=false
23951 <conf:(0.77)> lift:(1.97) lev:(0.05) [11822] conv:(2.7)

386. *phylum*=Arthropoda *hasGeospatialIssues*=false 122680 ==> taxonomicStatus=ACCEPTED
hasCoordinate=false 94584 <conf:(0.77)> lift:(1.3) lev:(0.1) [21812] conv:(1.78)

387. taxonomicStatus=ACCEPTED hasCoordinate=true *hasGeospatialIssues*=false 56067 ==>
taxonRank=SPECIES 43146 <conf:(0.77)> lift:(1.84) lev:(0.09) [19688] conv:(2.52)

388. preparations=Alcool 70 taxonRank=SPECIES *hasGeospatialIssues*=false 30987 ==>
hasCoordinate=true 23822 <conf:(0.77)> lift:(2.42) lev:(0.06) [13984] conv:(2.95)

389. preparations=Alcool 70 taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 38109 ==>
hasCoordinate=true 29164 <conf:(0.77)> lift:(2.41) lev:(0.08) [17065] conv:(2.91)

390. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> hasCoordinate=true 31394
<conf:(0.76)> lift:(2.41) lev:(0.08) [18352] conv:(2.89)

391. *stateProvince*=RJ *phylum*=Arthropoda 38657 ==> hasCoordinate=false 29520 <conf:(0.76)>
lift:(1.12) lev:(0.01) [3135] conv:(1.34)

392. preparations=Alcool 70 52374 ==> taxonomicStatus=ACCEPTED hasCoordinate=true 39781
<conf:(0.76)> lift:(2.54) lev:(0.11) [24110] conv:(2.91)

393. preparations=Alcool 70 hasCoordinate=true *hasGeospatialIssues*=false 31394 ==>
taxonRank=SPECIES 23822 <conf:(0.76)> lift:(1.81) lev:(0.05) [10687] conv:(2.41)

394. type=Collection *hasGeospatialIssues*=false 52583 ==> taxonRank=SPECIES 39710
<conf:(0.76)> lift:(1.81) lev:(0.08) [17710] conv:(2.38)

395. type=Collection 52831 ==> taxonRank=SPECIES 39891 <conf:(0.76)> lift:(1.8) lev:(0.08)
[17787] conv:(2.37)

396. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> taxonRank=SPECIES 30987
<conf:(0.75)> lift:(1.8) lev:(0.06) [13800] conv:(2.37)

397. type=Collection 52831 ==> taxonRank=SPECIES *hasGeospatialIssues*=false 39710
<conf:(0.75)> lift:(1.92) lev:(0.08) [18978] conv:(2.45)

398. *stateProvince*=RJ *phylum*=Arthropoda 38657 ==> hasCoordinate=false
hasGeospatialIssues=false 29012 <conf:(0.75)> lift:(1.11) lev:(0.01) [2969] conv:(1.31)

399. type=PhysicalObject 55597 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 41405
<conf:(0.74)> lift:(0.89) lev:(-0.02) [-4976] conv:(0.65)

400. preparations=Alcool 70 taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 38109 ==>
taxonRank=SPECIES 28323 <conf:(0.74)> lift:(1.78) lev:(0.06) [12378] conv:(2.26)

401. *stateProvince*=RJ *phylum*=Arthropoda taxonomicStatus=ACCEPTED 33419 ==>
hasCoordinate=false 24835 <conf:(0.74)> lift:(1.09) lev:(0.01) [2025] conv:(1.24)

402. type=Collection taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 48355 ==>
taxonRank=SPECIES 35865 <conf:(0.74)> lift:(1.77) lev:(0.07) [15634] conv:(2.25)

403. type=Collection taxonomicStatus=ACCEPTED 48580 ==> taxonRank=SPECIES 36026
<conf:(0.74)> lift:(1.77) lev:(0.07) [15701] conv:(2.25)

404. type=PhysicalObject *hasGeospatialIssues*=false 43924 ==> taxonRank=SPECIES
taxonomicStatus=ACCEPTED 32544 <conf:(0.74)> lift:(1.95) lev:(0.07) [15820] conv:(2.39)
405. stateProvince=RJ phylum=Arthropoda 38657 ==> taxonomicStatus=ACCEPTED
hasGeospatialIssues=false 28619 <conf:(0.74)> lift:(0.89) lev:(-0.02) [-3630] conv:(0.64)
406. phylum=Chordata 57290 ==> taxonRank=SPECIES taxonomicStatus=ACCEPTED 42358
<conf:(0.74)> lift:(1.94) lev:(0.09) [20546] conv:(2.38)
407. type=Collection taxonomicStatus=ACCEPTED 48580 ==> taxonRank=SPECIES
hasGeospatialIssues=false 35865 <conf:(0.74)> lift:(1.88) lev:(0.07) [16801] conv:(2.32)
408. phylum=Chordata *hasGeospatialIssues*=false 56652 ==> taxonRank=SPECIES
taxonomicStatus=ACCEPTED 41810 <conf:(0.74)> lift:(1.94) lev:(0.09) [20241] conv:(2.36)
409. hasCoordinate=false 153570 ==> phylum=Arthropoda 113097 <conf:(0.74)> lift:(1.22)
lev:(0.09) [20504] conv:(1.51)
410. preparations=Alcool 70 hasCoordinate=true 42661 ==> *hasGeospatialIssues*=false 31394
<conf:(0.74)> lift:(0.78) lev:(-0.04) [-8622] conv:(0.23)
411. type=PhysicalObject taxonRank=SPECIES taxonomicStatus=ACCEPTED
hasGeospatialIssues=false 32544 ==> phylum=Chordata 23908 <conf:(0.73)> lift:(2.89) lev:(0.07)
[15621] conv:(2.81)
412. hasCoordinate=false *hasGeospatialIssues*=false 151581 ==> phylum=Arthropoda 111138
<conf:(0.73)> lift:(1.22) lev:(0.09) [19745] conv:(1.49)
413. preparations=Alcool 70 taxonomicStatus=ACCEPTED hasCoordinate=true 39781 ==>
hasGeospatialIssues=false 29164 <conf:(0.73)> lift:(0.78) lev:(-0.04) [-8150] conv:(0.23)
414. type=PhysicalObject taxonRank=SPECIES *hasGeospatialIssues*=false 34773 ==>
phylum=Chordata 25400 <conf:(0.73)> lift:(2.87) lev:(0.07) [16546] conv:(2.76)
415. phylum=Chordata 57290 ==> taxonRank=SPECIES taxonomicStatus=ACCEPTED
hasGeospatialIssues=false 41810 <conf:(0.73)> lift:(2.04) lev:(0.09) [21337] conv:(2.38)
416. stateProvince=RJ phylum=Arthropoda taxonomicStatus=ACCEPTED 33419 ==>
hasCoordinate=false *hasGeospatialIssues*=false 24350 <conf:(0.73)> lift:(1.08) lev:(0.01) [1835]
conv:(1.2)
417. preparations=Alcool 70 52374 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false
38109 <conf:(0.73)> lift:(0.87) lev:(-0.02) [-5583] conv:(0.61)
418. stateProvince=RJ phylum=Arthropoda *hasGeospatialIssues*=false 33543 ==>
taxonomicStatus=ACCEPTED hasCoordinate=false 24350 <conf:(0.73)> lift:(1.22) lev:(0.02) [4452]
conv:(1.48)
419. hasCoordinate=true *hasGeospatialIssues*=false 59471 ==> taxonRank=SPECIES
taxonomicStatus=ACCEPTED 43146 <conf:(0.73)> lift:(1.91) lev:(0.09) [20503] conv:(2.26)
420. preparations=Alcool 70 taxonRank=SPECIES 35096 ==> taxonomicStatus=ACCEPTED
hasCoordinate=true 25417 <conf:(0.72)> lift:(2.42) lev:(0.07) [14915] conv:(2.54)
421. hasCoordinate=false 153570 ==> phylum=Arthropoda *hasGeospatialIssues*=false 111138
<conf:(0.72)> lift:(1.33) lev:(0.12) [27405] conv:(1.65)

422. taxonomicStatus=ACCEPTED hasCoordinate=false 133466 ==> *phylum*=Arthropoda 96382
<conf:(0.72)> lift:(1.2) lev:(0.07) [15911] conv:(1.43)

423. *stateProvince*=RJ hasCoordinate=false 41004 ==> *phylum*=Arthropoda 29520 <conf:(0.72)>
lift:(1.19) lev:(0.02) [4797] conv:(1.42)

424. taxonomicStatus=ACCEPTED hasCoordinate=false *hasGeospatialIssues*=false 131640 ==>
phylum=Arthropoda 94584 <conf:(0.72)> lift:(1.19) lev:(0.07) [15214] conv:(1.41)

425. *hasGeospatialIssues*=false 211052 ==> hasCoordinate=false 151581 <conf:(0.72)> lift:(1.05)
lev:(0.03) [7531] conv:(1.13)

426. *stateProvince*=RJ hasCoordinate=false *hasGeospatialIssues*=false 40486 ==>
phylum=Arthropoda 29012 <conf:(0.72)> lift:(1.19) lev:(0.02) [4601] conv:(1.4)

427. hasCoordinate=true 71431 ==> taxonRank=SPECIES 50968 <conf:(0.71)> lift:(1.71) lev:(0.09)
[21082] conv:(2.03)

428. *phylum*=Arthropoda 135660 ==> taxonomicStatus=ACCEPTED hasCoordinate=false 96382
<conf:(0.71)> lift:(1.2) lev:(0.07) [15911] conv:(1.41)

429. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> taxonomicStatus=ACCEPTED
hasCoordinate=true 29164 <conf:(0.71)> lift:(2.37) lev:(0.07) [16872] conv:(2.42)

430. *phylum*=Arthropoda taxonomicStatus=ACCEPTED hasCoordinate=false
hasGeospatialIssues=false 94584 ==> taxonRank=FAMILY 67137 <conf:(0.71)> lift:(2.11) lev:(0.16)
[35378] conv:(2.29)

431. taxonomicStatus=ACCEPTED hasCoordinate=false 133466 ==> *phylum*=Arthropoda
hasGeospatialIssues=false 94584 <conf:(0.71)> lift:(1.3) lev:(0.1) [21812] conv:(1.56)

432. *stateProvince*=RJ hasCoordinate=false 41004 ==> *phylum*=Arthropoda
hasGeospatialIssues=false 29012 <conf:(0.71)> lift:(1.3) lev:(0.03) [6654] conv:(1.55)

433. type=PhysicalObject 55597 ==> taxonRank=SPECIES 39269 <conf:(0.71)> lift:(1.69) lev:(0.07)
[16008] conv:(1.98)

434. *stateProvince*=RJ taxonomicStatus=ACCEPTED hasCoordinate=false 35268 ==>
phylum=Arthropoda 24835 <conf:(0.7)> lift:(1.17) lev:(0.02) [3570] conv:(1.34)

435. taxonomicStatus=ACCEPTED hasCoordinate=true 67323 ==> taxonRank=SPECIES 47338
<conf:(0.7)> lift:(1.68) lev:(0.09) [19171] conv:(1.96)

436. taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 187707 ==> hasCoordinate=false
131640 <conf:(0.7)> lift:(1.03) lev:(0.02) [3524] conv:(1.06)

437. *stateProvince*=RJ taxonomicStatus=ACCEPTED hasCoordinate=false *hasGeospatialIssues*=false
34773 ==> *phylum*=Arthropoda 24350 <conf:(0.7)> lift:(1.16) lev:(0.02) [3384] conv:(1.32)

438. *phylum*=Arthropoda taxonomicStatus=ACCEPTED hasCoordinate=false 96382 ==>
taxonRank=FAMILY 67357 <conf:(0.7)> lift:(2.08) lev:(0.16) [34994] conv:(2.21)

439. type=PhysicalObject taxonomicStatus=ACCEPTED 52408 ==> taxonRank=SPECIES 36547
<conf:(0.7)> lift:(1.67) lev:(0.06) [14620] conv:(1.92)

440. *phylum*=Arthropoda 135660 ==> *taxonomicStatus*=ACCEPTED *hasCoordinate*=false *hasGeospatialIssues*=false 94584 <conf:(0.7)> lift:(1.19) lev:(0.07) [15214] conv:(1.37)
441. *phylum*=Arthropoda *taxonomicStatus*=ACCEPTED *hasCoordinate*=false 96382 ==> *taxonRank*=FAMILY *hasGeospatialIssues*=false 67137 <conf:(0.7)> lift:(2.18) lev:(0.16) [36377] conv:(2.24)
442. *type*=PhysicalObject *taxonomicStatus*=ACCEPTED *hasCoordinate*=true 36148 ==> *hasGeospatialIssues*=false 25171 <conf:(0.7)> lift:(0.74) lev:(-0.04) [-8735] conv:(0.2)
443. *type*=PhysicalObject *hasCoordinate*=true 38124 ==> *hasGeospatialIssues*=false 26478 <conf:(0.69)> lift:(0.74) lev:(-0.04) [-9282] conv:(0.2)
444. *stateProvince*=RJ *taxonomicStatus*=ACCEPTED *hasCoordinate*=false 35268 ==> *phylum*=Arthropoda *hasGeospatialIssues*=false 24350 <conf:(0.69)> lift:(1.27) lev:(0.02) [5120] conv:(1.47)
445. *type*=PhysicalObject *taxonomicStatus*=ACCEPTED 52408 ==> *hasCoordinate*=true 36148 <conf:(0.69)> lift:(2.17) lev:(0.09) [19510] conv:(2.2)
446. *preparations*=Alcool 70 *hasGeospatialIssues*=false 41079 ==> *taxonRank*=SPECIES *taxonomicStatus*=ACCEPTED 28323 <conf:(0.69)> lift:(1.81) lev:(0.06) [12683] conv:(1.99)
447. *type*=PhysicalObject *taxonRank*=SPECIES *hasGeospatialIssues*=false 34773 ==> *phylum*=Chordata *taxonomicStatus*=ACCEPTED 23908 <conf:(0.69)> lift:(2.89) lev:(0.07) [15639] conv:(2.44)
448. *type*=PhysicalObject 55597 ==> *hasCoordinate*=true 38124 <conf:(0.69)> lift:(2.16) lev:(0.09) [20473] conv:(2.17)
449. *preparations*=Alcool 70 *hasCoordinate*=true 42661 ==> *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 29164 <conf:(0.68)> lift:(0.82) lev:(-0.03) [-6425] conv:(0.52)
450. *type*=Collection *hasGeospatialIssues*=false 52583 ==> *taxonRank*=SPECIES *taxonomicStatus*=ACCEPTED 35865 <conf:(0.68)> lift:(1.79) lev:(0.07) [15845] conv:(1.95)
451. *type*=Collection 52831 ==> *taxonRank*=SPECIES *taxonomicStatus*=ACCEPTED 36026 <conf:(0.68)> lift:(1.79) lev:(0.07) [15911] conv:(1.95)
452. *type*=Collection 52831 ==> *taxonRank*=SPECIES *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 35865 <conf:(0.68)> lift:(1.9) lev:(0.08) [16985] conv:(2)
453. *preparations*=Alcool 70 *taxonRank*=SPECIES 35096 ==> *hasCoordinate*=true *hasGeospatialIssues*=false 23822 <conf:(0.68)> lift:(2.57) lev:(0.06) [14545] conv:(2.29)
454. *preparations*=Alcool 70 52374 ==> *taxonRank*=SPECIES 35096 <conf:(0.67)> lift:(1.6) lev:(0.06) [13183] conv:(1.76)
455. *type*=PhysicalObject *taxonRank*=SPECIES *taxonomicStatus*=ACCEPTED 36547 ==> *phylum*=Chordata 24440 <conf:(0.67)> lift:(2.63) lev:(0.07) [15134] conv:(2.25)
456. *taxonomicStatus*=ACCEPTED 200789 ==> *hasCoordinate*=false 133466 <conf:(0.66)> lift:(0.97) lev:(-0.02) [-3578] conv:(0.95)
457. *hasCoordinate*=true 71431 ==> *taxonRank*=SPECIES *taxonomicStatus*=ACCEPTED 47338 <conf:(0.66)> lift:(1.74) lev:(0.09) [20142] conv:(1.84)

458. type=PhysicalObject taxonRank=SPECIES 39269 ==> *phylum*=Chordata 25967 <conf:(0.66)> lift:(2.6) lev:(0.07) [15968] conv:(2.2)
459. type=PhysicalObject hasCoordinate=true 38124 ==> taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 25171 <conf:(0.66)> lift:(0.79) lev:(-0.03) [-6633] conv:(0.49)
460. *stateProvince*=RJ *hasGeospatialIssues*=false 61516 ==> hasCoordinate=false 40486 <conf:(0.66)> lift:(0.96) lev:(-0.01) [-1500] conv:(0.93)
461. type=PhysicalObject 55597 ==> taxonRank=SPECIES taxonomicStatus=ACCEPTED 36547 <conf:(0.66)> lift:(1.73) lev:(0.07) [15379] conv:(1.81)
462. taxonomicStatus=ACCEPTED 200789 ==> hasCoordinate=false *hasGeospatialIssues*=false 131640 <conf:(0.66)> lift:(0.97) lev:(-0.02) [-3629] conv:(0.95)
463. preparations=Alcool 70 taxonomicStatus=ACCEPTED 48751 ==> taxonRank=SPECIES 31956 <conf:(0.66)> lift:(1.57) lev:(0.05) [11559] conv:(1.69)
464. preparations=Alcool 70 hasCoordinate=true 42661 ==> taxonRank=SPECIES 27919 <conf:(0.65)> lift:(1.56) lev:(0.04) [10070] conv:(1.68)
465. type=PhysicalObject taxonRank=SPECIES taxonomicStatus=ACCEPTED 36547 ==> *phylum*=Chordata *hasGeospatialIssues*=false 23908 <conf:(0.65)> lift:(2.6) lev:(0.07) [14705] conv:(2.16)
466. type=PhysicalObject 55597 ==> taxonomicStatus=ACCEPTED hasCoordinate=true 36148 <conf:(0.65)> lift:(2.17) lev:(0.09) [19512] conv:(2)
467. hasCoordinate=true 71431 ==> taxonRank=SPECIES *hasGeospatialIssues*=false 46260 <conf:(0.65)> lift:(1.65) lev:(0.08) [18229] conv:(1.72)
468. type=PhysicalObject hasCoordinate=true 38124 ==> taxonRank=SPECIES 24684 <conf:(0.65)> lift:(1.55) lev:(0.04) [8733] conv:(1.65)
469. type=PhysicalObject taxonRank=SPECIES 39269 ==> *phylum*=Chordata *hasGeospatialIssues*=false 25400 <conf:(0.65)> lift:(2.57) lev:(0.07) [15512] conv:(2.12)
470. *phylum*=Arthropoda taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 105509 ==> taxonRank=FAMILY 68128 <conf:(0.65)> lift:(1.92) lev:(0.15) [32701] conv:(1.87)
471. *stateProvince*=RJ *phylum*=Arthropoda 38657 ==> taxonomicStatus=ACCEPTED hasCoordinate=false 24835 <conf:(0.64)> lift:(1.08) lev:(0.01) [1904] conv:(1.14)
472. taxonomicStatus=ACCEPTED hasCoordinate=true 67323 ==> taxonRank=SPECIES *hasGeospatialIssues*=false 43146 <conf:(0.64)> lift:(1.63) lev:(0.07) [16727] conv:(1.69)
473. *stateProvince*=RJ taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 54335 ==> hasCoordinate=false 34773 <conf:(0.64)> lift:(0.94) lev:(-0.01) [-2312] conv:(0.88)
474. preparations=Alcool 70 taxonomicStatus=ACCEPTED hasCoordinate=true 39781 ==> taxonRank=SPECIES 25417 <conf:(0.64)> lift:(1.53) lev:(0.04) [8773] conv:(1.61)
475. type=PhysicalObject *hasGeospatialIssues*=false 43924 ==> *phylum*=Chordata 28052 <conf:(0.64)> lift:(2.51) lev:(0.07) [16868] conv:(2.06)

476. type=PhysicalObject taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 41405 ==> *phylum*=Chordata 26395 <conf:(0.64)> lift:(2.5) lev:(0.07) [15852] conv:(2.06)
477. type=PhysicalObject taxonomicStatus=ACCEPTED *hasCoordinate*=true 36148 ==> taxonRank=SPECIES 23027 <conf:(0.64)> lift:(1.52) lev:(0.04) [7903] conv:(1.6)
478. type=PhysicalObject *hasCoordinate*=true 38124 ==> preparations=Alcool 70 24275 <conf:(0.64)> lift:(2.74) lev:(0.07) [15400] conv:(2.11)
479. *phylum*=Arthropoda taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 105509 ==> taxonRank=FAMILY *hasCoordinate*=false 67137 <conf:(0.64)> lift:(2.05) lev:(0.15) [34342] conv:(1.89)
480. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> disposition=Na Colecao 26087 <conf:(0.64)> lift:(5.31) lev:(0.09) [21170] conv:(2.41)
481. type=PhysicalObject taxonomicStatus=ACCEPTED *hasCoordinate*=true 36148 ==> preparations=Alcool 70 22872 <conf:(0.63)> lift:(2.72) lev:(0.06) [14457] conv:(2.09)
482. type=PhysicalObject taxonRank=SPECIES taxonomicStatus=ACCEPTED 36547 ==> *hasCoordinate*=true 23027 <conf:(0.63)> lift:(1.98) lev:(0.05) [11424] conv:(1.84)
483. *stateProvince*=RJ *phylum*=Arthropoda 38657 ==> taxonomicStatus=ACCEPTED *hasCoordinate*=false *hasGeospatialIssues*=false 24350 <conf:(0.63)> lift:(1.08) lev:(0.01) [1733] conv:(1.12)
484. type=PhysicalObject taxonRank=SPECIES 39269 ==> *hasCoordinate*=true 24684 <conf:(0.63)> lift:(1.98) lev:(0.05) [12217] conv:(1.84)
485. *hasCoordinate*=false 153570 ==> *phylum*=Arthropoda taxonomicStatus=ACCEPTED 96382 <conf:(0.63)> lift:(1.2) lev:(0.07) [16055] conv:(1.28)
486. preparations=Alcool 70 taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 38109 ==> disposition=Na Colecao 23841 <conf:(0.63)> lift:(5.23) lev:(0.09) [19280] conv:(2.35)
487. type=PhysicalObject 55597 ==> taxonRank=SPECIES *hasGeospatialIssues*=false 34773 <conf:(0.63)> lift:(1.59) lev:(0.06) [12956] conv:(1.62)
488. *hasCoordinate*=false *hasGeospatialIssues*=false 151581 ==> *phylum*=Arthropoda taxonomicStatus=ACCEPTED 94584 <conf:(0.62)> lift:(1.19) lev:(0.07) [15297] conv:(1.27)
489. *hasGeospatialIssues*=false 211052 ==> taxonomicStatus=ACCEPTED *hasCoordinate*=false 131640 <conf:(0.62)> lift:(1.05) lev:(0.03) [6448] conv:(1.08)
490. type=PhysicalObject taxonRank=SPECIES 39269 ==> *phylum*=Chordata taxonomicStatus=ACCEPTED 24440 <conf:(0.62)> lift:(2.62) lev:(0.07) [15102] conv:(2.02)
491. type=PhysicalObject taxonomicStatus=ACCEPTED 52408 ==> taxonRank=SPECIES *hasGeospatialIssues*=false 32544 <conf:(0.62)> lift:(1.58) lev:(0.05) [11978] conv:(1.6)
492. *hasCoordinate*=false 153570 ==> *phylum*=Arthropoda taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 94584 <conf:(0.62)> lift:(1.31) lev:(0.1) [22570] conv:(1.38)
493. *stateProvince*=RJ 66709 ==> *hasCoordinate*=false 41004 <conf:(0.61)> lift:(0.9) lev:(-0.02) [-4526] conv:(0.82)

494. preparations=Alcool 70 52374 ==> taxonRank=SPECIES taxonomicStatus=ACCEPTED 31956
<conf:(0.61)> lift:(1.6) lev:(0.05) [12015] conv:(1.59)

495. *phylum*=Arthropoda taxonomicStatus=ACCEPTED 117690 ==> taxonRank=FAMILY 71716
<conf:(0.61)> lift:(1.81) lev:(0.14) [32199] conv:(1.7)

496. type=PhysicalObject taxonRank=SPECIES 39269 ==> *phylum*=Chordata
taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 23908 <conf:(0.61)> lift:(2.59) lev:(0.07)
[14675] conv:(1.96)

497. type=PhysicalObject taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 41405 ==>
hasCoordinate=true 25171 <conf:(0.61)> lift:(1.91) lev:(0.05) [12026] conv:(1.74)

498. *stateProvince*=RJ 66709 ==> hasCoordinate=false *hasGeospatialIssues*=false 40486
<conf:(0.61)> lift:(0.9) lev:(-0.02) [-4455] conv:(0.83)

499. *stateProvince*=RJ hasCoordinate=false 41004 ==> *phylum*=Arthropoda
taxonomicStatus=ACCEPTED 24835 <conf:(0.61)> lift:(1.16) lev:(0.02) [3387] conv:(1.21)

500. type=Collection taxonRank=SPECIES 39891 ==> hasCoordinate=true 24127 <conf:(0.6)>
lift:(1.91) lev:(0.05) [11462] conv:(1.73)

501. *phylum*=Arthropoda hasCoordinate=false *hasGeospatialIssues*=false 111138 ==>
taxonRank=FAMILY 67218 <conf:(0.6)> lift:(1.8) lev:(0.13) [29900] conv:(1.68)

502. *phylum*=Arthropoda hasCoordinate=false *hasGeospatialIssues*=false 111138 ==>
taxonRank=FAMILY taxonomicStatus=ACCEPTED 67137 <conf:(0.6)> lift:(1.8) lev:(0.13) [29927]
conv:(1.68)

503. hasCoordinate=true 71431 ==> taxonRank=SPECIES taxonomicStatus=ACCEPTED
hasGeospatialIssues=false 43146 <conf:(0.6)> lift:(1.69) lev:(0.08) [17619] conv:(1.62)

504. type=PhysicalObject hasCoordinate=true 38124 ==> taxonRank=SPECIES
taxonomicStatus=ACCEPTED 23027 <conf:(0.6)> lift:(1.59) lev:(0.04) [8512] conv:(1.56)

505. type=Collection taxonRank=SPECIES *hasGeospatialIssues*=false 39710 ==> hasCoordinate=true
23951 <conf:(0.6)> lift:(1.9) lev:(0.05) [11344] conv:(1.72)

506. type=PhysicalObject *hasGeospatialIssues*=false 43924 ==> hasCoordinate=true 26478
<conf:(0.6)> lift:(1.9) lev:(0.06) [12533] conv:(1.72)

507. *stateProvince*=RJ hasCoordinate=false *hasGeospatialIssues*=false 40486 ==>
phylum=Arthropoda taxonomicStatus=ACCEPTED 24350 <conf:(0.6)> lift:(1.15) lev:(0.01) [3173]
conv:(1.2)

508. type=PhysicalObject *hasGeospatialIssues*=false 43924 ==> *phylum*=Chordata
taxonomicStatus=ACCEPTED 26395 <conf:(0.6)> lift:(2.53) lev:(0.07) [15950] conv:(1.91)

509. type=Collection taxonRank=SPECIES 39891 ==> hasCoordinate=true *hasGeospatialIssues*=false
23951 <conf:(0.6)> lift:(2.27) lev:(0.06) [13407] conv:(1.84)

510. type=PhysicalObject hasCoordinate=true 38124 ==> preparations=Alcool 70
taxonomicStatus=ACCEPTED 22872 <conf:(0.6)> lift:(2.77) lev:(0.06) [14611] conv:(1.96)

511. preparations=Alcool 70 52374 ==> hasCoordinate=true *hasGeospatialIssues*=false 31394
<conf:(0.6)> lift:(2.27) lev:(0.08) [17550] conv:(1.84)

512. preparations=Alcool 70 taxonomicStatus=ACCEPTED 48751 ==> hasCoordinate=true
hasGeospatialIssues=false 29164 <conf:(0.6)> lift:(2.26) lev:(0.07) [16278] conv:(1.83)
513. hasCoordinate=true 71431 ==> preparations=Alcool 70 42661 <conf:(0.6)> lift:(2.57)
lev:(0.12) [26033] conv:(1.9)
514. *phylum*=Arthropoda hasCoordinate=false 113097 ==> taxonRank=FAMILY 67438 <conf:(0.6)>
lift:(1.78) lev:(0.13) [29463] conv:(1.65)
515. preparations=Alcool 70 hasCoordinate=true 42661 ==> taxonRank=SPECIES
taxonomicStatus=ACCEPTED 25417 <conf:(0.6)> lift:(1.56) lev:(0.04) [9174] conv:(1.53)
516. *stateProvince*=RJ taxonomicStatus=ACCEPTED 59211 ==> hasCoordinate=false 35268
<conf:(0.6)> lift:(0.87) lev:(-0.02) [-5145] conv:(0.79)
517. *phylum*=Arthropoda hasCoordinate=false 113097 ==> taxonRank=FAMILY
taxonomicStatus=ACCEPTED 67357 <conf:(0.6)> lift:(1.78) lev:(0.13) [29491] conv:(1.64)
518. *phylum*=Arthropoda hasCoordinate=false 113097 ==> taxonRank=FAMILY
hasGeospatialIssues=false 67218 <conf:(0.59)> lift:(1.86) lev:(0.14) [31123] conv:(1.68)
519. *stateProvince*=RJ hasCoordinate=false 41004 ==> *phylum*=Arthropoda
taxonomicStatus=ACCEPTED hasGeospatialIssues=false 24350 <conf:(0.59)> lift:(1.27) lev:(0.02)
[5122] conv:(1.31)
520. type=Collection taxonomicStatus=ACCEPTED 48580 ==> hasCoordinate=true 28841
<conf:(0.59)> lift:(1.87) lev:(0.06) [13418] conv:(1.68)
521. *phylum*=Arthropoda hasCoordinate=false 113097 ==> taxonRank=FAMILY
taxonomicStatus=ACCEPTED hasGeospatialIssues=false 67137 <conf:(0.59)> lift:(1.86) lev:(0.14)
[31092] conv:(1.68)
522. type=Collection taxonomicStatus=ACCEPTED hasGeospatialIssues=false 48355 ==>
hasCoordinate=true 28621 <conf:(0.59)> lift:(1.86) lev:(0.06) [13269] conv:(1.67)
523. preparations=Alcool 70 52374 ==> taxonRank=SPECIES hasGeospatialIssues=false 30987
<conf:(0.59)> lift:(1.51) lev:(0.05) [10435] conv:(1.49)
524. taxonomicStatus=ACCEPTED hasCoordinate=true 67323 ==> preparations=Alcool 70 39781
<conf:(0.59)> lift:(2.54) lev:(0.11) [24110] conv:(1.88)
525. type=Collection taxonomicStatus=ACCEPTED 48580 ==> hasCoordinate=true
hasGeospatialIssues=false 28621 <conf:(0.59)> lift:(2.23) lev:(0.07) [15780] conv:(1.79)
526. *stateProvince*=RJ taxonomicStatus=ACCEPTED 59211 ==> hasCoordinate=false
hasGeospatialIssues=false 34773 <conf:(0.59)> lift:(0.87) lev:(-0.02) [-5116] conv:(0.79)
527. type=PhysicalObject taxonRank=SPECIES 39269 ==> taxonomicStatus=ACCEPTED
hasCoordinate=true 23027 <conf:(0.59)> lift:(1.96) lev:(0.05) [11277] conv:(1.69)
528. taxonomicStatus=ACCEPTED 200789 ==> *phylum*=Arthropoda 117690 <conf:(0.59)> lift:(0.97)
lev:(-0.01) [-3371] conv:(0.96)
529. type=PhysicalObject 55597 ==> taxonRank=SPECIES taxonomicStatus=ACCEPTED
hasGeospatialIssues=false 32544 <conf:(0.59)> lift:(1.64) lev:(0.06) [12676] conv:(1.55)

530. type=Collection 52831 ==> hasCoordinate=true 30908 <conf:(0.59)> lift:(1.84) lev:(0.06) [14135] conv:(1.64)

531. type=Collection *hasGeospatialIssues*=false 52583 ==> hasCoordinate=true 30667 <conf:(0.58)> lift:(1.84) lev:(0.06) [13973] conv:(1.64)

532. *hasGeospatialIssues*=false 211052 ==> *phylum*=Arthropoda 122680 <conf:(0.58)> lift:(0.96) lev:(-0.02) [-4569] conv:(0.95)

533. preparations=Alcool 70 taxonomicStatus=ACCEPTED 48751 ==> taxonRank=SPECIES *hasGeospatialIssues*=false 28323 <conf:(0.58)> lift:(1.48) lev:(0.04) [9192] conv:(1.45)

534. type=Collection 52831 ==> hasCoordinate=true *hasGeospatialIssues*=false 30667 <conf:(0.58)> lift:(2.2) lev:(0.07) [16703] conv:(1.75)

535. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> disposition=Na Colecao taxonomicStatus=ACCEPTED 23841 <conf:(0.58)> lift:(5.3) lev:(0.09) [19346] conv:(2.12)

536. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> taxonRank=SPECIES hasCoordinate=true 23822 <conf:(0.58)> lift:(2.56) lev:(0.06) [14516] conv:(1.84)

537. *stateProvince*=RJ 66709 ==> *phylum*=Arthropoda 38657 <conf:(0.58)> lift:(0.96) lev:(-0.01) [-1563] conv:(0.94)

538. *phylum*=Arthropoda taxonomicStatus=ACCEPTED 117690 ==> taxonRank=FAMILY *hasGeospatialIssues*=false 68128 <conf:(0.58)> lift:(1.81) lev:(0.14) [30567] conv:(1.62)

539. type=PhysicalObject *hasGeospatialIssues*=false 43924 ==> *phylum*=Chordata taxonRank=SPECIES 25400 <conf:(0.58)> lift:(2.84) lev:(0.07) [16459] conv:(1.89)

540. type=PhysicalObject taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 41405 ==> *phylum*=Chordata taxonRank=SPECIES 23908 <conf:(0.58)> lift:(2.84) lev:(0.07) [15480] conv:(1.88)

541. *phylum*=Chordata taxonRank=SPECIES taxonomicStatus=ACCEPTED 42358 ==> type=PhysicalObject 24440 <conf:(0.58)> lift:(2.34) lev:(0.06) [13973] conv:(1.78)

542. preparations=Alcool 70 taxonomicStatus=ACCEPTED hasCoordinate=true 39781 ==> type=PhysicalObject 22872 <conf:(0.57)> lift:(2.33) lev:(0.06) [13042] conv:(1.77)

543. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> type=Collection 23562 <conf:(0.57)> lift:(2.44) lev:(0.06) [13916] conv:(1.79)

544. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> *phylum*=Mollusca 23555 <conf:(0.57)> lift:(5.34) lev:(0.09) [19147] conv:(2.09)

545. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> type=Collection *phylum*=Mollusca 23555 <conf:(0.57)> lift:(5.35) lev:(0.09) [19148] conv:(2.09)

546. type=PhysicalObject *hasGeospatialIssues*=false 43924 ==> taxonomicStatus=ACCEPTED hasCoordinate=true 25171 <conf:(0.57)> lift:(1.92) lev:(0.05) [12028] conv:(1.64)

547. *phylum*=Arthropoda taxonomicStatus=ACCEPTED 117690 ==> taxonRank=FAMILY hasCoordinate=false 67357 <conf:(0.57)> lift:(1.84) lev:(0.14) [30776] conv:(1.61)

548. *phylum*=Chordata taxonRank=SPECIES taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 41810 ==> type=PhysicalObject 23908 <conf:(0.57)> lift:(2.31) lev:(0.06) [13576] conv:(1.76)

549. *phylum*=Arthropoda taxonomicStatus=ACCEPTED 117690 ==> taxonRank=FAMILY
hasCoordinate=false *hasGeospatialIssues*=false 67137 <conf:(0.57)> lift:(1.84) lev:(0.14) [30672]
conv:(1.61)
550. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> type=Collection disposition=Na
Colecao 23427 <conf:(0.57)> lift:(5.35) lev:(0.08) [19044] conv:(2.08)
551. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> disposition=Na Colecao
phylum=Mollusca 23420 <conf:(0.57)> lift:(5.35) lev:(0.08) [19038] conv:(2.08)
552. preparations=Alcool 70 *hasGeospatialIssues*=false 41079 ==> type=Collection disposition=Na
Colecao *phylum*=Mollusca 23420 <conf:(0.57)> lift:(5.35) lev:(0.08) [19038] conv:(2.08)
553. preparations=Alcool 70 hasCoordinate=true 42661 ==> type=PhysicalObject 24275
<conf:(0.57)> lift:(2.3) lev:(0.06) [13733] conv:(1.75)
554. *phylum*=Chordata taxonRank=SPECIES 45798 ==> type=PhysicalObject 25967 <conf:(0.57)>
lift:(2.29) lev:(0.07) [14650] conv:(1.74)
555. *stateProvince*=RJ *hasGeospatialIssues*=false 61516 ==> taxonomicStatus=ACCEPTED
hasCoordinate=false 34773 <conf:(0.57)> lift:(0.95) lev:(-0.01) [-1717] conv:(0.94)
556. *phylum*=Chordata taxonRank=SPECIES taxonomicStatus=ACCEPTED 42358 ==>
type=PhysicalObject *hasGeospatialIssues*=false 23908 <conf:(0.56)> lift:(2.89) lev:(0.07) [15639]
conv:(1.85)
557. *stateProvince*=RJ taxonomicStatus=ACCEPTED 59211 ==> *phylum*=Arthropoda 33419
<conf:(0.56)> lift:(0.94) lev:(-0.01) [-2281] conv:(0.91)
558. taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 187707 ==> *phylum*=Arthropoda
105509 <conf:(0.56)> lift:(0.93) lev:(-0.03) [-7665] conv:(0.91)
559. taxonRank=SPECIES hasCoordinate=false *hasGeospatialIssues*=false 42032 ==>
phylum=Chordata 23622 <conf:(0.56)> lift:(2.21) lev:(0.06) [12919] conv:(1.7)
560. *phylum*=Chordata taxonRank=SPECIES *hasGeospatialIssues*=false 45214 ==>
type=PhysicalObject 25400 <conf:(0.56)> lift:(2.27) lev:(0.06) [14227] conv:(1.72)
561. preparations=Alcool 70 hasCoordinate=true 42661 ==> taxonRank=SPECIES
hasGeospatialIssues=false 23822 <conf:(0.56)> lift:(1.42) lev:(0.03) [7081] conv:(1.38)
562. hasCoordinate=true 71431 ==> preparations=Alcool 70 taxonomicStatus=ACCEPTED 39781
<conf:(0.56)> lift:(2.57) lev:(0.11) [24304] conv:(1.77)
563. preparations=Alcool 70 52374 ==> taxonomicStatus=ACCEPTED hasCoordinate=true
hasGeospatialIssues=false 29164 <conf:(0.56)> lift:(2.23) lev:(0.07) [16113] conv:(1.69)
564. *phylum*=Arthropoda *hasGeospatialIssues*=false 122680 ==> taxonRank=FAMILY 68217
<conf:(0.56)> lift:(1.66) lev:(0.12) [27024] conv:(1.5)
565. *phylum*=Arthropoda *hasGeospatialIssues*=false 122680 ==> taxonRank=FAMILY
taxonomicStatus=ACCEPTED 68128 <conf:(0.56)> lift:(1.66) lev:(0.12) [27054] conv:(1.5)
566. *phylum*=Chordata taxonRank=SPECIES 45798 ==> type=PhysicalObject
hasGeospatialIssues=false 25400 <conf:(0.55)> lift:(2.84) lev:(0.07) [16459] conv:(1.81)

567. taxonRank=SPECIES taxonomicStatus=ACCEPTED 85664 ==> hasCoordinate=true 47338
<conf:(0.55)> lift:(1.74) lev:(0.09) [20142] conv:(1.53)

568. type=Collection taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 48355 ==>
phylum=Chordata 26507 <conf:(0.55)> lift:(2.15) lev:(0.06) [14194] conv:(1.65)

569. *phylum*=Arthropoda *hasGeospatialIssues*=false 122680 ==> taxonRank=FAMILY
hasCoordinate=false 67218 <conf:(0.55)> lift:(1.76) lev:(0.13) [29086] conv:(1.52)

570. taxonRank=SPECIES hasCoordinate=true 50968 ==> preparations=Alcool 70 27919
<conf:(0.55)> lift:(2.35) lev:(0.07) [16055] conv:(1.7)

571. taxonRank=SPECIES hasCoordinate=false 43168 ==> *phylum*=Chordata 23631 <conf:(0.55)>
lift:(2.15) lev:(0.06) [12639] conv:(1.65)

572. *phylum*=Arthropoda *hasGeospatialIssues*=false 122680 ==> taxonRank=FAMILY
taxonomicStatus=ACCEPTED hasCoordinate=false 67137 <conf:(0.55)> lift:(1.76) lev:(0.13) [29054]
conv:(1.52)

573. taxonRank=SPECIES hasCoordinate=false 43168 ==> *phylum*=Chordata
hasGeospatialIssues=false 23622 <conf:(0.55)> lift:(2.17) lev:(0.06) [12752] conv:(1.65)

574. type=Collection taxonomicStatus=ACCEPTED 48580 ==> *phylum*=Chordata 26545
<conf:(0.55)> lift:(2.15) lev:(0.06) [14175] conv:(1.64)

575. type=Collection 52831 ==> taxonomicStatus=ACCEPTED hasCoordinate=true 28841
<conf:(0.55)> lift:(1.82) lev:(0.06) [13033] conv:(1.54)

576. type=Collection taxonomicStatus=ACCEPTED 48580 ==> *phylum*=Chordata
hasGeospatialIssues=false 26507 <conf:(0.55)> lift:(2.17) lev:(0.06) [14275] conv:(1.65)

577. *stateProvince*=RJ *hasGeospatialIssues*=false 61516 ==> *phylum*=Arthropoda 33543
<conf:(0.55)> lift:(0.9) lev:(-0.02) [-3546] conv:(0.87)

578. type=PhysicalObject *hasGeospatialIssues*=false 43924 ==> *phylum*=Chordata
taxonRank=SPECIES taxonomicStatus=ACCEPTED 23908 <conf:(0.54)> lift:(2.89) lev:(0.07) [15639]
conv:(1.78)

579. type=Collection *hasGeospatialIssues*=false 52583 ==> taxonomicStatus=ACCEPTED
hasCoordinate=true 28621 <conf:(0.54)> lift:(1.82) lev:(0.06) [12887] conv:(1.54)

580. type=Collection *hasGeospatialIssues*=false 52583 ==> *phylum*=Chordata 28600 <conf:(0.54)>
lift:(2.14) lev:(0.07) [15211] conv:(1.63)

581. type=Collection 52831 ==> *phylum*=Chordata 28639 <conf:(0.54)> lift:(2.13) lev:(0.07)
[15187] conv:(1.63)

582. type=Collection 52831 ==> taxonomicStatus=ACCEPTED hasCoordinate=true
hasGeospatialIssues=false 28621 <conf:(0.54)> lift:(2.17) lev:(0.07) [15456] conv:(1.64)

583. taxonRank=SPECIES 94136 ==> hasCoordinate=true 50968 <conf:(0.54)> lift:(1.71) lev:(0.09)
[21082] conv:(1.49)

584. type=Collection 52831 ==> *phylum*=Chordata *hasGeospatialIssues*=false 28600 <conf:(0.54)>
lift:(2.15) lev:(0.07) [15297] conv:(1.63)

585. preparations=Alcool 70 52374 ==> taxonRank=SPECIES taxonomicStatus=ACCEPTED
hasGeospatialIssues=false 28323 <conf:(0.54)> lift:(1.51) lev:(0.04) [9606] conv:(1.4)
586. taxonomicStatus=ACCEPTED hasCoordinate=true 67323 ==> type=PhysicalObject 36148
<conf:(0.54)> lift:(2.17) lev:(0.09) [19512] conv:(1.63)
587. taxonRank=SPECIES taxonomicStatus=ACCEPTED hasCoordinate=true 47338 ==>
preparations=Alcool 70 25417 <conf:(0.54)> lift:(2.31) lev:(0.06) [14398] conv:(1.66)
588. taxonRank=SPECIES taxonomicStatus=ACCEPTED hasGeospatialIssues=false 80405 ==>
hasCoordinate=true 43146 <conf:(0.54)> lift:(1.69) lev:(0.08) [17619] conv:(1.47)
589. preparations=Alcool 70 hasCoordinate=true 42661 ==> type=PhysicalObject
taxonomicStatus=ACCEPTED 22872 <conf:(0.54)> lift:(2.3) lev:(0.06) [12935] conv:(1.65)
590. hasCoordinate=true 71431 ==> type=PhysicalObject 38124 <conf:(0.53)> lift:(2.16) lev:(0.09)
[20473] conv:(1.61)
591. phylum=Chordata taxonRank=SPECIES 45798 ==> type=PhysicalObject
taxonomicStatus=ACCEPTED 24440 <conf:(0.53)> lift:(2.29) lev:(0.06) [13772] conv:(1.64)
592. preparations=Alcool 70 52374 ==> taxonRank=SPECIES hasCoordinate=true 27919
<conf:(0.53)> lift:(2.35) lev:(0.07) [16055] conv:(1.66)
593. phylum=Chordata hasGeospatialIssues=false 56652 ==> hasCoordinate=false 30141
<conf:(0.53)> lift:(0.78) lev:(-0.04) [-8525] conv:(0.68)
594. phylum=Arthropoda 135660 ==> taxonRank=FAMILY 71924 <conf:(0.53)> lift:(1.58) lev:(0.12)
[26373] conv:(1.41)
595. taxonomicStatus=ACCEPTED hasCoordinate=false hasGeospatialIssues=false 131640 ==>
taxonRank=FAMILY 69624 <conf:(0.53)> lift:(1.58) lev:(0.11) [25422] conv:(1.41)
596. phylum=Chordata taxonRank=SPECIES hasGeospatialIssues=false 45214 ==>
type=PhysicalObject taxonomicStatus=ACCEPTED 23908 <conf:(0.53)> lift:(2.27) lev:(0.06) [13376]
conv:(1.63)
597. stateProvince=RJ 66709 ==> taxonomicStatus=ACCEPTED hasCoordinate=false 35268
<conf:(0.53)> lift:(0.89) lev:(-0.02) [-4302] conv:(0.86)
598. phylum=Arthropoda 135660 ==> taxonRank=FAMILY taxonomicStatus=ACCEPTED 71716
<conf:(0.53)> lift:(1.58) lev:(0.12) [26296] conv:(1.41)
599. hasCoordinate=true hasGeospatialIssues=false 59471 ==> preparations=Alcool 70 31394
<conf:(0.53)> lift:(2.27) lev:(0.08) [17550] conv:(1.63)
600. stateProvince=RJ taxonomicStatus=ACCEPTED hasGeospatialIssues=false 54335 ==>
phylum=Arthropoda 28619 <conf:(0.53)> lift:(0.87) lev:(-0.02) [-4141] conv:(0.84)
601. hasGeospatialIssues=false 211052 ==> phylum=Arthropoda hasCoordinate=false 111138
<conf:(0.53)> lift:(1.05) lev:(0.02) [5052] conv:(1.05)
602. phylum=Chordata 57290 ==> hasCoordinate=false 30150 <conf:(0.53)> lift:(0.77) lev:(-0.04) [-
8952] conv:(0.67)

603. *phylum*=Chordata 57290 ==> *hasCoordinate*=false *hasGeospatialIssues*=false 30141
<conf:(0.53)> lift:(0.78) lev:(-0.04) [-8454] conv:(0.69)
604. *taxonomicStatus*=ACCEPTED 200789 ==> *phylum*=Arthropoda *hasGeospatialIssues*=false 105509 <conf:(0.53)> lift:(0.96) lev:(-0.02) [-3969] conv:(0.96)
605. *taxonRank*=SPECIES *hasGeospatialIssues*=false 88292 ==> *hasCoordinate*=true 46260
<conf:(0.52)> lift:(1.65) lev:(0.08) [18229] conv:(1.43)
606. *taxonomicStatus*=ACCEPTED *hasCoordinate*=false 133466 ==> *taxonRank*=FAMILY 69846
<conf:(0.52)> lift:(1.56) lev:(0.11) [25031] conv:(1.39)
607. *phylum*=Chordata *taxonRank*=SPECIES *hasGeospatialIssues*=false 45214 ==>
hasCoordinate=false 23622 <conf:(0.52)> lift:(0.77) lev:(-0.03) [-7237] conv:(0.66)
608. *phylum*=Chordata *taxonRank*=SPECIES 45798 ==> *type*=PhysicalObject
taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 23908 <conf:(0.52)> lift:(2.84) lev:(0.07)
[15480] conv:(1.71)
609. *taxonomicStatus*=ACCEPTED *hasCoordinate*=false 133466 ==> *taxonRank*=FAMILY
hasGeospatialIssues=false 69624 <conf:(0.52)> lift:(1.63) lev:(0.12) [27028] conv:(1.42)
610. *preparations*=Alcool 70 *taxonomicStatus*=ACCEPTED 48751 ==> *taxonRank*=SPECIES
hasCoordinate=true 25417 <conf:(0.52)> lift:(2.3) lev:(0.06) [14373] conv:(1.62)
611. *stateProvince*=RJ 66709 ==> *taxonomicStatus*=ACCEPTED *hasCoordinate*=false
hasGeospatialIssues=false 34773 <conf:(0.52)> lift:(0.89) lev:(-0.02) [-4256] conv:(0.87)
612. *taxonomicStatus*=ACCEPTED *hasCoordinate*=true *hasGeospatialIssues*=false 56067 ==>
preparations=Alcool 70 29164 <conf:(0.52)> lift:(2.23) lev:(0.07) [16113] conv:(1.6)
613. *phylum*=Chordata *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 52902 ==>
hasCoordinate=false 27512 <conf:(0.52)> lift:(0.76) lev:(-0.04) [-8595] conv:(0.66)
614. *taxonRank*=SPECIES *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 80405 ==>
phylum=Chordata 41810 <conf:(0.52)> lift:(2.04) lev:(0.09) [21337] conv:(1.55)
615. *taxonRank*=SPECIES *hasCoordinate*=true *hasGeospatialIssues*=false 46260 ==> *type*=Collection
23951 <conf:(0.52)> lift:(2.21) lev:(0.06) [13088] conv:(1.59)
616. *phylum*=Chordata *taxonRank*=SPECIES 45798 ==> *hasCoordinate*=false 23631 <conf:(0.52)>
lift:(0.76) lev:(-0.03) [-7627] conv:(0.66)
617. *phylum*=Chordata *taxonRank*=SPECIES 45798 ==> *hasCoordinate*=false
hasGeospatialIssues=false 23622 <conf:(0.52)> lift:(0.77) lev:(-0.03) [-7231] conv:(0.67)
618. *hasCoordinate*=true *hasGeospatialIssues*=false 59471 ==> *type*=Collection 30667
<conf:(0.52)> lift:(2.2) lev:(0.07) [16703] conv:(1.58)
619. *type*=PhysicalObject 55597 ==> *phylum*=Chordata 28651 <conf:(0.52)> lift:(2.02) lev:(0.06)
[14494] conv:(1.54)
620. *taxonRank*=SPECIES *hasCoordinate*=true *hasGeospatialIssues*=false 46260 ==>
preparations=Alcool 70 23822 <conf:(0.51)> lift:(2.21) lev:(0.06) [13053] conv:(1.58)

621. type=PhysicalObject taxonomicStatus=ACCEPTED 52408 ==> *phylum*=Chordata 26959
<conf:(0.51)> lift:(2.02) lev:(0.06) [13614] conv:(1.53)
622. *phylum*=Chordata taxonomicStatus=ACCEPTED 53504 ==> hasCoordinate=false 27521
<conf:(0.51)> lift:(0.75) lev:(-0.04) [-8997] conv:(0.65)
623. *phylum*=Chordata taxonomicStatus=ACCEPTED 53504 ==> hasCoordinate=false
hasGeospatialIssues=false 27512 <conf:(0.51)> lift:(0.76) lev:(-0.04) [-8533] conv:(0.67)
624. taxonRank=SPECIES hasGeospatialIssues=false 88292 ==> *phylum*=Chordata 45214
<conf:(0.51)> lift:(2.01) lev:(0.1) [22732] conv:(1.53)
625. taxonomicStatus=ACCEPTED hasCoordinate=true hasGeospatialIssues=false 56067 ==>
type=Collection 28621 <conf:(0.51)> lift:(2.17) lev:(0.07) [15456] conv:(1.56)
626. taxonomicStatus=ACCEPTED hasCoordinate=false hasGeospatialIssues=false 131640 ==>
phylum=Arthropoda taxonRank=FAMILY 67137 <conf:(0.51)> lift:(1.6) lev:(0.11) [25056] conv:(1.39)
627. hasCoordinate=true 71431 ==> type=PhysicalObject taxonomicStatus=ACCEPTED 36148
<conf:(0.51)> lift:(2.17) lev:(0.09) [19510] conv:(1.55)
628. *phylum*=Chordata hasGeospatialIssues=false 56652 ==> type=Collection 28600 <conf:(0.5)>
lift:(2.15) lev:(0.07) [15297] conv:(1.55)
629. taxonomicStatus=ACCEPTED hasCoordinate=false 133466 ==> *phylum*=Arthropoda
taxonRank=FAMILY 67357 <conf:(0.5)> lift:(1.58) lev:(0.11) [24693] conv:(1.37)
630. type=PhysicalObject 55597 ==> *phylum*=Chordata hasGeospatialIssues=false 28052
<conf:(0.5)> lift:(2) lev:(0.06) [14053] conv:(1.51)
631. type=Collection hasGeospatialIssues=false 52583 ==> *phylum*=Chordata
taxonomicStatus=ACCEPTED 26507 <conf:(0.5)> lift:(2.12) lev:(0.06) [14003] conv:(1.54)
632. taxonomicStatus=ACCEPTED hasGeospatialIssues=false 187707 ==> *phylum*=Arthropoda
hasCoordinate=false 94584 <conf:(0.5)> lift:(1) lev:(0) [232] conv:(1)
633. *phylum*=Chordata taxonomicStatus=ACCEPTED 53504 ==> type=PhysicalObject 26959
<conf:(0.5)> lift:(2.04) lev:(0.06) [13738] conv:(1.52)
634. taxonRank=SPECIES taxonomicStatus=ACCEPTED 85664 ==> hasCoordinate=true
hasGeospatialIssues=false 43146 <conf:(0.5)> lift:(1.91) lev:(0.09) [20503] conv:(1.48)
635. type=PhysicalObject taxonomicStatus=ACCEPTED 52408 ==> *phylum*=Chordata
hasGeospatialIssues=false 26395 <conf:(0.5)> lift:(2) lev:(0.06) [13199] conv:(1.51)
636. taxonomicStatus=ACCEPTED hasCoordinate=false 133466 ==> *phylum*=Arthropoda
taxonRank=FAMILY hasGeospatialIssues=false 67137 <conf:(0.5)> lift:(1.66) lev:(0.12) [26672]
conv:(1.4)
637. preparations=Alcool 70 taxonomicStatus=ACCEPTED 48751 ==> type=PhysicalObject 24517
<conf:(0.5)> lift:(2.04) lev:(0.06) [12470] conv:(1.51)
638. preparations=Alcool 70 52374 ==> disposition=Na Colecao 26339 <conf:(0.5)> lift:(4.2)
lev:(0.09) [20071] conv:(1.77)

639. taxonRank=SPECIES 94136 ==> taxonomicStatus=ACCEPTED hasCoordinate=true 47338
<conf:(0.5)> lift:(1.68) lev:(0.09) [19171] conv:(1.41)

640. *phylum*=Arthropoda 135660 ==> taxonRank=FAMILY *hasGeospatialIssues*=false 68217
<conf:(0.5)> lift:(1.58) lev:(0.11) [24921] conv:(1.37)

641. *stateProvince*=RJ 66709 ==> *phylum*=Arthropoda *hasGeospatialIssues*=false 33543
<conf:(0.5)> lift:(0.92) lev:(-0.01) [-2829] conv:(0.91)

642. type=Collection 52831 ==> *phylum*=Chordata taxonomicStatus=ACCEPTED 26545 <conf:(0.5)>
lift:(2.11) lev:(0.06) [13982] conv:(1.53)

643. *phylum*=Arthropoda 135660 ==> taxonRank=FAMILY taxonomicStatus=ACCEPTED
hasGeospatialIssues=false 68128 <conf:(0.5)> lift:(1.58) lev:(0.11) [24892] conv:(1.37)

644. type=Collection 52831 ==> *phylum*=Chordata taxonomicStatus=ACCEPTED
hasGeospatialIssues=false 26507 <conf:(0.5)> lift:(2.13) lev:(0.06) [14085] conv:(1.54)

645. *phylum*=Chordata taxonomicStatus=ACCEPTED *hasGeospatialIssues*=false 52902 ==>
type=Collection 26507 <conf:(0.5)> lift:(2.13) lev:(0.06) [14085] conv:(1.53)

646. *stateProvince*=RJ 66709 ==> *phylum*=Arthropoda taxonomicStatus=ACCEPTED 33419
<conf:(0.5)> lift:(0.96) lev:(-0.01) [-1474] conv:(0.96)

647. *phylum*=Chordata 57290 ==> type=PhysicalObject 28651 <conf:(0.5)> lift:(2.02) lev:(0.06)
[14494] conv:(1.51)

APENDICE 14 RESULTADOS DA ANÁLISE DE CLASSIFICAÇÃO APRIORI NO ATRIBUTO “*phylum*”

Resultados classificação apriori *Phylum*

1. taxonRank=FAMILY taxonomicStatus=ACCEPTED hasCoordinate=false 69846 ==> *phylum*=Arthropoda 67357 conf:(0.96)
2. taxonRank=FAMILY hasCoordinate=false 69935 ==> *phylum*=Arthropoda 67438 conf:(0.96)
3. taxonRank=FAMILY taxonomicStatus=ACCEPTED hasCoordinate=false hasGeospatialIssues=false 69624 ==> *phylum*=Arthropoda 67137 conf:(0.96)
4. taxonRank=FAMILY hasCoordinate=false hasGeospatialIssues=false 69713 ==> *phylum*=Arthropoda 67218 conf:(0.96)
5. taxonRank=FAMILY 75549 ==> *phylum*=Arthropoda 71924 conf:(0.95)
6. taxonRank=FAMILY taxonomicStatus=ACCEPTED 75331 ==> *phylum*=Arthropoda 71716 conf:(0.95)
7. taxonRank=FAMILY taxonomicStatus=ACCEPTED hasGeospatialIssues=false 71709 ==> *phylum*=Arthropoda 68128 conf:(0.95)
8. taxonRank=FAMILY hasGeospatialIssues=false 71808 ==> *phylum*=Arthropoda 68217 conf:(0.95)
9. preparations=Alcool 70 disposition=Na Colecao hasGeospatialIssues=false 26087 ==> *phylum*=Mollusca 23420 conf:(0.9)
10. preparations=Alcool 70 disposition=Na Colecao 26339 ==> *phylum*=Mollusca 23629 conf:(0.9)
11. disposition=Na Colecao hasGeospatialIssues=false 26673 ==> *phylum*=Mollusca 23789 conf:(0.89)
12. disposition=Na Colecao 26927 ==> *phylum*=Mollusca 23998 conf:(0.89)
13. hasCoordinate=false 153570 ==> *phylum*=Arthropoda 113097 conf:(0.74)
14. hasCoordinate=false hasGeospatialIssues=false 151581 ==> *phylum*=Arthropoda 111138 conf:(0.73)
15. taxonomicStatus=ACCEPTED hasCoordinate=false 133466 ==> *phylum*=Arthropoda 96382 conf:(0.72)
16. stateProvince=RJ hasCoordinate=false 41004 ==> *phylum*=Arthropoda 29520 conf:(0.72)
17. taxonomicStatus=ACCEPTED hasCoordinate=false hasGeospatialIssues=false 131640 ==> *phylum*=Arthropoda 94584 conf:(0.72)

18. *stateProvince*=RJ *hasCoordinate*=false *hasGeospatialIssues*=false 40486 ==> *phylum*=Arthropoda 29012 conf:(0.72)
19. *stateProvince*=RJ *taxonomicStatus*=ACCEPTED *hasCoordinate*=false 35268 ==> *phylum*=Arthropoda 24835 conf:(0.7)
20. *stateProvince*=RJ *taxonomicStatus*=ACCEPTED *hasCoordinate*=false *hasGeospatialIssues*=false 34773 ==> *phylum*=Arthropoda 24350 conf:(0.7)
21. *taxonomicStatus*=ACCEPTED 200789 ==> *phylum*=Arthropoda 117690 conf:(0.59)
22. *hasGeospatialIssues*=false 211052 ==> *phylum*=Arthropoda 122680 conf:(0.58)
23. *stateProvince*=RJ 66709 ==> *phylum*=Arthropoda 38657 conf:(0.58)
24. *preparations*=Alcool 70 *hasGeospatialIssues*=false 41079 ==> *phylum*=Mollusca 23555 conf:(0.57)
25. *stateProvince*=RJ *taxonomicStatus*=ACCEPTED 59211 ==> *phylum*=Arthropoda 33419 conf:(0.56)
26. *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 187707 ==> *phylum*=Arthropoda 105509 conf:(0.56)
27. *taxonRank*=SPECIES *hasCoordinate*=false *hasGeospatialIssues*=false 42032 ==> *phylum*=Chordata 23622 conf:(0.56)
28. *taxonRank*=SPECIES *hasCoordinate*=false 43168 ==> *phylum*=Chordata 23631 conf:(0.55)
29. *stateProvince*=RJ *hasGeospatialIssues*=false 61516 ==> *phylum*=Arthropoda 33543 conf:(0.55)
30. *stateProvince*=RJ *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 54335 ==> *phylum*=Arthropoda 28619 conf:(0.53)
31. *taxonRank*=SPECIES *taxonomicStatus*=ACCEPTED *hasGeospatialIssues*=false 80405 ==> *phylum*=Chordata 41810 conf:(0.52)
32. *taxonRank*=SPECIES *hasGeospatialIssues*=false 88292 ==> *phylum*=Chordata 45214 [conf:\(0.51\)](#)

APENDICE 15 – RESULTADOS PART “Ano codificado”

Matrix de confusão

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s <--	classified as
3593	4540	1871	70	1334	0	530	74	0	0	0	0	0	0	0	0	0	0	0	a = 1981-90
1221	17998	1357	19	840	0	535	31	0	0	0	0	0	0	0	0	0	0	0	b = 2001-10
2144	7097	5034	57	1178	0	291	119	0	0	0	0	0	0	0	0	0	0	0	c = 1991-200
570	1842	513	125	401	0	518	196	0	0	0	0	0	0	0	0	0	0	0	d = 1971-80
1170	1074	176	36	3471	0	593	80	0	0	0	0	0	0	0	0	0	0	0	e = 1941-50
198	2161	90	2	1	0	89	4	0	0	0	0	0	0	0	0	0	0	0	f = 2011-17
635	1955	371	68	624	0	1646	118	0	0	0	0	0	0	0	0	0	0	0	g = 1951-60
386	1537	566	69	241	0	642	282	0	0	0	0	0	0	0	0	0	0	0	h = 1961-70
92	82	58	6	101	0	24	8	0	0	0	0	0	0	0	0	0	0	0	i = 1931-40
50	51	21	1	146	0	18	1	0	0	0	0	0	0	0	0	0	0	0	j = 1911-20
57	134	35	0	73	0	6	0	0	0	0	0	0	0	0	0	0	0	0	k = 1901-10
16	43	11	0	24	0	9	4	0	0	0	0	0	0	0	0	0	0	0	l = 1921-30
2	26	0	0	27	0	22	1	0	0	0	0	0	0	0	0	0	0	0	n = 1871-80
1	3	1	0	5	0	2	0	0	0	0	0	0	0	0	0	0	0	0	o = 1891-190
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	p = 1821-30
1	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	q = 1881-90
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	r = 1811-20
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	s = 1851-60
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	s = 1861-70

Lista de regras

1. `scientificName` = Amphipoda: 1981-90 (661.0/374.0)
2. `scientificName` = Characidium Reinhardt 1867: 2001-10 (650.0/418.0)
3. `scientificName` = Astyanax Baird & Girard 1854: 2001-10 (631.0/438.0)
4. `genus` = Astyanax: 2001-10 (1507.69/1046.58)
5. `genus` = Hypostomus: 2001-10 (940.35/606.09)
6. `genus` = Characidium: 1941-50 (872.23/539.37)
7. `genus` = Geophagus: 2001-10 (728.16/481.82)
8. `class` = Polychaeta AND `hasCoordinate` = false: 1991-2000 (984.01/29.87)
9. `class` = Polychaeta AND `taxonRank` = SPECIES: 2001-10 (736.17/384.7)
10. `class` = Polychaeta: 2001-10 (1198.42/568.05)
11. `family` = Characidae AND `hasCoordinate` = true: 2001-10 (1881.29/837.11)
12. `family` = Loricariidae AND `hasCoordinate` = true: 2001-10 (1431.59/917.2)
13. `family` = Gonyleptidae: 2001-10 (2418.11/1472.29)
14. `phylum` = Arthropoda AND `hasGeospatialIssues` = false AND `class` = Malacostraca AND `stateProvince` = RJ: 2001-10 (3561.39/1387.14)
15. `phylum` = Echinodermata AND `class` = Asteroidea: 1981-90 (1181.82/790.7)
16. `class` = Arachnida: 2001-10 (1190.92/663.21)
17. `phylum` = Arthropoda AND `order` = Amphipoda AND `hasGeospatialIssues` = true: 1991-2000 (1220.14/490.33)
18. `phylum` = Echinodermata AND `class` = Ophiuroidea: 2001-10 (1085.69/683.86)
19. `phylum` = Arthropoda AND `class` = Malacostraca AND `stateProvince` = BA: 1991-2000 (1263.88/396.05)

20. *phylum* = Chordata AND *hasCoordinate* = true AND *stateProvince* = MT: 2001-10 (1205.05/304.83)
21. *phylum* = Chordata AND *hasCoordinate* = true AND *stateProvince* = GO: 1991-2000 (1117.07/449.15)
22. *phylum* = Arthropoda AND *class* = Maxillopoda: 1981-90 (957.19/488.87)
23. *phylum* = Chordata AND *hasCoordinate* = true AND *order* = Siluriformes: 2001-10 (1243.99/730.11)
24. *phylum* = Chordata AND *stateProvince* = MG: 1981-90 (1536.89/883.41)
25. *phylum* = Arthropoda AND *order* = Decapoda AND *taxonRank* = SPECIES AND *georeferenceProtocol* = Convertido dos dados originais : 1981-90 (811.48/555.02)
26. *phylum* = Arthropoda AND *order* = Decapoda AND *taxonRank* = SPECIES: 1991-2000 (796.5/519.82)
27. *phylum* = Arthropoda AND *order* = Isopoda: 1971-80 (2043.3/1583.15)
28. *phylum* = Bryozoa: 2001-10 (771.6/163.5)
29. *phylum* = Arthropoda AND *hasGeospatialIssues* = false AND *georeferenceProtocol* = GOOGLE EARTH: 1991-2000 (726.92/411.05)
30. *phylum* = Arthropoda AND *hasGeospatialIssues* = true: 1991-2000 (2067.32/1024.84)
31. *phylum* = Chordata AND *stateProvince* = ES: 1941-50 (1452.54/511.94)
32. *phylum* = Mollusca AND *dynamicProperties* = Convenio: 2001-10 (2943.9/814.56)
33. *phylum* = Chordata AND *stateProvince* = BA: 1941-50 (973.04/582.89)
34. *phylum* = Echinodermata: 2001-10 (1135.21/837.21)
35. *phylum* = Chordata AND *stateProvince* = RJ AND *order* = Perciformes AND *hasCoordinate* = false: 1941-50 (868.16/600.4)
36. *phylum* = Chordata AND *stateProvince* = AM: 1941-50 (652.98/249.08)
37. *phylum* = Arthropoda: 2001-10 (645.83/168.39)
38. *class* = Actinopterygii AND *stateProvince* = RJ AND *order* = Cyprinodontiformes: 2001-10 (630.71/323.57)
39. *class* = Actinopterygii AND *stateProvince* = RJ AND *order* = Siluriformes: 1941-50 (625.79/435.18)
40. *class* = Actinopterygii AND *stateProvince* = RJ AND *hasCoordinate* = true: 1981-90 (1184.8/707.63)

- 41.class = Actinopterygii AND *stateProvince* = SP: 1941-50 (1248.25/771.24)
- 42.class = Actinopterygii AND order = Characiformes: 1981-90 (1292.98/842.23)
- 43.class = Actinopterygii AND order = Perciformes: 1941-50 (1068.33/711.81)
- 44.class = Actinopterygii: 1991-2000 (2368.23/1715.83)
- 45.family = Muricidae: 1951-60 (691.3/527.04)
- 46.family = Veneridae: 1951-60 (661.91/408.66)
- 47.family = Bulimulidae: 1951-60 (650.76/525.51)
- 48.class = Gastropoda AND georeferenceSources = ORIGINAL: 2001-10
(2705.51/1242.02)
- 49.class = Bivalvia AND dynamicProperties = Doacao: 1951-60 (2104.96/1364.17)
- 50.class = Bivalvia: 2001-10 (1511.38/1144.65)
- 51.order = Neogastropoda AND dynamicProperties = Doacao: 1951-60
(1633.81/1209.97)
- 52.order = Littorinimorpha AND dynamicProperties = Doacao: 1951-60
(1144.27/867.61)
- 53.order = Neogastropoda: 1981-90 (1095.64/824.65)
- 54.order = Stylommatophora AND dynamicProperties = Coleta: 2001-10
(772.82/544.06)
- 55.order = Stylommatophora AND taxonRank = GENUS: 1951-60 (771.74/618.17)
- 56.order = Littorinimorpha: 2001-10 (772.42/581.17)
- 57.*phylum* = Mollusca: 2001-10 (3860.75/2379.71)
- 58.: 1961-70 (665.82/485.74)

APENDICE 16 – RESULTADOS PART DO ATRIBUTO “*phylum*”

Matrix de confusão

a	b	c	d	e	f	g	h	i	j	k	l	m	<--	classifiedas
731	13	91	0	0	0	0	0	0	0	0	0	0	0	a=Bryozoa
0	135653	7	0	0	0	0	0	0	0	0	0	0	0	b=Arthropoda
0	28	56971	291	0	0	0	0	0	0	0	0	0	0	c=Chordata
0	2	268	23869	0	0	0	0	0	0	0	0	0	0	d=Mollusca
0	0	6	0	0	0	0	0	0	0	0	0	0	0	e=Tardigrada
0	171	12	0	0	0	0	0	0	0	0	0	0	0	f=Platyhelminthes
0	2	0	0	0	0	3171	0	0	0	0	0	0	0	g=Annelida
0	15	67	0	0	0	0	0	0	0	0	0	0	0	h=Sipuncula
0	27	2	0	0	0	0	0	0	0	0	0	0	0	i=Nemertea
0	0	32	0	0	0	0	0	0	0	0	0	0	0	j=Nematomorpha
0	213	66	0	0	0	0	0	0	0	3131	0	0	0	k=Echinodermata
0	17	16	0	0	0	0	0	0	0	0	0	0	0	l=Brachiopoda
0	96	1	0	0	0	0	0	0	0	0	0	0	0	m=Nematoda

1. class = Aves: Chordata (28763.4/94.1)
2. class = Actinopterygii: Chordata (27876.7/91.2)
3. class = Malacostraca AND taxonRank = SPECIES AND order = Decapoda:
Arthropoda (4762.54)
4. class = Malacostraca AND taxonRank = FAMILY: Arthropoda (4385.6/0.6)
5. class = Malacostraca AND taxonRank = SPECIES: Arthropoda (4516.77/7.3)
6. class = Malacostraca AND taxonRank = GENUS: Arthropoda (3228.3/0.3)
7. class = Insecta AND taxonRank = FAMILY: Arthropoda (65006.93/3.93)
8. class = Insecta AND taxonRank = UNRANKED: Arthropoda (12250.0)
9. class = Insecta AND taxonRank = GENUS: Arthropoda (8509.08/2.08)
10. class = Insecta AND taxonRank = SPECIES: Arthropoda (4787.79/17.79)
11. class = Malacostraca AND scientificName = Amphipoda: Arthropoda (858.0)
12. class = Insecta AND taxonRank = SUBSPECIES: Arthropoda (2817.0)
13. class = Arachnida AND taxonRank = GENUS: Arthropoda (6919.73/1.73)
14. class = Arachnida AND taxonRank = SPECIES AND order = Araneae:
Arthropoda (4484.0)
15. class = Insecta: Arthropoda (3225.43/44.1)
16. class = Arachnida AND taxonRank = SPECIES: Arthropoda (2627.46/10.46)
17. class = Gastropoda AND taxonRank = SPECIES AND order = Neogastropoda:
Mollusca (4920.68)
18. class = Gastropoda AND taxonRank = SPECIES: Mollusca (9146.34/27.02)
19. class = Bivalvia: Mollusca (5561.35/163.2)

- 20.class = Gastropoda: Mollusca (3727.48/109.38)
- 21.class = Arachnida: Arthropoda (2722.98/57.3)
- 22.class = Malacostraca: Arthropoda (1549.51/32.6)
- 23.class = Polychaeta AND taxonRank = SPECIES: Annelida (1527.76/5.76)
- 24.class = Maxillopoda: Arthropoda (1512.92/35.53)
- 25.class = Polychaeta AND taxonRank = GENUS: Annelida (911.29/0.29)
- 26.class = Asteroidea: Echinodermata (1229.96/44.96)
- 27.class = Ophiuroidea: Echinodermata (1128.24/41.24)
- 28.class = Echinoidea: Echinodermata (891.59/32.59)
- 29.type = Collection AND hasCoordinate = true: Mollusca (1019.45/206.45)
- 30.class = Polychaeta: Annelida (769.36/31.32)
- 31.class = Gymnolaemata: Bryozoa (762.06/29.21)
- 32.hasCoordinate = true: Arthropoda (1751.39/258.56)
- 33.: Chordata (817.9/430.69)

APENDICE 17 – RESULTADOS PART DO ATRIBUTO “stateProvince”

Matrix de confusão

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	<-	classified as
5942		0	703	9	263	0	0	0	0	0	0	0	411	712	4003	0	0	0	0	285	481	0	0	0	0	0	0	a=RJ
2526		0	45	0	50	0	0	0	0	0	0	0	106	38	272	0	0	0	0	7	68	0	0	0	0	0	0	b=PR
9221		0	1272	8	140	0	0	0	0	0	0	0	86	445	1079	0	0	0	0	18	222	0	0	0	0	0	0	c=ES
976		0	93	12	144	0	0	0	0	0	0	0	1	20	11	0	0	0	0	0	11	0	0	0	0	0	0	d=RN
9232		0	226	22	499	0	0	0	0	0	0	0	270	425	1118	0	0	0	0	134	189	0	0	0	0	0	0	e=BA
243		0	0	0	3	0	0	0	0	0	0	0	7	7	5	0	0	0	0	0	22	0	0	0	0	0	0	f=SE
992		0	38	2	4	0	0	0	0	0	0	0	17	65	78	0	0	0	0	7	67	0	0	0	0	0	0	g=AL
1958		0	148	1	34	0	0	0	0	0	0	0	442	167	1126	0	0	0	0	50	162	0	0	0	0	0	0	h=SP
366		0	18	0	57	0	0	0	0	0	0	0	8	50	322	0	0	0	0	0	18	0	0	0	0	0	0	i=PE
5891		0	110	1	13	0	0	0	0	0	0	0	114	282	941	0	0	0	0	22	44	0	0	0	0	0	0	j=SC
674		0	3	2	45	0	0	0	0	0	0	0	10	142	154	0	0	0	0	0	17	0	0	0	0	0	0	k=PB
490		0	9	0	3	0	0	0	0	0	0	0	10	4	102	0	0	0	0	5	12	0	0	0	0	0	0	l=DF
9111		0	144	0	78	0	0	0	0	0	0	0	842	243	2010	0	0	0	0	146	193	0	0	0	0	0	0	m=MG
8285		0	153	0	34	0	0	0	0	0	0	0	100	1093	4549	0	0	0	0	30	228	0	0	0	0	0	0	n=PA
6353		0	155	0	55	0	0	0	0	0	0	0	202	412	8726	0	0	0	0	43	193	0	0	0	0	0	0	o=AM
634		0	29	0	3	0	0	0	0	0	0	0	107	119	561	0	0	0	0	0	27	0	0	0	0	0	0	p=AP
2204		0	63	6	121	0	0	0	0	0	0	0	68	61	75	0	0	0	0	1	63	0	0	0	0	0	0	q=CE
906		0	112	0	11	0	0	0	0	0	0	0	21	11	176	0	0	0	0	8	37	0	0	0	0	0	0	r=MS
2649		0	37	0	63	0	0	0	0	0	0	0	46	30	451	0	0	0	0	43	50	0	0	0	0	0	0	s=RS
3672		0	230	0	8	0	0	0	0	0	0	0	405	204	1038	0	0	0	0	592	167	0	0	0	0	0	0	t=GO
6578		0	421	0	25	0	0	0	0	0	0	0	443	382	2078	0	0	0	0	273	1305	0	0	0	0	0	0	u=MT
542		0	38	0	4	0	0	0	0	0	0	0	142	89	1852	0	0	0	0	5	57	0	0	0	0	0	0	v=RO
593		0	79	3	8	0	0	0	0	0	0	0	9	30	46	0	0	0	0	3	20	0	0	0	0	0	0	w=MA
491		0	13	0	16	0	0	0	0	0	0	0	2	29	553	0	0	0	0	0	3	0	0	0	0	0	0	x=AC
239		0	9	0	1	0	0	0	0	0	0	0	13	1	19	0	0	0	0	4	6	0	0	0	0	0	0	y=PI
362		0	21	0	5	0	0	0	0	0	0	0	10	48	35	0	0	0	0	2	33	0	0	0	0	0	0	z=TO
480		0	7	0	4	0	0	0	0	0	0	0	18	142	19	0	0	0	0	1	18	0	0	0	0	0	0	aa=RR

1. scientificName = Chrysomelidae: RJ (14326.0/10028.0)
2. scientificName = Acrididae: AM (9205.0/7088.0)
3. scientificName = Nymphalidae: AM (7295.0/5523.0)
4. scientificName = Riodinidae: AM (5290.0/3951.0)
5. scientificName = Romaleidae: AM (4567.0/2256.0)
6. scientificName = Tachinidae: RJ (4442.0/2483.0)
7. scientificName = Stratiomyidae: RJ (2831.0/1446.0)
8. scientificName = Orthoptera: RJ (2311.0/1377.0)
9. scientificName = Buprestidae: PA (1929.0/1593.0)
10. scientificName = Pieridae: RJ (1914.0/1201.0)
11. scientificName = Lycaenidae: RJ (1369.0/741.0)
12. scientificName = Tettigoniidae: RJ (1015.0/326.0)
13. scientificName = Sarcophagidae: RJ (992.0/588.0)
14. scientificName = Charidotis Boheman 1854: RJ (877.0/648.0)
15. scientificName = Amphipoda: RJ (777.0/263.0)
16. scientificName = Blaberidae: RJ (693.0/423.0)
17. scientificName = Miridae: MG (686.0/565.0)
18. scientificName = Cetoniidae: RJ (665.0/488.0)
19. scientificName = Characidium Reinhardt 1867: RJ (651.0/496.0)
20. scientificName = Astyanax Baird & Girard 1854: MG (640.0/523.0)

- 21.scientificName = Calliphoridae: RJ (607.0/324.0)
- 22.scientificName = Simuliidae: RJ (569.0/119.0)
- 23.scientificName = Geophagus brasiliensis Quoy & Gaimard 1824 : RJ (554.0/269.0)
- 24.scientificName = Chelymorpha Chevrolat 1836: RJ (550.0/444.0)
- 25.genus = Astyanax: RJ (1562.86/975.84)
- 26.genus = Hypostomus: RJ (990.75/796.17)
- 27.genus = Characidium: RJ (914.02/650.58)
- 28.genus = Thamnophilus: RJ (785.38/635.21)
- 29.genus = Amazilia: SP (611.6/456.34)
- 30.genus = Prepona: PA (592.42/410.03)
- 31.genus = Hoplias: RJ (587.91/454.82)
- 32.genus = Trichomycterus: RJ (581.14/331.33)
- 33.genus = Eusarcus: RJ (576.62/384.0)
- 34.family = Gonyleptidae AND taxonRank = GENUS: RJ (2821.52/1778.22)
- 35.family = Gonyleptidae AND taxonRank = SPECIES: RJ (1446.98/922.71)
- 36.family = Lycaenidae: RJ (2250.94/1475.06)
- 37.family = Nymphalidae AND taxonRank = SPECIES: RJ (1612.08/1016.1)
- 38.family = Heptapteridae: RJ (995.22/704.1)
- 39.family = Salticidae: BA (948.63/602.57)
- 40.family = Poeciliidae: RJ (932.84/423.56)
- 41.family = Nymphalidae AND taxonRank = SUBSPECIES: RJ (892.07/670.07)
- 42.family = Sclerosomatidae: ES (884.13/644.38)
- 43.family = Veneridae: RJ (883.77/306.82)
- 44.family = Muricidae: RJ (786.48/357.63)
- 45.family = Limoniidae: RJ (783.22/431.58)
- 46.family = Ropalomeridae: ES (744.57/477.23)
- 47.family = Carabidae: RJ (730.16/499.94)
- 48.family = Pyramidellidae: RJ (704.55/349.63)
- 49.family = Eunicidae: BA (668.49/508.59)
- 50.family = Columbelloidae: RJ (657.0/358.9)
- 51.family = Anostomidae: MG (646.86/506.31)
- 52.family = Tachinidae: RJ (628.77/372.7)
- 53.family = Balanidae: RJ (614.44/178.52)

- 54. family = Loricariidae AND ano codificado = 1991-2000: RJ (602.38/369.05)
- 55. family = Pieridae: RJ (1830.93/1112.36)
- 56. family = Loricariidae AND ano codificado = 2001-10: RJ (596.05/423.09)
- 57. family = Loricariidae: RJ (1265.03/775.74)
- 58. family = Characidae AND ano codificado = 2001-10: MT (1096.6/467.15)
- 59. family = Acrididae: AM (1911.86/1430.11)
- 60. family = Araneidae: RJ (1185.9/830.88)
- 61. family = Thraupidae: RJ (964.15/588.82)
- 62. family = Psittacidae AND hasCoordinate = false: RJ (946.86/548.47)
- 63. family = Bulimulidae: RJ (755.78/582.27)
- 64. family = Characidae AND ano codificado = 1981-90: GO (700.11/497.53)
- 65. family = Trochilidae: RJ (2562.09/2031.08)
- 66. family = Chrysomelidae AND taxonRank = UNRANKED: RJ (1845.03/1232.03)
- 67. family = Tyrannidae: RJ (1332.24/1035.51)
- 68. family = Cosmetidae: AM (871.89/724.57)
- 69. family = Cichlidae AND hasCoordinate = false: AM (859.11/681.8)
- 70. family = Palaemonidae: RJ (698.42/447.76)
- 71. family = Ramphastidae: MT (691.99/555.84)
- 72. family = Cichlidae: MT (690.08/481.44)
- 73. family = Accipitridae: RJ (650.39/519.27)
- 74. family = Characidae AND hasCoordinate = false: MG (606.42/510.08)
- 75. family = Thamnophilidae AND hasCoordinate = true: RJ (1667.31/1391.2)
- 76. family = Furnariidae: RJ (3272.83/2569.73)
- 77. family = Chrysomelidae: SC (1625.1/1385.11)
- 78. family = Thamnophilidae: PA (1406.85/1092.47)
- 79. family = Pipridae: PA (966.77/711.89)
- 80. family = Cotingidae: RJ (829.73/661.05)
- 81. family = Psittacidae: ES (734.63/615.8)
- 82. family = Picidae AND hasCoordinate = true: ES (693.0/589.96)
- 83. family = Picidae: MT (661.89/543.7)
- 84. family = Tinamidae: MT (611.97/499.45)
- 85. family = Philosciidae: RJ (596.38/357.86)
- 86. family = Triphoridae: ES (587.1/335.44)
- 87. family = Nymphalidae: AM (584.03/408.79)

- 88.family = Cuculidae: RJ (576.51/425.03)
- 89.family = Thomisidae: RJ (575.7/376.07)
- 90.family = Caprimulgidae: MT (574.61/416.22)
- 91.family = Columbidae: RJ (569.62/384.91)
- 92.family = Callichthyidae: RJ (568.88/404.45)
- 93.family = Pimelodidae: MG (548.56/442.34)
- 94.family = Olividae: RJ (546.12/303.85)
- 95.family = Gonyleptidae: ES (532.07/375.49)
- 96.family = Mytilidae: RJ (518.45/187.5)
- 97.order = Cardiida: RJ (1078.35/390.38)
- 98.order = Tanaidacea: RJ (994.05/480.4)
- 99.order = Trochida: RJ (890.88/470.0)
- 100. order = Amphipoda AND ano codificado = 2001-10: RJ (851.76/229.75)
- 101. order = Amphipoda AND taxonRank = SPECIES: RJ (951.34/447.25)
- 102. order = Decapoda: RJ (4840.53/2425.48)
- 103. order = Perciformes: RJ (3435.64/2078.09)
- 104. order = Littorinimorpha: RJ (2882.87/1656.8)
- 105. order = Isopoda: RJ (2796.6/1386.75)
- 106. order = Neogastropoda AND georeferenceSources = GEOLOCATE: RJ
(1892.82/1298.51)
- 107. order = Araneae AND taxonRank = SPECIES: RJ (2794.98/1858.2)
- 108. order = Araneae AND taxonRank = GENUS: RJ (1490.42/978.77)
- 109. order = Diptera: RJ (3209.54/2013.79)
- 110. order = Coleoptera AND taxonomicStatus = SYNONYM: RJ
(2005.88/1422.21)
- 111. order = Hymenoptera: RJ (1414.28/1150.12)
- 112. order = Amphipoda: RJ (1389.87/807.54)
- 113. order = Characiformes AND ano codificado = 1991-2000: GO
(1089.87/671.66)
- 114. order = Blattodea: RJ (1421.28/1134.35)
- 115. order = Araneae: RJ (965.69/642.63)
- 116. order = Neogastropoda: ES (885.03/538.96)
- 117. order = Dermaptera: RJ (868.26/479.94)
- 118. order = Cephalaspidea: RJ (836.62/408.73)

119. order = Coleoptera: RJ (779.81/633.72)
120. order = Lepidoptera: RJ (753.73/579.97)
121. order = Amphilepidida: RJ (752.66/431.13)
122. order = Cheilostomatida: RJ (708.42/282.74)
123. order = Lepetellida: RJ (659.65/289.27)
124. order = Gymnotiformes: RJ (646.82/529.72)
125. order = Characiformes AND ano codificado = 2001-10: MT
(637.26/316.5)
126. order = Stylommatophora: RJ (2718.39/1873.28)
127. order = Orthoptera: RJ (2223.87/1697.47)
128. order = Passeriformes: RJ (2125.71/1460.52)
129. order = Characiformes: MG (1238.38/1005.28)
130. order = Piciformes: PA (900.45/728.4)
131. order = Siluriformes: RJ (891.75/701.79)
132. order = Scorpiones: RJ (626.3/508.43)
133. order = Pelecaniformes: RJ (621.9/406.34)
134. order = Arcida: RJ (611.88/301.22)
135. order = Hemiptera: RJ (609.81/398.72)
136. order = Phyllodocida: RJ (574.59/400.07)
137. order = Unionida: RJ (568.33/434.28)
138. class = Asteroidea: RJ (1169.16/261.17)
139. class = Bivalvia AND dynamicProperties = Doacao: RJ (962.59/523.85)
140. : RJ (17694.17/9803.24)

APENDICE 18 – MATRIZES DE CONFUSÃO DOS RESULTADOS J48

A) Matrix de confusão ano codificado

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	<-	classified
5943	3004	1334	339	452		83	645	182	4	8	2	16	0	0	0	0	0	0	0	a=1981-90
1868	16853	1866	203	359		297	407	135	5	2	5	0	1	0	0	0	0	0	0	b=2001-10
2202	4280	8168	159	359		148	464	121	2	5	4	7	1	0	0	0	0	0	0	c=1991-20
759	903	410	972	178		47	631	260	1	4	0	0	0	0	0	0	0	0	0	d=1971-80
657	1136	295	177	3548		25	629	110	8	8	5	1	1	0	0	0	0	0	0	e=1941-50
139	902	226	45	9	1083		115	26	0	0	0	0	0	0	0	0	0	0	0	f=2011-17
677	899	419	287	432		30	2429	225	8	4	4	2	0	1	0	0	0	0	0	g=1951-60
502	620	423	345	168		27	792	839	3	0	2	1	0	0	0	1	0	0	0	h=1961-70
62	90	18	23	72		1	28	30	42	1	0	4	0	0	0	0	0	0	0	i=1931-40
58	60	13	5	102		0	27	3	2	13	0	4	1	0	0	0	0	0	0	j=1911-20
77	93	31	13	37		3	17	6	4	0	15	9	0	0	0	0	0	0	0	k=1901-10
25	18	5	7	13		0	9	12	0	1	1	16	0	0	0	0	0	0	0	l=1921-30
7	12	6	4	24		0	24	0	0	1	0	0	0	0	0	0	0	0	0	m=1821-30
1	4	1	0	4		0	2	0	0	0	0	0	0	0	0	0	0	0	0	n=1891-19
0	0	0	0	1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	o=1821-30
2	0	1	0	0		0	0	2	0	0	0	1	0	0	0	0	0	0	0	p=1881-90
0	1	0	0	0		0	0	1	0	0	0	0	0	0	0	0	0	0	0	q=1811-20
1	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	r=1851-60
0	0	0	0	0		0	0	1	0	0	0	0	0	0	0	0	0	0	0	s=1861-70

B) Matrix de confusão stateProvince

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	<-	classified	
55790	114	1992	183	1276	1	93	898	148	212	25	3	778	600	3527	40	72	39	181	144	511	54	7	11	2	5	3		a=RI	
1778	417	99	4	101	1	5	270	9	165	5	0	145	53	250	3	3	3	70	83	45	3	0	0	0	0	0		b=PR	
7754	19	5373	101	771	2	28	227	122	61	2	11	146	445	855	19	28	15	22	103	312	60	2	5	0	2	6		c=ES	
766	0	385	680	176	2	7	12	37	11	0	5	9	12	9	3	57	5	3	1	24	0	2	0	0	1	0		d=HN	
5705	40	1192	146	5207	4	57	148	148	38	29	9	550	250	1040	19	154	12	12	65	282	43	7	1	2	6	9		e=BA	
130	1	11	4	41	25	3	7	2	0	1	0	36	9	5	0	0	3	1	2	6	0	0	0	0	0	0		f=SE	
920	1	157	38	286	4	283	22	27	5	7	1	53	85	87	14	37	6	2	11	104	9	3	1	1	5	11		g=AL	
7802	129	377	25	235	2	8	2907	35	169	16	5	487	214	1006	6	20	6	68	142	195	26	2	3	2	1	0		h=SP	
1694	13	328	60	442	5	28	53	484	7	27	1	36	64	314	1	28	2	8	6	33	1	2	0	0	0	1		i=PE	
4111	70	174	15	62	0	6	232	14	1099	12	0	176	302	914	5	3	1	81	85	47	1	2	6	0	0	0		j=SC	
312	2	26	16	109	4	6	8	35	1	128	2	9	135	146	0	56	0	5	3	41	0	1	0	1	1	0		k=PB	
324	2	16	0	22	0	1	20	0	0	2	68	23	15	87	0	0	2	2	29	22	0	0	0	0	0	0		l=DF	
6334	40	395	4	354	0	21	553	16	92	4	7	3273	230	1787	3	28	24	31	221	331	9	3	3	1	3	0		m=MG	
3633	11	236	14	163	0	34	43	32	32	14	4	121	4130	4916	118	42	24	5	167	576	68	17	20	1	9	22		n=PA	
2774	4	157	1	63	0	5	67	4	23	0	4	115	1107	10930	103	6	17	6	95	476	87	8	70	1	1	15		o=AM	
668	3	56	9	51	0	5	11	5	13	0	0	55	301	697	293	3	1	0	39	120	10	10	0	1	2	7		p=AP	
1141	5	116	61	288	3	31	31	35	3	18	0	39	76	77	5	564	4	2	27	118	3	7	1	2	0	5		q=CE	
388	3	40	5	41	0	2	9	6	0	1	0	66	49	177	4	2	185	9	51	237	4	0	0	0	2	1		r=MS	
1524	63	59	7	55	0	4	135	10	105	1	0	124	42	419	0	2	14	707	37	60	0	0	0	1	0	0		s=RS	
1696	17	144	6	95	1	6	163	4	21	0	8	396	181	949	23	21	25	5	1913	616	14	0	1	1	9	1		t=GO	
2993	13	221	3	161	0	22	118	22	24	3	3	447	700	2063	41	34	105	3	20	541	3838	81	1	15	2	17		u=MT	
585	0	55	1	34	0	10	27	0	3	0	1	99	326	1933	8	2	2	1	62	252	305	2	11	2	1	1	8		v=RO
292	0	48	7	72	0	9	5	12	2	2	0	12	82	50	7	21	4	0	37	75	8	37	1	5	3	0		w=MA	
289	2	12	0	9	0	1	9	1	1	0	0	4	54	571	5	2	0	2	5	22	2	1	114	0	0	1		x=AC	
106	1	6	12	31	0	2	1	2	2	0	1	8	5	31	1	1	2	0	30	6	2	2	0	40	0	0		y=PI	
113	6	26	0	40	0	6	3	5	0	0	0	17	78	21	1	7	6	2	29	95	0	1	2	1	56	1		z=TO	
174	0	20	0	30	0	0	0	7	2	2	0	0	4	138	83	19	5	1	0	8	52	12	2	0	0	5	125		aa=RR

C) Matrix de confusão phylum

a	b	c	d	e	f	g	h	i	j	k	l	m	<--	classified
778	0	57	0	0	0	0	0	0	0	0	0	0		a=Bryozoa
0	135351	309	0	0	0	0	0	0	0	0	0	0		b=Arthrop
0	0	57290	0	0	0	0	0	0	0	0	0	0		c=Chordat
0	3	0	24136	0	0	0	0	0	0	0	0	0		d=Mollusc
0	0	6	0	0	0	0	0	0	0	0	0	0		e=Tardigr
0	29	154	0	0	0	0	0	0	0	0	0	0		f=Platyhel
0	1	1	0	0	0	3171	0	0	0	0	0	0		g=Annelid
0	0	80	0	0	0	0	2	0	0	0	0	0		h=Sipuncu
0	0	29	0	0	0	0	0	0	0	0	0	0		i=Nemert
0	0	6	0	0	0	0	0	0	26	0	0	0		j=Nemato
0	0	0	0	0	0	0	0	0	0	3410	0	0		k=Echinod
0	1	27	0	0	0	0	0	0	0	0	5	0		l=Brachiop
0	1	35	0	0	0	0	0	0	0	0	0	61		m=Nemat