

Franklin Alves de Oliveira

**Visualização de coleções científicas digitais de
biodiversidade: um *framework* em Altair,
Python**

Rio de Janeiro, Brasil

2021

Franklin Alves de Oliveira

Visualização de coleções científicas digitais de biodiversidade: um *framework* em Altair, Python

Dissertação apresentada à Escola de Matemática Aplicada da Fundação Getúlio Vargas para a obtenção do Título de Mestre em Modelagem Matemática, na Área de Análise da Informação.

Fundação Getúlio Vargas - FGV

Escola de Matemática Aplicada - EMap

Programa de Pós-Graduação em Modelagem Matemática

Orientador(a): Asla Medeiros e Sá

Rio de Janeiro, Brasil

2021

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pelo Sistema de Bibliotecas/FGV

Oliveira, Franklin Alves de

Visualização de coleções científicas digitais de biodiversidade: um framework em altair, python / Franklin Alves de Oliveira – 2021.
128 f.

Dissertação (mestrado) - Fundação Getulio Vargas, Escola de Matemática Aplicada.

Orientador: Asla Medeiros e Sá.

Inclui bibliografia.

1. Visualização da informação. 2. História natural - Catálogos e coleções. 3. Preservação digital. 4. Interação homem-máquina. I. Sá, Asla Medeiros. II. Fundação Getulio Vargas. Escola de Matemática Aplicada. III. Título.

CDD – 001.4226

Elaborada por Kelly Ayala – CRB-7/7007

FRANKLIN ALVES DE OLIVEIRA

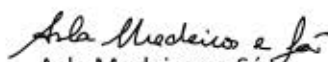
“VISUALIZAÇÃO DE COLEÇÕES CIENTÍFICAS DIGITAIS DE BIODIVERSIDADE: UM FRAMEWORK EM ALTAIR, PYTHON”.

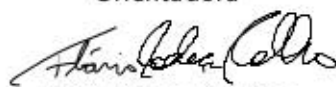
Dissertação apresentado(a) ao Curso de Mestrado em Modelagem Matemática do(a) Escola de Matemática Aplicada para obtenção do grau de Mestre em Modelagem Matemática.


Data da defesa: 30/03/2021

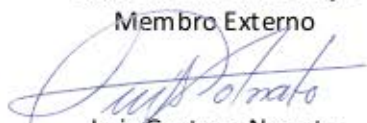
ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

Presidente da Comissão Examinadora: Profª Asla Medeiros e Sá


Asla Medeiros e Sá
Orientadora

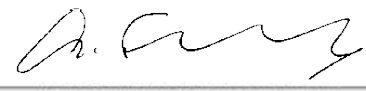

Flávio Codeço Coelho
Membro Interno


Cristiana Silveira Serejo
Membro Externo


Luiz Gustavo Nonato
Membro Externo

Nos termos da Lei nº 13.979 de 06/02/20 - DOU nº 27 de 07/02/20 e Portaria MEC nº 544 de 16/06/20 - DOU nº 114 de 17/06/20 que dispõem sobre a suspensão temporária das atividades acadêmicas presenciais e a utilização de recursos tecnológicos face ao COVID-19, as apresentações das defesas de Tese e Dissertação, de forma excepcional, serão realizadas de forma remota e síncrona, incluindo-se nessa modalidade membros da banca e discente.


César Leopoldo Camacho Manco
Diretor


Antonio de Araujo Freitas Junior
Pró-Reitor de Ensino, Pesquisa e Pós-Graduação FGV

Antonio Freitas, PhD
Pró-Reitor de Ensino, Pesquisa e Pós-Graduação
Fundação Getúlio Vargas

Instrução Normativa nº 01/19, de 09/07/19 - Pró-Reitoria FGV

Em caso de participação de Membro(s) da Banca Examinadora de forma não-presencial*, o Presidente da Comissão Examinadora assinará o documento como representante legal, delegado por esta I.N.

*Skype, Videoconferência, Apps de vídeo etc

Este trabalho é dedicado à minha namorada, Ione, aos meus pais, Rellyton e Irraelita, e à minha irmã, Lillyan. Não seria nada sem vocês.

Agradecimentos

À professora Asla, pela orientação, paciência e todo empenho dedicado à este trabalho, que foi de suma importância desde sua concepção.

À equipe do Museu Nacional, em especial aos setores de Herpetologia (Paulo Passos, Manoela Woitovicz, Pedro Pinna e Ronaldo Fernandes), Polychaeta (Joana Zanol, Camila Messias e Monique Santos) e Carcinologia (representado por Cristiana Serejo), pela parceria fundamental para viabilizar este trabalho. Agradeço também à Beatriz Santos pelo trabalho de alta qualidade em ajustar as bases de dados. Muito obrigado a todos vocês por toda dedicação e paciência durante as reuniões, em particular por me mostrar um mundo novo que é a pesquisa em biodiversidade.

Meu agradecimento especial à Cristiana Serejo, vice-diretora do Museu Nacional e diretora adjunta de coleções do MN, que nos ajudou a firmar essa parceria com muita alegria e atenção. Obrigado por acreditar neste trabalho e nos contagiar com seu bom humor desde o princípio.

À meus pais, Rellyton e Irraelita, por todo carinho e compreensão. O apoio de vocês foi fundamental não só durante a minha trajetória no segundo mestrado, mas por toda minha vida. Não seria nada sem vocês. Muito obrigado!

À minha namorada, Ione, por todo seu incentivo e paciência ao longo desses dois anos. Não foi fácil ficar em cidades diferentes por tanto tempo, mas conseguimos passar por isso juntos. Obrigado por estar sempre ao meu lado!

À minha irmã, Lillyan, por sempre trazer alegria à nossa família e cuidar bem dos nossos pais enquanto estive longe. A primeira memória que tenho na vida é do seu nascimento e sou muito grato por você nos proporcionar ainda mais felicidade com o nascimento da minha sobrinha, Heloísa. Bem-vinda à nossa família!

Aos meus colegas de mestrado por toda convivência, pelas trocas de conhecimento e por caminharem junto comigo ao longo do curso. Obrigado!

À todo corpo docente da EMap, pela excelência de ensino.

À equipe do FGV CERI pela oportunidade de aplicar meus conhecimentos em um ambiente de pesquisa integrado ao mercado. Aprendi muitas coisas e fui inspirado por pessoas de muita competência. Obrigado e muito sucesso a todos nós!

"Logic will get you from A to Z; imagination will get you everywhere."

Albert Einstein

Resumo

Coleções científicas de biodiversidade têm o compromisso de ser um registro permanente da herança natural, constituídas de espécimes ou objetos relacionados ao seu domínio. Todo material é catalogado, recebendo um número de registro, permitindo que seja incorporado ao acervo. Coleções digitais tipicamente contêm uma versão digitalizada dos metadados correspondentes a cada item do inventário e podem, adicionalmente, conter arquivos multimídia tais como textos, registros fotográficos ou outros registros associados ao item, quando pertinente. Garantir a qualidade desses registros é uma tarefa complexa e de fundamental relevância. Fatores como o grande volume de dados e a interdependência entre múltiplas variáveis dificultam determinar até que ponto esses dados estão completos, corretos e se, de fato, fornecem uma boa cobertura geográfica, temporal e taxonômica das espécies correspondentes. Não obstante, garantir a acurácia no registro de espécimes é uma tarefa que se inicia no campo de coleta, passando pelo seu registro, que muitas vezes é escrito em uma folha de papel ou anotado por meio de planilhas e dispositivos eletrônicos, até o processo de tombo do registro na coleção e futura publicação. O presente trabalho propõe a construção de um *framework* que visa guiar a aplicação de princípios e técnicas de Visualização da Informação ao contexto de coleções científicas digitais de história natural, visando fornecer um conjunto de representações visuais cuja finalidade é a de facilitar a verificação de qualidade dos registros por um especialista. Aliando a flexibilidade da linguagem de programação Python para tratamento de dados e as vantagens de se empregar uma gramática declarativa de gráficos, construiu-se propostas gráficas empregando-se a biblioteca Altair. Tal conjunto de técnicas e propostas gráficas visa atender demandas de especialistas de domínio - curadores e pesquisadores do campo de biodiversidade - fornecendo recursos visuais para a identificação de possíveis inconsistências em suas bases de dados auxiliando, por exemplo, no processo de preparação dos dados para publicação. Com pequenos ajustes, o mesmo conjunto de visualizações também pode servir à finalidade de exibir os dados da coleção em publicações científicas ou diretamente ao público não-especializado.

Palavras-chave: Coleções Digitais, Biodiversidade, Visualização

Abstract

Scientific collections of biodiversity are committed to being a permanent record of natural heritage, consisting of specimens or objects related to their field domain. All material is cataloged, receiving a registration number, allowing it to be incorporated into the collection. Digital collections typically contain a digitized version of metadata corresponding to each item in the inventory and, additionally, may contain multimedia files such as texts, photographic records or other records associated with the item, when relevant. Ensuring the quality of these records is a complex and fundamentally important task. Factors such as the large volume of data and the interdependence between multiple variables make it difficult to determine to what extent these informations are complete, correct and whether, in fact, they provide a good geographical, temporal and taxonomic coverage of the corresponding species. Nevertheless, ensuring accuracy in specimen registration is a task that begins in the collection field, passing through its registration, which is often written on a sheet of paper or annotated through spreadsheets and electronic devices, until the process of surveying the registers into the collection and future publication. The present work proposes the construction of a *framework* that aims to guide the application of Information Visualization principles and techniques to the context of digital scientific collections of natural history, aiming to provide a set of visual representations whose purpose is to facilitate the quality check of records by an expert. Combining the flexibility of the Python programming language for data processing and the advantages of using a declarative grammar of graphics, graphical proposals were built using the Altair library. Such set of techniques and graphic proposals aims to meet the demands of domain experts - curators and researchers in the field of biodiversity - providing visual resources for the identification of possible inconsistencies in their databases, assisting, for example, in the process of preparing data for publication. With minor adjustments, the same set of visualizations also serves the purpose of displaying collection data in scientific publications or directly to non-specialized audiences.

Keywords: Digital Collections, Biodiversity, Visualization

Lista de Figuras

Figura 1 – Contabilização da prata à disposição do governador Sumério.	22
Figura 2 – A Terra esférica desenhada por Ptolomeu.	23
Figura 3 – Movimentos planetários mostrados como inclinações cíclicas ao longo do tempo.	23
Figura 4 – <i>Pipeline</i> de InfoVis.	25
Figura 5 – Níveis de granularidade em coleções de objetos.	29
Figura 6 – Diferentes localidades de postos de saúde no Reino Unido.	30
Figura 7 – Diferenças nos canais de codificação.	31
Figura 8 – Quarteto de Anscombe.	36
Figura 9 – <i>Framework</i> de análise visual proposto por Liu et al. (2018).	37
Figura 10 – Total de artigos analisados de cada uma das fontes: GBIF e BDJ. . . .	41
Figura 11 – Quantidade de artigos que apresentam recursos gráficos para ilustrar a cobertura temporal, taxonômica e geográfica, respectivamente.	42
Figura 12 – Distribuição espacial de registros de diferentes fontes de dados.	44
Figura 13 – Distribuição geográfica dos espécimes depositados na coleção do Museu de Herpetologia da Universidade de Antioquia.	45
Figura 14 – Locais de coletas de répteis na bacia do rio Volga.	46
Figura 15 – Distribuição de dados de ocorrência para <i>T. infestans</i>	47
Figura 16 – Distribuição da espécie <i>Hydrophilidae</i> na área de estudo.	47
Figura 17 – Mapa das estações amostradas por Thor e outras embarcações durante as expedições principais de 1908-1909, 1910 e expedições adicionais de 1905-1906 e 1911-1912.	48
Figura 18 – Mapa da área de estudo de Santos et al. (2017).	49
Figura 19 – Distribuição geográfica de espécimes na Península Ibérica, Baleares, Canarias, Madeira e Açores.	49
Figura 20 – Distribuição temporal dos registros de espécimes nas coleções do Museu Herpetológico da Universidade de Antioquia.	51
Figura 21 – Distribuição do número de espécimes coletados por ano na coleção de serpentes do Museu Paranaense Emílio Goeldi, Pará - Brasil.	52
Figura 22 – Descrição de espécies e subespécies de répteis ao longo de 260 anos. . .	52
Figura 23 – Inventário histórico de 73.316 registros de pequenos mamíferos no conjunto de dados de ' <i>pellet sampling</i> ' que incluem o ano de coleta. . .	53
Figura 24 – Cobertura temporal do Herbário UNEX.	53
Figura 25 – Padrão de acumulação de registros de plantas Togolesas coletadas por mês.	54

Figura 26 – Cobertura temporal do <i>dataset</i> apresentado por Pérez-Luque et al. (2016).	55
Figura 27 – Sistema moderno de classificação biológica.	56
Figura 28 – Principais gêneros na coleção MGC-Cormof.	57
Figura 29 – Cobertura taxonômica (percentuais por família) da URJC GB.	57
Figura 30 – Cobertura taxonômica por Classe, Família e Ordem da base de dados Sinfonevada.	58
Figura 31 – Distribuição taxonômica por Filo, Ordem e Família (se conhecida). . .	59
Figura 32 – Distribuição taxonômica de espécies entre famílias de musgo no conjunto de dados.	59
Figura 33 – Fachada do Museu Nacional (MNRJ) e seu entorno.	61
Figura 34 – Exemplares do acervo de répteis do setor de Herpetologia do Museu Nacional.	63
Figura 35 – Exemplares da coleção de crustáceos do setor de Carcinologia do Museu Nacional.	64
Figura 36 – Espécimes do acervo de poliquetas do setor de Polychaeta do Museu Nacional.	65
Figura 37 – Subdivisões do canal de cor.	73
Figura 38 – Seleção de esquema de cores quantitativo a partir de uma dada categoria. .	76
Figura 39 – Paletas de cores criadas com o Colorgorical ^a	79
Figura 40 – Exemplo de atribuição de cores para 4 elementos a partir das 9 cores sugeridas pelo Color Crafter ^{®a}	82
Figura 41 – Exemplo de atribuição de cores para uma quantidade N grande de elementos usando a ferramenta Color Crafter ^{®a}	83
Figura 42 – Paletas de cores discretas criadas com o Color Crafter [®] para a coleção de répteis.	84
Figura 43 – Paletas de cores discretas criadas com o Color Crafter ^{®a} para a coleção de poliquetas.	84
Figura 44 – Cobertura geográfica da coleção de répteis (MNRJ).	88
Figura 45 – Cobertura geográfica da coleção de poliquetas (MNRJ).	89
Figura 46 – Cobertura geográfica da América Latina - coleção de répteis (MNRJ). .	90
Figura 47 – Evolução temporal por continente da coleção de poliquetas (MNRJ). .	91
Figura 48 – Evolução temporal por país da coleção de poliquetas (MNRJ).	92
Figura 49 – Evolução temporal por Estados brasileiros da coleção de poliquetas (MNRJ).	93
Figura 50 – Altitude de coleta por gênero para a coleção de répteis (MNRJ).	94
Figura 51 – Profundidade de coleta por famílias para a coleção de poliquetas (MNRJ). .	95
Figura 52 – Volume de espécimes coletados por ano para diferentes Ordens da coleção de poliquetas (MNRJ).	96

Figura 53 – Materiais-tipo coletados por ano para a coleção de poliquetas (MNRJ).	97
Figura 54 – Volume de espécimes coletados por ano para diferentes Infraordens da coleção de crustáceos (MNRJ).	98
Figura 55 – Materiais-tipo coletados por ano, distinguidos por gênero, para a coleção de crustáceos (MNRJ).	99
Figura 56 – Crescimento acumulado da quantidade de espécimes por família na coleção de répteis (MNRJ).	100
Figura 57 – Crescimento acumulado da quantidade de espécimes por país na coleção de répteis (MNRJ).	101
Figura 58 – Crescimento acumulado da quantidade de espécimes por Estados brasileiros na coleção de répteis (MNRJ).	102
Figura 59 – Quantidade de espécimes capturados pelos 50 coletores mais expressivos, distinguidos por família, para a coleção de crustáceos (MNRJ).	103
Figura 60 – Quantidade de espécimes identificados ou reclassificados pelos 50 determinadores mais expressivos, distinguidos por família, para a coleção de crustáceos (MNRJ).	104
Figura 61 – Contribuição acumulada de coletores para a coleção de crustáceos (MNRJ).	105
Figura 62 – Contribuição acumulada de determinadores para a coleção de crustáceos (MNRJ).	106
Figura 63 – Identificando inconsistências nos eixos X e Y - exemplo para a coleção de crustáceos (MNRJ).	108
Figura 64 – Identificando padrões anômalos nos dados - exemplo para a coleção de poliquetas (MNRJ).	110
Figura 65 – Identificando inconsistências por meio da <i>tooltip</i> - exemplo para a coleção de répteis (MNRJ).	111
Figura 66 – Localização geográfica dos exemplares coletados em 1987 para a coleção de répteis (MNRJ).	112
Figura 67 – Localização geográfica dos holótipos coletados na América Latina para a coleção de poliquetas (MNRJ).	113
Figura 68 – Distribuição dos holótipos por gênero na coleção de crustáceos (MNRJ).	114

Lista de Tabelas

Tabela 1 – Estatísticas descritivas das coleções com acesso cedido pelo Museu Nacional ^a	66
---	----

Lista de Quadros

Quadro 1 – Matriz criativa.	72
Quadro 2 – Representações gráficas criadas a partir da matriz criativa.	73

Lista de abreviaturas e siglas

BDJ	Biodiversity Data Journal
CONABIO	Comision Nacional para el Conocimiento y Uso de la Biodiversidad
DwC-A	Darwin Core Archives
GBIF	Global Biodiversity Information Facility
GPS	Global Positioning System
IABIN	Inter-American Biodiversity Information Network
InfoVis	Information Visualization
MNRJ	Museu Nacional, Universidade Federal do Rio de Janeiro
OBIS	Ocean Biodiversity Information System
PBR	Primary Biodiversity Records
SiBBr	Sistema de Informação sobre a Biodiversidade Brasileira

Sumário

1	INTRODUÇÃO	17
2	REVISÃO DE LITERATURA	20
2.1	Visualização da Informação	20
2.1.1	Contextualização	22
2.1.2	Conceitos, princípios e boas práticas	24
2.1.3	Visualizando coleções digitais de biodiversidade	32
2.1.4	Visualização como ferramenta de apoio ao tratamento de dados	35
2.2	Biodiversidade	38
2.2.1	Cobertura geográfica	43
2.2.2	Cobertura temporal	50
2.2.3	Cobertura taxonômica	56
3	COLEÇÕES DE DADOS DE BIODIVERSIDADE MNRJ	61
3.1	Pré-tratamento dos dados	66
4	CONSTRUÇÃO DO <i>FRAMEWORK</i> EM APOIO À CURADORIA DE COLEÇÕES CIENTÍFICAS BIOLÓGICAS DIGITAIS	69
4.1	<i>Design thinking</i>	69
4.2	Representação visual	74
4.2.1	Atribuição do canal de cor	77
4.3	Interatividade	86
4.4	Propostas gráficas	87
4.4.1	Cobertura geográfica	87
4.4.1.1	Relação espaço-temporal	89
4.4.1.2	Dados de altitude e profundidade	93
4.4.2	Cobertura taxonômica	95
4.4.3	Cobertura temporal	99
4.4.4	Contribuição de indivíduos para as coleções	102
5	ANÁLISE DE RESULTADOS	107
5.1	Visualização como ferramenta de apoio aos curadores	107
5.2	Criação de novas visualizações a partir da interatividade	112
6	CONCLUSÃO	115
	REFERÊNCIAS	118

APÊNDICE A – ORDENAMENTO DE CORES	127
---	-----

1 Introdução

Nas últimas décadas, o advento da Internet e o aumento do poder computacional revolucionaram a forma de criar, armazenar e recuperar dados. No mundo moderno, informações são constantemente criadas e consumidas de diversas formas, por exemplo, lendo o jornal, recebendo e enviando e-mails, interagindo com aplicativos de serviços, navegando em *websites*, etc. Cotidianamente, temos que lidar com um volume de informações maior do que conseguimos processar. Nesse contexto, técnicas para representar dados de forma a facilitar sua manipulação e compreensão são de extrema importância.

Conforme Keim (2002) destaca, para que a exploração de dados seja efetiva, é importante que o aspecto humano seja integrado à esse processo, combinando sua flexibilidade, criatividade e conhecimento à grande capacidade de armazenamento e poder de processamento dos computadores. Não obstante, Visualização da Informação (InfoVis) é uma área de pesquisa em crescente relevância, uma vez que tem como objetivo auxiliar seus usuários a explorar, compreender e analisar dados de forma visual, permitindo a extração de valiosos *insights* (SHIRAVI; SHIRAVI; GHORBANI, 2011). Diversas são suas aplicações em contextos como: esportes (PERIN et al., 2018), negócios (LIU; CAO; LV, 2008; KO et al., 2012), dados textuais (CUI et al., 2011), elementos audiovisuais (CHEN; LU; HU, 2012; PRETORIUS et al., 2011), mecanismos de busca (NOCAJ; BRANDES, 2012), etc. Essas e outras aplicações são amplamente comentadas em Liu et al. (2014).

No domínio da Biodiversidade, iniciativas de digitalização realizadas por grandes instituições como museus, herbários, universidades e instituições de pesquisa, dentre outras, deram origem à uma vasta coleção de registros que, agora, estão disponíveis *online*. Tais registros dão suporte a estudos que buscam, por exemplo, compreender a distribuição, mudanças no tempo, interações entre diferentes espécies, presença ou ausência de certos animais em um dado bioma, etc. Em particular, o impacto sofrido pela biodiversidade em meio a décadas de mudanças climáticas não precedidas é um exemplo de discussão frequentemente fomentada em diferentes debates a nível mundial.

Nesse cenário, essas instituições têm sua importância revelada dado que foram responsáveis por documentar e armazenar dados de biodiversidade levantados por pesquisadores taxonomistas ao longo do tempo. Essa documentação baseia-se em registros primários de biodiversidade (PBR¹), que contém informações como a data e o local de coleta, a identificação taxonômica e o taxonomista responsável pelos espécimes coletados. Iniciativas como o *Global Biodiversity Information Facility* (GBIF)², *Ocean Biodiversity Information System* (OBIS), *Inter-American Biodiversity Information Network* (IABIN),

¹ *Primary Biodiversity Records*.

² Oferece acesso livre à mais de 1,08 bilhão de *datasets* de biodiversidade.

Comision Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) e Sistema de Informação sobre a Biodiversidade Brasileira³ (SiBBr), juntas viabilizam acesso à bilhões de registros de biodiversidade.

Para a comunidade científica, a qualidade desses conjuntos de dados é de suma importância para garantir a precisão dos estudos desenvolvidos. De fato, essa é uma preocupação existente na literatura desde o trabalho de Thomas e Mintz (1998). No entanto, garantir a qualidade desses registros não é trivial por motivos que vão desde o grande volume de dados à interdependência entre múltiplas variáveis. Nesse contexto, ferramentas de visualização podem ser úteis ao fornecer mapas mentais de informações (LIU et al., 2014), auxiliando seus usuários na exploração, compreensão e análise de dados, ao mesmo tempo em que reforça aspectos da cognição humana. Muitas são as ferramentas disponíveis que podem ser empregadas na produção de visualizações de dados biológicos. Wang et al. (2015) faz uma boa síntese de ferramentas *open source*⁴ que podem ser utilizadas para visualização de dados no campo da Bioinformática. Entretanto, ainda não há um consenso na literatura a respeito do quanto os sistemas de visualização existentes são capazes de auxiliar nas tarefas de triagem e diagnóstico visual dos dados, além de produzir a um especialista informações pertinentes à melhoria da qualidade dos registros analisados.

Assim, o presente trabalho tem como objetivo sugerir um conjunto técnicas e abordagens de visualização de dados voltadas para bancos de dados de coleções biológicas⁵, visando atender conceitos como granularidade, *data-ink ratio* e consistência global. Tais abordagens constituem um *framework* desenhado a partir da necessidade de especialistas do campo da Biologia em identificar e corrigir problemas de qualidade nos seus conjuntos de dados. Adicionalmente, também é proposto um conjunto de técnicas e visualizações providas de recursos dinâmicos e interativos que têm o potencial de elevar o grau de imersão dos gráficos e facilitar a identificação de possíveis inconsistências na base de dados, além de permitir ao usuário final mais possibilidades para navegação e detalhamento da coleção analisada. Tais recursos de interatividade também possibilitam a geração de um novo conjunto de gráficos estáticos, ao aplicar filtros e seleções ao conjunto de dados original, podendo ser facilmente adaptados para apresentação em artigos científicos, livros e demais ambientes de veiculação impressa.

Vale ressaltar que as ferramentas desenvolvidas neste trabalho tem o propósito de servir à comunidade de biólogos, pesquisadores e curadores de coleções científicas de biodiversidade não só no sentido de consumir diretamente seus produtos, mas também

³ <<https://ipt.sibbr.gov.br/mnrj/>>

⁴ Bibliotecas e *frameworks* desenvolvidos em linguagens de programação com código aberto, como Python e JavaScript, que podem ser livremente utilizadas, i.e., sem necessidade de aquisição de licenças específicas para uso proprietário.

⁵ Isto é, metadados (*Primary Biodiversity Records* - PBR) associados a coleções biológicas, digitalizados em planilhas eletrônicas. Normalmente, seguem o padrão *DwC* (*Darwin Core*).

de permitir sua utilização em aplicações diárias associadas à melhoria de qualidade dos registros analisados. Ainda, o mesmo conjunto de visualizações, com pequenos ajustes, pode servir ao propósito de apresentar as coleções de dados a entusiastas e ao público em geral, por exemplo, em ambiente *web*. Isto posto, foram adotadas linguagens de programação e padrões de escrita de código que permitem um alto nível de flexibilidade ao usuário no que diz respeito à realização de pequenas modificações e adaptação dos *scripts* à coleções de dados de natureza semelhante.

O *framework* apresentado neste trabalho foi desenvolvido em parceria com o Museu Nacional (MN)⁶, instituição autônoma integrante do Fórum de Ciência e Cultura da Universidade Federal do Rio de Janeiro (UFRJ). Criado por D. João VI em 1818, o Museu Nacional foi responsável por coletar e registrar dados em coleções de ciências naturais e antropológicas ao longo de mais de 200 anos de existência. Nessa parceria, foi concedido acesso a três grandes coleções de dados do tipo PBR: Crustáceos⁷, Répteis⁸ e Poliquetas⁹. Como produto, priorizou-se o uso de técnicas de visualização de dados com os objetivos de averiguar a qualidade dos registros e diagnosticar possíveis inconsistências para, finalmente, apresentar esses dados de forma eficaz ao público interessado.

Os capítulos subsequentes deste estudo estão dispostos conforme a seguinte estrutura: O capítulo 2 faz uma revisão extensa, mas não exaustiva, dos principais conceitos, princípios e boas práticas da comunidade de Visualização da Informação relevantes para a execução deste trabalho. Além disso, apresenta soluções existentes na literatura de InfoVis para visualização de coleções biológicas digitais, assim como as abordagens gráficas usadas em artigos de biodiversidade para retratar suas bases de dados; O capítulo 3 apresenta as coleções de dados biológicos empregadas nesse estudo, enquanto o capítulo 4 sintetiza as técnicas, ferramentas e abordagens empregadas para propor um *framework* capaz de visualizar coleções científicas biológicas digitais de forma a potencializar a detecção de inconsistências nas bases de dados; O capítulo 5 expõe a maneira pela qual os gráficos propostos foram úteis para a detecção de problemas de qualidade nos dados durante o desenvolvimento deste trabalho; O capítulo 6 encerra este estudo ressaltando suas principais contribuições para ambas as literaturas, os ganhos observados na aplicação das técnicas sugeridas e sugestões para trabalhos futuros que busquem dar continuidade à essa discussão ou estendê-la a diferentes contextos.

⁶ Mais informações sobre essa instituição podem ser encontradas em <<http://www.museunacional.ufrj.br/>>.

⁷ Também conhecidos como Carcinos. Esses registros pertencem ao setor de Carcinologia do Museu Nacional (MNRJ).

⁸ Pertencente ao setor de Herpetologia do Museu Nacional (MNRJ).

⁹ Pertencente ao setor de Polychaeta do Museu Nacional (MNRJ).

2 Revisão de Literatura

Este capítulo inicia-se com a apresentação de trabalhos relevantes à literatura de InfoVis, seus principais conceitos e aplicações, além dos pontos pelos quais cada um destes trabalhos inspirou a criação e desenvolvimento do presente tópico. Em seguida, serão apresentados *data papers* relevantes ao campo de Biodiversidade, destacando as principais abordagens gráficas e técnicas de visualização adotadas nestes trabalhos para apresentar e destacar certas características de suas coleções de dados biológicos.

2.1 Visualização da Informação

Visualização da Informação é uma área do conhecimento focada no estudo da transformação e apresentação de dados¹ em representações visuais de modo a revelar padrões que propiciam o surgimento de *insights* cognitivos. Nessa área, são abordados diversos aspectos multidisciplinares. Conhecimentos de psicologia e percepção humana são combinados com estruturas matemáticas e algoritmos com a finalidade de fornecer ao mundo científico novas maneiras de apresentar e extrair *insights* a partir de um conjunto de informações.

Adicionalmente, InfoVis é uma área de pesquisa cujas aplicações podem ser rastreadas até o surgimento da escrita (FRIENDLY, 2006) e que visa auxiliar seus usuários nas tarefas de exploração, compreensão e análise de informações por meio de exploração visual progressiva e iterativa (LIU et al., 2014). Revisões extensivas, mas não exaustivas, de trabalhos e avanços recentes nesse campo estão disponíveis em Liu et al. (2014), Zudalova, Adriaansen e vanLiere (2009) e Geisler (1998). Por ser uma área que engloba desenvolvimentos e aplicações em diferentes domínios, e por buscar reforçar o raciocínio dedutivo, seu potencial tem sido cada vez mais reforçado em contextos atuais.

O *boom* do *big data* incitou uma crescente demanda por métodos eficazes para exploração de dados. Para melhor aproveitar a quantidade massiva de informações geradas e armazenadas a todo momento, inúmeras áreas de pesquisa buscaram modernizar seus métodos de mineração e análise de dados, tirando proveito de um maior poder computacional sobretudo na automatização de tarefas.

No entanto, conforme Keim (2002) destaca, para que se tenha maior proveito na extração de conhecimento a partir de dados brutos, é fundamental que o aspecto humano seja integrado à esse processo, combinando suas características cognitivas e conhecimento prévio ao grande poder de processamento dos computadores. Bekri e Peinsipp-Byma (2016)

¹ Ou informações, no sentido amplo.

também chama a atenção para importância do conceito de *user-in-the-loop* e destaca que, em muitos casos, conhecimento de domínio acessado por meio de especialistas contribui significativamente para uma melhor performance de algoritmos de gerenciamento de qualidade de dados. Não obstante, há um crescente interesse em definir a melhor maneira de combinar usuários e métodos computacionais para se extrair melhor performance no manuseio e tratamento de dados. Nesse sentido, cada vez mais esforços científicos direcionam-se ao campo da visualização de informações como alternativa viável e eficaz para se combinar a cognição humana ao amplo poder de processamento das máquinas.

Em particular, milhões de coleções de registros de biodiversidade tiveram acesso facilitado em plataformas *online* como resultado de esforços de digitalização efetuados por grandes instituições como museus, herbários, universidades e demais instituições de pesquisa. Tais coleções são de fundamental importância à inúmeras pesquisas em ciências biológicas. Artigos como Arts, Wal e Adams (2015), Marx (2013), Reichman, Jones e Schildhauer (2011), Kelling et al. (2009), Howe et al. (2008), dentre outros, destacam os potenciais desafios que a era do *Big Data* trouxe para pesquisas científicas nessa área, além de apontar maneiras pelas quais o emprego da tecnologia pode trazer benefícios.

Paralelamente ao avanço da tecnologia, pesquisas em biologia estão passando por uma rápida transformação, buscando agregar e sintetizar novas maneiras de interagir com grandes conjuntos de informações digitais (GURALNICK; HILL, 2009). Para assegurar a precisão e a veracidade dos estudos conduzidos, é de fundamental importância que essas informações tenham alto nível de qualidade. Contudo, para as instituições responsáveis por manter e fazer a curadoria desses conjuntos de dados, acessar e verificar seus atributos qualitativos nem sempre é uma tarefa trivial. Fatores como grande escala, interdependência entre variáveis e desafios em verificar se os registros estão corretos, completos, e fornecem uma boa cobertura dificultam o dia a dia de curadores e pesquisadores.

Até o momento em que este trabalho está sendo escrito, não há um consenso a respeito de até que ponto soluções de visualização existentes podem auxiliar no processo de avaliação e melhoria da qualidade dos dados. No entanto, o ferramental proposto na literatura de InfoVis é um recurso valioso que tem o potencial de auxiliar nesse processo, fornecendo soluções engenhosas para apresentação, exploração e análise dessas informações. Não obstante, o presente estudo visa preencher essa lacuna ao propor uma solução na forma de *framework*, desenhada a partir de demandas de especialistas de domínio, com o objetivo de auxiliar nas tarefas de triagem, diagnóstico e correção de dados ao explorar visualmente as coleções científicas digitais sob diferentes óticas.

Isto posto, antes de tratar das principais definições e aplicações, a subseção 2.1.1 traz um breve contexto histórico de visualização da informação; a subseção 2.1.2 visa apresentar importantes conceitos, princípios e boas práticas desenvolvidos em InfoVis e sua relevância ao desenvolvimento do presente estudo; a subseção 2.1.3 é voltada à apresentação

de soluções existentes na literatura que exploram diferentes técnicas, conceitos e práticas com o objetivo de introduzir novas propostas de visualização para coleções de dados de biodiversidade; por fim, a subseção 2.1.4 traz breves comentários sobre trabalhos que se propõem a guiar a aplicação de soluções visuais na gestão da qualidade de dados visando melhorar a interação homem-máquina.

2.1.1 Contextualização

Ao contrário do que muitos podem imaginar, o ato representar dados de maneira visual é muito anterior ao surgimento dos primeiros microcomputadores. Friendly (2006) traz uma boa contextualização histórica da visualização de dados e destaca que a representação gráfica de informações quantitativas tem raízes profundas, podendo ser rastreadas até períodos tão antigos quanto o surgimento da escrita. Apesar de ter sido reconhecida como disciplina em meados dos anos 80 pela Xerox PARC², há algumas gravações em pedra da antiga Mesopotâmia, por exemplo, que mostram representações de dados tabulares e que têm mais de 2 mil anos de idade. A figura 1 ilustra uma dessas representações.

Figura 1 – Contabilização da prata à disposição do governador Sumério.



Fonte: <<https://www.interaction-design.org/literature/article/information-visualization-a-brief-pre-20th-century-history>>. Acessado em 29/01/2021.

Nota: A imagem foi disponibilizada para domínio público pelo autor, Gavin.collins.

No entanto, pode-se afirmar que o uso dessas representações era especialmente raro e seu entendimento bastante limitado (FRIENDLY, 2006). Com o desenvolvimento de ciências como cartografia e astronomia, o uso de representações visuais começou a ser mais frequentemente empregado. A figura 2 exibe uma recriação, feita no século 15, de um mapa da Terra esférica desenhado por Ptolomeu em algum momento entre os anos 100 e 165 depois de Cristo. Esta certamente é uma das visualizações conhecidas de maior longevidade.

² Xerox Palo Alto Research Center foi uma divisão de pesquisa da empresa Xerox Corporation, com sede em Palo Alto, Califórnia - EUA.

Figura 2 – A Terra esférica desenhada por Ptolomeu.

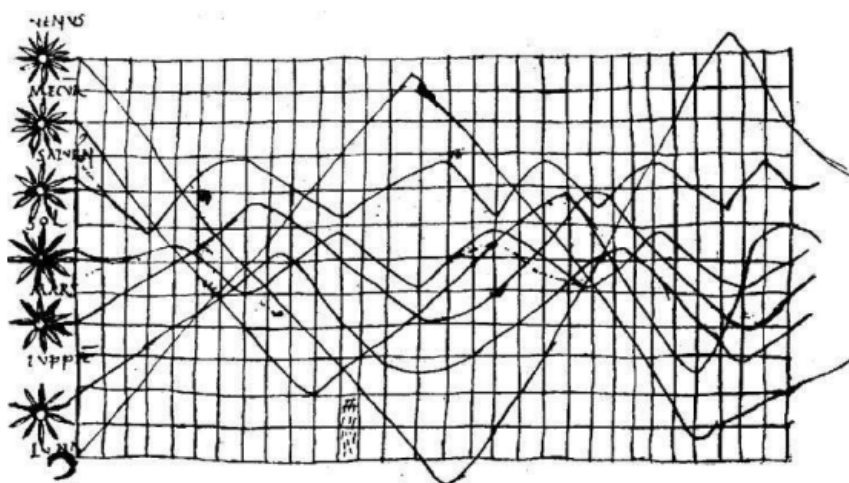


Fonte: <<https://www.interaction-design.org/literature/article/information-visualization-a-brief-pre-20th-century-history>>. Acessado em 29/01/2021.

Nota: A imagem foi disponibilizada para domínio público e sua autoria é de SCEhardt.

A figura 3, por sua vez, mostra uma representação gráfica usada para ilustrar movimentos planetários que aparece na obra *In Somnium Scipionus*, do século X.

Figura 3 – Movimentos planetários mostrados como inclinações cíclicas ao longo do tempo.



Fonte: Funkhouser (1936, apud Friendly, 2006).

Nota: O desenho foi criado por um astrônomo desconhecido e aparece em um apêndice aos comentários de A. T. Macrobius na obra *In Somnium Scipionus* de Cícero, do séc. 10.

Com a evolução das diferentes áreas do conhecimento científico, a visualização da informação passou a ser cada vez mais adotada como abordagem eficaz para apresentar dados e destacar padrões. Foi mais especificamente no século XIX que técnicas estatísticas de análise de dados tornaram-se populares (FRIENDLY, 2006) e, com elas, inúmeras representações gráficas empregadas atualmente. Nos tempos modernos, o acesso a novas tecnologias tornou a aplicação desse tipo de recurso mais rápido e acessível aos mais diversos públicos. Hoje, habilidades artísticas manuais podem ser substituídas por programação de computadores, ou apenas manuseio de *softwares* específicos.

O trabalho seminal de Leland Wilkinson nos anos 80, que culminou na publicação do livro *The Grammar of Graphics* em 1999 (WILKINSON, 1999), estabeleceu um conjunto de regras estruturais, sintetizadas em um *framework*, que visa descrever e estabelecer componentes essenciais à construção de gráficos. Em sua abordagem original, Wilkinson (1999) propõe um sistema de sete componentes gráficos independentes (dados, estética, escala, objetos geométricos, estatísticas, facetas e sistema de coordenadas) que representam diferentes aspectos do sistema visual de forma tal que toda combinação desses componentes é relevante.

Wilkinson (1999) guiou a implementação de diversos sistemas e *frameworks* de visualização, inspirando trabalhos como Wickham (2010), e ferramentas como *ggplot2*³, D3.js⁴, Tableau®, Vega e Vega-Lite. Vale ressaltar que, com exceção do Tableau®, todas as ferramentas citadas têm licença de *software* livre e estão disponíveis a todo interessado que tenha disposição de mesclar criatividade com algumas linhas de código.

VanderPlas et al. (2018) desenvolveu uma biblioteca declarativa de visualização estatística⁵, baseada na gramática de visualização interativa Vega-Lite, implementada na linguagem de programação Python que, por sua vez, tem características desejáveis como sintaxe limpa, diversas bibliotecas otimizadas para os mais diversos usos e amplo suporte da comunidade (OLIPHANT, 2007). Ao unir uma linguagem de programação flexível com princípios e estruturas de gramática de gráficos, a biblioteca de visualização Altair revela um potencial que será explorado no decorrer deste trabalho.

2.1.2 Conceitos, princípios e boas práticas

Na literatura moderna de InfoVis, conhecimentos milenares ainda podem ser facilmente identificados e são somados à novas linhas de pesquisa que, aliadas ao uso da tecnologia, buscam auxiliar seus usuários em tarefas como exploração, compreensão e análise de informações, ao mesmo tempo em que reforça aspectos da cognição humana. Bons *surveys* como Liu et al. (2014), Zudalova, Adriaansen e vanLiere (2009) e Geisler

³ Biblioteca desenvolvida para a linguagem de programação R.

⁴ Biblioteca de visualização criada por Mike Bostock, (BOSTOCK; HEER; OGIEVETSKY, 2019), para a linguagem JavaScript.

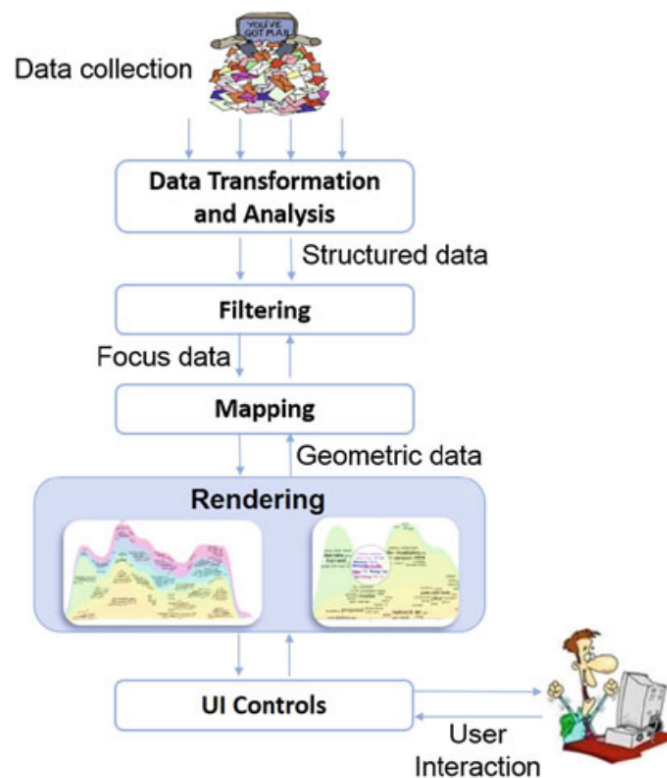
⁵ Subconjunto da visualização de dados com foco em modelagem estatística.

(1998) fazem um esforço bem sucedido de reunir e sintetizar pesquisas e avanços recentes nessa área.

Liu et al. (2014) classifica trabalhos modernos de visualização nas seguintes categorias: metodologias empíricas, interações de usuários, *frameworks* e aplicações. A primeira categoria engloba estudos que buscam prover fundamentação teórica a um grande número de aplicações de diversos domínios, enquanto a segunda categoria incorpora estudos que analisam maneiras pelas quais usuários interagem com a visualização por meio de diferentes canais de *input*. Há também trabalhos que se propõem a desenhar sistemas e *frameworks* para ampla aplicação de técnicas de visualização e, por fim, enquadram-se as mais diversas aplicações de InfoVis em diferentes domínios.

Em geral, modelos, sistemas e *frameworks* de visualização seguem o *pipeline* ilustrado na figura 4, descrito por Liu et al. (2014) em cinco etapas:

Figura 4 – *Pipeline* de InfoVis.



Fonte: (LIU et al., 2014)

1. **Transformação e análise de dados:** esse módulo geralmente consiste na extração de um subconjunto de dados estruturados do conjunto de *inputs*. Se a coleção de dados excede a capacidade de memória do computador em questão, técnicas de redução de dimensionalidade podem ser aplicadas. Além disso, técnicas de suavização, redução de ruído e interpolação podem se fazer relevantes em algumas

coleções de dados estruturados. Em conjuntos de informações não estruturadas, algumas técnicas de *data mining* como clusterização podem ser empregadas para se obter uma certa estruturação associada ao conjunto de dados original.

2. **Filtragem:** porções de dados com interesse específico são selecionadas a partir do *output* do módulo anterior para que sejam visualizados.
3. **Mapeamento:** as informações pré-selecionadas, então, são mapeadas em primitivas geométricas (pontos, linhas, barras, etc.) e atributos (ex: cor, posição, forma, tamanho).
4. **Renderização:** nessa etapa, dados geométricos são transformados em imagens. Em outras palavras, as primitivas geográficas são desenhadas na tela de acordo com suas proporções (altura, largura e quantidade de pixels).
5. **Interação:** após a realização das etapas anteriores, os usuários podem então interagir com a imagem gerada a partir de controles de interface, por exemplo: mouse, teclado e *touch screen*.

Para que esse *pipeline* seja bem empregado de forma a gerar *outputs* satisfatórios, são necessários diversos tipos de conhecimentos específicos. Por exemplo, pode-se dizer que, para que a seleção e transformação inicial dos dados seja bem empregada, é necessária *expertise* de domínio para que, ao final do processo, as informações exibidas sejam fidedignas à sua versão original. Sem esse tipo de especialidade, distorções podem ser criadas e levar à conclusões errôneas ou equivocadas. Por esse motivo, a ferramenta desenvolvida neste estudo contou com a colaboração de especialistas de domínio para o manuseio e filtragem de dados de forma que suas necessidades pudessem ser melhor atendidas.

Ainda, no que diz respeito ao mapeamento e renderização das informações selecionadas em primitivas gráficas, conhecimentos e princípios avançados em InfoVis são de fundamental importância. Kirk (2016), de maneira ímpar, conduz o leitor em um passo a passo didático que tem como objetivo apresentar os principais conceitos, técnicas e abordagens de visualização de dados além de organizar, de maneira simples e intuitiva, o processo criativo e os elementos essenciais à formulação de uma solução customizada. Seus ensinamentos foram amplamente empregados neste trabalho e, certamente, a leitura dessa obra é recomendada a qualquer indivíduo interessado em conhecer mais a respeito dessa ampla área de conhecimento. Em especial, Kirk (2016) é fortemente recomendado àqueles que almejam construir sua própria ferramenta de visualização, independentemente da área de aplicação.

O sucesso de um modelo de visualização é, em grande parte, associado à sua capacidade de permitir facilmente a identificação de padrões⁶. Nesse quesito, a clássica obra de Bertin (1983) traz uma rica discussão acerca de como técnicas tão simples quanto reordenar um eixo categórico podem revelar padrões previamente não acessíveis nos dados. O principal objetivo de seu método de análise matricial pode ser sintetizado na máxima: "simplificar sem destruir", codificando visualmente os valores das células e agrupando linhas e colunas similares (PERIN; DRAGICEVIC; FEKETE, 2014). Seus conhecimentos atemporais são empregados sempre que possível no contexto aqui explorado.

Perin, Dragicevic e Fekete (2014) também merece posição de destaque uma vez que revisitam a obra original de Bertin (1983) e oferecem à comunidade científica uma nova forma interativa de criar visualizações tabulares com sua ferramenta Bertifier. Perin, Dragicevic e Fekete (2014), ainda, apresentam uma extensa bibliografia das obras de Jacques Bertin e outros trabalhos que exploram seu legado.

A última camada do *pipeline* genérico para criação de soluções de visualização incorpora técnicas de interação, criando um novo conjunto de possibilidades pelas quais os usuários podem consumir tais produtos. Foley et al. (1996) define tais técnicas como maneiras de realizar tarefas genéricas na interação homem-máquina, por meio de mecanismos físicos de *input/output*. Yi et al. (2007) estende essa definição para o universo de visualização como um conjunto de *features* que fornecem aos usuários a habilidade de, direta ou indiretamente, manipular e interpretar diferentes representações. De acordo com essa definição, imagens estáticas ou animações pré-carregadas não são dotadas de tais técnicas de interação. Alternativamente, a presença de um menu que permite filtrar os dados visualizados se enquadra perfeitamente neste conceito.

Yi et al. (2007) classifica diversas técnicas de interação nas seguintes categorias:

- **Selecionar:** permite marcar itens de interesse.
- **Explorar:** habilidade de examinar um subconjunto diferente de casos.
- **Reconfigurar:** fornece aos usuários diferentes perspectivas alternando o arranjo espacial das representações.
- **Codificar:** permite aos usuários alterar a representação visual dos dados em termos de cor, tamanho, forma, etc.
- **Abstrair/elaborar:** possibilita o ajuste do nível de abstração da representação dos dados. Normalmente, pode-se navegar em diferentes níveis a partir de um cenário de *overview* até casos individuais (ex: *tooltips*).

⁶ Em alguns contextos, destacar a ausência de padrão significativo também pode ser relevante. Por exemplo, um *scatterplot* pode ilustrar facilmente que dada variável é um ruído branco.

- **Filtrar:** permite escolher o conjunto de dados em exibição baseado em condições específicas.
- **Conectar:** refere-se à técnicas de interação que são usadas para destacar associações entre itens já representados na tela ou mostrar dados ocultos que são relevantes para um item específico (ex: *brushing*).

Adicionalmente, Liu et al. (2014) traz uma classificação mais recente do estado-da-arte das técnicas de interação, englobando a categorização proposta por Yi et al. (2007), em apenas duas categorias:

1. **WIMP** (*Windows, Icons, Mouse, Pointer*): refere-se ao conjunto de técnicas que tiram proveito do paradigma clássico da interação homem-máquina (i.e., por meio de mouse, teclado, ícones, etc.). Os tipos mais comuns são seleção, filtragem e *brushing*.
2. **post-WIMP:** engloba técnicas voltadas para tirar proveito das novas formas de interação com dispositivos eletrônicos desenvolvidas nos últimos anos (ex: telas sensíveis ao toque).

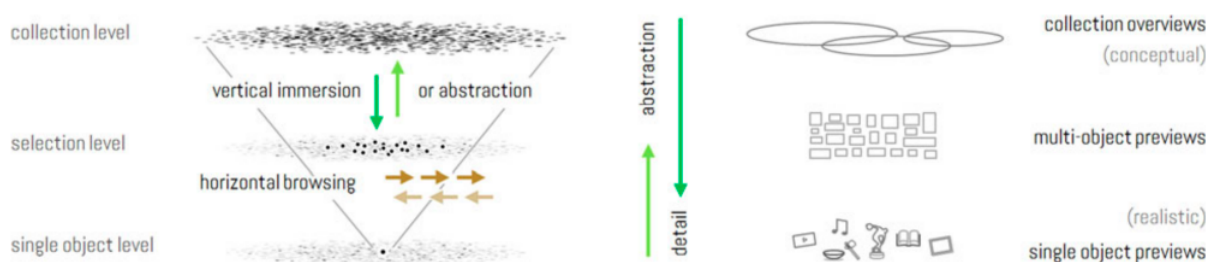
Vale ressaltar que as visualizações propostas neste estudo fazem uso de elementos de interação para, além das tarefas elementares, permitirem a realização de tarefas de seleção, filtragem e detalhamento de pontos de interesse por meio de *tooltips*. O emprego desses elementos visa auxiliar especialistas de domínio a realizar, em certo grau, tarefas mais complexas como identificar possíveis anomalias em suas coleções digitais e informações que apontem para o registro exato na planilha de dados. Por exemplo, exibir o número de catálogo do referido exemplar em uma *tooltip* restringe a área de busca nas planilhas, fazendo com que a correção seja facilmente efetuada. No entanto, todos os elementos utilizados se enquadram no paradigma WIMP tradicional, ficando como sugestão para trabalhos futuros o desenvolvimento de uma ferramenta, com base nas recomendações aqui propostas, cuja utilização seja otimizada para interfaces sensíveis ao toque, como em celulares e *tablets*.

No contexto de coleções digitais de herança cultural, Windhager et al. (2018) propõe uma categorização de sistemas de visualização, tomando os seguintes critérios: dados, usuários, tarefas, granularidade e codificação visual. De forma a complementar a discussão traçada até então, serão abordados breves comentários a respeito das categorias não mencionadas previamente adaptando, quando necessário, à presente aplicação. Bases de dados do tipo PBR (*Primary Basic Records*) são, em sua maioria, metadados associados ao evento de coleta de diferentes espécies, identificadas por seu nome científico, acompanhado da sua classificação taxonômica, o local e a data de coleta, podendo incluir

outras informações, como o nome do taxonomista que contribuiu para sua documentação. Normalmente, usuários desses *datasets* são especialistas.

Granularidade visual denota o nível de agregação em que os dados estão sendo apresentados (WINDHAGER et al., 2018). A figura 5 ilustra esse conceito.

Figura 5 – Níveis de granularidade em coleções de objetos.



Fonte: (WINDHAGER et al., 2018)

A apresentação de um único objeto (*single-object preview*) é um nível em que se representa os itens individuais da coleção, com uma maior riqueza de detalhes. O nível acima, multi-objeto (*multi-object previews*) provê uma apresentação de uma seleção de objetos, enquanto o nível mais alto (*collection overview*) representa uma visão geral da coleção. Por meio de elementos interativos, é possível transitar entre níveis de granularidade (imersão vertical) ou dentro de um mesmo nível granular (navegação horizontal). Em linha com esse conceito, Shneiderman (2003) propõe o seguinte mantra: "*overview first, zoom and filter, then details on demand*", que foi empregado nas etapas de desenvolvimento deste estudo.

Vale pontuar, ainda, que a codificação visual de dados espaço-temporais é amplamente discutida na literatura. Complementarmente, a apresentação de dados hierárquicos normalmente, segue duas propostas: Dendogramas⁷ ou *Treemaps* (WINDHAGER et al., 2018).

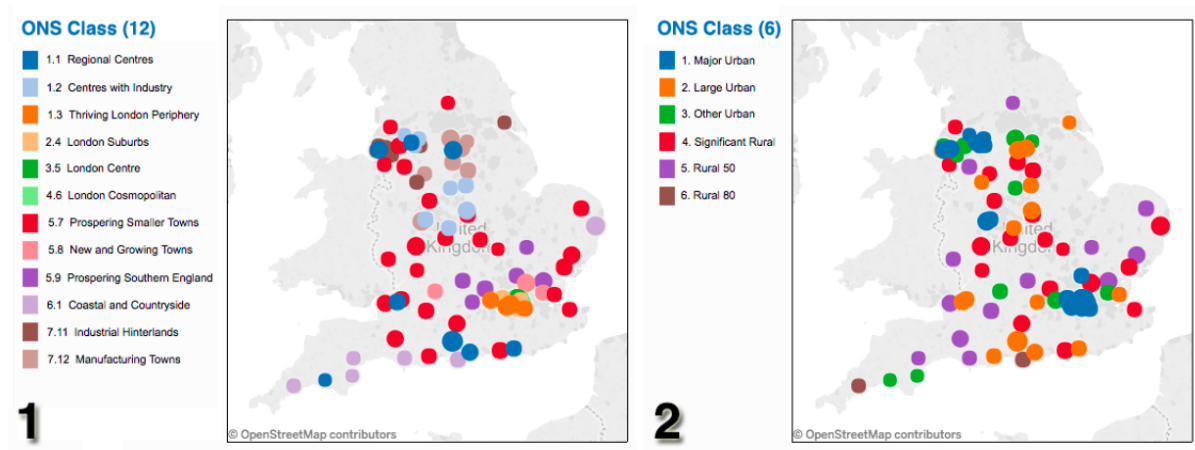
Na presença de tantos conceitos e fatores a serem considerados, pode-se afirmar que o design e a implementação de soluções visuais efetivas é uma tarefa complexa e que depende fortemente do conjunto de dados e da aplicação (HEER; CARD; LANDAY, 2005; FEKETE, 2004). Diferentes tipos de dados têm diferentes características e padrões de interesse que requerem um conjunto de ferramentas especializadas para visualizá-los (LIU et al., 2014). Ainda, é importante entender as necessidades e expectativas do público-alvo, além de conhecer suas habilidades técnicas e conhecimento de domínio, para que se desenvolva um desenho de interface efetivo (COOPER; REIMANN et al., 2003). Nesse sentido, a presente aplicação buscou adotar um design centrado no seu grupo principal de usuários, i.e., pesquisadores e curadores de coleções digitais de biodiversidade, empregando um processo de avaliação iterativo em todas as etapas de desenvolvimento, requerimento

⁷ Se apresentado no formato radial, este é denominado *Sunburst*

deste tipo de abordagem (NORMAN, 2013). Vale destacar que a implementação do *framework* aqui proposto contou com a colaboração de equipes voluntárias de especialistas de domínio ao longo de todo seu desenvolvimento. Especificamente, eram conduzidas reuniões periódicas nas quais soluções visuais eram avaliadas ou diretamente sugeridas pela equipe de biólogos⁸.

Por fim, ao se trabalhar com um conjunto de visualizações, é importante considerar o *trade-off* entre efetividade local e consistência global (QU; HULLMAN, 2016). Criar e desenvolver gráficos independentemente, mesmo à luz das práticas e conceitos aqui apresentados, pode trazer algumas dificuldades cognitivas a seus usuários devido à algumas inconsistências que aparecem nos campos de codificação quando estes gráficos são apresentados em conjunto. Para exemplificar como esse tipo de inconsistência global pode acontecer, Qu e Hullman (2016) traz os mapas exibidos na figura 6.

Figura 6 – Diferentes localidades de postos de saúde no Reino Unido.



Fonte: (QU; HULLMAN, 2016)

Nota: As mesmas cores representam dados diferentes nas duas visualizações.

Note que, nos mapas 1 e 2 da figura 6, as mesmas cores são usadas para representar informações diferentes, exigindo que os usuários mantenham o significado de cada cor na memória enquanto analisam todo o conjunto. Para reduzir a ocorrência desse tipo de inconsistência, Qu e Hullman (2016) propõe duas simples restrições de alto nível:

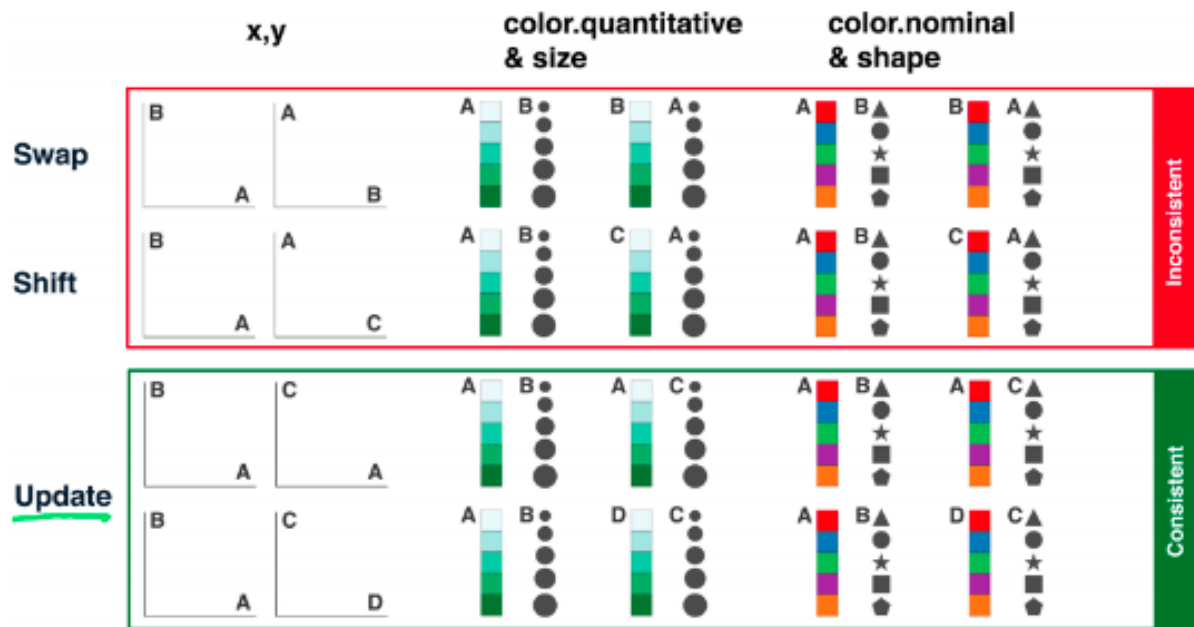
- R1.** Os mesmos campos de dados devem ser codificados da mesma forma.
- R2.** Campos diferentes devem ser codificados de maneiras distintas.

Adicionalmente, Qu e Hullman (2016) define três tipos de diferenças entre os canais de codificação que podem ser detectadas ao se comparar quaisquer pares de visualizações

⁸ Curadores das coleções científicas de biodiversidade trabalhadas neste estudo, apresentadas no capítulo 3.

presentes em um conjunto: Troca (*swap*), deslocamento (*shift*) e atualização (*update*). Vide a figura 7 para uma sintetização de cada um desses tipos.

Figura 7 – Diferenças nos canais de codificação.



Fonte: (QU; HULLMAN, 2016)

Nota: Troca (*swap*), deslocamento (*shift*) e atualização (*update*) para os eixos X e Y e os pares: cor quantitativa e tamanho; cor nominal e forma.

Uma troca significa que dois canais codificam os mesmos campos, mas trocaram sua posição. Um deslocamento significa que dois canais de codificação têm apenas uma variável em comum, mas essa variável não é codificada da mesma maneira. Se dois canais de codificação têm apenas um campo em comum e esse campo é mapeado pelo mesmo canal, tem-se uma situação de *update*⁹. Qu e Hullman (2016) ressaltam que um cenário de troca ou mudança (*shift*) entre dois pares de canais prejudica a consistência global porque codificam os mesmos campos de dados de maneiras diferentes. Em contrapartida, uma situação de atualização codifica os mesmos campos com os mesmos canais e campos diferentes com canais diferentes. Portanto, não deteriora a consistência do conjunto, embora ainda possa ter inconsistências de menor nível (QU; HULLMAN, 2016).

Ainda, Qu e Hullman (2016) provê mecanismos para se negociar *trade-offs* entre otimização local (apenas uma visualização) e as restrições de consistência global (múltiplas visualizações) e destacam que a reutilização dos eixos X e Y é um tipo de inconsistência entre visualizações com a qual a maioria do público está acostumado. Quando dois eixos X, por exemplo, codificam campos diferentes, os indivíduos tendem a examinar as anotações do eixo. Todavia, essa prática para tamanho e cor sobrecarrega os espectadores com uma

⁹ Se dois pares de codificação não têm nenhum campo em comum, Qu e Hullman (2016) também define essa situação como *update*.

carga cognitiva maior devido à necessidade de “desaprender” um conjunto de mapeamentos a cada vez que eles fazem a transição entre as visualizações¹⁰.

Tendo em vista que o conjunto de visualizações proposto neste estudo têm caráter complementar ao fornecer diferentes ângulos de visão sobre as coleções de dados biológicos, buscou-se empregar as recomendações dadas por Qu e Hullman (2016), especialmente mantendo-se fixa a paleta de cores usada em cada um dos gráficos referentes à mesma base de dados. Embora pequenas inconsistências ainda possam estar presentes, o conjunto de gráficos resultante é considerado adequado para integração em sistemas, construção de *dashboards* ou apresentação em produções acadêmicas que se beneficiarão de maior fluidez na sua leitura, além de evitar possíveis confusões aos especialistas no momento de executar as tarefas de manutenção e curadoria das coleções.

2.1.3 Visualizando coleções digitais de biodiversidade

Esta subseção será direcionada à apresentação de soluções de estado-da-arte presentes na literatura para visualização de dados biológicos. Tais ferramentas vão além da simples apresentação de dados, servindo também ao propósito de exploração, favorecendo pesquisadores no meio científico (FOX; HENDLER, 2011). As últimas décadas foram marcadas por diversas iniciativas de digitalização de coleções de biodiversidade por parte de grandes instituições como museus, herbários, universidades e outras instituições de pesquisa. Paralelamente à esse processo, podemos destacar o surgimento de diversas aplicações, muitas em ambiente *web*¹¹, que objetivam disponibilizar acesso a consultas, análises e explorações visuais dessas coleções.

Para citar algumas, AntMaps (JANICKI et al., 2016) oferece uma aplicação *web*¹² otimizada para visualizar e interagir com uma base de dados pré-processada de ocorrências de diferentes espécies de formigas, observadas em diversas localizações ao redor do mundo. O projeto *Map of Life* (JETZ; MCPHERSON; GURALNICK, 2012) provê uma aplicação *online* que visa servir como uma interface global e colaborativa que permite consultar mais de 370 milhões de registros de distribuição de espécies, além de possibilitar o armazenamento de novas informações.

HerbariaViz (AUER et al., 2011) é uma interface voltada à exposição de dados provenientes da observação de espécies da flora do Estado da Califórnia, EUA. Essa ferramenta objetiva três quesitos: (1) desenvolver um método eficiente para lidar com grandes volumes de dados espaço-temporais; (2) implementar estratégias para simbolizar dados agregados de forma a favorecer a exploração das dimensões espaço, tempo e espécie; (3) criar uma interface intuitiva a ponto de facilitar interações do tipo cliente-servidor.

¹⁰ Os usuários precisam consultar as legendas e guardar na memória diferentes valores toda vez que analisar um gráfico diferente.

¹¹ Aplicativos que são baseados na estrutura cliente/servidor, cujo acesso é realizado por meio da *internet*.

¹² Baseada em Leaflet.js, D3.js e jQuery.

GBIF-MAPA (FLEMONS et al., 2007) foi desenvolvida com a ambição de garantir acesso efetivo às coleções de biodiversidade disponíveis na plataforma GBIF. Do seu desenvolvimento, pode-se destacar os seguintes *milestones*: garantir acesso rápido e ágil à uma grande quantidade de *datasets*; a construção de uma interface de mapeamento flexível e de fácil manuseio.

Adicionalmente, pode-se citar projetos como AquaMaps, Rebioma e Movebank, que apresentam recursos visuais para busca ou análise dos seus dados de estudo. AquaMaps (READY et al., 2010) é um sistema de modelagem da distribuição global de espécies marinhas que permite *inputs* de conhecimento especializado sobre habitats. Sua interface de busca permite navegação por um mapa que, a partir da pré-seleção de determinadas espécies, exibe locais com a maior probabilidade destas serem encontradas. Ao clicar em um ponto com certas coordenadas de latitude e longitude, é exibida uma lista de todos os animais catalogados em suas bases de dados nas proximidades do ponto escolhido, contendo seus nomes científico e popular, além de imagens e um mapa apontando as regiões em que cada animal pode ser encontrado com alta probabilidade.

Rebioma (TSIKY, 2018) desenvolveu um portal de dados *online*¹³ para fornecer fácil acesso à informações de qualidade e atualizadas sobre ocorrências de espécies em Madagascar. Sua interface oferece visualizações simples, como gráficos de barras e pizza, para ilustrar a frequência de ocorrência de espécies. Dados geoespaciais são integrados ao *Open Street Maps*. Ainda, Movebank¹⁴ é uma ferramenta de gerenciamento e arquivamento de dados de rastreamento de animais baseada na *web*, apresentando uma interface de busca que alia dados geoespaciais à ferramenta *Open Street Maps* para, dada uma localidade, exibir uma lista de trabalhos que reportaram espécies na região selecionada.

Vale ressaltar que as ferramentas citadas até então têm em comum a característica de permitir a exploração apenas de *datasets* pré-carregados, seja diretamente na ferramenta, ou em alguma plataforma. Isto é, não oferecem ao pesquisador a possibilidade de analisar seu próprio conjunto de dados de interesse. A seguir, serão apresentadas ferramentas integradas ao ambiente *web* que possibilitam, em alguma instância, carregar uma base de dados própria.

Não obstante, Vesper (GRAHAM; KENNEDY, 2014) é uma ferramenta integrada ao *browser*, construída com base na biblioteca D3.js e recebe *inputs* de conjuntos de dados compatíveis com *Darwin Core Archives* (DwC-A). Há uma versão disponível *online* em que o usuário pode carregar e explorar um conjunto de dados próprio por meio de uma interface gráfica, sem a necessidade de conhecimentos de programação. Vesper também oferece recursos para verificação de *missing data* e suas soluções gráficas englobam representações do tipo *Sunburst* e *Icicle*, fornecendo uma visão geral da estrutura hierárquica

¹³ Acessível em <<http://data.rebioma.net/>>.

¹⁴ <<https://www.movebank.org/>>

de dados biológicos. Além disso, oferece integração com a aplicação *Open Street Maps* para visualização de dados espaciais. Variáveis de outros tipos podem ser visualizadas em histogramas. Vale destacar que é oferecido certo nível de imersão vertical por meio de elementos de interação.

BIDDSAT (OTEGUI; ARIÑO, 2012) trata-se de um ambiente *web* construído sobre a ferramenta Google Charts® para os conjuntos de dados do GBIF. Adicionalmente, oferece acesso a um conjunto limitado de coleções pré-carregadas e apenas editores registrados podem carregar um conjunto de dados de interesse particular. BIDDSAT produz gráficos simples¹⁵ com configurações pré-carregadas a partir de informações espaciais e temporais. Também oferece a funcionalidade de construir uma *treemap* para visualizar a taxonomia das espécies com um certo nível de agregação. Como diferencial, permite comparações entre diferentes versões de uma mesma coleção ao longo do tempo.

Há também um conjunto de soluções que, apesar de não apresentarem nenhum nível de integração com ambientes de navegação (*browsers*), oferecem funcionalidades integradas ao ambiente *desktop* ou estendem o potencial de certas linguagens de programação para o manuseio de dados PBR. Nesse contexto, Krona (ONDOV; BERGMAN; PHILLIPPY, 2015) é uma solução *desktop* que permite a produção de um gráfico interativo do tipo *Sunburst*, compatível com visualização em *browser*. Krona Tools, que é um conjunto de *scripts* destinados à criação de gráficos do Krona, é compatível com conjuntos de dados de vários meta-agregadores, além de texto bruto e arquivos XML. Foi desenvolvida para os sistemas operacionais MacOS e Linux, não estando disponível para Windows.

Não se pode deixar de citar bibliotecas criadas para a linguagem de programação R que oferecem um conjunto de extensões desenvolvido para se trabalhar com dados biológicos. O pacote *taxa* (FOSTER; CHAMBERLAIN; GRÜNWALD, 2018) fornece funcionalidades completas para se construir taxonomias do zero, bem como manipular estruturas de dados usando métodos de filtragem e classificação que são "taxonomicamente cientes". É altamente personalizável e há muitos métodos para construir ou importar dados taxonômicos de arquivos estáticos ou bancos de dados. A biblioteca *metacoder* (FOSTER; SHARPTON; GRÜNWALD, 2017) provê uma opção atrativa para visualizar estruturas de Árvores a partir de dados pré-processados, por exemplo, usando o pacote *taxa*. Ainda, *bdvis* (BARVE; OTEGUI, 2016) é um pacote R que provê um conjunto de funções para criar visualizações básicas para investigar, por exemplo, a presença de lacunas no inventário e a extensão das coberturas taxonômica, temporal e geográfica.

Por fim, vale destacar que o *framework* aqui proposto não tem a intenção de competir com nenhuma das soluções apresentadas ao longo dessa discussão, mas sim servir às necessidades de biólogos e curadores em identificar e corrigir erros de anotações em suas coleções, podendo também complementá-las no sentido de expor, sob diferentes ângulos,

¹⁵ Gráfico de barras, dispersão, pizza, dentre outros.

informações digitais de biodiversidade ao público interessado.

2.1.4 Visualização como ferramenta de apoio ao tratamento de dados

Nesta subseção, será desenvolvida uma discussão a respeito do uso de modelos de visualização para acessar a qualidade dos registros em um *dataset*. Até o momento em que este texto está sendo escrito, não se conhece sistemas de visualização de dados desenvolvidos especificamente para auxiliar na tarefa de melhorar a qualidade de dados biológicos do tipo PBR. O presente estudo objetiva preencher essa lacuna. Não obstante, serão expostos fundamentos, conceitos e práticas discutidos na literatura de forma a fundamentar o desenvolvimento do *framework* aqui proposto.

Na era do *big data*, o desenvolvimento cada vez mais intensivo de abordagens centradas em dados elevou a importância de se assegurar a qualidade e usabilidade das informações (LIU et al., 2018; SONG; SZAFIR, 2018; MCCURDY; GERDES; MEYER, 2018; FAN; GEERTS, 2012). Nas etapas de coleta e pré-processamento de dados, podem ocorrer inconsistências como duplicações, imprecisões¹⁶, perda de registros ou, até mesmo, aplicação de transformações irreversíveis¹⁷ nas informações originais (DOMADIYA; RAO, 2013; MODI; RAO; PATEL, 2010; WU; CHIANG; CHEN, 2006).

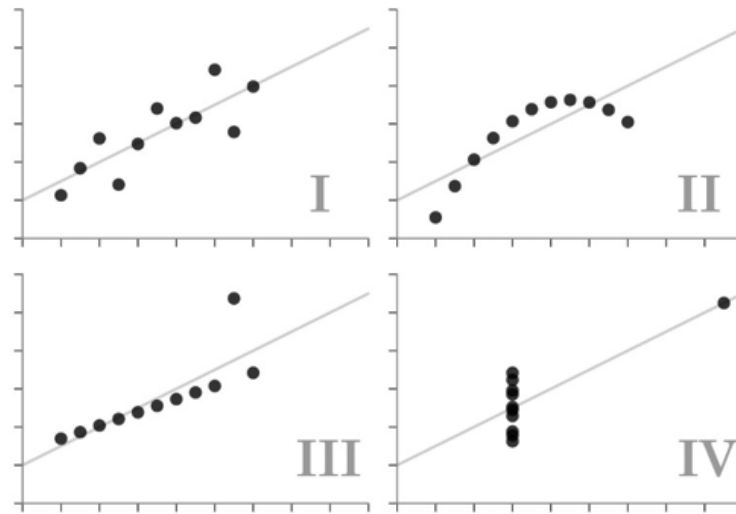
Em muitas aplicações, o processo de limpeza dos dados não pode ser completamente automatizado devido à ambiguidade de alguns erros (LIU et al., 2018) e a necessidade de se acessar o conhecimento de especialistas de domínio a fim de verificar os resultados de ajustes de qualidade e extrair melhor performance de algoritmos (BEKRI; PEINSIPP-BYMA, 2016). O conceito de *human-in-the-loop* têm sido cada vez mais explorado na literatura como forma de combinar o conhecimento de especialistas ao poder de processamento dos computadores para tarefas associadas ao tratamento de dados (SONG; SZAFIR, 2018; MCCURDY; GERDES; MEYER, 2018; GSCHWANDTNER et al., 2014).

Nesse contexto, a importância do campo de Visualização da Informação é fundamentada na provisão de modelos mentais de informações (LIU et al., 2014). Um simples exemplo, na figura 8, pode ilustrar a eficácia do uso de recursos visuais para representar conjuntos de dados de forma a identificar diferenças em suas composições. Nesse exemplo clássico de InfoVis, é fácil ver que não se pode confiar cegamente em estatísticas descritivas para analisar *datasets*. De fato, todos os conjuntos de exemplo na figura 8 possuem os mesmos valores de média, desvio-padrão e correlação. O uso desse tipo de recurso visual têm o potencial de facilitar, por exemplo, a detecção de *outliers* ao reforçar aspectos da cognição humana e acessar o conhecimento prévio de analistas.

¹⁶ Anotações erradas devido a erros de digitação é um problema recorrente em dados do tipo PBR onde, muitas vezes, esses registros são feitos à mão diretamente do campo de coleta ou, ainda, resgatados a partir de livros antigos onde a escrita pode estar comprometida pela ação de intempéries.

¹⁷ Por exemplo, em alguns casos, aplica-se uma bijeção (mapeamento de nomes para códigos de identificação) para preservar a identidade de membros integrantes da base de dados.

Figura 8 – Quarteto de Anscombe.



Fonte: (ANSCOMBE, 1973)

Nota: Cada conjunto de dados possui exatamente as mesmas estatísticas descritivas (média, desvio-padrão e correlação). No entanto, os *datasets* são completamente diferentes.

Nesse sentido, Broeck et al. (2005) propõe um *framework* baseado em três etapas: Triagem¹⁸ → Diagnóstico → Correção. Na etapa de triagem, Broeck et al. (2005) destaca a importância de se investigar tipos básicos de inconformidades, como falta ou excesso de dados, presença de *outliers*, padrões incomuns em distribuições e resultados não esperados em análises preliminares. Ressalta ainda que métodos de triagem não precisam ser restringidos à métodos estatísticos. Muitos *outliers* são detectados devido à experiência prévia do investigador, estudos-piloto, evidências na literatura ou, apenas, bom senso (BROECK et al., 2005), realçando diretamente a relevância do conceito de *human-in-the-loop*.

A fase de diagnóstico busca investigar a verdadeira natureza das inconsistências identificadas na etapa de triagem. Exemplos de possíveis diagnósticos, para cada registro, são: errôneo, extremo verdadeiro, falso positivo¹⁹ ou idiopático (para o caso de nenhuma explicação ter sido encontrada, mas ainda restar alguma suspeita). Após a identificação de erros e valores inconsistentes, o pesquisador deve decidir o que fazer a respeito de tais observações. Na última etapa do processo, as opções limitam-se a corrigir, excluir ou manter os registros inalterados.

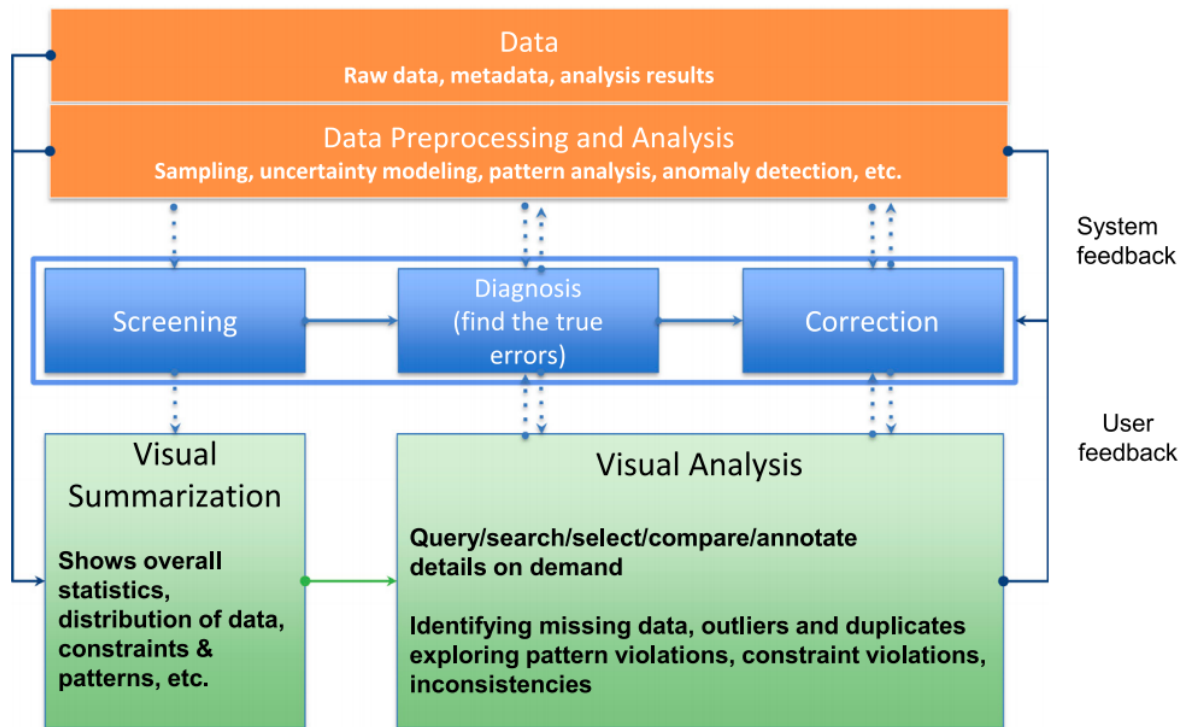
Liu et al. (2018) estende o *framework* proposto por Broeck et al. (2005) de forma a incorporar ferramentas de análise visual para investigação e melhoria da qualidade dos dados. Seu principal objetivo é auxiliar o usuário a detectar potenciais problemas em

¹⁸ Tradução direta do termo original *Screening*.

¹⁹ Isto é, a expectativa anterior estava incorreta. Nesse caso, testa-se a hipótese de que o ponto investigado é, de fato, um *outlier*. Um falso positivo, nesse contexto, indica que esse ponto foi previamente apontado como inconsistente de forma indevida.

suas bases de dados e prover métodos eficientes e convenientes à realização de ajustes de qualidade à luz de conhecimento de domínio e experiências prévias do analista. A figura 9 ilustra as camadas que compõem o *framework* desenvolvido em Liu et al. (2018): (1) manuseio de dados - na cor laranja; (2) visualização - representada na cor verde - conectada à camada (3) interação - em azul.

Figura 9 – *Framework* de análise visual proposto por Liu et al. (2018).



Fonte: (LIU et al., 2018)

Nota: Camadas do *framework* de análise visual proposto por Liu et al. (2018): (1) manuseio de dados - na cor laranja; (2) visualização - representada na cor verde - conectada à camada (3) interação - em azul.

Cada um desses módulos foi, respectivamente, desenhado para (LIU et al., 2018): (1) extração de *insights*; (2) representação intuitiva e interpretação de dados; (3) facilitar exploração e análise segundo o conceito de *human-in-the-loop*. A partir dessa estrutura, pode-se inferir o seguinte *pipeline*:

1. A partir dos dados coletados como *input*, é realizada uma etapa de pré-processamento com o objetivo de revelar padrões, descobrir *outliers* ou, até mesmo, recuperar dados ausentes.
2. Com base nos resultados da etapa anterior, é acionada uma etapa de interação incorporando os módulos originalmente propostos em Broeck et al. (2005), com a adição de soluções visuais.

- 2.1 Triagem: nesse módulo, o usuário é capaz de ilustrar uma visão geral da coleção de dados, medidas estatísticas e determinados padrões por meio de gráficos intuitivos.
- 2.2 Diagnóstico: por intermédio das visualizações, o usuário é capaz de identificar potenciais inconsistências (*missing data*, duplicações, etc).
- 2.3 Correção: Finalmente, a correção dos dados pode ser efetuada por métodos interativos.

Nesse ponto, torna-se clara a importância de sistemas de visualização no suporte à interpretação de dados e tomada de decisão (LIU et al., 2018). Essa discussão pode ser estendida facilmente ao contexto de coleções de biodiversidade. Não obstante, este estudo têm como objetivo elaborar um *framework* de visualização para dados biológicos do tipo PBR, à luz dos trabalhos de Broeck et al. (2005) e Liu et al. (2018), voltado para a melhoria de qualidade sob a supervisão de *experts*. No decorrer dos próximos capítulos, são apresentados métodos e aplicações de conceitos relevantes que podem servir de base ao desenvolvimento de sistemas de visualização voltados à auxiliar especialistas de domínio nas tarefas de manutenção e curadoria de seus *datasets*.

2.2 Biodiversidade

Pesquisadores estimam que o primeiro ser vivo no planeta Terra surgiu a aproximadamente 3,8 bilhões de anos. Dessa data até os dias de hoje, uma série de mutações, transformações e adaptações aconteceram de forma a dar suporte à vida como a conhecemos hoje. Grande parte do entendimento que temos da nossa trajetória como raça humana se deve a uma série de esforços científicos que se propuseram a estudar a evolução e a mecânica da vida na Terra. Em particular, estudos em biodiversidade são responsáveis por catalogar e descrever as diferentes variedades de formas de vida, seja em um dado ecossistema, um certo bioma ou região geográfica. Esses estudos são importantes não só para compreender a trajetória evolucionária dos seres vivos, mas também os aspectos que mantêm o equilíbrio na coabitação e interação entre diferentes espécies.

Nesse contexto, grandes instituições de história natural (museus, herbários, universidades, instituições de pesquisa, etc.) têm um papel crucial uma vez que foram as maiores responsáveis pela documentação e armazenamento de dados levantados por taxonomistas a respeito de uma ampla diversidade de espécimes ao longo do tempo. Registros primários de biodiversidade (PBR²⁰) contêm informações como data e local de coleta, identificação taxonômica, o taxonomista responsável por aquele exemplar, dentre outras. Conforme

²⁰ *Primary Biodiversity Records.*

Drew (2011) destaca, tais coleções de dados apresentam registros precisos que podem ser usados para reconstruir o histórico de diferentes espécies em uma extensão que vai além do tempo de carreira de um único pesquisador. Várias décadas de esforço em catalogar e armazenar estas informações deram origem a grandes coleções de dados que dão suporte a inúmeras pesquisas em diferentes áreas do conhecimento biológico.

Muitas são as formas nas quais estas coleções beneficiam a comunidade científica (SUAREZ; TSUTSUI, 2004). Por exemplo, esses dados podem ser utilizados para auxiliar no monitoramento de espécies invasivas (SRIVASTAVA et al., 2020; BARBET-MASSIN et al., 2018; PAINI et al., 2016; MULLER, 2015; SARNAT, 2008), na gestão de parques nacionais (JACOBS et al., 2018), no mapeamento de doenças patógenas e parasíticas (PINTO et al., 2010), revelar histórico de doenças (PERSING et al., 1990; Marshall III et al., 1994), na avaliação do impacto da poluição (MODICA; LANUZA; GARCÍA-CASTRILLO, 2020) e de mudanças ambientais (ROBBIRT et al., 2011; LISTER; GROUP et al., 2011), no mapeamento processos evolucionários (HOLMES et al., 2016) e também são cruciais no ensino e formação de novos pesquisadores e cientistas (MONFILS et al., 2017). Tarli, Grandcolas e Pellens (2018) e Suarez e Tsutsui (2004) trazem boas discussões a respeito da importância das coleções de museus para a sociedade e toda a comunidade científica.

Iniciativas de digitalização como Sistema de Informação sobre a Biodiversidade Brasileira²¹ (SiBBR), *Inter-American Biodiversity Information Network* (IABIN), *Comision Nacional para el Conocimiento y Uso de la Biodiversidad* (CONABIO), *Global Biodiversity Information Facility* (GBIF) e várias outras tiveram enfoque em disponibilizar grande parte dessas coleções em um ambiente *online*. Esse processo de democratização da informação tem o potencial de trazer implicações de larga escala para a maneira como as pessoas interagem e fazem uso destas coleções (ADDIS et al., 2005). Drew (2011) enfatiza que essas bases de dados digitais constituem uma mudança do que as coleções de história natural representam e a quem servem. Destaca ainda que, com a democratização do acesso à internet, essas coleções podem alcançar segmentos sociais restritos geograficamente ou economicamente.

No entanto, garantir a qualidade desses registros é uma preocupação presente na literatura, discutida desde o trabalho de Thomas e Mintz (1998). Fatores como o grande volume de dados e a interdependência entre múltiplas variáveis dificultam determinar até que ponto esses dados estão completos, corretos e se, de fato, fornecem uma boa cobertura geográfica, temporal e taxonômica das espécies em questão. Garantir a acurácia no registro de espécimes é uma tarefa que inicia-se no campo de coleta e perpetua-se até o momento em que aquele novo dado é gravado, seja escrevendo em uma folha de papel ou fazendo uso de planilhas ou dispositivos eletrônicos²². No entanto, por motivos que vão desde a

²¹ <<https://ipt.sibbr.gov.br/mnrj/>>

²² A maneira como esses registros são efetuados foi modificando-se ao longo dos anos, mas ainda nos

grande extensão temporal desses registros até dificuldades de acesso ao local de coleta, raridade de certas espécies e questões orçamentárias referentes à grandes expedições, a existência de uma ferramenta capaz de verificar e disponibilizar os registros existentes de forma a possibilitar a detecção de erros e incongruências é de importância direta.

Não obstante, o presente trabalho se propõe a aplicar princípios e técnicas advindos da literatura de Visualização da Informação ao contexto de coleções científicas digitais de história natural, visando fornecer um conjunto de técnicas e visualizações capazes de expor tais coleções de forma tal que a qualidade dos registros seja facilmente acessada por um especialista. Uma vez que possíveis incongruências sejam resolvidas, as visualizações propostas neste estudo podem facilmente ser adaptadas para servir ao propósito de apresentar a coleção ao público em geral, destacando suas riquezas em termos de cobertura geográfica, temporal e entre diferentes espécies. Além disso, também destacar a contribuição de cada pesquisador e coletor para a constituição da coleção é uma abordagem inédita trazida na presente aplicação. Adicionalmente, visando esse propósito, a presença de elementos dinâmicos e interativos possibilitam maior grau de interação do usuário com a coleção e facilita a geração de diferentes visualizações estáticas a partir da aplicação de filtros e seleções no conjunto de dados original, podendo ser publicadas em veículos de circulação impressa, como artigos científicos.

Vale destacar que o *framework* desenvolvido no presente estudo destina-se à exibição e tratamento de coleções de dados do tipo PBR, tendo seus produtos melhor aplicados à *data papers*²³, i.e., documentos revisados pela comunidade científica que descrevem um conjunto de dados de biodiversidade. Alternativamente a documentos de pesquisas tradicionais, *data papers* não objetivam descrever possíveis hipóteses e conclusões, mas sim relatar dados e as circunstâncias de sua coleta. Apesar de o escopo deste trabalho estar delimitado à apresentação de bases de dados de biodiversidade, alguns de seus produtos podem ser facilmente ajustados para auxiliar no processo de inferência quanto à características de determinadas espécies²⁴.

Com o intuito de documentar as principais abordagens utilizadas por *data papers* para retratar suas bases de dados, foram coletados 227 artigos de biodiversidade publicados nas plataformas *Biodiversity Data Journal* (BDJ²⁵) e *Global Biodiversity Information Facility* (GBIF²⁶) que, por sua vez, permitem acesso livre e gratuito a todo seu conteúdo. Dos 39 artigos provenientes da plataforma GBIF, 2 estavam armazenados na plataforma BDJ e, portanto, foram retirados para evitar que estes sejam contados em duplicidade.

dias atuais, muitos registros são feitos em folhas de papel diretamente do campo de coleta que, muitas vezes, é de difícil acesso.

²³ Para mais informações a respeito de *data papers*, vide <<https://www.gbif.es/en/datos-biodiversidad/participa-en-gbif-es/data-papers/>>.

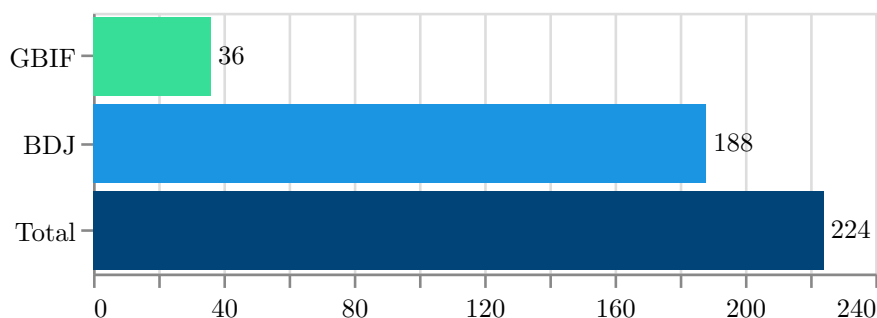
²⁴ No entanto, esse exercício é deixado a cargo de trabalhos futuros.

²⁵ <<https://bdj.pensoft.net/>>

²⁶ <<https://www.gbif.es/>>

Ainda, um dos artigos (ESTRADA-PENÑA; FUENTE, 2016) não tem o objetivo de apresentar uma coleção de dados, mas sim de usá-la para auxiliar na criação de uma rede de interações entre espécies e, portanto, foge ao escopo do presente estudo. A figura 10 sintetiza o total de *data papers* restantes segundo suas fontes.

Figura 10 – Total de artigos analisados de cada uma das fontes: GBIF e BDJ.



Fonte: Elaboração própria.

Após a coleta, cada um dos 224 artigos foram analisados buscando-se averiguar se algum recurso de visualização de dados²⁷ foi utilizado com o propósito de ilustrar a cobertura geográfica, taxonômica ou temporal de suas respectivas coleções de dados. Foi constatado que 182 dos 224 artigos (aproximadamente 81%) fazem uso de algum tipo de recurso visual para expressar certas características de suas coleções, enquanto os outros 19% (42 dos 224 artigos) fazem uso de elementos textuais, como listas e tabelas, ou fotografias para registrar e ilustrar os espécimes, seu local ou período de coleta²⁸.

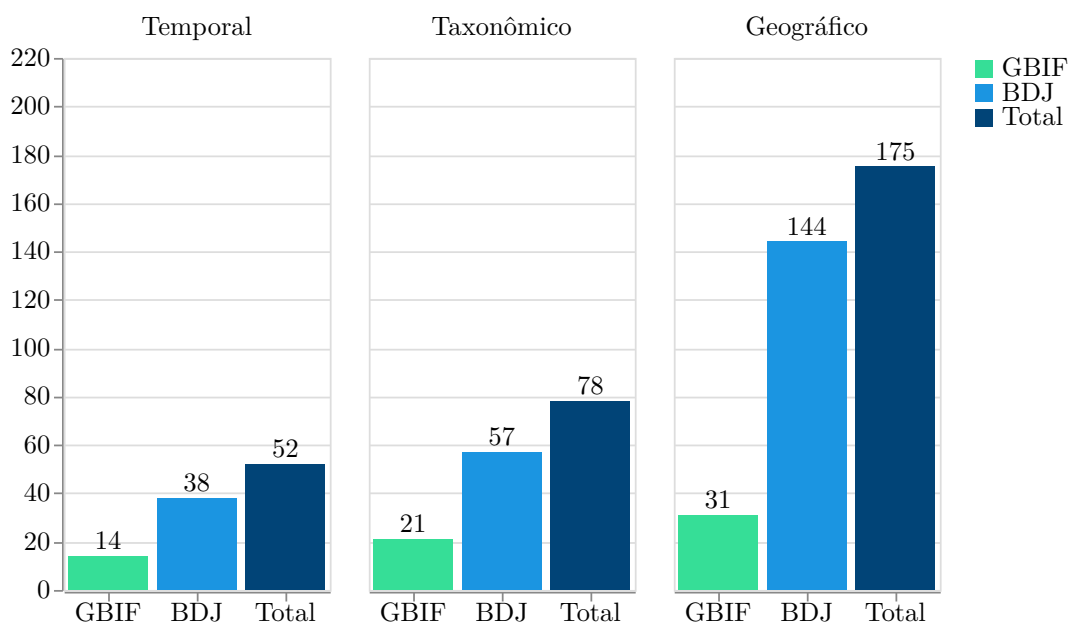
A figura 11 ilustra a distribuição de artigos que faz uso de elementos gráficos para apresentar características geográficas, temporais ou taxonômicas das coleções. Foi considerado que um dado artigo apresenta elementos gráficos para ilustrar a cobertura geográfica de suas coleções se este incorpora pelo menos um elemento visual como, por exemplo, gráficos, mapas ou imagens de satélite que retratem a localização exata dos espécimes coletados ou, ao menos, delimite sua região de estudo. De maneira análoga, considerou-se que um *data paper* emprega recursos visuais para ilustrar a distribuição temporal e taxonômica, respectivamente, se apresentar pelo menos uma imagem que represente o número de espécimes coletados ao longo do tempo²⁹ e a contagem, proporção ou hierarquia de exemplares na coleção segundo algum nível de taxonomia, como espécie, ordem, família, etc.

²⁷ Gráficos, imagens, figuras, etc. Aqui não foram consideradas fotografias dos espécimes ou do local de coleta por não conter nenhum critério matemático para mapear dados em código visual.

²⁸ O fato de um *data paper* não utilizar gráficos para apresentar suas coleções está associado a fatores como a raridade dos espécimes, localização geográfica específica ou período temporal restrito que, por sua vez, estão associados a uma menor quantidade de exemplares na coleção.

²⁹ Embora a unidade de tempo possa variar (meses, anos ou por amostragem), foi considerado se o recurso visual apresentado representa o crescimento do número de exemplares na coleção ao longo do tempo.

Figura 11 – Quantidade de artigos que apresentam recursos gráficos para ilustrar a cobertura temporal, taxonômica e geográfica, respectivamente.



Fonte: Elaboração própria.

Nota: Os três tipos de cobertura não são mutuamente exclusivas, i.e., há artigos que trazem visualizações de dados para mais de um tipo de distribuição; Cada tipo de cobertura foi analisada separadamente, i.e., todas as contagens apresentadas são em relação ao total de 224 artigos coletados.

Como podemos ver facilmente na figura 11, aproximadamente 78% dos artigos analisados apresentam visualizações para dados geográficos de suas coleções. Isso pode estar associado ao fato de a localização ser um forte indicativo para se inferir possíveis características dos espécimes, o que pode ser mais facilmente compreendido se levarmos em consideração fatores adaptativos e evolucionários. Em outras palavras, espécimes coletados nas regiões dos trópicos tendem a ser visivelmente distintos de seres coletados em regiões polares de forma que, para um especialista, o local de coleta de um exemplar é um forte indicativo de quais características ele pode apresentar e em que posição ele pode situar-se na árvore taxonômica. É importante ressaltar que isso não significa que, para 22% desses artigos, o local de coleta dos exemplares seja, de alguma forma, menos relevante e não mereça destaque. Fatores como a raridade dos espécimes analisados ou o processo de coleta³⁰ podem justificar a escolha de alguns *data papers* não retratarem tal característica de suas bases de dados graficamente. Mais comentários sobre a utilização de recursos gráficos que fazem menção à localização geográfica estão presentes na subseção 2.2.1.

De maneira semelhante, pode-se pontuar que aproximadamente 35% dos 224 *papers* incorporaram algum tipo de abordagem visual para destacar a taxonomia dos seus exemplares. Geralmente, esses artigos fizeram uso de gráficos de barras e pizza para

³⁰ Algumas coleções foram compostas por amostragem em postos de coleta específicos e, portanto, alguns artigos optaram por apresentar essa informação de forma textual.

ilustrar a proporção de espécimes em um certo nível da árvore taxonômica, como espécie, gênero, ordem, família, dentre outros. A subseção 2.2.3 detalha a maneira em que esses artigos retrataram tal informação. No entanto, vale destacar que a maioria dos *data papers* analisados optou por apresentar essa informação de forma textual, estruturada em listas e tabelas.

Por fim, nota-se também na figura 11 que 52 dos 224 artigos (aprox. 23%) usam gráficos de barras ou linhas, por exemplo, para retratar o crescimento de suas coleções ao longo do tempo. A subseção 2.2.2 detalha os elementos gráficos trazidos por esses artigos para apresentar essa informação. Não obstante, notou-se que a maioria dos documentos optou por trazer essa informação de forma textual, muitas vezes destacando apenas o intervalo de tempo no qual os espécimes foram coletados.

Dando continuidade à discussão iniciada nessa seção, as próximas subseções trazem breves comentários a respeito dos *data papers* de biodiversidade coletados que fazem uso de algum tipo de recurso para visualização de dados, visando apresentar suas coleções biológicas e evidenciar alguns de seus aspectos. A subseção 2.2.1 aborda a maneira pela qual os artigos buscaram retratar a coleta e distribuição dos espécimes em termos de distribuição geográfica; a subseção 2.2.2 destaca as escolhas mais comuns com o intuito de ilustrar a evolução temporal de suas coleções; a subseção 2.2.3 finaliza esta seção ilustrando como a variedade de espécies em termos taxonômicos é apresentada visualmente.

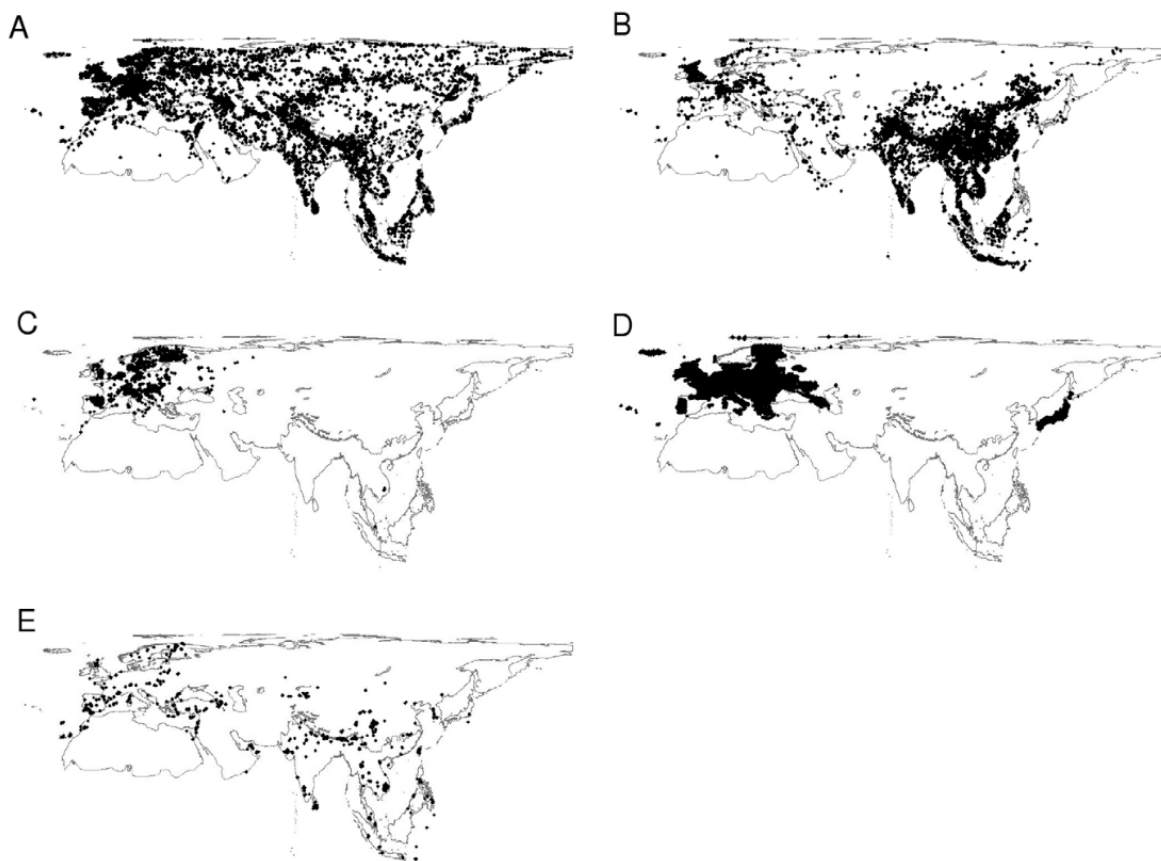
2.2.1 Cobertura geográfica

Esta subseção tem como objetivo apontar as principais escolhas adotadas por *data papers* de biodiversidade para retratar o aspecto geográfico de suas coleções de dados. Não obstante, são apresentados exemplos retirados diretamente dos artigos originais e, sempre que pertinente, comentados possíveis pontos em que tais visualizações poderiam ser enriquecidas levando em consideração o *framework* desenvolvido no presente trabalho.

Em geral, o local de coleta de espécimes é registrado por meio de técnicas de georreferenciamento, i.e., pontos em coordenadas de latitude e longitude que, quando projetados sobre um mapa em um plano bidimensional, podem nos fornecer uma ilustração eficaz para apresentar onde estes exemplares foram encontrados. Vamos iniciar essa discussão apresentando um exemplo de aplicação gráfica simples e bem-sucedida executada por Boakes et al. (2010), ilustrada na figura 12, que apesar de não fazer uso de muitos recursos visuais (como cor, forma, tamanho, etc.), cumpre bem seu objetivo de mostrar a abrangência espacial de diferentes fontes de dados contendo registros de diferentes espécies de pássaros. Boakes et al. (2010) se propôs a investigar a presença de viés geográfico em diferentes coleções de dados e, portanto, não se enquadra na categoria de *data paper* por não ter o objetivo bem definido de apresentar uma coleção de dados, mas sim buscar fazer inferência sobre a natureza dos dados encontrados em diferentes tipos de coleções. Apesar

disso, vale ser mencionado para exemplificar o potencial desse tipo de recurso visual para destacar e revelar padrões relevantes ao mesmo tempo em que apresenta a totalidade dos registros presentes nos *datasets*³¹.

Figura 12 – Distribuição espacial de registros de diferentes fontes de dados.



Fonte: (BOAKES et al., 2010).

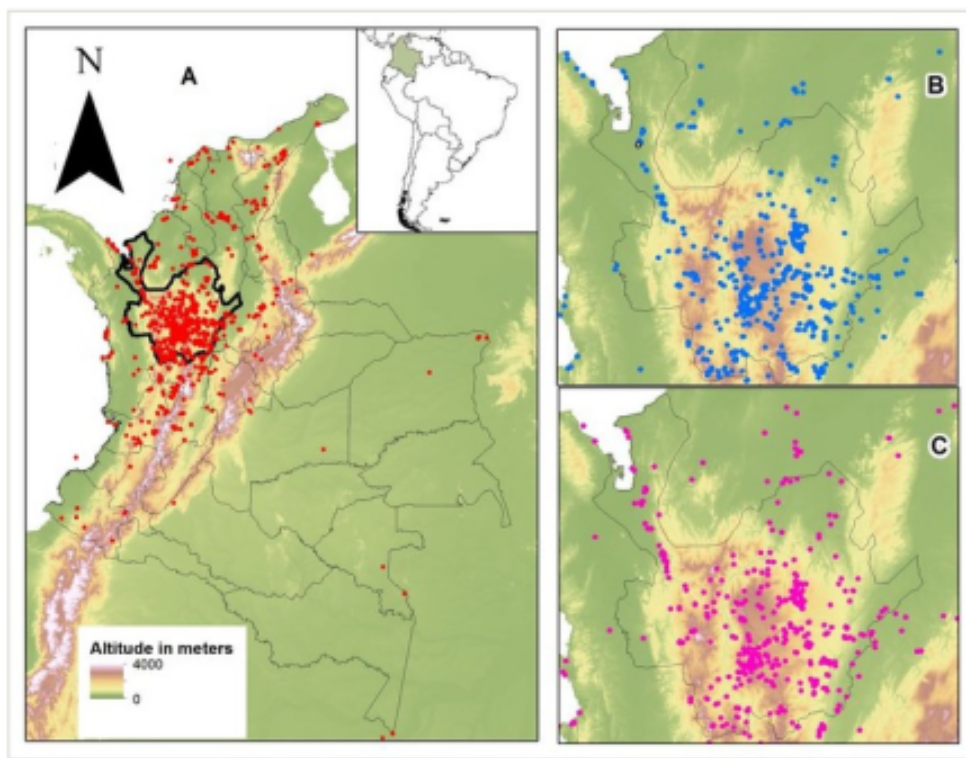
Nota: Distribuição espacial de registros de diferentes espécies de pássaros, segundo as seguintes fontes: A) Museus; B) Literatura; C) Anilhagem; D) Atlas; E) Relatórios de sites de viagem.

Adicionalmente, essa é a representação mais usual empregada por estudos das mais diversas áreas do conhecimento quando se faz necessária a apresentação de dados geoespaciais. No caso de pesquisas em biodiversidade, o registro da localização, muitas vezes, é feita pelo pesquisador no momento em que o espécime é recolhido em campo por meio de aparelhos de georreferenciamento que marcam, por exemplo, as coordenadas de GPS. Quando disponível, a localização exata de coleta de cada um dos espécimes é destacada nos *data papers* conforme apresentado nas figuras 13 e 14, a seguir.

Na figura 13 os pontos de interesse são apresentados sobre o desenho de um mapa, fidedigno às proporções reais, porém com menos detalhes que uma imagem de satélite

³¹ Com esse tipo de recurso, Boakes et al. (2010) concluiu que coleções de Museus são as mais completas no sentido de abrangência geográfica e, portanto, são as que apresentam menor viés.

Figura 13 – Distribuição geográfica dos espécimes depositados na coleção do Museu de Herpetologia da Universidade de Antioquia.



Fonte: (ORTIZ-YUSTY et al., 2015)

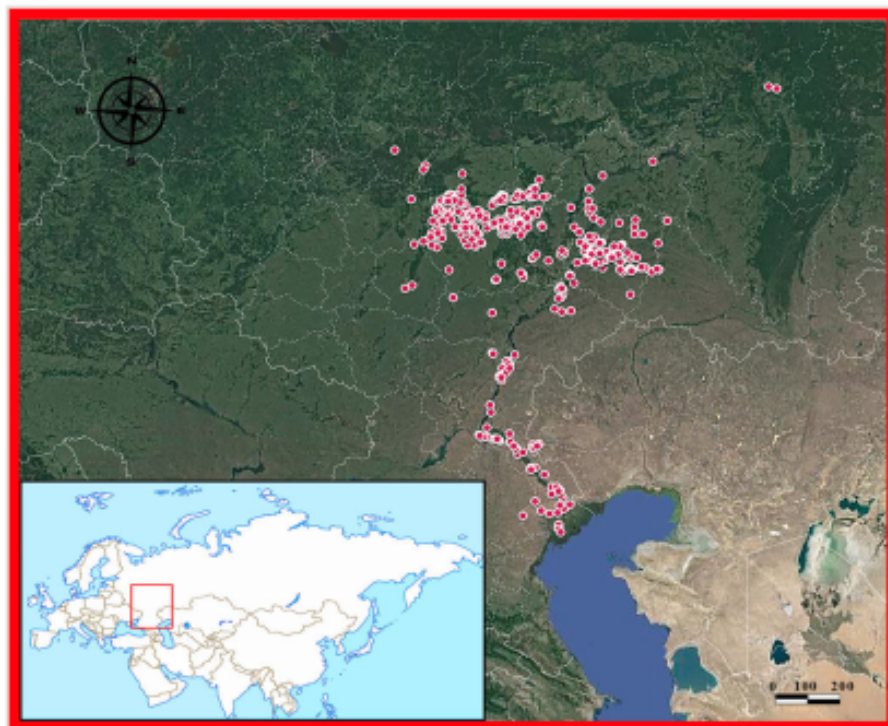
Nota: A) Todos os espécimes (Anfíbios e Répteis); B) Anfíbios; C) Répteis; todos coletados na província de Antioquia (localizada no Noroeste da Colômbia).

real. Nessa abordagem são adicionados apenas detalhes suficientes para facilitar a leitura e interpretação do gráfico, chamando a atenção do leitor para os pontos pertencentes à coleção. O mesmo tipo de exercício é realizado na figura 14, porém com uma imagem real gerada por satélite em ferramentas como Google Maps³² e Google Earth³³, onde mais informações referentes à vegetação, terreno e outros elementos existentes na região são adicionadas. Ambos os gráficos poderiam se beneficiar em adotar uma escala de cores que mapeasse, por exemplo, elementos taxonômicos desses espécimes, como ordens ou famílias. Essa estratégia poderia revelar padrões ainda não explorado nesses gráficos, como por hipótese, destacar que exemplares de uma mesma família foram todos coletados em localidades próximas. Assim, o presente estudo usa estratégias semelhantes para explorar possíveis padrões geográficos em relação a taxonomia de determinados exemplares das coleções.

³² <<https://www.google.com.br/maps/>>

³³ <<https://www.google.com.br/intl/pt-BR/earth/>>

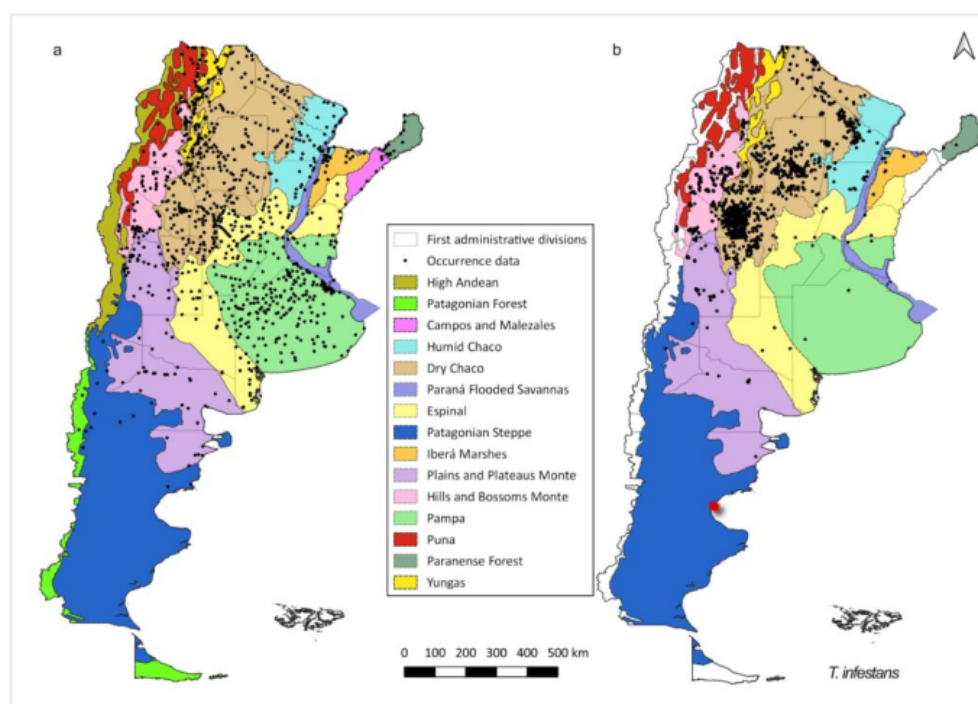
Figura 14 – Locais de coletas de répteis na bacia do rio Volga.



Fonte: (BAKIEV et al., 2020)

Nota: As localidades de coleta dos répteis são mostradas nos pontos em vermelho. O mapa foi produzido na ferramenta Google Maps® pelos autores do artigo original.

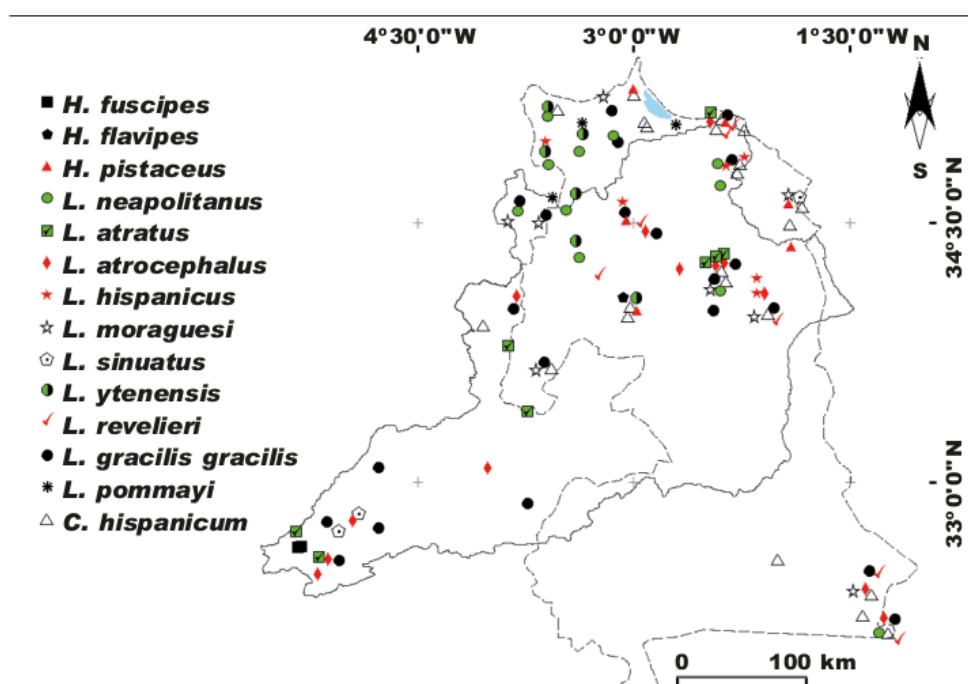
Alguns *data papers* como Ceccarelli et al. (2020) e Mabrouki et al. (2018), além de apresentar os pontos de coleta em suas coleções, buscam fazer uso também de esquemas de cores para destacar certos padrões por biomas (figura 15) e por diferentes gêneros taxonômicos (figura 16). No entanto, nota-se que não há um padrão na definição dos elementos gráficos empregados e, muitas vezes, são escolhidas certas representações que parecem não considerar recomendações de boas práticas fornecidas pela literatura de InfoVis. Na figura 15, o emprego das cores está bem definido e nota-se uma maior concentração de ocorrências em um certo bioma para registros após o ano 2000, i.e., os pontos estão mais concentrados na região de cor marrom no gráfico do lado direito. Já na figura 16, forma e cor representam, juntas, diferentes gêneros dos animais coletados da espécie *Hydrophilidae*. Porém, para o leitor, é necessária mais atenção para identificar a qual gênero cada um dos pontos no gráfico pertencem. Além disso, não parece haver um critério bem definido para atribuir uma certa forma a um dado espécime, o que impossibilita a um não-especialista saber se, por exemplo, animais representados na forma de um triângulo vermelho são mais semelhantes a animais codificados como um triângulo preto, em comparação a um dado animal apresentado na forma de uma bola verde.

Figura 15 – Distribuição de dados de ocorrência para *T. infestans*.

Fonte: (CECCARELLI et al., 2020)

Nota: (a) registros pré-2000; (b) registros pós-2000. Os dados de ocorrência de espécies são mostrados como pontos pretos. As áreas coloridas representam ecorregiões.

Figura 16 – Distribuição da espécie Hydrophilidae na área de estudo.

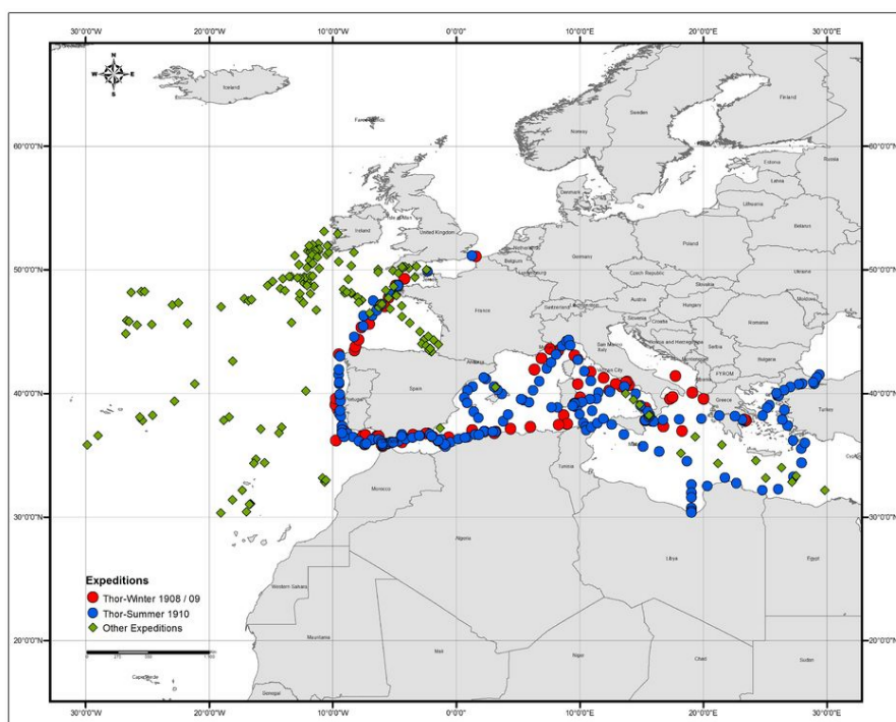


Fonte: (MABROUKI et al., 2018)

Nota: Cor e forma representam diferentes gêneros da espécie Hydrophilidae.

Há casos ainda nos quais a informação de onde os espécimes foram coletados não está disponível com esse nível de detalhamento, mas sim agregadas por estação de coleta ou por região (país, estado, cidade ou microrregião³⁴). Essa característica pode estar diretamente associada ao método de amostragem empregado. Como exemplo, podemos citar os casos em que os exemplares foram coletados em grandes expedições, conforme apresentado na figura 17; casos em que o estudo se propõe a catalogar espécies encontradas em um certo parque ou reserva nacional (figura 18); ou situações em que a localização geográfica é registrada apenas em termos de Estado ou País (uma alternativa factível, adotada em alguns artigos, é apresentar a densidade de espécimes coletados, como na figura 19), ou simplesmente não está disponível³⁵.

Figura 17 – Mapa das estações amostradas por Thor e outras embarcações durante as expedições principais de 1908-1909, 1910 e expedições adicionais de 1905-1906 e 1911-1912.



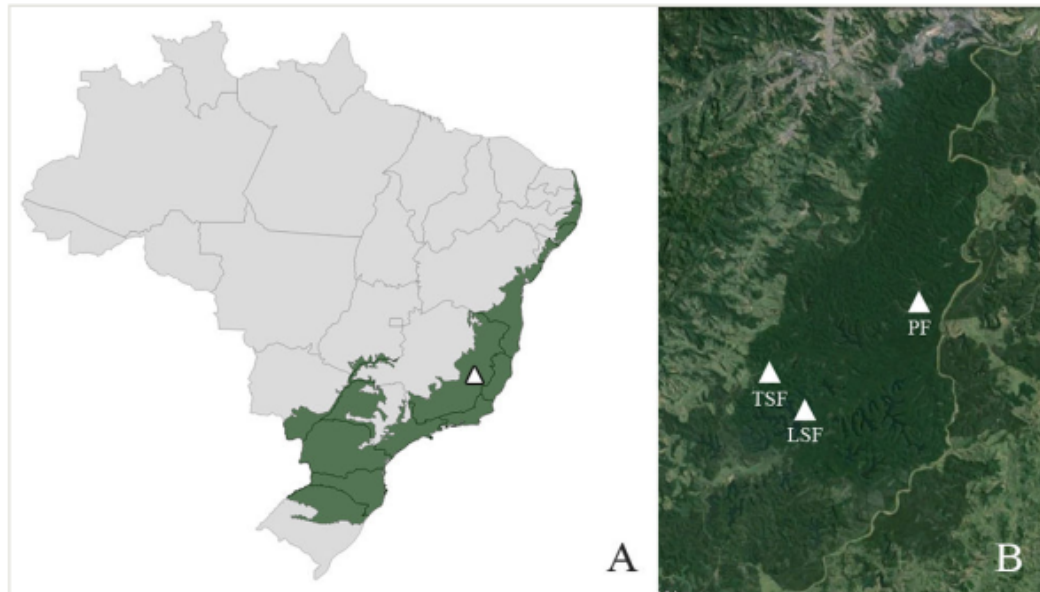
Fonte: (MAVRAKI et al., 2016)

Nota: Expedições são mapeadas por cor; Vermelho: Thor-Inverno 1908/09; Azul: Thor-Verão 1910; Verde: Outras expedições.

³⁴ Caso exista, por exemplo, uma certa margem de erro na determinação da localização exata dos objetos de estudo.

³⁵ Essa dificuldade pode surgir devido ao fato de que registros muito antigos, por exemplo do começo do século XX, eram feitos de forma completamente manual diretamente do campo de coleta pelo pesquisador, usando papel e caneta. Por motivos como: os registros terem se perdido ao longo do tempo; o pesquisador não registrou uma localização precisa (não haviam padrões bem definidos para se fazer esses registros em campo) ou, ainda, um pesquisador dos dias atuais pode não ter compreendido o registro original no momento de fazer a digitalização dessas informações

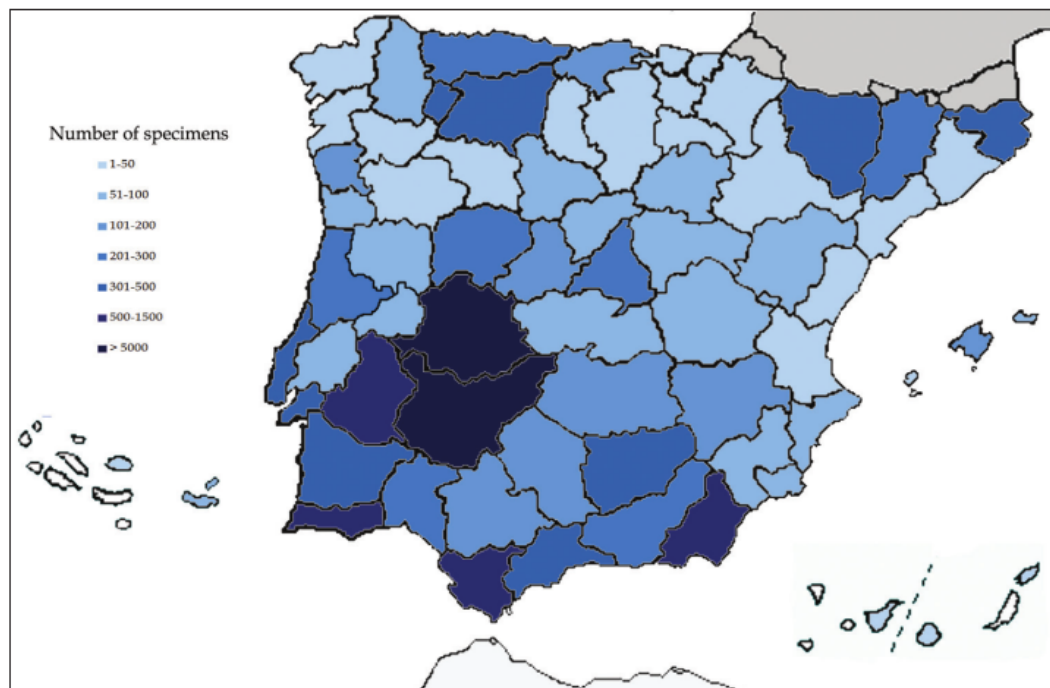
Figura 18 – Mapa da área de estudo de Santos et al. (2017).



Fonte: (SANTOS et al., 2017)

Nota: A. Mata Atlântica brasileira (verde) com a localização do Parque Estadual do Rio Doce em destaque (triângulo branco); B. Mapa da área de estudo, com os três postos de amostragem em destaque. PF: floresta primária; TSF: floresta secundária alta; LSF: secundário baixo floresta.

Figura 19 – Distribuição geográfica de espécimes na Península Ibérica, Baleares, Canarias, Madeira e Açores.



Fonte: (ESPINOSA; LÓPEZ, 2013)

Nota: O *dataset* apresentado nesse estudo tem registros que datam desde 1911, o que pode justificar a escolha gráfica utilizada caso a localização geográfica dos registros mais antigos não possa ser estabelecida com alto nível de precisão.

Vale ressaltar ainda que, além de apresentações gráficas semelhantes às evidenciadas nessa subseção, há *data papers* que fazem uso de abordagens menos usuais para representar a distribuição geográfica, como gráficos de pizza e barras.

Em suma, esta subseção é concluída apontando que, dos 224 *papers* analisados, nem sempre a maneira pela qual estes artigos apresenta a distribuição espacial de suas bases de dados é facilmente compreensível e, em muitos casos, os canais de representação da informação (como cor, forma, tamanho, etc.) parecem não refletir as boas práticas amplamente recomendadas pela literatura de InfoVis. Nesse sentido, o presente trabalho visa contribuir para a literatura de biodiversidade ao apontar um conjunto de técnicas e propostas gráficas sugeridas por estudos de visualização da informação que, portanto, têm o potencial de representar esse tipo de dado geoespacial de maneira compreensível à especialistas e não-especialistas, além de possibilitar o uso de recursos visuais para identificar e corrigir possíveis inconsistências nos registros de localização nas coleções por seus mantenedores.

2.2.2 Cobertura temporal

Esta subseção traz alguns exemplos de artigos que fazem uso de recursos gráficos para ilustrar a cobertura temporal de certas coleções biológicas, seja no sentido de apresentar o volume de espécimes coletados a cada ano, ou no sentido de destacar um dado padrão temporal referente à coleta determinadas espécies³⁶. Conforme ilustrado na figura 11, aproximadamente 23% dos *papers* analisados (52 de 224) trazem esse tipo de visualização em seus documentos. Isso não significa que os demais artigos não abordam a cobertura temporal de suas coleções, mas que apenas não fazem uso de recursos gráficos para tal. Uma abordagem comum é reportar apenas o intervalo de tempo no qual a coleção foi composta de forma textual.

Conforme citado previamente, este estudo tem enfoque em produções literárias do tipo *data papers*, que usualmente buscam apresentar uma dada coleção representando, sob diferentes ângulos, características inerentes aos seus exemplares. Uma maneira comum de retratar o crescimento das bases de dados é ilustrar o número de espécimes coletados a cada ano, por exemplo. Usualmente, essa informação é demonstrada por meio de um gráfico de linhas ou barras em que o tempo é codificado no eixo X³⁷, enquanto o total de espécimes coletados em um certo período é representado no eixo Y³⁸. Para corroborar tal afirmação, serão exibidos a seguir alguns exemplos retirados diretamente dos artigos analisados para comunicar a abrangência temporal de suas coleções.

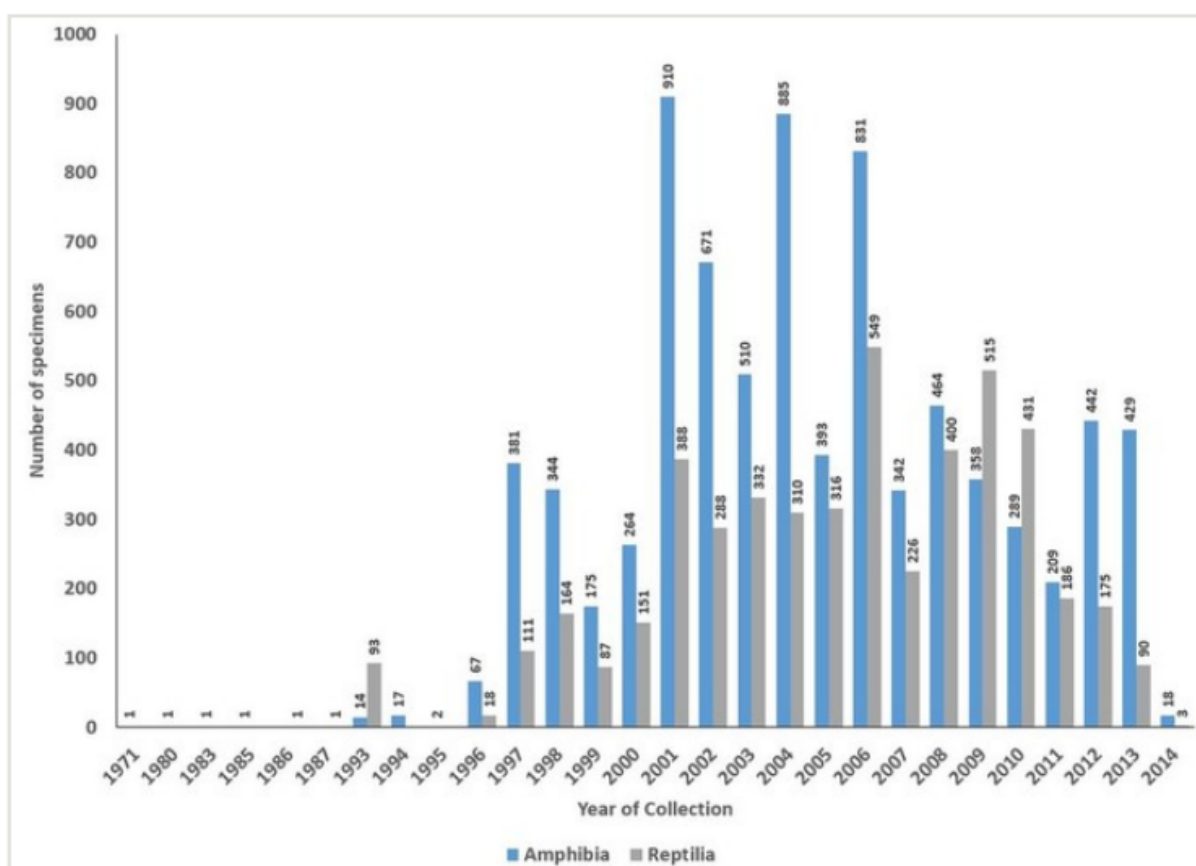
³⁶ Comumente, esse tipo de recurso é usado para mostrar o impacto de expedições para o volume de exemplares presentes na coleção.

³⁷ A depender do tempo de amostragem, a unidade temporal relevante pode variar. As mais comuns são por ano, mês, número de amostras ou períodos de coleta.

³⁸ As unidades mais comumente utilizadas são contagens por período ou acumuladas, em termos absolutos ou na base logarítmica.

Como primeiro exemplo, Ortiz-Yusty et al. (2015) apresenta, em seu artigo, a coleção do Museu Herpetológico da Universidade de Antioquia. A figura 20 exibe a representação gráfica utilizada para destacar a distribuição temporal dessas coleções de anfíbios e répteis. Vale destacar o esforço de Ortiz-Yusty et al. (2015) para diferenciar o volume de coleta de diferentes espécies codificando-as no canal de cor. A representação por barras é bastante comum e diversos outros artigos fazem uso de gráficos semelhantes³⁹, para citar alguns: Nualart et al. (2017), Pando et al. (2016), Camacho, Dorda e Rey (2014), García-Sánchez e Cabezudo (2013), dentre outros.

Figura 20 – Distribuição temporal dos registros de espécimes nas coleções do Museu Herpetológico da Universidade de Antioquia.

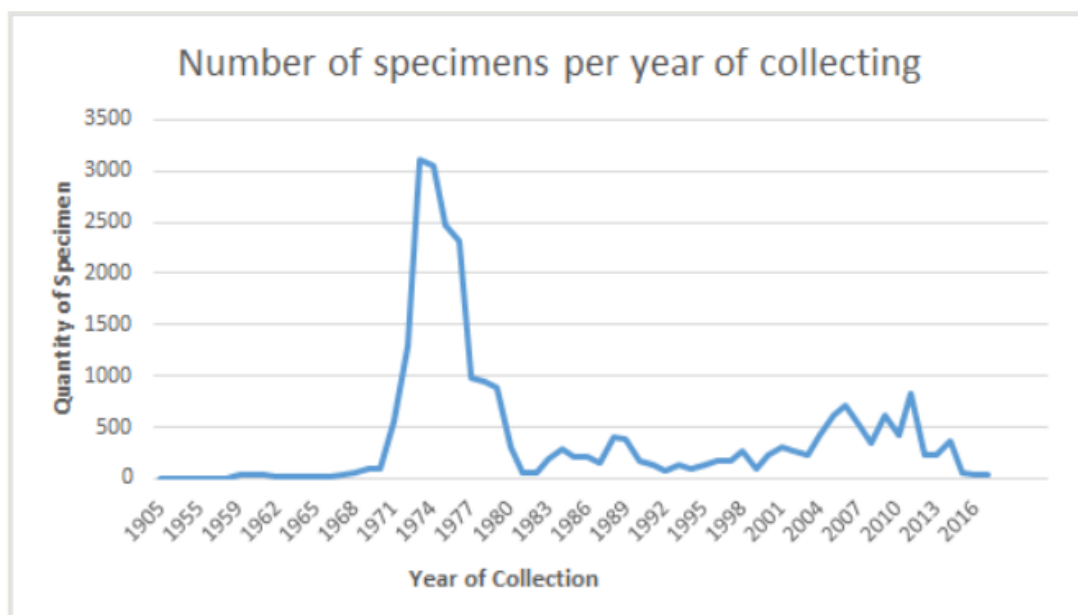


Fonte: (ORTIZ-YUSTY et al., 2015)

Prudente et al. (2019) e Uetz e Stylianou (2018) retratam o mesmo tipo de informação usando propostas gráficas distintas: gráfico em linha e *scatterplot*, respectivamente. Essas propostas são apresentadas nas figuras 21 e 22. Ainda, podemos citar Escribano et al. (2016) que, além de usar um gráfico de barras simples (não muito diferente do exibida na figura 20), combina um gráfico de linhas para mostrar o número acumulado de espécies na coleção, buscando despertar certa curiosidade no leitor quanto à abrangência taxonômica de sua base de dados em relação ao número de espécies distintas (vide figura 23).

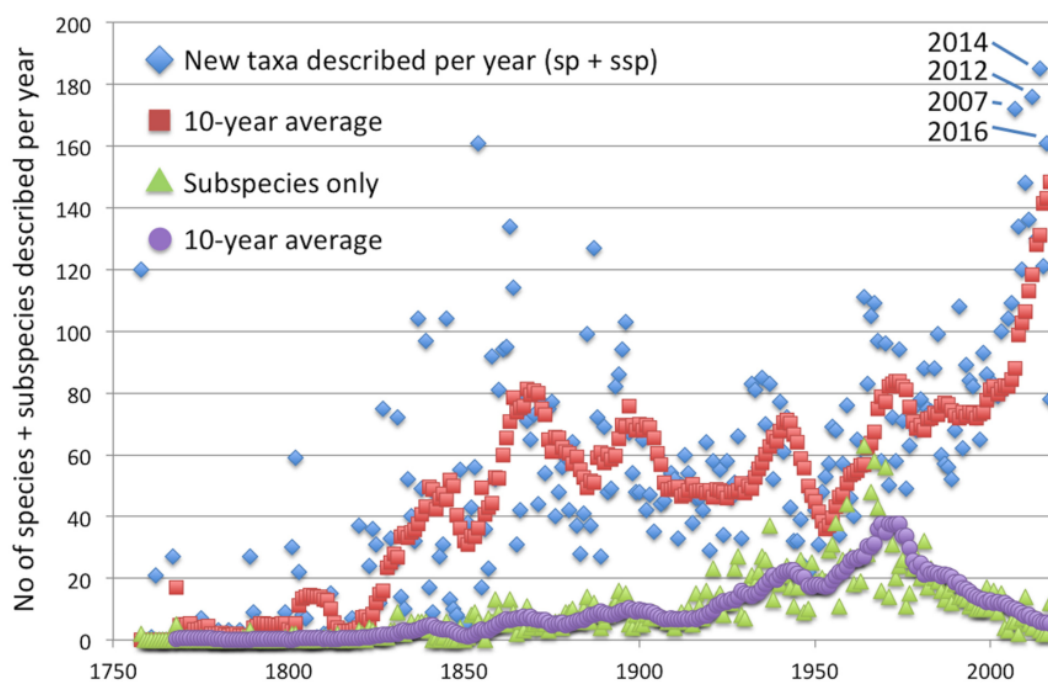
³⁹ Com ou sem codificação por cor para diferenciar taxas de crescimento por algum nível taxonômico, ambiente de coleta, ou outra informação relevante à identificação das espécies.

Figura 21 – Distribuição do número de espécimes coletados por ano na coleção de serpentes do Museu Paranaense Emílio Goeldi, Pará - Brasil.



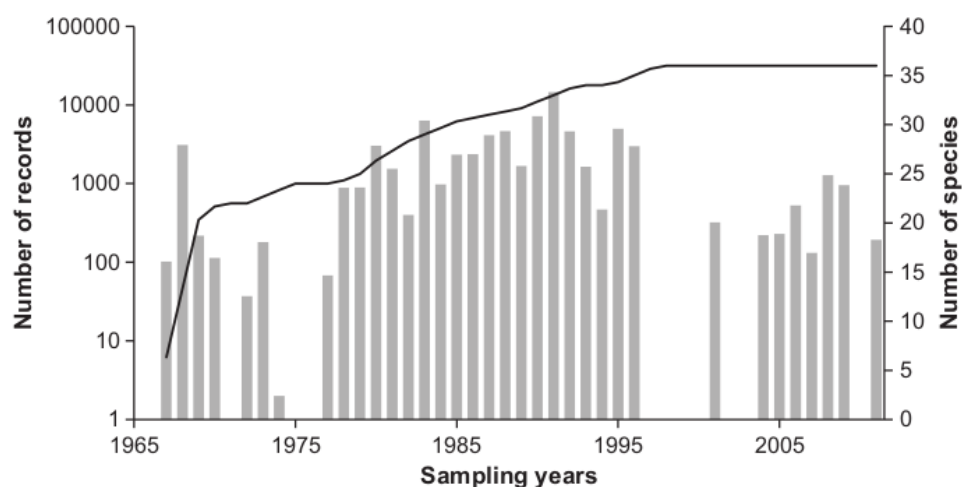
Fonte: (PRUDENTE et al., 2019)

Figura 22 – Descrição de espécies e subespécies de répteis ao longo de 260 anos.



Fonte: (UETZ; STYLIANOU, 2018)

Figura 23 – Inventário histórico de 73.316 registros de pequenos mamíferos no conjunto de dados de 'pellet sampling' que incluem o ano de coleta.

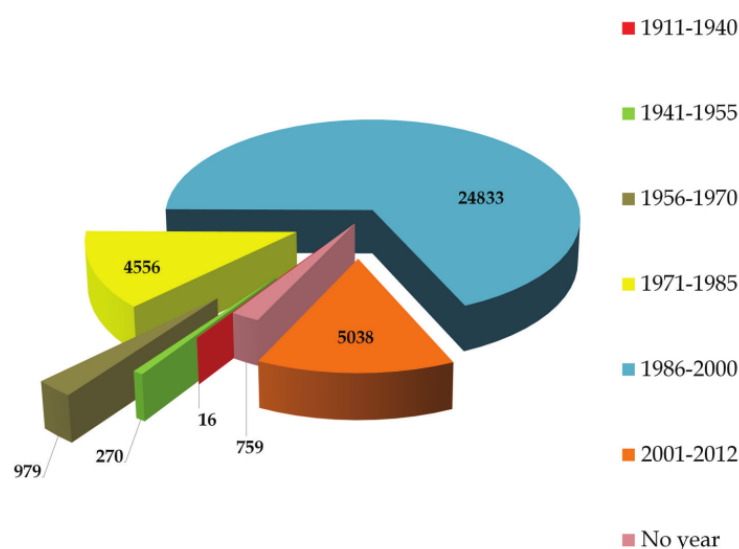


Fonte: (ESCRIBANO et al., 2016)

Nota: As barras cinzas representam o número de registros coletados a cada ano (em escala logarítmica). A linha preta é a curva de acumulação de espécies.

Dentro do grupo de visualizações empregado para destacar evolução temporal, Espinosa e López (2013) usa uma abordagem não muito usual. Essa representação é apresentada na figura 24.

Figura 24 – Cobertura temporal do Herbário UNEX.



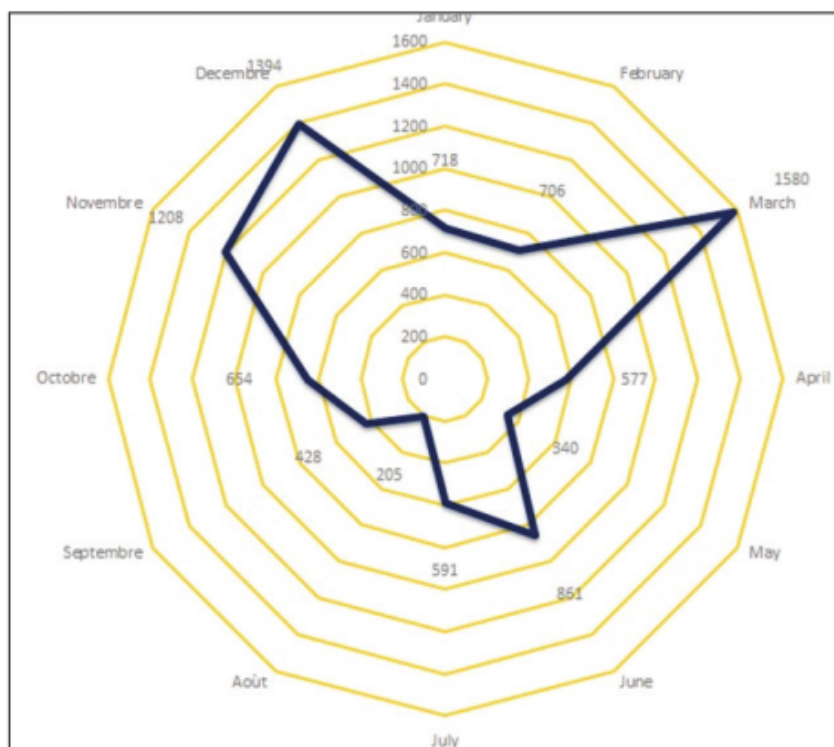
Fonte: (ESPINOSA; LÓPEZ, 2013)

Por vezes, se faz relevante analisar a presença de sazonalidade na coleta dos espécimes. Esse tipo de análise pode evidenciar, por exemplo, efeitos do clima⁴⁰ ou de

⁴⁰ Condições climáticas adversas podem dificultar a ida dos pesquisadores a campo, refletindo um certo

meses de expedição refletidos no total de espécimes coletados por período. Em geral, quando apresentada, a sazonalidade é investigada por meio de propostas gráficas comuns, bastante semelhante às ilustradas nas figuras 20 e 21. Como exemplo, podemos citar Nualart et al. (2017) e Merino-Sáinz, Anadón e Torralba-Burrial (2013). No entanto, a implementação feita por Radji et al. (2018), mostrada na figura 25, merece destaque.

Figura 25 – Padrão de acumulação de registros de plantas Togolesas coletadas por mês.



Fonte: (RADJI et al., 2018)

Todos os artigos citados nesta subseção se assemelham no fato de que fazem uso de recursos visuais para exibir um retrato temporal da coleção como um todo. Geralmente, buscam transmitir ao leitor um retrato imediato do volume desta coleção ao longo do tempo. Para isso, empregam o uso de métricas de agregação, como contagens, médias, somas, etc, sacrificando uma maior granularidade em prol da visão do todo. Agora, será apresentada uma visualização empregada por Pérez-Luque et al. (2016) que, diferente dos demais estudos, busca apresentar o universo temporal da seu conjunto de dados com um maior nível de granularidade, ao mesmo tempo em que procura transmitir uma visão global (vide figura 26).

Note que, nessa visualização, o número de registros é codificado no tamanho de cada ponto, enquanto o nível taxonômico está representado no eixo X, e o eixo Y apresenta o tempo em meses. Contrastar a figura 26 com as demais nessa seção é um bom exercício para destacar a importância de se selecionar um nível de granularidade adequado para responder padrão na coleta dos exemplares.

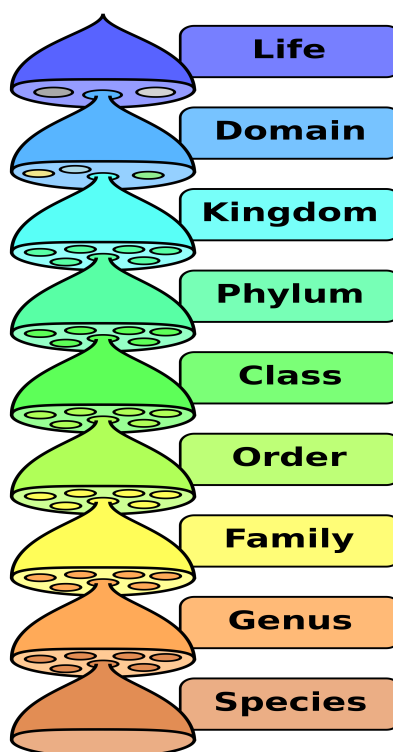
quantidades coletadas a cada período⁴¹.

Em linha com a discussão apresentada nessa subseção, o presente estudo propõe um novo conjunto de visualizações que têm o potencial de ilustrar a evolução temporal de quantidades agregadas de dados, assim como em maior nível de granularidade destacando também um dado nível taxonômico, como espécie, ordem ou família.

2.2.3 Cobertura taxonômica

Por definição, taxonomia é o estudo científico da nomeação, definição e classificação de grupos de organismos biológicos baseado em características em comum. Naturalmente, a classificação de seres segue uma estrutura hierárquica conforme exibida na figura 27. De maneira simplificada, espécimes que se situam no mesmo grupo na parte inferior da árvore tendem a ser muito mais parecidos do que espécimes situados em um mesmo grupo da parte superior. Em outras palavras, pode-se dizer que certos grupos taxonômicos são mais informativos que outros no sentido de apresentar um maior número de restrições quanto às características que seus elementos possuem.

Figura 27 – Sistema moderno de classificação biológica.

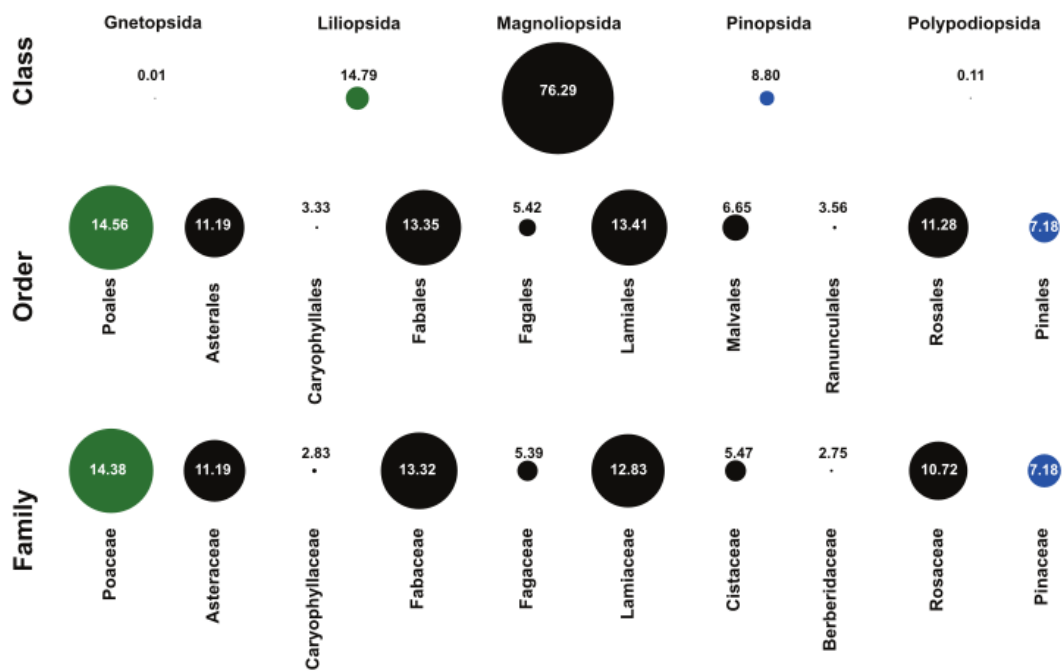


Fonte: <https://en.wikipedia.org/wiki/File:Biological_classification_L_Pengo_vflip.svg>, acessado em 19/01/2021.

⁴¹ Em outras palavras, é mais difícil inferir o número de animais coletados a partir do tamanho de um dado ponto. Porém, há uma maior riqueza de detalhes ao incorporar a taxonomia e uma unidade temporal relevantes à análise desses espécimes.

Esses grupos de visualização analisam cortes transversais na árvore taxonômica em um nível determinado pelo pesquisador especialista. No entanto, há abordagens que buscam conciliar a estrutura hierárquica com uma agregação quantitativa dos registros da coleção. Pérez-Luque et al. (2014) traz uma abordagem inusitada ao combinar gráficos de pizza com diferentes níveis hierárquicos (vide figura 30).

Figura 30 – Cobertura taxonômica por Classe, Família e Ordem da base de dados Sinfonevada.

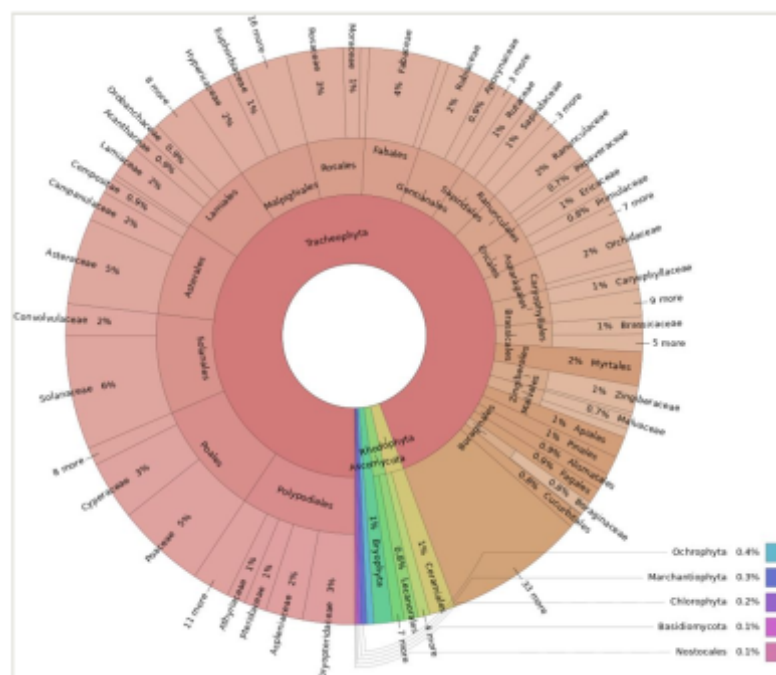


Fonte: (PÉREZ-LUQUE et al., 2014)

Nota: O tamanho dos círculos é proporcional ao número de registros da base de dados. Números indicam a porcentagem de registros. A classe está codificada no canal de cor.

Nessa linha, Dillen et al. (2019) apresenta uma proposta gráfica mais conhecida, do tipo *sunburst* (vide figura 31). Por fim, Zheleznova et al. (2020), Zheleznova et al. (2019) e Pando et al. (2016) utilizam *treemaps* para apresentar esse tipo de informação. A figura 32 ilustra o gráfico exposto em Zheleznova et al. (2019).

Figura 31 – Distribuição taxonômica por Filo, Ordem e Família (se conhecida).



Fonte: (DILLEN et al., 2019)

Nota: Uma versão interativa no formato HTML foi disponibilizada pelos autores do artigo original como material suplementar.

Figura 32 – Distribuição taxonômica de espécies entre famílias de musgo no conjunto de dados.



Fonte: (ZHELEZNOVA et al., 2019)

Nota: Visualização gerada usando o pacote "treemap" para a linguagem R.

A ideia de representar a hierarquia filogenética de uma coleção de dados PBR graficamente se faz muito pertinente ao possibilitar, instantaneamente, a visualização de um sumário da coleção, permitindo identificar, por exemplo, quais nós da cadeia hierárquica são mais proeminentes e quais são mais raros, fornecendo uma ideia clara ao usuário do que é encontrado em toda a coleção e, ao mesmo tempo, chamando a atenção para possíveis pontos de interesse, o que pode despertar a curiosidade para uma exploração mais detalhada. Não obstante, este trabalho apresenta um conjunto de visualizações que buscam detalhar pontos taxonômicos de interesse da coleção biológica, apontados por especialistas. Quando possível, também são empregados recursos dinâmicos para fornecer ao usuário um maior nível de exploração, facilitando a identificação de possíveis inconsistências ao nível taxonômico selecionado. Ainda, alterar as visualizações para navegar entre diferentes níveis taxonômicos pode ser feito sem grandes custos.

3 Coleções de Dados de Biodiversidade

MNRJ

Este trabalho foi desenvolvido em parceria com o Museu Nacional (MN), instituição autônoma integrante do Fórum de Ciência e Cultura da Universidade Federal do Rio de Janeiro (UFRJ), cuja sede é ilustrada na figura 33. Criado em 1818 por meio de um decreto de D. João VI, foi a primeira instituição museológica e de pesquisa do Brasil (Museu Nacional, 2020). Ao longo de seus mais de 200 anos de existência, o MN foi responsável por coletar e registrar dados em diversas coleções de ciências naturais e antropológicas.

Figura 33 – Fachada do Museu Nacional (MNRJ) e seu entorno.



Fonte: <<http://www.museunacional.ufrj.br/dir/omuseu/omuseu.html>> - Acessado em 16/02/2021

Nota: Fotografia por Roberto da Silva.

Desde os anos 80, a instituição vem realizando diversos esforços voltados à digitalização de seus registros, por meio de planilhas Excel[®] e Programas de Gerenciamento de Coleções (Museu Nacional, 2020). Tais registros incluem metadados associados às suas coleções, como *Primary Biodiversity Records* (PBR), contendo o nome do espécime/objeto, local de coleta, identificador/qualificador, datas relevantes e, em alguns casos, também podem incorporar elementos fotográficos ou sonoros. Seu acervo é considerado um dos maiores e mais importantes da América Latina, com 36 coleções de diferentes tipologias, totalizando cerca de 20 milhões de itens (Museu Nacional, 2020). Parte desse acervo pode ser acessado em bancos de dados *online* como o Sistema de Informação sobre a Biodiversidade Brasileira¹ (SiBBr); o *Global Biodiversity Facility*² (GBIF) e o SpeciesLink³.

O Museu Nacional possui 29 coleções científicas vinculadas a seis departamentos: Antropologia, Botânica, Entomologia, Geologia e Paleontologia, Invertebrados e Vertebrados. O desenvolvimento deste estudo contou com a colaboração ativa dos departamentos de Invertebrados e Vertebrados, mais especificamente dos setores de Herpetologia⁴, Carcinologia⁵ e Polychaeta⁶. Ao longo dessa parceria, foi cedido acesso à algumas de suas coleções em formato eletrônico: répteis, crustáceos e poliquetas, respectivamente. A seguir, serão feitos comentários acerca de cada uma dessas coleções.

A coleção de répteis possui mais de 30 mil exemplares, sendo uma das maiores do mundo quanto à fauna neotropical e uma das coleções mais consultadas da América Latina (Museu Nacional, 2020). Répteis são um grupo de animais vertebrados tetrápodes (possuem quatro membros⁷) e ectotérmicos (não possuem mecanismo interno para controle da temperatura corporal). Entre seus espécimes mais conhecidos, estão serpentes, lagartos, tartarugas e crocodilos. A figura 34 ilustra alguns dos espécimes encontrados nessa coleção.

¹ <<https://ipt.sibbr.gov.br/mnrj/>>

² <<https://www.gbif.org>>

³ <<http://splink.cria.org.br/>>

⁴ Vinculado ao departamento de Vertebrados. É direcionado ao estudo da diversidade de anfíbios e répteis, como sapos, cobras e lagartos.

⁵ Vinculado ao departamento de Invertebrados. Destina-se ao estudo de crustáceos como caranguejos, camarões e lagostas.

⁶ Vinculado ao departamento de Invertebrados. É voltado ao estudo da biodiversidade de poliquetas, como minhocas marinhas e sanguessugas.

⁷ Serpentes, apesar de não possuírem membros, são consideradas tetrapoda pois seus membros sofreram um processo de regressão e foram perdidos ao longo do tempo.

Figura 34 – Exemplos do acervo de répteis do setor de Herpetologia do Museu Nacional.

(a) *Micrurus carvalhoi*(b) *Atractus ronnie*(c) *Phyllopezus lutzae*(d) *Liolemus lutzae*

Fonte: <<http://www.herpetologiamuseunacional.com.br/galeria.html>> - Acessado em 16/02/2021.

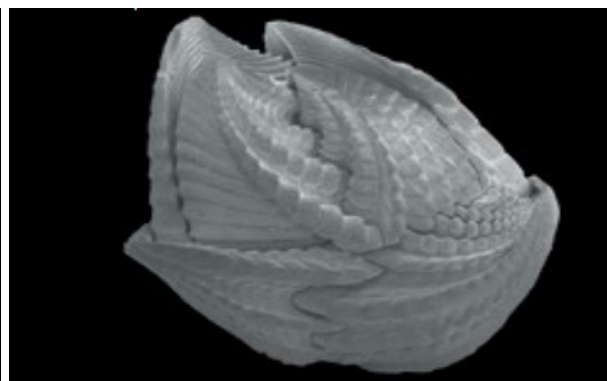
Nota: Fotografia por: (a) Thiago Silva-Soares (b) Daniel Loebmann; (c) Edelman de Melo Gonçalves; (d) Davor Vrcibradic.

O acervo de crustáceos compreende animais invertebrados pertencentes ao filo dos artrópodes (i.e., possuem carapaça rígida), podendo ser encontrados no mar e em água doce. Seus espécimes mais comuns são caranguejos, siris, camarões e lagostas. A figura 35 expõe alguns exemplares que compõem essa coleção.

Figura 35 – Exemplos da coleção de crustáceos do setor de Carcinologia do Museu Nacional.



(a) *Rochinia cf. umbonata*



(b) *Costatoverruca flavidula*



(c) *Epimeria rotunda*



(d) *Lepechinella hirsuta*

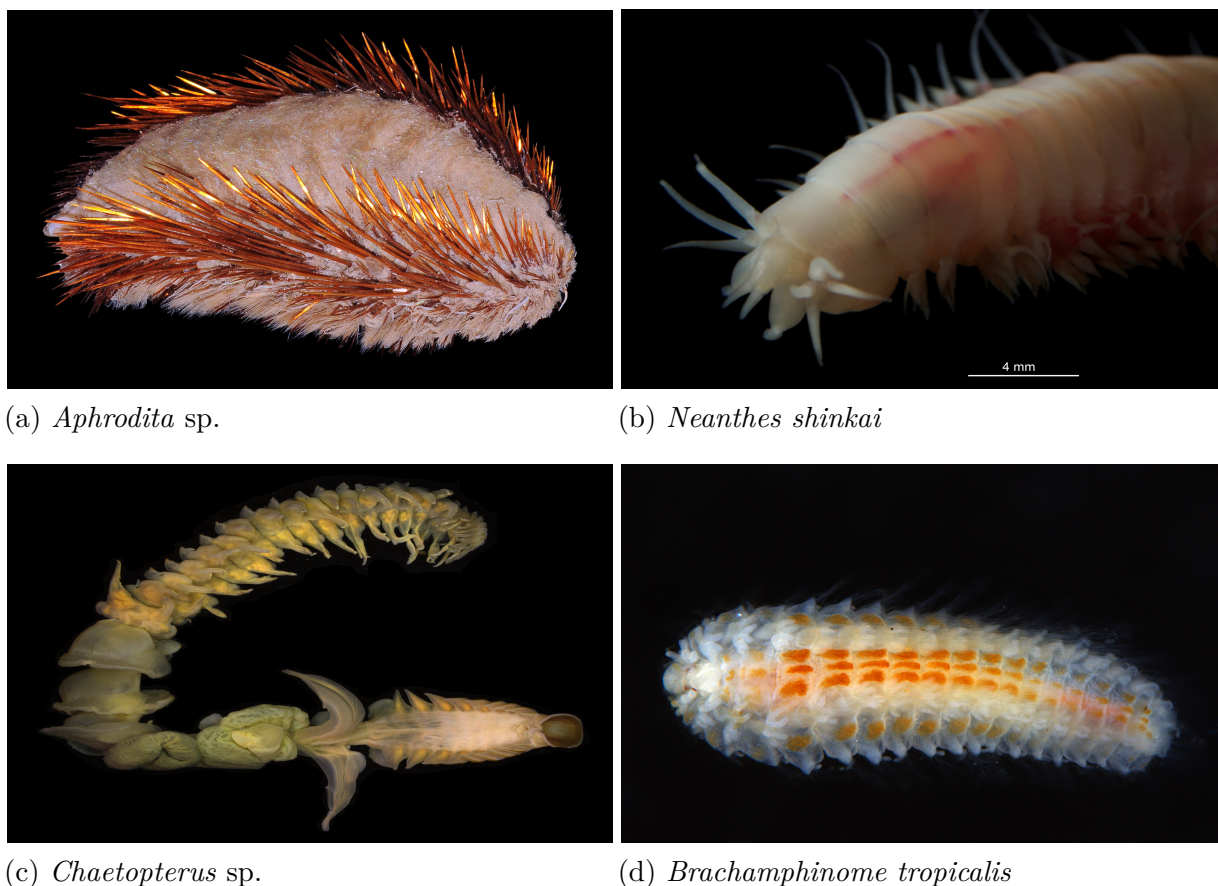
Fonte: (Museu Nacional, 2020)

Nota: (a) Caranguejo; (b) Craca de profundidade; (c) Anfípode; (d) Anfípode; todos localizados na Bacia de Campos, RJ.

Polychaeta, do grego “muitas cerdas”, são vermes segmentados, exclusivamente marinhos, pertencentes ao filo Annelida⁸. A figura 36 ilustra alguns dos animais presentes na coleção de poliquetas do Museu Nacional.

⁸ Esse filo engloba animais de corpo segmentado, que podem viver em ambientes diversos como solo úmido, água doce e salgada. A classe poliqueta refere-se, em específico, à vermes aquáticos encontrados exclusivamente em ambientes marinhos.

Figura 36 – Espécimes do acervo de poliquetas do setor de Polychaeta do Museu Nacional.



Fonte: (Museu Nacional, 2020); Imagens cedidas pelo departamento de Polychaeta.

Os metadados associados a cada um dos exemplares que constituem o acervo do Museu Nacional, em particular para as coleções aqui trabalhadas, estão registrados em planilhas eletrônicas. Isso possibilita novos meios de se explorar as características dessas coleções com o emprego de técnicas de análise e visualização de dados. Este trabalho direciona-se ao problema de melhoria qualitativa desses registros propondo um *framework* de soluções visuais que busquem evidenciar possíveis inconsistências nas bases de dados. É importante citar que todos os *scripts* relevantes para a construção de tais ferramentas visuais, assim como os arquivos gráficos gerados como *outputs*, estão disponíveis em um repositório público publicado na plataforma GitHub, acessível pelo link: <<https://github.com/Franklin-oliveira/Dissertacao-EMAp>>. O capítulo 4 apresenta as soluções aqui propostas para endereçar essa questão. No momento, serão apresentadas algumas estatísticas descritivas na tabela 1 dos conjuntos de dados a serem trabalhados.

Tabela 1 – Estatísticas descritivas das coleções com acesso cedido pelo Museu Nacional^a

–	Departamento:	Vertebrados	Invertebrados	Invertebrados
	Sector:	Herpetologia	Carcinologia	Polychaeta
	Nome da coleção:	Répteis	Crustáceos ^b	Poliquetas
–	Total de registros ^c	23132	8323	6923
Cobertura Taxonômica ^d	Ordens distintas	3	1	10
	Famílias distintas	52	108	59
	Gêneros distintos	278	350	298
	Espécies distintas	751	592	391
Cobertura Temporal	Ano de coleta	1824 - 2016	1871 - 2018	1900 - 2019
	Ano de determinação ^e	1915 - 2019	1905 - 2020	1986 - 2020
Cobertura Geográfica	Continentes	4	5	6
	Países	30	23	29
	Estados brasileiros ^f	27	27	14

Fonte: Elaboração própria.

Nota: Apesar de os dados terem sido sujeitos à uma etapa de pré-tratamento (vide seção 3.1), há a possibilidade de ainda existirem erros de digitação nos respectivos campos (a serem identificados e corrigidos pela equipe em um momento posterior à elaboração deste trabalho). ^aInstituição autônoma integrante do Fórum de Ciência e Cultura da Universidade Federal do Rio de Janeiro (UFRJ); ^bDa coleção de crustáceos, foram considerados apenas espécimes da ordem Decapoda (registros revisados e trabalhados pela equipe de Carcinologia no decorrer deste estudo); ^cCalculados com base em números de catálogos distintos; ^dA quantidade de registros distintos em cada campo foi calculada com base em suas respectivas colunas identificadas nas bases de dados, separadamente; ^eAno em que os espécimes foram identificados ou tiveram sua classificação revisada por um curador especializado; ^fConsiderando todas as unidades da federação (i.e., incluindo os 26 Estados mais o Distrito Federal).

3.1 Pré-tratamento dos dados

Muitos dos registros de coleções científicas de biodiversidade advém de períodos anteriores ao advento de microcomputadores e *internet*. Informações antigas costumam estar registradas em livros físicos, escritas em letra cursiva pelo pesquisador responsável, muitas vezes diretamente do campo de coleta. No processo de digitalização, esses registros são transcritos, um a um, para o formato de planilha eletrônica Excel[®]. Nesse processo, algumas inconsistências podem surgir, como:

1. Perda de informações devido à ação de intempéries sobre o livro físico;
2. O curador responsável pela digitalização das informações pode não compreender a escrita cursiva do registro original;
3. Em meio à digitação de tantos registros, podem acontecer erros comuns de anotação. Nesse cenário, dois erros comuns podem ser apontados:
 - a) Confusão de caracteres com grafia semelhante. Por exemplo, é fácil confundir os caracteres I (i maiúsculo) com l (letra "ele"minúscula) ou o número 1; letra O com o número 0; letra S com o número 5, dentre outros;

- b) Breves momentos de desatenção podem acarretar na troca de um caractere por outro cuja tecla situa-se próxima à que se deseja pressionar⁹. Por exemplo, ao tentar pressionar a tecla "A", pode-se acidentalmente pressionar a tecla "S"(imediatamente ao lado). Ainda, em especial para a língua portuguesa, que faz pleno uso de acentuação, 'João' pode acidentalmente ser escrito como "Joçao"ou "Joáo"¹⁰.

Com o intuito de corrigir erros ocasionais de digitação (3.a e 3.b), as três bases de dados foram sujeitas à uma etapa de pré-processamento. Além disso, essa etapa tem como objetivo padronizar informações para que estas sejam consistentemente expostas visualmente. Isto posto, foi realizado o seguinte processo:

- Colunas textuais, i.e., que contém nomes de pessoas, de lugares e classificação taxonômica, foram controladas para a presença de caracteres indevidos como espaços¹¹, símbolos e acentos incorretos.
- Campos de data, depois que seu formato foi identificado (i.e., associado a um dos padrões conhecidos de ano-mês-dia, dia/mês/ano, etc.), foram varridos para se extrair duas informações relevantes para as visualizações: Ano e mês.
- Campos numéricos, como altitude e profundidade, foram tratados para certificar-se de que todos os seus registros estavam representados na mesma unidade de medida (neste caso, em metros). Ainda, foram retirados quaisquer caracteres exceto números (i.e., espaços, letras e símbolos). Por fim, tomou-se o cuidado de representar o separador decimal de maneira correta¹²
- Coordenadas pontuais de latitude e longitude obedeceram o seguinte processo: 1. Todos os valores foram representados com o mesmo separador de casas decimais (neste caso, vírgula ',' foi substituída por ponto '.' quando necessário); 2. Foram eliminados quaisquer caracteres de caráter não-numérico (como símbolos, letras e espaços); 3. Coordenadas representadas em formatos diferentes (i.e., DMS e UTM), foram convertidas para a convenção decimal; 4. Uma vez que todos os pares de coordenadas estavam representados no mesmo formato, foi verificada a presença de pontos que extrapolam os valores extremos (i.e., valores fora do intervalo $[-90^{\circ}, +90^{\circ}]$ para latitude e $[-180^{\circ}, +180^{\circ}]$ para longitude).

⁹ Embora se possa argumentar que a ocorrência desse tipo de erro seja comum a qualquer processo no qual o *input* de dados é feito à mão, a troca de caracteres, em particular, depende do *layout* de teclado utilizado. Os exemplos que citei aqui são defensáveis para o *layout* ABNT-2 da língua portuguesa.

¹⁰ No *layout* ABNT-2, a combinação de teclas para produzir o caractere "ã" é semelhante à "Ç" e "''" (acento agudo).

¹¹ Por exemplo, a variável *string* 'Maria ', para o computador, é diferente de 'Maria'.

¹² Para o intérprete Python, isso significa empregar o ponto '.' como separador de casas decimais.

É importante destacar que todo esse processo de manuseio de dados foi realizado empregando-se a linguagem de programação Python 3¹³, e todo o código desenvolvido está disponível em um repositório público na plataforma GitHub¹⁴. Em linha com as observações dadas por Oliphant (2007), essa escolha está fundamentada nos seguintes fatores: 1. Python tem se revelado cada vez mais como principal escolha adotada pela comunidade científica por ser uma linguagem de alta flexibilidade, com sintaxe limpa e facilmente adaptável para lidar com bases de dados de diferentes tipos e tamanhos; 2. Trata-se de uma ferramenta *open source*, isto é, de uso e acesso livres de qualquer direito autoral, além de apresentar amplo suporte da comunidade; 3. É defensável como uma linguagem de programação de baixo custo de entrada para o público leigo, i.e., há uma enorme gama de cursos, tutoriais e fóruns de perguntas e respostas disponíveis na internet, além diversas bibliotecas otimizadas para as mais diversas aplicações. Essas são características desejáveis visando a manutenção e continuidade desse recurso a longo prazo pelas equipes de curadoria do Museu Nacional.

Finalmente, uma vez controlada a presença de pequenos erros de digitação e diferenças de formatação, as coleções estavam, então, prontas para a etapa de codificação e exposição visual. Vale ressaltar que, finalizada esta etapa, as bases de dados não estão completamente livre de erros. O principal objetivo da ferramenta aqui proposta é auxiliar nos processos de detecção, triagem e diagnóstico de dados apontando a presença de erros qualitativos nas coleções, cuja dificuldade de identificação extrapola à de simples erros de digitação. Os capítulos 4 e 5 buscam apresentar como o *framework* aqui proposto foi desenhado e apontar como suas visualizações podem auxiliar na detecção de inconsistências nas bases de dados a partir de elementos visuais.

¹³ Mais especificamente, a versão 3.7.4. No entanto, todo o código aqui desenvolvido deve funcionar em todas as versões de Python 3 ou superiores, desde que haja compatibilidade com as bibliotecas empregadas.

¹⁴ Link do repositório: <<https://github.com/Franklin-oliveira/Dissertacao-EMAp>>.

4 Construção do *framework* em apoio à curadoria de coleções científicas biológicas digitais

Este capítulo visa expor processos de criação e o emprego de técnicas e práticas recomendadas na literatura pertinentes à concepção de um *framework* de visualização capaz de auxiliar o público especialista a identificar e solucionar inconsistências em suas coleções científicas de biodiversidade. Vale destacar que todas as visualizações desenvolvidas nesse estudo foram incorporadas ao dia a dia de curadores e mantenedores de bases de dados biológicas visando servir ao propósito de facilitar a identificação e correção de possíveis inconsistências. Especificamente, eram conduzidas reuniões periódicas em que discutia-se a maneira pela qual as ferramentas gráficas lhes eram úteis, além de coletar sugestões diretas das equipes voluntárias do Museu Nacional (MNRJ) para execução de melhorias. Então, o conjunto de gráficos eram utilizados pelas equipes para auxiliar no processo de melhoria qualitativa de suas coleções.

A discussão pertinente à construção das ferramentas visuais será segmentada da seguinte forma: a seção 4.1 desenvolve uma maneira de classificar ou propor novas visualizações a partir de um raciocínio criativo; a seção 4.2 trata da atribuição dos canais de codificação visual para implementação do conjunto de visualizações. Em particular, a subseção 4.2.1 aborda as nuances envolvidas na definição dos mapas de cores; a seção 4.3 apresenta brevemente os elementos de interatividade aderidos aos conjuntos de gráficos; a seção 4.4 finaliza esse capítulo apresentando as propostas gráficas criadas a partir do raciocínio desenvolvido nas seções anteriores.

4.1 *Design thinking*

Esta seção inicia-se apresentando um conjunto de questões, amplamente utilizado em áreas como jornalismo, investigações policiais e pesquisas científicas, que têm o potencial de descrever completamente um evento, situação ou estória. Tais perguntas foram originadas a partir do estudo da ética e da retórica, apresentadas pioneiramente por Aristóteles como "as sete circunstâncias"¹. São elas: O quê?, Onde?, Quando?, Quem?, Por que?, Como? e Quanto?. Além de direcionar a confecção de soluções de visualização, esses questionamentos podem também configurar um mecanismo para classificação de soluções visuais existentes na literatura voltadas, em particular, a conjuntos de dados biológicos.

¹ Em inglês, são conhecidas como "*Five W's and one H*".

Coleções digitais de biodiversidade são constituídas por registros do tipo PBR (*Primary Biodiversity Records*), que são metadados associados à coleta de espécimes, contendo informações a respeito do local e data de coleta, classificação taxonômica, podendo também incorporar outras informações como o nome do taxonomista que contribuiu para a sua documentação. Consumidores dessas informações são, geralmente, especialistas de domínio e a exploração desses dados está presente em seu cotidiano, seja para a confecção de estudos científicos ou para a manutenção e curadoria desses registros. A importância dessas coleções é amplamente reconhecida na literatura (SUAREZ; TSUTSUI, 2004) e muitas são suas aplicações nos mais diversos contextos, em especial, para avanços científicos das ciências biológicas. A seção 2.2 do capítulo 2 aborda mais aspectos sobre a relevância e diferentes aplicações desses dados no contexto de biologia.

Nesse sentido, quando se está analisando dados dessa natureza, as perguntas essenciais têm o potencial de guiar o processo de exploração de dados. Um simples exercício pode mostrar como tais questionamentos podem ser facilmente adaptados a esse contexto:

- **O quê?** Normalmente refere-se a um conjunto de elementos associados à identificação do espécime em questão. Por exemplo, seu nome científico e sua classificação taxonômica (Reino, Filo, Ordem, Gênero, Espécie, etc)². Para especialistas, essas informações revelam um conjunto de características associadas aos exemplares que têm o potencial de, a priori, destacar pontos associados ao seu *habitat*, período de descoberta³ e também, possivelmente, o pesquisador especializado no estudo desse grupo taxonômico.
- **Onde?** Indaga, literalmente, a localização geográfica em que um dado exemplar foi coletado. Normalmente, essa informação é gravada na forma de coordenadas de GPS (latitude e longitude) diretamente do campo de coleta com o auxílio de aparelhos de georreferenciamento, quando disponível. A partir dessas coordenadas, consegue-se recuperar outras informações como País, Estado e cidade mais próxima. Contudo, devido à grande extensão temporal destas coleções, técnicas de georreferenciamento não estavam disponíveis no momento em que alguns espécimes foram coletados. Nesses casos, aplica-se o processo inverso, isto é, aproxima-se (com um certo nível de precisão) as coordenadas de GPS a partir de informações como província e município de coleta. A acurácia desses registros, muitas vezes, é verificada em conjunto com a classificação taxonômica, determinando se um dado espécime é, possivelmente, nativo de uma certa região a partir de suas características adaptativas.

² A figura 27 ilustra o sistema de classificação biológica moderna, apresentada na segunda parte deste capítulo.

³ Especialistas possuem o conhecimento necessário para determinar a partir de que ponto no tempo um exemplar foi reconhecido e classificado, por exemplo, reconhecendo o artigo científico que apresentou a descoberta desse espécime à comunidade.

- **Quando?** Diz respeito às datas de coleta e identificação dos espécimes. Especialistas de domínio possuem conhecimento para corresponder o período de publicação de certos artigos científicos à exemplares presentes em suas coleções.
- **Quem?** Essa pergunta direciona-se às pessoas envolvidas nos processos de coleta, classificação e revisão dos registros digitais da coleção (coletores, taxonomistas, pesquisadores e curadores, respectivamente). Profissionais do campo das ciências biológicas tendem a especializar-se no estudo e identificação de certos táxons, evidenciando uma alta correlação entre as pessoas envolvidas no ciclo de pesquisa e elementos como a classificação e a localidade do objeto de estudo. Além disso, a duração da carreira profissional destes especialistas pode coincidir com o período de tempo dedicado ao estudo de certas espécies, refletindo um padrão a ser explorado nas bases de dados.
- **Por que?**, nesse contexto, não abre grande possibilidade de exploração pois sua resposta está associada à motivação original do pesquisador/coletor em captar o espécime na natureza. Por esse motivo, não será amplamente empregada no desenvolvimento do raciocínio criativo apresentado no decorrer desta seção.
- **Como?** essa questão faz referência ao método de amostragem utilizado para coleta dos espécimes e varia amplamente com o tipo de ser recolhido na natureza. Técnicas de amostragem de plantas diferem completamente de métodos de captura de insetos que, por sua vez, são completamente distintos da coleta de espécimes marinhos. No entanto, essa pergunta pode ser parcialmente respondida, de forma mais genérica, por outras perguntas essenciais, como O quê?, Onde? e Quando?. Por essa razão, não será diretamente abordada no raciocínio indutivo desenvolvido no restante desta seção.
- **Quanto?** Essa pergunta pode ser compreendida como uma indagação a respeito do volume de espécimes coletados com certa característica de interesse. Sua resposta está associada ao cálculo de certas agregações quantitativas, como contagens, soma, média, dentre outras, em relação a um conjunto fixo de características.

Visando melhor servir ao público interessado⁴ nessas informações, tais questionamentos podem guiar o *design* e implementação de ferramentas de visualização, elevando seu grau de eficácia. Soluções visuais, por *default*, têm a capacidade de apresentar mapas mentais de informações (LIU et al., 2014) e codificar múltiplos campos das bases de dados em diferentes dimensões, nos eixos X e Y, e em elementos estéticos, como cor, tamanho e forma. Dessa maneira, um único gráfico pode permitir usuários a explorar diferentes ângulos de visão das bases de dados a partir de combinações das perguntas essenciais.

⁴ Especialistas de domínio e usuários casuais.

O quadro 1 traz uma matriz criativa que ilustra como esses questionamentos podem ser combinados de forma a destacar os elementos explorados nas ferramentas existentes ou, idealmente, guiar a confecção de um novo instrumento visual que objetive realçar tais aspectos, à luz do processo construtivo apresentado em Keck, Groh e Vosough (2020).

Quadro 1 – Matriz criativa.

–	O quê?	Onde?	Quando?	Quem?	Tipo ^a	Canais de Codificação ^b
O quê?	–				Nominal	X, Y, <i>color.nominal</i> , <i>shape</i>
Onde?		–			Nominal	X, Y, <i>color.nominal</i> , <i>shape</i>
Quando?			–		Ordinal	X, Y, <i>color.quantitative</i> , <i>size</i>
Quem?				–	Nominal	X, Y, <i>color.nominal</i> , <i>shape</i>
Quanto? ^c					Quantitativo	X, Y, <i>size</i> , <i>color.quantitative</i>

Fonte: Elaboração própria.

Nota: Matriz criada a partir das perguntas essenciais: O quê?, Onde?, Quando?, Quem? e Quanto?.

^aClassifica o tipo de informação a ser codificada em uma das três categorias: Nominal, Ordinal e Quantitativa. ^bApresenta os canais de codificação visual (segundo a definição do Vega-Lite) classificados (da esquerda para a direita) por sua efetividade em codificar o respectivo tipo de dado, com base em Mackinlay (1986) e Cleveland e McGill (1984). ^cRepresenta qualquer tipo de agregação quantitativa feita sobre as informações essenciais (ex.: contagens, soma, média).

A matriz criativa (quadro 1) ilustra um mapa mental de como as questões essenciais podem ser combinadas para criar diferentes representações gráficas. Adicionalmente, são exibidos os tipos de dados comumente associados à esses questionamentos, assim como os canais de codificação adequados para representá-los visualmente. Canal de codificação pode ser compreendido como um recurso visual que pode ser alocado a diferentes campos das bases de dados (SATYANARAYAN et al., 2016). Em linha com Qu e Hullman (2016), são consideradas as maneiras pelas quais campos nominais, ordinais e quantitativos podem ser codificados nos seguintes canais de codificação visuais mais comuns⁵: X, Y, cor (*color*), tamanho (*size*) e forma (*shape*). Adicionalmente, é proposto um ordenamento (da esquerda para a direita) destes canais segundo sua eficácia em representar a informação desejada (MACKINLAY, 1986; CLEVELAND; MCGILL, 1984). Note que os eixos X e Y são as formas mais eficazes e podem ser empregadas para todo tipo de informação. O tamanho dos símbolos, normalmente, codifica apenas dados quantitativos, enquanto sua forma representa apenas dados nominais (QU; HULLMAN, 2016). Finalmente, o canal de cor é dividido conforme a figura 37.

Isto posto, um processo criativo pode ser desenhado a partir da matriz exposta no quadro 1 ao combinar as diferentes questões e codificar os campos associados na base de dados de forma a criar representações gráficas distintas. O quadro 2 expõe algumas das combinações exploradas ao longo deste trabalho.

Note que cada uma das perguntas essenciais pode estar associada à um ou mais

⁵ Para facilitar uma associação direta com o trabalho de Qu e Hullman (2016), é adotada a nomenclatura em inglês sempre que se fizer pertinente.

Figura 37 – Subdivisões do canal de cor.



Fonte: (QU; HULLMAN, 2016)

Nota: É adotada a definição de canal de codificação visual do Vega-Lite, apresentada em Satyanarayan et al. (2016); *color.quantitative* é ordenado e pode codificar dados ordinais e quantitativos, enquanto *color.nominal* é não-ordenado e codifica, apenas, dados nominais.

Quadro 2 – Representações gráficas criadas a partir da matriz criativa.

X	Y	<i>color.nominal</i>	<i>color.quantitative</i>	<i>shape</i>	<i>size</i>
Quando?	Quem?	O quê?	–	–	Quanto? ^a
Quando?	Onde?	O quê?	–	–	Quanto?
Quando?	Quanto?	Onde?	–	–	–
Quando?	Quanto?	O quê?	–	–	–
Quando? (Ano)	Quando? (Mês)	–	Agregação (Soma)	–	–
Quando?	Quem?	O quê? (Ordem)	–	O quê? (Tipo ^b)	Quanto? (Contagem)
Onde? (Longitude)	Onde? (Latitude)	O quê? (Família)	–	O quê? (Tipo)	–
O quê? (Gênero)	Onde? (Altitude ^c)	O quê? (Família)	–	–	–

Fonte: Elaboração própria.

Nota: O quadro ilustra como as perguntas essenciais podem guiar a codificação de diferentes informações e dar origem a representações gráficas distintas, à luz do processo construtivo presente em Keck, Groh e Vosough (2020). De fato, essas combinações, dentre outras, foram exploradas ao longo do desenvolvimento deste trabalho. OBS¹: uma pergunta pode estar associada à um ou mais campos da base de dados (ex: a pergunta onde? pode ser respondida em coordenadas de latitude e longitude). OBS²: Os exemplos ilustrados nesse quadro não exaurem todo o universo de combinações possíveis desses elementos (novas combinações, necessariamente, acarretam em novas soluções visuais). ^aRepresenta qualquer tipo de agregação quantitativa dos dados (ex: contagens, soma, média). ^bTrata-se de um nível de classificação taxonômica. ^cPara algumas espécies, se faz relevante analisar a altitude ou profundidade em que esta foi encontrada.

campos das coleções de dados. Por exemplo, a pergunta "Onde?" pode ser respondida por meio de coordenadas de latitude e longitude, assim como "O quê?" está diretamente associada aos campos que contêm os diferentes níveis de classificação taxonômica (leia-se Reino, Família, Gênero, Ordem, Espécie, etc.). Vale ressaltar que diferentes campos associados à mesma questão podem estar presentes no mesmo gráfico, elevando assim o nível de detalhamento ao qual determinada informação é exposta. Ainda, é importante atentar para o fato de que o quadro 2 não exaure todo o conjunto de formas possíveis pelas quais se pode combinar tais elementos aos diferentes canais de codificação. A exploração

desse raciocínio criativo é fortemente incentivada para trabalhos futuros.

A partir desse processo, originam-se novas ideias de exploração visual para as coleções digitais de dados PBR. No entanto, para sua implementação, é necessário o uso de uma ferramenta de linguagem declarativa de gráficos. Nesse sentido, Wang et al. (2015) busca guiar o público interessado (biólogos e desenvolvedores) à visualização de dados biológicos, apresentando um conjunto de bibliotecas e instrumentos *open source*. O presente estudo optou pela aplicação da ferramenta gráfica Altair por possibilitar o desenvolvimento de um ambiente integrado, tirando proveito da flexibilidade da linguagem de programação Python para as tarefas de tratamento, análise e exposição de dados.

Por fim, é importante ressaltar que o *design* criativo apresentado até este ponto foi periodicamente discutido com especialistas de domínio, de forma a selecionar e aprimorar as propostas mais adequadas às suas necessidades. As demais subseções deste capítulo apresentam detalhes técnicos referentes à implementação do *framework*, além de apresentar as visualizações obtidas ao fim desse processo.

4.2 Representação visual

Na seção 4.1, foi desenvolvido um raciocínio criativo a partir do qual pode-se criar diferentes propostas de representação visual, em particular, para coleções científicas de biodiversidade. No que se refere à implementação dessas propostas em sua melhor forma⁶, é necessário o uso de uma ferramenta de gramática interativa de gráficos. Wang et al. (2015) busca guiar desenvolvedores à escolha de ferramentas *open source* para criação de visualizações na área de bioinformática.

Dentre opções como *ggplot2*⁷, D3.js⁸, Tableau®, Vega e Vega-Lite, optou-se pelo uso da ferramenta Altair, desenvolvida por VanderPlas et al. (2018), que pode ser compreendida como um *wrapper* de Vega-Lite criado para funcionar junto à linguagem de programação Python. Dessa forma, é possível aliar os benefícios de uma linguagem de programação flexível à estruturas de gramática interativa de gráficos. Além disso, é importante salientar que essa escolha também visa facilitar a manutenção e continuidade a longo prazo dos recursos aqui criados pelas equipes de curadoria do Museu Nacional (MNRJ).

Isto posto, é adotada a definição de canal de codificação visual do Vega-Lite, apresentada em Satyanarayan et al. (2016), que pode ser compreendido como um recurso visual que permite ser alocado a diferentes campos das bases de dados. São eles: X, Y, cor (*color*), tamanho (*size*) e forma (*shape*). Mackinlay (1986) e Cleveland e McGill (1984) sugerem um ordenamento destes canais de codificação segundo sua efetividade em

⁶ Com pleno emprego de canais de codificação visual e elementos de interatividade.

⁷ Biblioteca desenvolvida para a linguagem de programação R.

⁸ Biblioteca de visualização criada por Mike Bostock, (BOSTOCK; HEER; OGIEVETSKY, 2019), para a linguagem JavaScript.

apresentar dados de acordo com seu tipo (nominal, ordinal e quantitativo), apresentado no quadro 1, presente na seção 4.1 deste mesmo capítulo.

Uma vez que, para todos os tipos de dados, os canais de codificação visual mais eficazes são X e Y, buscou-se priorizar sua alocação aos campos da base de dados mais relevantes para responder ao conjunto de perguntas essenciais selecionado. Para se definir as colunas mais relevantes à cada pergunta, acessou-se o conhecimento prévio dos especialistas. Esse processo pode ser melhor compreendido analisando-se as seguintes situações:

- (1) Caso deseje-se representar a contribuição de cada coletor ao longo do tempo, é fácil associar os campos mais relevantes como aqueles que contém o nome dos responsáveis pela coleta dos espécimes e suas respectivas datas de coleta.
- (2) Supondo que se queira investigar a profundidade na qual certos espécimes marinhos foram encontrados, pode-se facilmente selecionar a coluna que apresenta a profundidade na qual o espécime foi coletado, assim como aqueles campos que contém informações da sua classificação taxonômica como os mais relevantes à essa aplicação.
- (3) Se a intenção é visualizar pontos de coleta em um mapa, é induzida a seleção dos campos correspondentes às coordenadas de latitude e longitude.

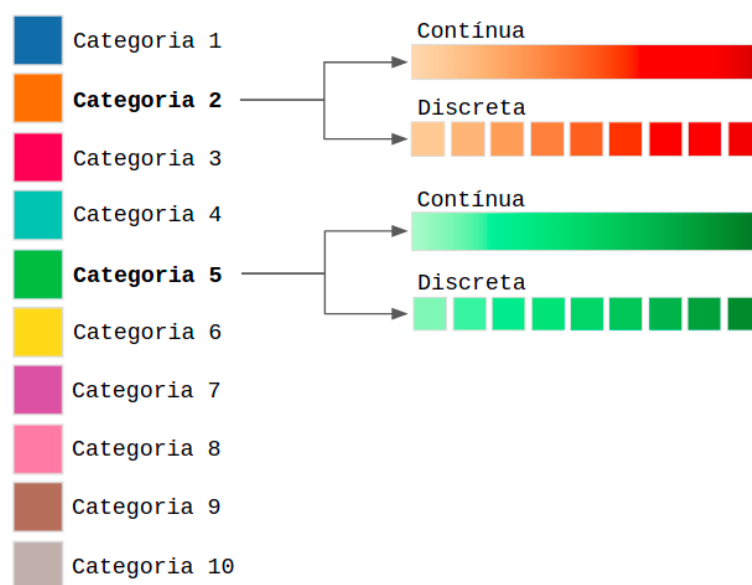
Esse tipo de conhecimento prévio é de fundamental importância para se criar propostas de visualização eficazes, como as ilustradas no quadro 2. Tratando, ainda, dos demais canais de codificação, a cor recebeu atenção especial e é tratada em detalhes na subseção 4.2.1. Dados quantitativos, nesse contexto, são oriundos de operações de agregação estatísticas (i.e., contagens, médias, somas, etc.) e foram retratados em tamanhos diferentes. A forma (*shape*) foi designada para representar informações nominais com poucas categorias, pois sua capacidade em representar dados desse tipo é inferior ao canal da cor (MACKINLAY, 1986; CLEVELAND; MCGILL, 1984).

Por fim, vale ressaltar que, buscando empregar as recomendações trazidas por Qu e Hullman (2016) para se manter um alto nível de consistência global ao trabalhar com conjuntos de visualizações, foram adotadas as seguintes medidas:

- O canal tamanho (*size*) foi reservado apenas para codificar dados quantitativos, nesse caso, quando foi empregado algum tipo de agregação estatística como contagem, soma, média, etc. Representar dados ordinais dessa forma, mesmo que em outro gráfico, pode despejar uma maior carga cognitiva sobre o usuário.

- Períodos de tempo, quando retratados, foram codificados sempre no eixo X. Por ser do tipo ordinal, a ordem intuitiva dessa informação sempre foi respeitada nas visualizações propostas⁹.
- O canal de cor foi reservado para uso apenas com variáveis nominais, atendendo a necessidade reportada por especialistas em representar a taxonomia dos espécimes visualizados nos diferentes gráficos. Mesmo se necessária a representação de informações quantitativas por meio da cor, recomenda-se a seleção (filtragem) de uma categoria nominal¹⁰ e a adoção de um esquema de cores quantitativo cujas graduações remetam à mesma cor da categoria analisada. Se necessário representar tal informação para múltiplas categorias, recomenda-se que sejam elaborados gráficos separados (um gráfico para cada categoria retratada). Esse processo é exemplificado na figura 38. A seção 4.2 aborda, em detalhes, a estratégia usada para se definir o mapeamento de cores empregado neste trabalho.
- Por fim, buscou-se manter os canais de *color.nominal* e forma (*shape*) constantes para um mesmo grupo de visualizações.

Figura 38 – Seleção de esquema de cores quantitativo a partir de uma dada categoria.



Fonte: Elaboração própria.

Nota: Imagem criada a partir dos esquemas de cores disponíveis nas ferramentas Vega e Vega-Lite, listados em <<https://vega.github.io/vega/docs/schemes/>> (acessado em 19/02/2021).

⁹ Foi convencionado que anos, meses, etc. seriam representados do mais antigo para o mais novo, sempre da esquerda para a direita.

¹⁰ Filtrar apenas informações dessa categoria para visualização.

4.2.1 Atribuição do canal de cor

O canal de cor foi predominantemente empregado para mapear informações referentes à pergunta essencial "O quê?" a qual, por sua vez, faz referência à classificação taxonômica dos espécimes. Ao representar táxons dessa forma, é possível integrar essas informações às visualizações que alocam os eixos X e Y para outros campos, como local, período e pessoas envolvidas no histórico de coleta e identificação dos espécimes, o que acrescenta uma nova dimensão para análise visual. Essa estratégia possibilita a identificação de padrões que, de outra forma, poderiam não ser identificados visualmente. Ainda, o objeto da legenda de cores possibilita a atribuição de seletores interativos, o que abre uma nova gama de possibilidades para exploração visual de táxons permitindo, por exemplo, a aplicação de filtros para facilitar comparações ou realçar padrões pertinentes a um subconjunto de táxons.

Contudo, por se tratar de uma estrutura hierárquica, toda a árvore taxonômica não poderia ser diretamente atribuída à cor. Assim, se fez necessário selecionar um nível taxonômico informativo, i.e., que atribui variabilidade o suficiente aos exemplares¹¹ e, ao mesmo tempo, apresenta um número restrito de categorias para que, a partir de sua cor, os espécimes sejam diferenciáveis ao olho humano¹².

Ao curso de reuniões periódicas e após a execução de alguns testes¹³, especialistas optaram por mapear cores ao nível de classificação taxonômica de Famílias. Por apresentar maior variabilidade, o campo de Famílias também favorece maior imersão horizontal ao acrescentar mais possibilidades de filtragem por meio de elementos interativos. Isto posto, a escolha de mapeamentos de cores foi feita empregando-se a ferramenta Colorgorical, construída por Gramazio, Laidlaw e Schloss (2017), que faz a pré-seleção de paletas de cores categóricas otimizadas com base nos seguintes parâmetros:

- **Distância perceptual** (*perceptual distance*): uma maior distância perceptual favorece a seleção de cores mais facilmente discrimináveis ao olho humano¹⁴.
- **Diferença de nomes** (*name difference*): um incremento nesse parâmetro beneficia a escolha de paletas de cores com nomes menos usuais.¹⁵ Gramazio, Laidlaw e Schloss (2017) destaca que esse parâmetro é semelhante à distância perceptual, mas pode fornecer resultados diferentes em certas áreas do espaço de cores.

¹¹ Selecionar um nível muito elevado da árvore taxonômica, por exemplo o "Filo", não seria informativo pois todos os pontos do gráfico estariam marcados da mesma cor.

¹² Usar uma paleta de cores com uma quantidade muito grande de categorias pode romper os limites da cognição humana em detectar padrões e diferenciar pares de cores.

¹³ Inicialmente, testou-se a codificação de cor para diferentes níveis taxonômicos, como Ordem e Infraordem. No entanto, a escolha de mapear famílias no canal de cor foi consensual.

¹⁴ A ferramenta Colorgorical mapeia a acuidade humana de cor usando a métrica CIEDE2000 no espaço de cores CIE Lab. Mais detalhes em Gramazio, Laidlaw e Schloss (2017).

¹⁵ O Colorgorical calcula essa métrica usando a função de diferença de nomes de Heer e Stone.

- **Preferência de pares** (*pair preference*): atribuir maior peso a esse critério propicia a escolha de cores que, em média, são consideradas esteticamente mais preferíveis quando apresentadas conjuntamente. Esse parâmetro é baseado em Schloss e Palmer (2011).
- **Exclusividade de nome** (*name uniqueness*): esse critério favorece a seleção de cores cujos nomes são exclusivos (i.e., identificados mais facilmente)¹⁶. Gramazio, Laidlaw e Schloss (2017) destaca que algumas cores, como o vermelho, são prontamente nomeadas favorecidas por esse critério, enquanto outras são nomeadas de forma menos óbvia e, portanto, são ignoradas.

Na presente aplicação, priorizou-se a escolha de paletas de cores perceptualmente distinguíveis para, assim, maximizar a identificação taxonômica dos espécimes visualizados mesmo que, para tal, fosse necessário o emprego de cores esteticamente não agradáveis quando visualizadas lado a lado. Assim, os parâmetros *perceptual distance*, *name difference* e *name uniqueness* foram configurados para o maior peso possível, enquanto que a preferência de pares foi quase desconsiderada.

No entanto, devido à grande quantidade de famílias distintas em cada uma das bases (108, 59 e 52 para crustáceos, poliquetas e répteis, respectivamente - vide tabela 1 no capítulo 3 para mais detalhes), a seleção de um conjunto de cores dessa ordem de grandeza comprometeria a diferenciabilidade dos espécimes a partir de sua cor, i.e., dificilmente poderia-se selecionar um mapeamento de cores eficaz nesses casos¹⁷. Vale pontuar que compreende-se como esquema de cores eficaz aqueles que usam cores suficientemente discrimináveis ao mesmo tempo em que se adequam bem aos dados e são esteticamente agradáveis (SMART; WU; SZAFIR, 2019).

Assim, a fim de reduzir, inicialmente, a dimensionalidade do problema de seleção de cores, o conhecimento de domínio de especialistas foi acessado de forma a agrupar tais famílias segundo algum critério de relevância. Tal critério revelou-se na forma de um nível hierárquico superior, mais especificamente, nos níveis de Infraordem, Subordem e Ordem para as bases de crustáceos, répteis e poliquetas, respectivamente. Vale destacar que alguns grupos foram subdivididos para melhor atender a necessidade dos curadores em retratar maior proximidade de características entre alguns subconjuntos de táxons. Isto posto, para cada base de dados, foi mapeada uma cor para cada um dos grupos sugeridos empregando-se a ferramenta Colorgorical. O resultado dessa etapa é exposto na figura 39.

¹⁶ Assim como a diferença de nomes, também é baseado na função de Heer e Stone.

¹⁷ Não obstante, a ferramenta Colorgorical não permite a seleção de paletas de cores com muitos elementos. Durante a etapa de testes dessa ferramenta, conseguiu-se gerar esquemas de cores com, no máximo, 25 elementos. Contudo, ao se tentar criar paletas muito grandes, muitas vezes uma mensagem de erro foi exibida e foi obtido um número de cores inferior ao limite observado.

É importante citar que algumas cores foram alteradas para melhor atender demandas dos curadores em retratar maior nível de semelhança entre alguns subconjuntos de táxons, ou melhor retratar os espécimes em questão realçando ou, até mesmo, identificando padrões visuais. As cores modificadas dizem respeito aos grupos: Serpentes - Scolecophidia, para a base de répteis, e Sabellida, para a coleção de poliquetas. Contudo, tomou-se o devido cuidado para preservar o alto nível de distância perceptual herdado da proposta de cores fornecida pelo Colorgorical.

Figura 39 – Paletas de cores criadas com o Colorgorical^a.



Fonte: Elaboração própria.

Nota: Paletas de cores criadas pela ferramenta Colorgorical (GRAMAZIO; LAIDLAW; SCHLOSS, 2017).

A sugestão dos grupos foi feita pela equipe de curadores do Museu Nacional responsável pelas respectivas coleções trabalhadas. Os seguintes níveis taxonômicos foram tomados como referência, mas não necessariamente seguidos à risca (alguns grupos podem ter sido subdivididos para melhor refletir a semelhança entre alguns de seus táxons, segundo sugestão dos especialistas): Infraordem, Subordem e Ordem para as coleções de crustáceos, répteis e poliquetas, respectivamente. ^aAlgumas cores foram alteradas para melhor atender a necessidade de especialistas em refletir maior grau de semelhança entre dois grupos, ou melhor representar os táxons em questão. Foram elas: Serpentes - Scolecophidia, para a base de répteis, Sabellida para a coleção de poliquetas.

Nesse ponto, uma solução aceitável para criar uma proposta de mapeamento a nível de Família seria representar um esquema de cores em que, a cada elemento (i.e., família), seria atribuída a cor selecionada para o seu respectivo grupo. De fato, essa é uma abordagem semelhante à adotada por Hans Rosling em sua famosa visualização Gapminder, apresentada em 2007 em sua palestra para o TEDTalks¹⁸, em que a cada país é atribuído uma cor associada ao seu respectivo continente. A identificação exata do país

¹⁸ Apresentação disponível em <<https://www.youtube.com/watch?v=hVimVzgtD6w>>.

é deixada a cargo do elemento de interatividade *Tooltip*¹⁹. É importante pontuar que essa foi a abordagem empregada para a coleção de crustáceos, uma vez que esta possui quase o dobro de famílias distintas contabilizadas nas demais bases de dados. Essa abordagem alia as vantagens da redução de dimensionalidade na etapa de escolha de cores diferenciáveis ao olho humano (usando a ferramenta Colorgorical), ao mesmo tempo em que possibilita a aplicação de filtros mais variados por meio de elementos de interatividade associados ao objeto de sua legenda.

Considerando as bases de dados Répteis e Poliquetas, em particular, optou-se por realizar um passo adicional para a composição de paletas de cores em que fossem assinaladas cores únicas a cada família, sem comprometer a diferenciação de seus grupos, i.e., possibilitar distinção entre famílias dentro de um mesmo grupo, sem comprometer a distância perceptual estabelecida pelo algoritmo de seleção de cores do Colorgorical. Para tal, foi empregado o seguinte processo:

1. A cor de cada um dos grupos, previamente selecionada com a ferramenta Colorgorical, foi definida como ponto centroide para a criação de um novo esquema de cores em torno dessa cor referência.
2. Então, cada cor centroide foi inserida na ferramenta Color Crafter[®] que, por sua vez, automatiza a criação paletas de cores eficazes a partir do *input* de um único código hexadecimal de cor.
3. Por meio de uma abordagem algorítmica²⁰ que modela padrões existentes em paletas de cores criadas diretamente por designers, o Color Crafter[®] fornece como *output*, para cada cor inserida, uma lista de nove códigos hexadecimais referentes à uma paleta de cores discretizada cujas entradas foram selecionadas nas proximidades do ponto escolhido.

Vale destacar que a ferramenta permite a alteração de alguns parâmetros, como matiz (*hue*), luminância e saturação. Não obstante, são fornecidos algumas pré-configurações desses parâmetros, o que resulta na sugestão de diferentes propostas de paletas de cores, além de permitir a modificação de um esquema de cores selecionado. Aqui, a definição destes parâmetros, i.e., a escolha de um dos esquemas de cores sugeridos é deixada à cargo do usuário por também envolver preferências estéticas. No entanto, deve-se salientar que o manuseio desses parâmetros deve ser feito com cautela para que o esquema de cores resultante não invada a vizinhança definida pelas demais cores centroides.

¹⁹ Moldura flutuante que é exibida na tela quando passa-se o *mouse* sobre um elemento da interface, contendo uma explicação adicional sobre tal elemento

²⁰ Utilizando técnicas de normalização, clusterização (*K-Means clustering*) e construção de modelos. Tais modelos foram construídos a partir de um corpus de 222 paletas de cores projetadas por especialistas e os resultados são usados nessa ferramenta para criar novas paletas de maneira automática, imitando as práticas adotadas por projetistas.

Uma vez selecionado um conjunto de 9 saídas, considera-se as seguintes situações:

- 3.1 Caso o número de famílias em um grupo seja inferior a nove (número de cores fornecido a cada passo pelo Color Crafter[®]), são assinaladas uma cor à cada família, desconsiderando algumas cores intermediárias como uma tentativa de induzir uma maior distância perceptual entre esses táxons. Esse passo é exemplificado para um grupo formado por 4 famílias na figura 40.
- 3.2 Para $N = 9$ elementos, o mapeamento de cores é imediato, podendo ou não seguir um ordenamento relevante.
- 3.3 Para o caso em que o número de N elementos (ou famílias), onde N é superior ao número de cores fornecido pelo Color Crafter[®] para uma *seed*, para se criar a quantidade excedente de cores, é necessário o manuseio dos parâmetros da ferramenta de forma a criar uma nova paleta de cores, variante da primeira. A figura 41 ilustra esse processo.

Da modificação dos parâmetros de matiz, luminância e saturação, deriva-se uma nova paleta de cores com nove entradas. Recomenda-se que seja dada preferência à ajustes de luminância por, facilmente, gerar novas cores que podem ser facilmente agrupadas às da primeira paleta por um processo de interpolação. Se necessário, leves ajustes de saturação podem auxiliar nesse processo para elevar a distância perceptual das novas cores em relação às anteriores²¹.

É importante ressaltar que se deve ter o devido cuidado para que as novas cores não invadam o espaço definido pelas demais cores centroides²². Para que a distância perceptual definida pelo Colorgorical seja mantida (i.e., para facilitar a distinção de grupos a partir da cor de seus elementos), é importante que os espaços de cores induzidos a partir dos seus pontos centroides sejam disjuntos. Por esse motivo, alterações diretas de matiz (*hue*) são desencorajadas ao passo que elevam as chances de que isso aconteça.

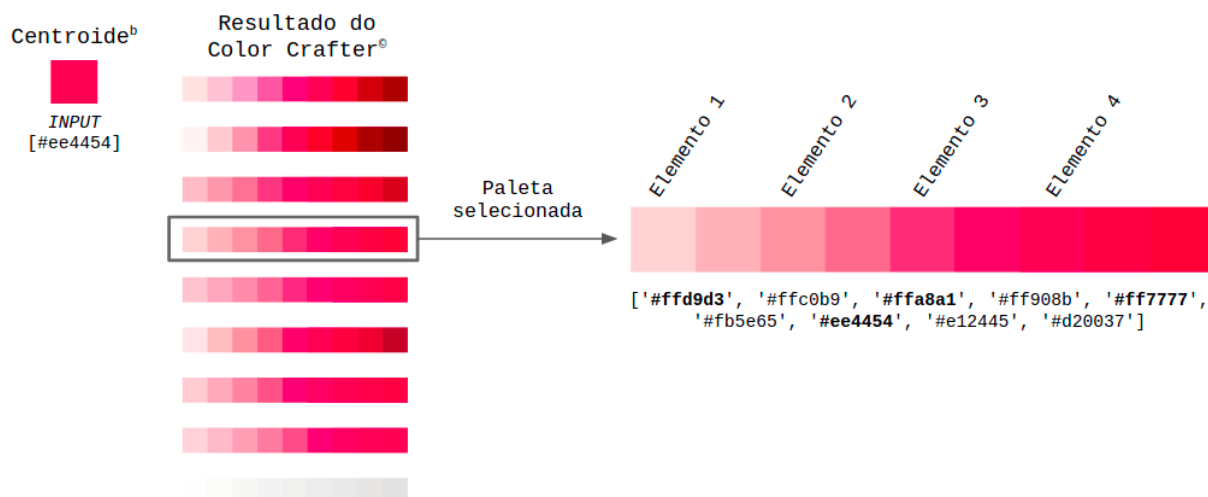
Uma vez obtido um novo conjunto derivado de cores, pode-se agregá-lo ao conjunto selecionado inicialmente para, assim, criar uma mapeamento de cores com mais entradas. Por fim, recomenda-se que a nova paleta estendida de cores seja ordenada de acordo com sua luminância, reproduzindo um resultado semelhante ao que seria dado como *output* pelo Color Crafter[®], segundo o apresentado em Smart, Wu e Szafir (2019). O ordenamento que produziu os

²¹ Ajustes na saturação devem ser feitos com cautela pois podem ocasionar na transição de tons mais "quentes" para tonalidades mais "frias", o que pode prejudicar a diferenciação entre grupos (misturar os espaços definidos pelas cores centroides).

²² Recomenda-se que seja dada prioridade à deslocamentos horizontais na curva C-L. Translações verticais também são toleráveis, mas devem ser feitos em menor medida e demandam mais atenção do usuário.

resultados mais satisfatórios foi feito a partir do cálculo do brilho percebido de uma cor em sua versão normalizada²³.

Figura 40 – Exemplo de atribuição de cores para 4 elementos a partir das 9 cores sugeridas pelo Color Crafter[®]^a.

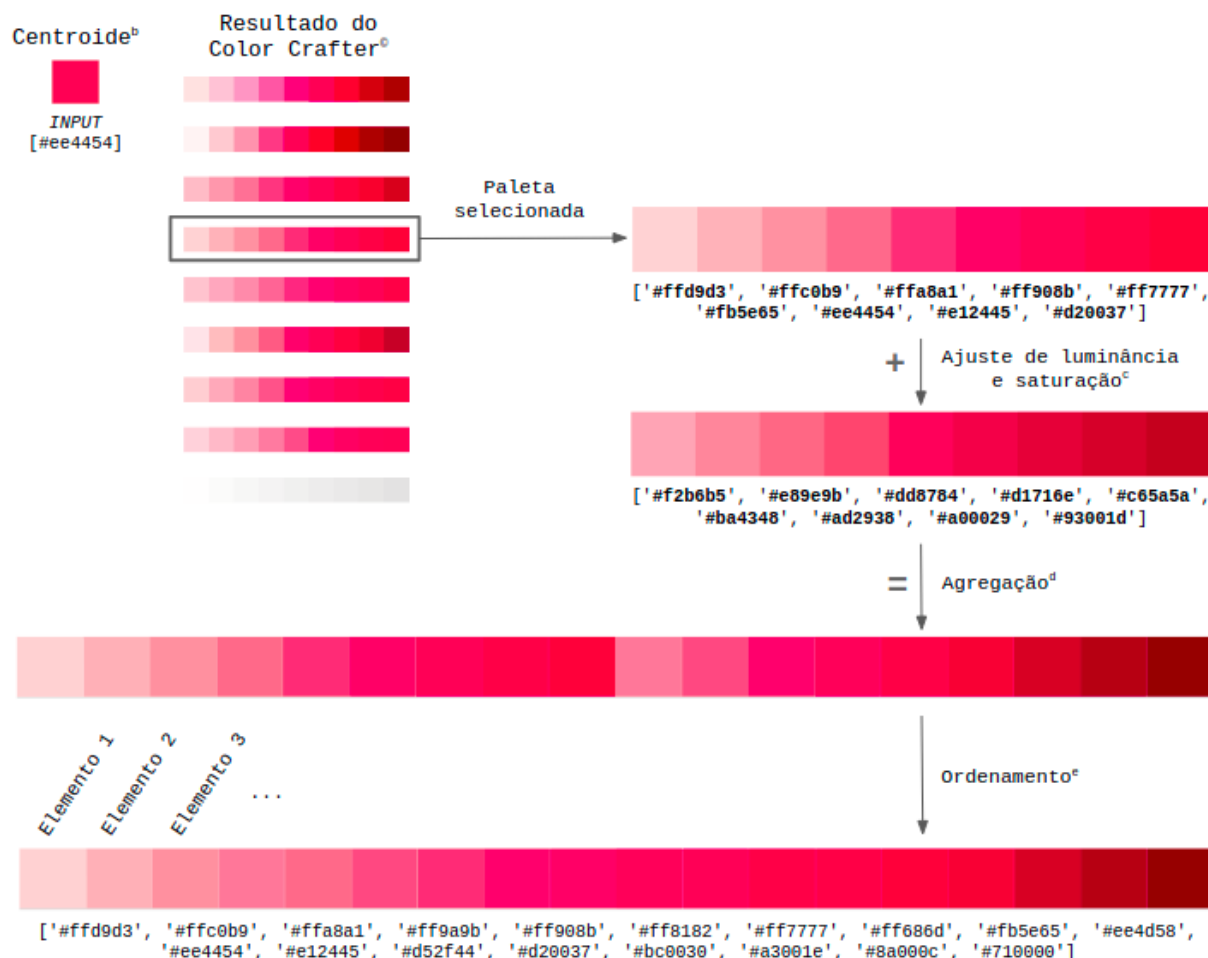


Fonte: Elaboração própria.

Nota: Neste exemplo, supõe-se o caso em que se quer atribuir uma graduação de cores para um grupo constituído por quatro elementos (nesse caso, famílias) a partir de sua cor centroide dentre de uma paleta de cores com nove entradas sugerida pelo Color Crafter[®]. O mesmo processo pode ser facilmente replicado para qualquer aplicação na qual se almeje atribuir cores a qualquer quantidade $N \leq 9$ elementos, empregando o algoritmo aqui desenvolvido. Para o caso em que $N < 9$, recomenda-se a eliminação de algumas cores intermediárias para, dessa forma, elevar a distância perceptual entre as diferentes categorias do mapa de cores. ^aDisponível em <http://cu-visualab.org/ColorCrafter/>; ^bA cor centroide é definida a partir da ferramenta Colorgorgical para o agrupamento sugerido. O código hexadecimal dessa cor referência é inserido no Color Crafter[®], gerando o resultado reproduzido na figura.

²³ Foram testados três formas de se fazer esse ordenamento a partir de formas distintas de se calcular luminância dado um código de cores RGB. São elas: a) $Y = 0.2126 \cdot R + 0.7152 \cdot G + 0.0722 \cdot B$ (usada em espaços de cores colorimétricos que usam primárias ITU-R BT.709); b) $Y = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$ (brilho percebido de uma cor - <https://www.w3.org/TR/AERT/#color-contrast>); c) $Y = \sqrt{0.299 \cdot R^2 + 0.587 \cdot G^2 + 0.114 \cdot B^2}$ (versão normalizada do brilho percebido de uma cor). Todos os códigos de cor fornecidos pelo Color Crafter[®] são hexadecimais. Para calcular tais métricas, primeiramente foi feita a conversão de todos eles para o formato RGB.

Figura 41 – Exemplo de atribuição de cores para uma quantidade N grande de elementos usando a ferramenta Color Crafter^{®a}.

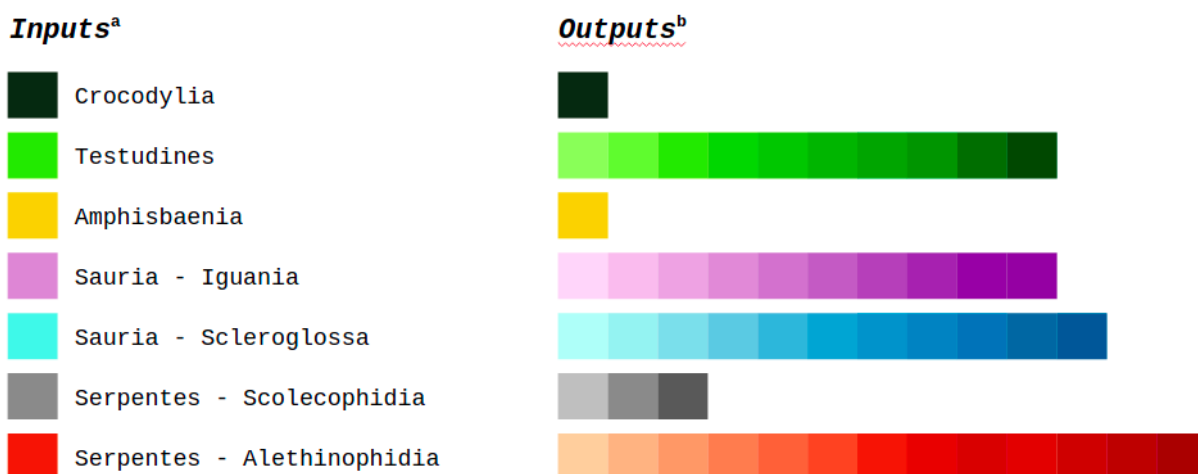


Fonte: Elaboração própria.

Nota: Neste exemplo, supõe-se o caso em que se deseja atribuir uma gradação de cores para N elementos, onde N é um número superior a 9 (quantidade de cores sugeridas pelo Color Crafter[®] a partir de uma *seed*). O mesmo processo pode ser facilmente replicado para qualquer aplicação na qual se almeje atribuir cores a qualquer quantidade $N > 9$ elementos, empregando o algoritmo aqui desenvolvido. ^aDisponível em <<http://cu-visualab.org/ColorCrafter/>>; ^bA cor centroeide é definida a partir da ferramenta Colorgorical para o agrupamento sugerido. O código hexadecimal dessa cor referência é inserido no Color Crafter[®], gerando o resultado reproduzido na figura; ^cO Color Crafter[®] permite a manipulação dos parâmetros: matiz (*hue*), luminância e saturação. No contexto aqui trabalhado, para que a vizinhança definida pelas demais cores centroeides não seja invadida, recomenda-se que seja dada preferência à ajustes de luminância e, se necessário, saturação (este último pode ser necessário para induzir mais variabilidade de cor, em especial para paletas muito saturadas); ^dA partir da manipulação dos *presets* da ferramenta, cria-se uma nova variação de cores a partir da primeira. Para criar uma versão estendida da paleta de cores selecionada, i.e., com mais categorias, basta agregar as novas entradas à lista anterior; ^eEm contextos em que a intensidade da cor é relevante, ou simplesmente para manter a consistência em sua apresentação, pode-se ordenar as cores da nova lista (no exemplo, com 18 entradas) de acordo com sua luminância (seguindo o ordenamento original proposto por Smart, Wu e Szafr (2019) e empregado na ferramenta Color Crafter[®]). Há mais de uma forma de se calcular a luminância de uma cor, mas os resultados mais satisfatórios foram apresentados pela versão normalizada do brilho percebido de uma cor ($\sqrt{0.299 \cdot R^2 + 0.587 \cdot G^2 + 0.114 \cdot B^2}$ - <<https://www.w3.org/TR/AERT/#color-contrast>>). O Apêndice A ilustra como o ordenamento foi feito para esse exemplo.

Com exceção da coleção de crustáceos, esse procedimento foi aplicado às bases de répteis e poliquetas, obtendo os resultados expostos nas figuras 42 e 43, respectivamente.

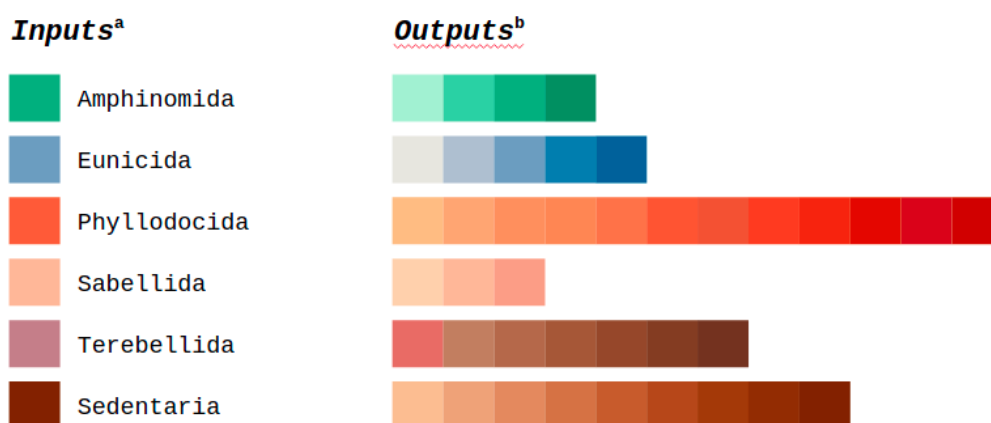
Figura 42 – Paletas de cores discretas criadas com o Color Crafter[®] para a coleção de répteis.



Fonte: Elaboração própria.

Nota: Paleta de cores resultante do processo apresentado no decorrer desta seção para a coleção de répteis. Feita empregando as ferramentas Colorgorical (GRAMAZIO; LAIDLAW; SCHLOSS, 2017), acessível em <<http://vrl.cs.brown.edu/color>>, e Color Crafter[®], desenvolvida por Smart, Wu e Szafr (2019) e disponível em <<http://cu-visualab.org/ColorCrafter/>>. ^aCores selecionadas para os sete grupos sugeridos pela equipe de curadores do departamento de Herpetologia do Museu Nacional (MNRJ), definidas como cores centroides a serem inseridas no Color Crafter[®]; ^besquemas de cores criados empregando o método apresentado nesta seção, exemplificado nas figuras 40 e 41.

Figura 43 – Paletas de cores discretas criadas com o Color Crafter^{®a} para a coleção de poliquetas.



Fonte: Elaboração própria.

Nota: Paleta de cores resultante do processo apresentado no decorrer desta seção para a coleção de poliquetas. Feita empregando as ferramentas Colorgorical (GRAMAZIO; LAIDLAW; SCHLOSS, 2017), acessível em <<http://vrl.cs.brown.edu/color>>, e Color Crafter[®], desenvolvida por Smart, Wu e Szafr (2019) e disponível em <<http://cu-visualab.org/ColorCrafter/>>. ^aCores selecionadas para os sete grupos sugeridos pela equipe de curadores do departamento de Polychaeta do Museu Nacional (MNRJ), definidas como cores centroides a serem inseridas no Color Crafter[®]; ^besquemas de cores criados empregando o método apresentado nesta seção, exemplificado nas figuras 40 e 41.

Adicionalmente, pode-se dizer que o processo exposto nessa seção pode ser facilmente replicado para quaisquer aplicações em que se tem um grande número de elementos a serem codificados no canal de cor, desde que estes possam ser agrupados segundo algum critério. No entanto, deve-se atentar a algumas limitações: 1. a presença de muitos grupos pode induzir a seleção de cores centroides em espaços de cores próximos, o que pode dificultar a expansão da paleta de cores pelo processo exemplificado na figura 41; 2. o mapeamento de cores para um grupo com muitos elementos pode ser desafiador dado que se torna mais difícil selecionar novas cores com diferença perceptível ao mesmo tempo em que se busca permanecer no interior da vizinhança definida pela cor centroide. No ponto em que se torna difícil identificar diferenças entre elementos pertencentes a um mesmo grupo, recomenda-se uma abordagem mais simples, como atribuir a todos os elementos a cor associada à seu grupo. Ainda, é importante enfatizar que, nas duas situações, talvez seja mais apropriado a codificação da informação desejada nos canais X e Y por serem mais eficazes em representar grandes quantidades de categorias.

Finalmente, é importante citar que, além da criação de um mapeamento de cores para o nível taxonômico de famílias, foi atribuída, quando necessário, um esquema de cores para localizações geográficas (para a pergunta essencial "Onde?", respondida pelos campos: continentes, países, regiões e estados brasileiros). Contudo, devido à alta carga cognitiva depositada pelas paletas de cores por família (por apresentar um grande número de elementos com cores únicas), para representar localizações geográficas, optou-se por empregar versões mais simples de esquemas de cores, selecionados em apenas um passo por meio da ferramenta Colorgorical (GRAMAZIO; LAIDLAW; SCHLOSS, 2017). Especificamente, selecionou-se um conjunto de cores para representar continentes e, quando necessário, associou-se cada país a cor do seu respectivo continente. O mesmo processo foi feito para se representar regiões e estados brasileiros.

Considerando que, toda vez que o mapa de cores é alterado, usuários vão se deparar com uma maior carga cognitiva devido à necessidade de "desaprender" um conjunto de atribuições (QU; HULLMAN, 2016), buscou-se manter a consistência global no conjunto de visualizações proposto à luz das recomendações dadas por Qu e Hullman (2016). Nesse sentido, manteve-se fixo o mapeamento de cores entre os gráficos pertencentes à mesma coleção. Ainda, como foi necessário o uso de mais de um mapeamento de cores, recomendou-se que gráficos com codificações de cor distintas não fossem apresentados lado a lado com a intenção de se reduzir a carga cognitiva depositada sobre o usuário, melhorando a fluidez de sua leitura e compreensão.

4.3 Interatividade

Essa seção traz uma breve exposição acerca dos elementos de interatividade empregados nos conjuntos de visualizações propostas. É importante mencionar que, durante as etapas de desenvolvimento, considerou-se o mantra: "*overview first, zoom and filter, then details on demand*", pioneiramente apresentado por Shneiderman (2003). Nesse sentido, buscou-se empregar elementos de interatividade de modo a permitir o tráfego entre diferentes níveis de granularidade e, ao mesmo tempo, prover navegação horizontal pela aplicação de filtros e seleção de diferentes objetos.

Deve-se ressaltar que é esperado que o público-alvo (i.e., especialistas do domínio da biodiversidade) interaja com as visualizações aqui construídas por meio do paradigma tradicional WIMP (definido em Liu et al. (2014)), isto é, por meio de elementos como janelas, ícones, *mouse* e ponteiros. Esse paradigma compreende ambientes *desktop* tradicionais em que *inputs* do usuário são coletados por meio de periféricos como teclado e *mouse* (em computadores e notebooks, por exemplo), e não abrange dispositivos modernos com interface sensível ao toque (i.e., celulares, tablets, etc.).

Isto posto, em cada visualização proposta neste estudo, foram empregados os seguintes elementos para:

- **Imersão vertical:** essa movimentação está presente por meio da implementação do elemento gráfico *tooltip*, que pode ser compreendido como uma moldura ou *card* flutuante que é exibido na tela no momento em que se passa o *mouse* sobre um elemento da interface, contendo uma informações adicionais sobre o elemento específico. Tendo em vista o objetivo de constituir um recurso visual capaz de auxiliar no processo de melhoria qualitativa das coleções digitais científicas de biodiversidade, o nível de detalhamento fornecido pela *tooltip* é de fundamental importância para guiar especialistas na detecção de tais inconsistências. Buscou-se sempre representar, desta forma, detalhamentos suficientes para possibilitar a identificação exata do registro inconsistente²⁴ ou para guiar a aplicação de filtros diretamente nas planilhas de dados, restringindo a área de busca²⁵.
- **Navegação horizontal:** esse tipo de interação está associado ao objeto da legenda de cores, permitindo a aplicação de filtros referentes às categorias codificadas

²⁴ Nesse caso, o número de catálogo do referido espécime é a informação mais exata que se pode apontar, possibilitando a localização imediata do registro desejado. Contudo, em visualizações que foram representadas agregações quantitativas (como contagens e somas), a capacidade de apresentar esta informação na forma de um *card* tornou-se, rapidamente, limitada por exigir a exibição de uma lista, muitas vezes, com dezenas ou centenas de números de catálogo. Nesses casos, foram apresentadas informações capazes de restringir a área de busca por meio da aplicação de filtros diretos na planilha de dados.

²⁵ Informações como a classificação taxonômica (Ordem, Família, Espécie, Gênero...), localização geográfica (Continente, País, Região ou Estado), além do período de coleta ou catalogação.

no canal de cor. Com isso, pode-se analisar padrões e possíveis inconsistências considerando apenas um subconjunto de categorias por vez, sem a interferência dos demais pontos. Métodos de filtragem podem auxiliar no processo de detecção, análise e diagnóstico de inconformidades nas bases de dados.

4.4 Propostas gráficas

Em linha com o que foi discutido no decorrer deste capítulo, a presente seção objetiva apresentar as propostas gráficas criadas a partir do raciocínio criativo desenvolvido na seção 4.1, codificando os canais visuais conforme a seção 4.2 e incorporando os elementos interativos mencionados na seção 4.3. Nesse sentido, buscou-se construir recursos visuais capazes de auxiliar na identificação de possíveis inconsistências nas bases de dados e, de certa forma, apontar informações suficientes para facilitar a adoção de medidas corretivas ao restringir a área de busca diretamente nas planilhas de dados ou, quando possível, apontar a identificação exata dos registros inconsistentes.

Isto posto, foram construídas propostas de visualização ao nível de granularidade definido por Windhager et al. (2018) como *multi-object previews*, que provê a apresentação de uma seleção de objetos, o mais próximo possível do nível de *collection overview*²⁶. Não obstante, todos os modelos de visualização aqui propostos, construídos por intermédio da ferramenta Altair para a linguagem de programação Python, são apresentados segundo a seguinte classificação, também adotada na seção 2.2 do capítulo 2: cobertura geográfica, taxonômica e temporal. Como novidade, o presente *framework* incorpora um conjunto de visualizações que têm enfoque nos indivíduos responsáveis pelo ciclo de pesquisa dos espécimes, como coletores e determinadores²⁷.

4.4.1 Cobertura geográfica

A extensão geográfica dos espécimes em cada uma das coleções foi retratada de forma semelhante ao que foi observado em *data papers* de biologia, por meio de um mapa. Compreende-se que essa visualização seja eficaz ao induzir no leitor noções que vão além do que é compreendido pela pergunta: *Onde os espécimes foram coletados?*. Ao visualizar um ponto no mapa, o leitor consegue, imediatamente, inferir o tipo de bioma no qual o animal foi coletado, o clima da região e, em alguns casos, o tipo de vegetação e relevo predominantes naquele ambiente. Nesse sentido, a figura 44 ilustra o mapa construído para a coleção de répteis, do departamento de Herpetologia do Museu Nacional (MNRJ).

²⁶ A visão geral da coleção, atingida tanto por meio de transformações visuais discretas, como glifos, quanto por abstrações geométricas, representando padrões encontrado nos dados (WINDHAGER et al., 2018)

²⁷ Pesquisador responsável pela identificação ou reclassificação taxonômica dos espécimes.

Figura 44 – Cobertura geográfica da coleção de répteis (MNRJ).



Fonte: Elaboração própria.

Nota: **Onde?** x **O quê?**: Representação das coordenadas de latitude e longitude do ponto de coleta de cada um dos espécimes presentes na coleção de répteis do departamento de Herpetologia do Museu Nacional (MNRJ). Mapeamento de cores por famílias e forma (*shape*) designado para material-tipo.

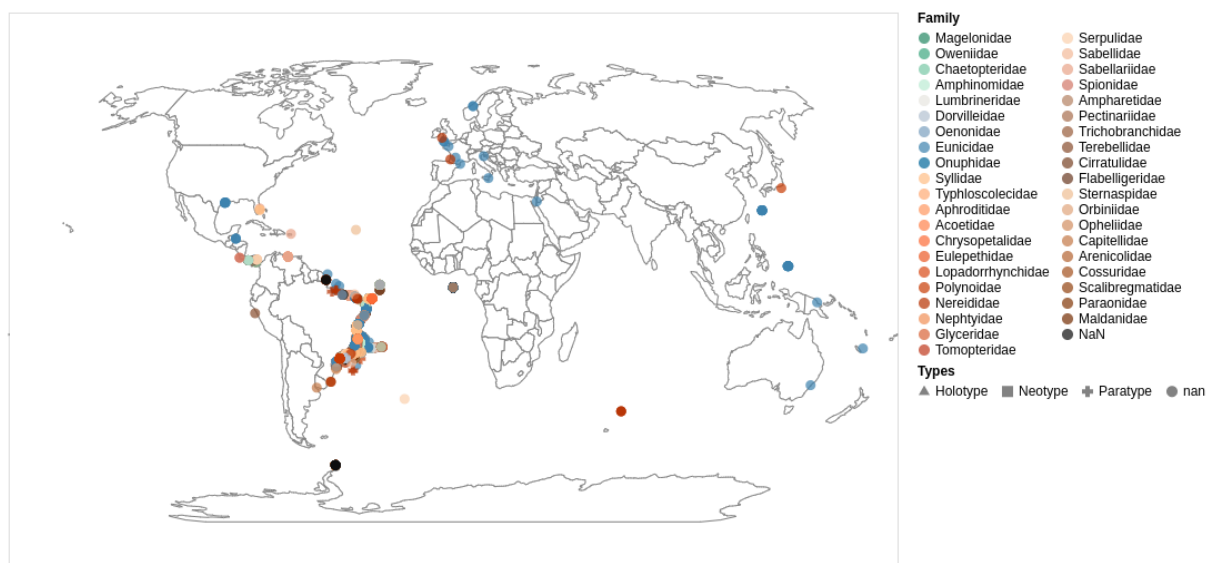
A proposta gráfica exposta na figura 44 busca evidenciar, além de *onde os espécimes foram coletados?*, *o que foi coletado?*. Essa última pergunta está associada à classificação taxonômica dos espécimes que, segundo sugestões de especialistas, é explorada em dois canais de codificação: cor, representando o nível taxonômico de famílias, e forma (*shape*) associando os materiais-tipo²⁸ presentes na coleção.

Nesse gráfico, inicialmente, tem-se um *overview* da distribuição geográfica de toda a coleção, permitindo destacar padrões gerais que, por exemplo, corroboram o fato de que répteis são animais terrestres, encontrados predominantemente nas placas continentais e, em alguns casos, em ilhas. O objeto da legenda contém seletores interativos que permitem filtrar quais famílias serão visualizadas por vez, favorecendo a exploração de padrões específicos a certos táxons. O mesmo se aplica à legenda de tipos. Para contrastar com a figura 44, a figura 45 exibe a mesma proposta gráfica para a coleção de poliquetas, animais exclusivamente marinhos.

Conforme pode ser inferido a partir das figuras 44 e 45, a maior parte dos exemplares da coleção foram captados na América Latina, mais especificamente no Brasil, que é o país sede da instituição. Assim, para explorar com mais detalhes a diversidade de espécimes coletados nesse continente, foi gerado o gráfico apresentado na figura 46 a seguir.

²⁸ Tipo é, basicamente, o que define a nomenclatura de um táxon. Para saber mais a respeito, vide o Código Internacional de Nomenclatura Zoológica (ICZN), disponível em <<https://www.iczn.org/the-code/the-international-code-of-zoological-nomenclature/the-code-online/>> - acessado em 23/02/2021.

Figura 45 – Cobertura geográfica da coleção de poliquetas (MNRJ).



Fonte: Elaboração própria.

Nota: **Onde?** x **O quê?**: Representação das coordenadas de latitude e longitude do ponto de coleta de cada um dos espécimes presentes na coleção de poliquetas do departamento de Polychaeta do Museu Nacional (MNRJ). Mapeamento de cores por famílias e forma (*shape*) designado para material-tipo.

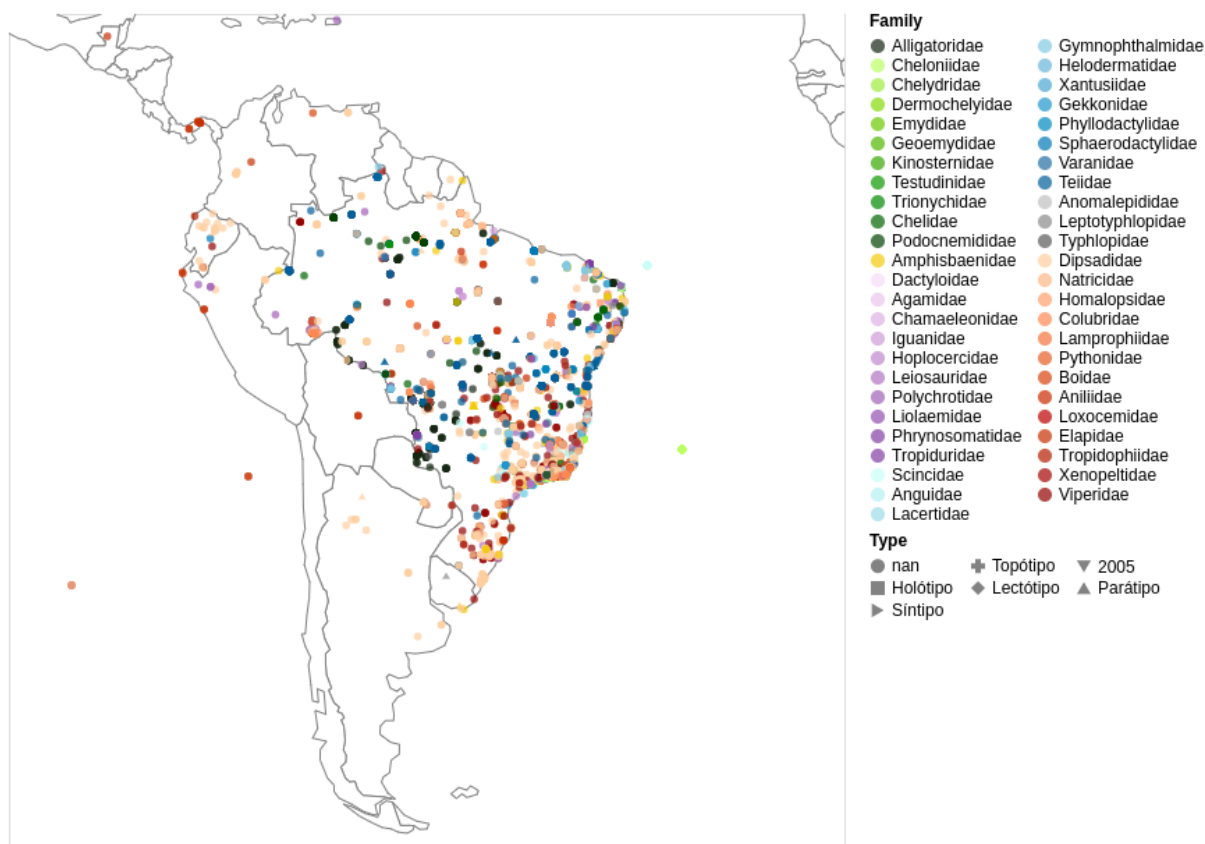
Por vezes, se faz relevante adicionar a dimensão temporal nesse contexto para, analisando a relação entre as dimensões espaço e tempo, concluir padrões pertinentes à trajetória de crescimento da coleção. Ainda, por localização geográfica, não compreende-se apenas coordenadas de latitude e longitude. De fato, a altitude (ou profundidade) em que os espécimes foram encontrados constitui informação relevante para caracterizar o exemplar em questão. No presente contexto, a altitude (ou profundidade) é informada a partir de uma medida em relação ao nível do mar, apresentada em metros. Essas propostas são exploradas nas subseções 4.4.1.1 e 4.4.1.2, respectivamente.

4.4.1.1 Relação espaço-temporal

A resposta à pergunta essencial "**Onde?**", principal foco desta seção, também é encontrada nos campos que contém registros de: continente, país, região e estado de coleta, além das coordenadas de latitude e longitude exploradas anteriormente. Nesse sentido, as visualizações apresentadas nesta subseção apresentam o local de coleta dos espécimes a partir de um dos rótulos citados, codificando essa informação no eixo Y, enquanto o tempo é representado no eixo X, mantendo-se o mesmo esquema de coloração por famílias. Ainda, o canal de codificação tamanho (*size*) foi mapeado para contagens de forma a responder: *Quantos animais foram coletados de uma certa família, em uma dada localização, em um certo ano?*²⁹. A figura 47 expõe essa proposta para o nível geográfico de continentes, considerando-se a coleção de poliquetas.

²⁹ O quê?, Onde? e Quando?

Figura 46 – Cobertura geográfica da América Latina - coleção de répteis (MNRJ).



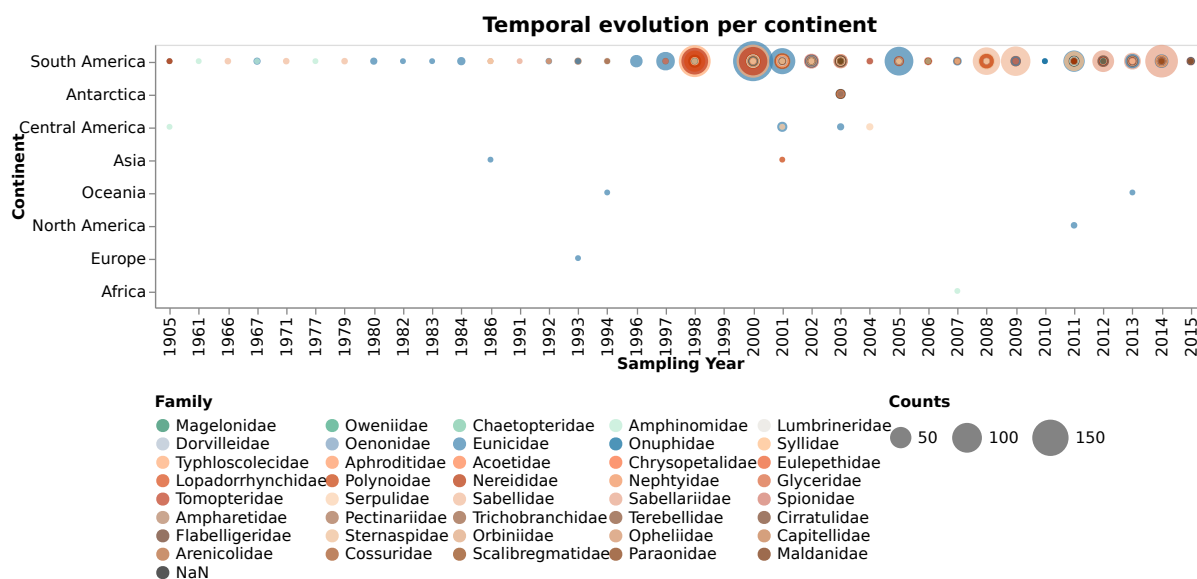
Fonte: Elaboração própria.

Nota: **Onde? x O quê?**: Representação das coordenadas de latitude e longitude, com enfoque na América Latina, do ponto de coleta de cada um dos espécimes presentes na coleção de répteis do departamento de Herpetologia do Museu Nacional (MNRJ). Mapeamento de cores por famílias e forma (*shape*) designado para material-tipo.

Como limitação dessa abordagem, deve-se apontar a presença de pontos sobrepostos, i.e., para um dado continente, em um certo ano, foram coletados exemplares pertencentes à mais de uma família taxonômica. Esse fator pode comprometer ligeiramente o *overview* dessa visualização quando todas as famílias são expostas ao mesmo tempo. No entanto, esse problema torna-se inexistente a partir do momento em que apenas uma família é selecionada, e parcialmente inexistente para um pequeno subconjunto de famílias visualizadas por vez. Apesar dessa limitação, esse tipo de gráfico traz novas possibilidades de exploração ao expor padrões espaço-temporais, além de permitir a análise individual ou comparação direta de famílias. É possível identificar, por exemplo, períodos em que há maior ênfase na captação de determinados táxons, podendo ser ligada à eventos como expedições.

Essa proposta gráfica foi explorada para mais dois níveis de agregação geográfica: países e estados brasileiros, representadas nas figuras 48 e 49, respectivamente.

Figura 47 – Evolução temporal por continente da coleção de poliquetas (MNRJ).



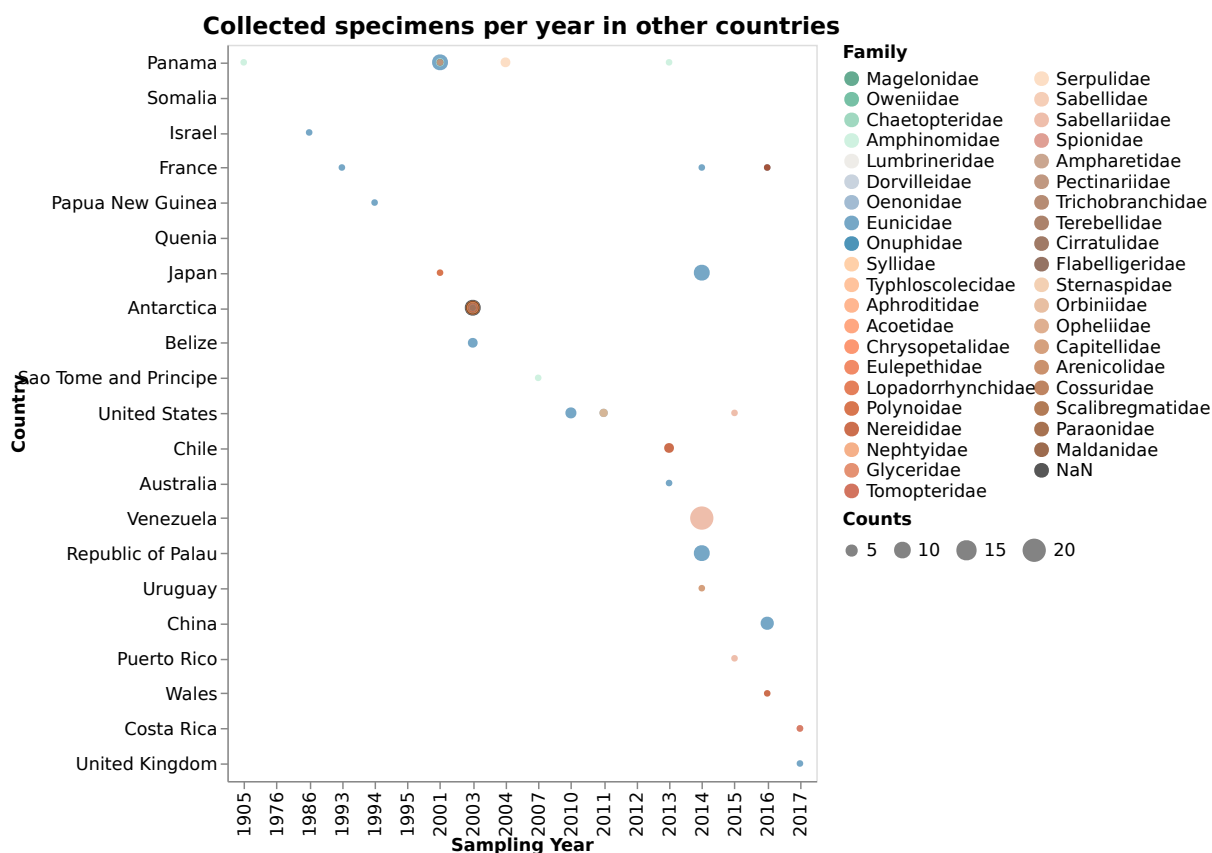
Fonte: Elaboração própria.

Nota: **Onde?** x **Quando?** x **O quê?** x **Quanto?**: Quantidade de exemplares coletados por ano em cada continente, colorido pelo nível taxonômico de famílias. O gráfico exibido refere-se à coleção de poliquetas, do departamento de Polychaeta do Museu Nacional (MNRJ).

Outra vantagem desse tipo de abordagem matricial é a possibilidade de se empregar as técnicas conhecidas como matrizes de Bertin (BERTIN, 1983), bem implementadas na ferramenta Bertifier (PERIN; DRAGICEVIC; FEKETE, 2014), que consistem em reordenar eixos não ordinais a fim de realçar padrões no conjunto de dados. No caso apresentado na figura 48, o eixo Y foi reordenado segundo dois critérios: primeiro ano de coleta registrado e maior número de contagens, isto é, na parte superior são exibidos países que apresentam os registros de coleta mais antigos e menor volume coletado. Desta forma, tanto o ano de coleta quando o volume de coleta associado a cada país aumenta de cima para baixo.

Já na figura 49, optou-se por representar no topo as regiões que mais contribuíram para a composição da coleção em termos de espécimes captados na natureza. Ainda, dentro de cada região, os Estados foram ordenados para que, no topo, fosse exibido a unidade da federação mais significativa para aquela região, resultando no padrão observado.

Figura 48 – Evolução temporal por país da coleção de poliquetas (MNRJ).

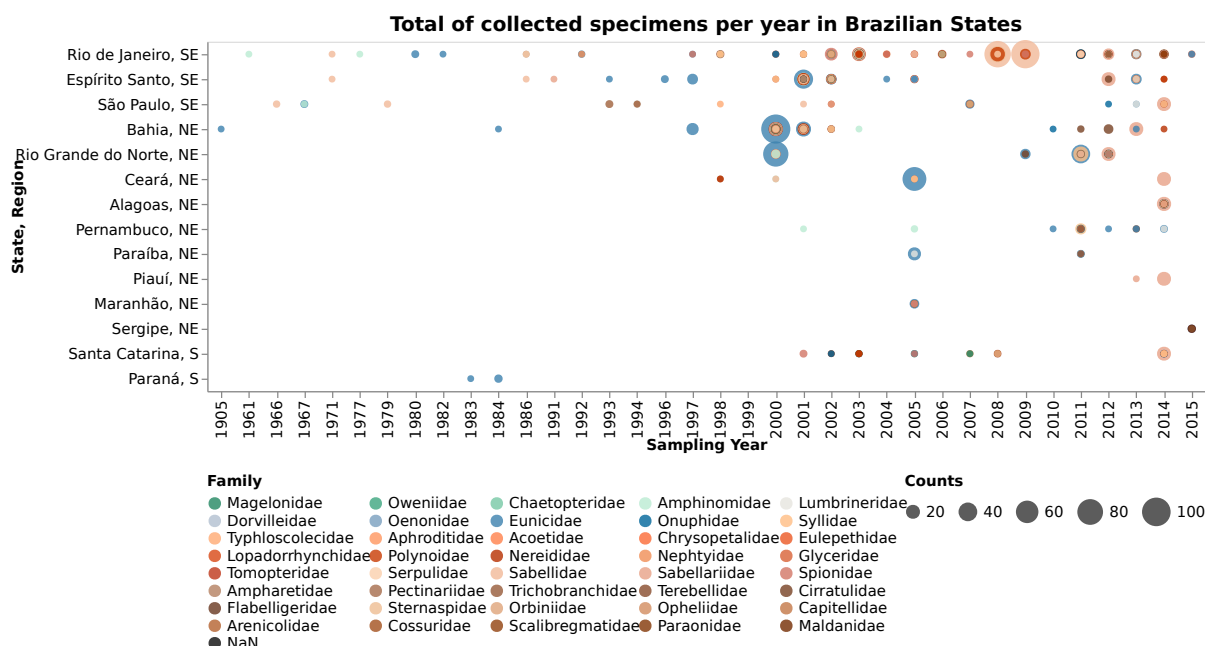


Fonte: Elaboração própria.

Nota: **Onde?** x **Quando?** x **O quê?** x **Quanto?**: Quantidade de exemplares coletados por ano em cada país presente na base de dados, colorido pelo nível taxonômico de famílias. O gráfico exibido refere-se à coleção de poliquetas, do departamento de Polychaeta do Museu Nacional (MNRJ).

Os exemplos mostrados nas figuras 48 e 49 servem o propósito de ilustrar como, a partir da representação matricial, pode-se agrupar conjuntos de pontos a partir de uma característica em comum, podendo favorecer o surgimento de padrões visuais.

Figura 49 – Evolução temporal por Estados brasileiros da coleção de poliquetas (MNRJ).



Fonte: Elaboração própria.

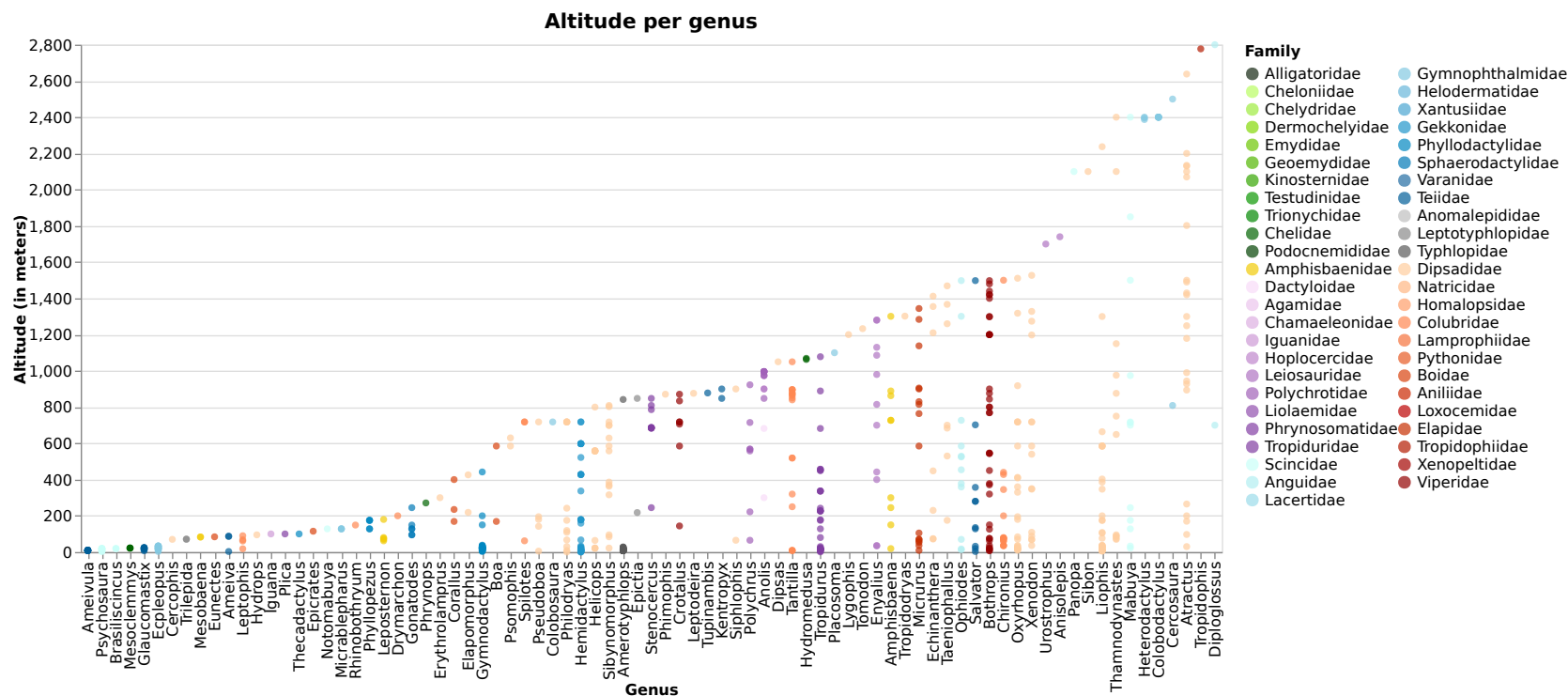
Nota: **Onde?** x **Quando?** x **O quê?** x **Quanto?**: Quantidade de exemplares coletados por ano em estados brasileiros, colorido pelo nível taxonômico de famílias. O gráfico exibido refere-se à coleção de poliquetas, do departamento de Polychaeta do Museu Nacional (MNRJ).

4.4.1.2 Dados de altitude e profundidade

Conforme mencionado previamente, a posição em que um exemplar foi coletado em relação ao nível do mar também é um fator importante à sua identificação, uma vez que releva um conjunto de possíveis características pertencente ao táxon ou ao ambiente no qual ele vive. Portanto, uma visualização capaz de expor a coleção de dados em termos de classificação taxonômica e altitude (ou profundidade) se faz relevante. Essa proposta é apresentada para a coleção de répteis na figura 50 que, por sua vez, são animais localizados acima do nível do mar.

Vale destacar que, mais uma vez, foi empregada a técnica de reordenamento, dessa vez no eixo nominal X, para que espécimes encontrados em localidades de altitude mais elevada fossem exibidos mais à direita no gráfico.

Figura 50 – Altitude de coleta por gênero para a coleção de répteis (MNRJ).

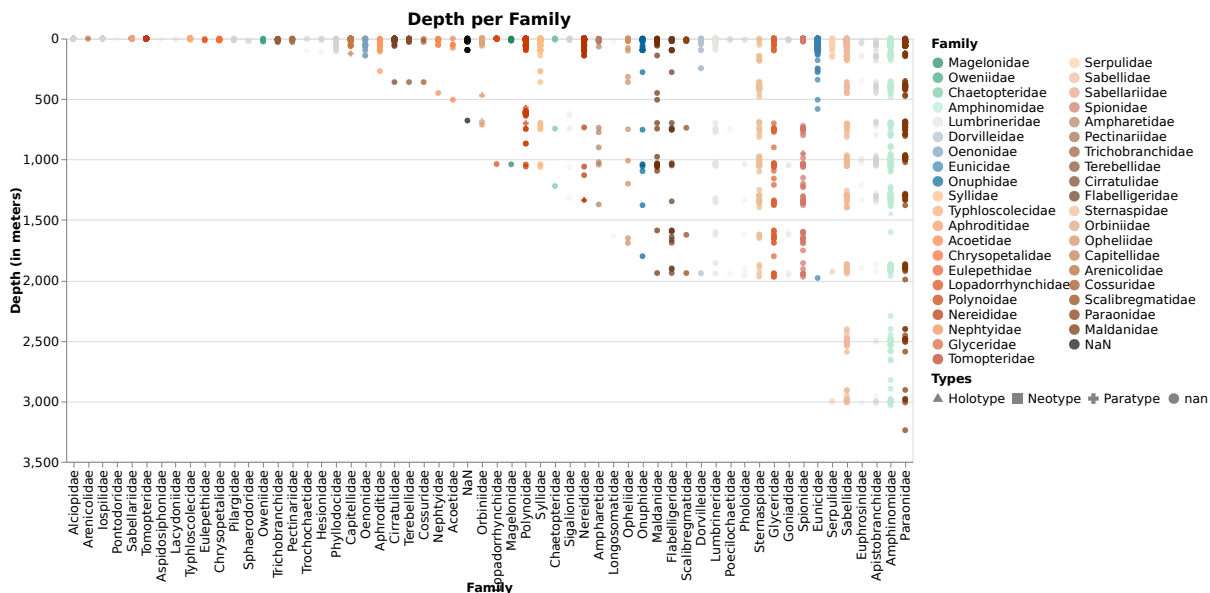


Fonte: Elaboração própria.

Nota: **Onde? x O quê?**: Altitude, em metros, na qual os espécimes foram coletados, distinguidos por gênero (eixo X) e coloridos por família. O gráfico exibido refere-se à coleção de répteis, pertencente ao departamento de Herpetologia do Museu Nacional (MNRJ).

De maneira análoga, a figura 51 apresenta a informação de profundidade em que os espécimes foram capturados para a coleção de poliquetas, que contém animais exclusivamente marinhos. O eixo X foi reordenado para que espécimes encontrados em regiões mais profundas fossem apresentados mais à direita.

Figura 51 – Profundidade de coleta por famílias para a coleção de poliquetas (MNRJ).



Fonte: Elaboração própria.

Nota: **Onde? x O quê?**: Profundidade na qual os espécimes foram encontrados, distinguível por famílias (eixo X e cor). O gráfico exibido refere-se à coleção de poliquetas, do departamento de Polychaeta do Museu Nacional (MNRJ).

Vale destacar que, se a coleção é composta por exemplares aquáticos e terrestres, altitude e profundidade podem ser exibidas no mesmo eixo, desde seja explicitada a convenção utilizada, i.e., que fique claro para o leitor que altitudes negativas significam medidas abaixo do nível do mar e, analogamente, profundidades negativas representam posições acima do nível do mar.

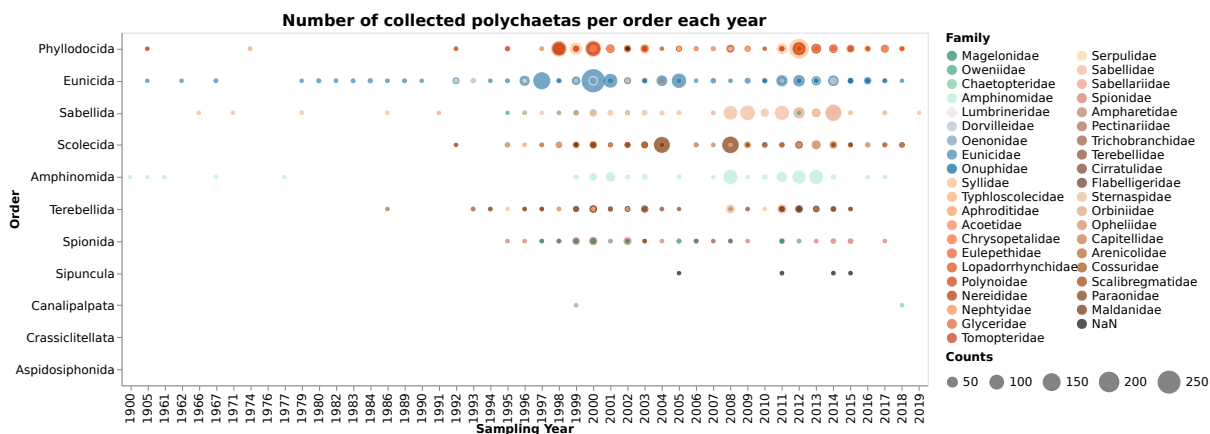
4.4.2 Cobertura taxonômica

Nesta seção serão apresentadas as propostas gráficas que visam explicitar visualmente a cobertura taxonômica das coleções biológicas em níveis selecionados por especialistas. Esse tipo de informação está diretamente associado à pergunta: *O que está presente na coleção?* (ou *Quais espécies são retratadas na base de dados?*). A pergunta essencial "**O quê?**", principal foco desta seção, pode ser respondida pela correspondência aos campos que contém informações de classificação taxonômica dos espécimes, como Reino, Filo, Ordem, Família, Gênero, Espécie, dentre outros³⁰. A figura 52 exibe uma

³⁰ Os campos correspondentes ao sistema moderno de classificação biológica são ilustrados na figura 27, na seção 2.2.3 do capítulo 2.

proposta gráfica construída para exibir o volume de espécimes coletados a cada ano para diferentes Ordens.

Figura 52 – Volume de espécimes coletados por ano para diferentes Ordens da coleção de poliquetas (MNRJ).



Fonte: Elaboração própria.

Nota: **O quê? x Quando? x Quanto?**: Quantidade de espécimes coletados a cada ano em diferentes Ordens, colorido pelo nível taxonômico de famílias. O gráfico exibido refere-se à coleção de poliquetas, do departamento de Polychaeta do Museu Nacional (MNRJ).

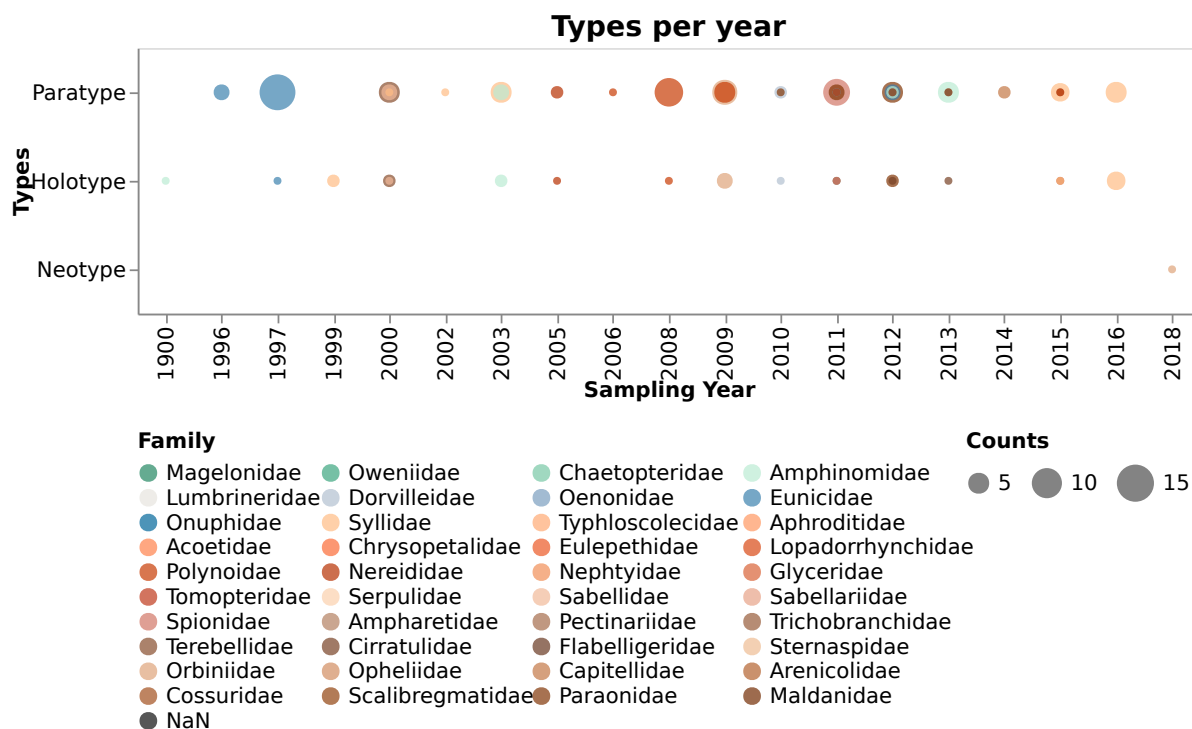
Além das desvantagens mencionadas anteriormente para esse tipo de abordagem gráfica, como a presença de pontos sobrepostos e visão limitada o *overview* da coleção, e das vantagens, como possibilitar o emprego das técnicas desenvolvidas por Bertin (1983) e facilitar a exploração individual ou comparação direta entre famílias, mais um ponto positivo pode ser apontado dado que, ao representar níveis taxonômicos superiores ao nível de Família no eixo Y, a partir da aplicação de filtros, espera-se que uma única família pertença a uma, e apenas uma, ordem³¹. Esse tipo de visualização expõe claramente a ocorrência desse tipo de padrão que, quando observado, demanda uma análise detalhada dos registros nas bases de dados a fim de verificar a presença de possíveis inconsistências.

Ainda, analisar diferentes níveis hierárquicos se revela uma tarefa extremamente simples nesse tipo de abordagem, bastando trocar a informação que se deseja codificar no eixo Y. Adicionalmente, deve-se apontar que, conforme observado por Qu e Hullman (2016), ao se trabalhar com um conjunto de visualizações, mudanças de codificação nos eixos X e Y são mais facilmente percebidas pelos leitores que, por sua vez, apresentam uma pré-disposição à leitura de anotações (rótulos, títulos de eixo, etc.) para entender o que foi alterado de uma visualização para outra, configurando um tipo de mudança que demanda menos carga cognitiva, contrastando com mudanças nos canais de codificação de cor, forma e tamanho.

³¹ Considerando a estrutura da árvore taxonômica, várias famílias podem derivar de uma mesma Ordem, mas uma única família não pode ser derivada de mais de uma ordem.

Com a finalidade de demonstrar como a interpretação da figura 52 é pouco modificada ao mudar o nível taxonômico codificado no eixo Y, a figura 53 traz essa mesma abordagem, também para a coleção de poliquetas do Museu Nacional (MNRJ), porém considerando apenas informações de material-tipo³².

Figura 53 – Materiais-tipo coletados por ano para a coleção de poliquetas (MNRJ).



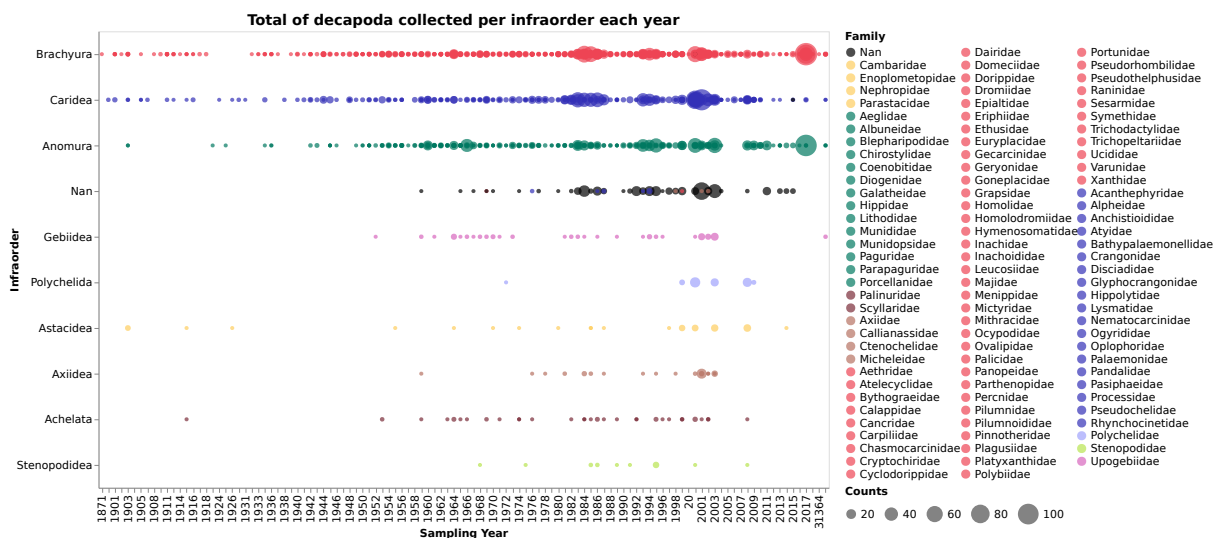
Fonte: Elaboração própria.

Nota: **O quê? x Quando? x Quanto?**: Quantidade de espécimes classificados como material-tipo coletados a cada ano, colorido pelo nível taxonômico de famílias. O gráfico exibido refere-se à coleção de poliquetas, do departamento de Polychaeta do Museu Nacional (MNRJ). Comparando com a figura 52, além do eixo Y, foram alterados apenas a escala de tamanhos (contagens) e a posição da legenda de cores.

Não obstante, até mesmo quando se altera a coleção de dados, salvo mudanças nos esquemas de cores e codificação de tamanho, uma vez que o leitor tenha sido apresentado previamente à essa proposta gráfica, a adaptação a mudanças de maior escala é facilitada. Para induzir esse experimento ao leitor, a figura 54 exhibe essa mesma abordagem, agora para o nível de Infraordens e considerando a coleção de crustáceos do Museu Nacional (MNRJ).

³² Um espécime é considerado material-tipo se, basicamente, é responsável pela nomenclatura de um táxon. Essa classificação é apresentada em detalhes no Código Internacional de Nomenclatura Zoológica (ICZN), disponível em <<https://www.iczn.org/the-code/the-international-code-of-zoological-nomenclature/the-code-online/>> - acessado em 23/02/2021.

Figura 54 – Volume de espécimes coletados por ano para diferentes Infraordens da coleção de crustáceos (MNRJ).

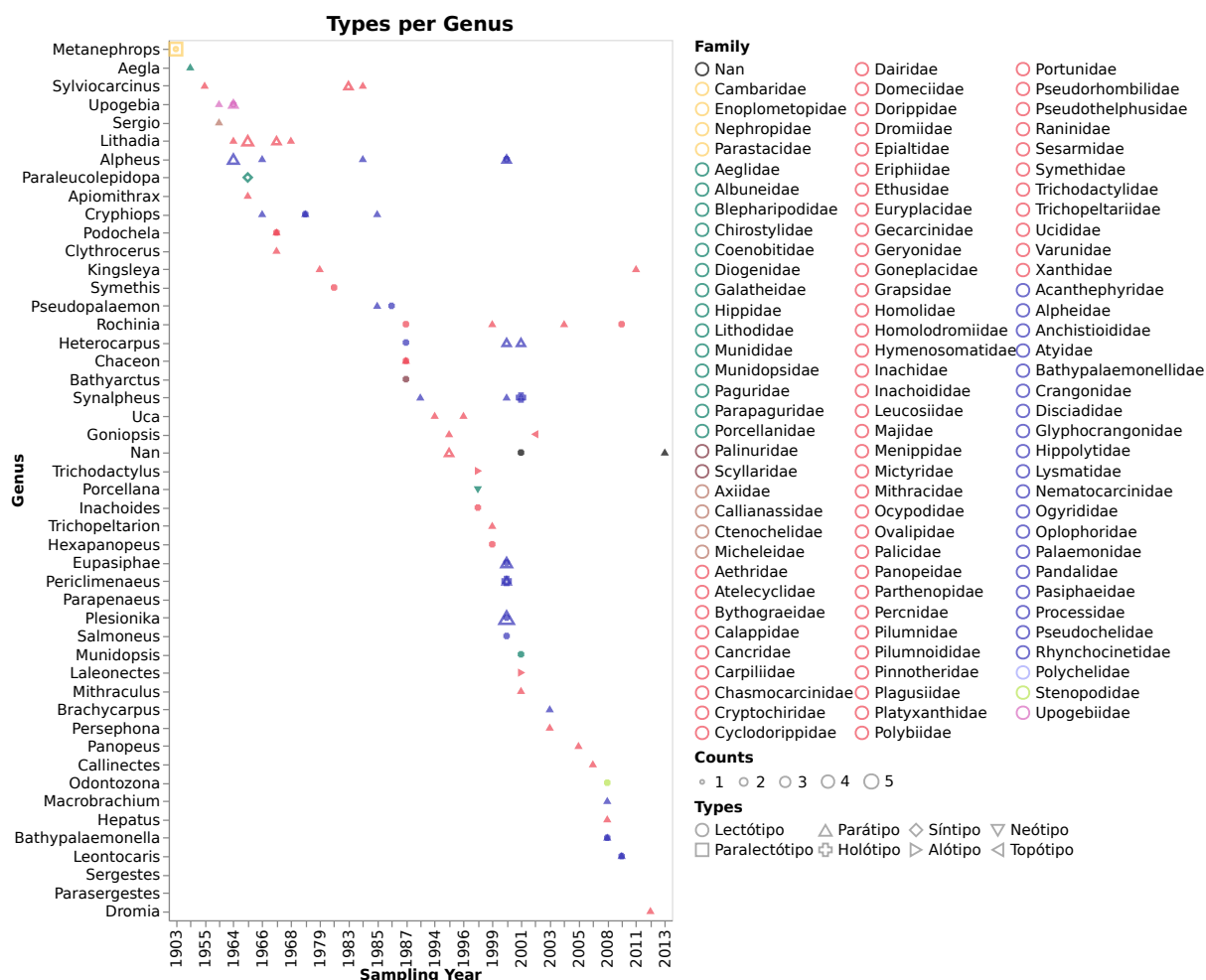


Fonte: Elaboração própria.

Nota: **O quê? x Quando? x Quanto?**: Quantidade de espécimes coletados a cada ano em diferentes Infraordens, colorido pelo nível taxonômico de famílias. O gráfico exibido refere-se à coleção de carcinos (crustáceos), do departamento de Carcinologia do Museu Nacional (MNRJ). Na figura, são contabilizados apenas espécimes da Ordem Decapoda, cujos registros foram revisados até a conclusão deste trabalho.

Por fim, devido à raridade da presença de materiais-tipo presente em coleções de biodiversidade, surgiu a necessidade de se explorar os diferentes tipos encontrados nas coleções aqui trabalhadas de forma a retratar essa riqueza taxonômica. Para isso, os tipos foram atribuídos às diferentes formas (*shapes*), resultando na proposta apresentada na figura 55.

Figura 55 – Materiais-tipo coletados por ano, distinguidos por gênero, para a coleção de crustáceos (MNRJ).



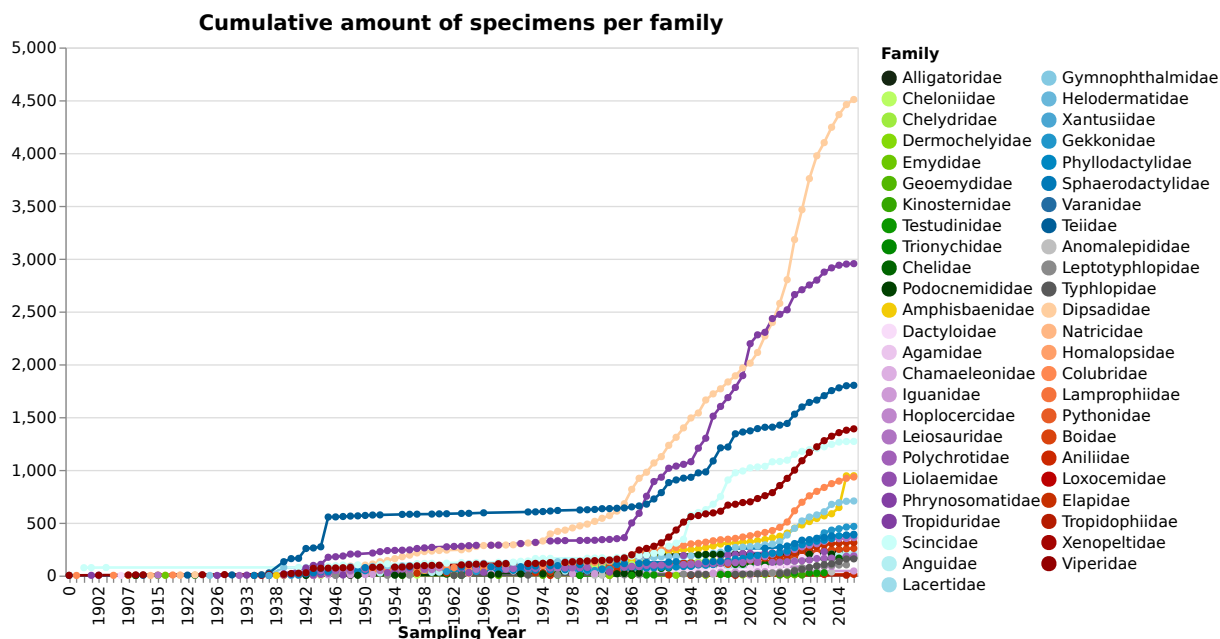
Fonte: Elaboração própria.

Nota: **O quê? x Quando? x Quanto?**: Quantidade de espécimes coletados a cada ano em diferentes Gêneros, colorido pelo nível taxonômico de famílias, e tipos representados em diferentes formas. O gráfico exibido refere-se à coleção de carcinos (crustáceos), do departamento de Carcinologia do Museu Nacional (MNRJ). Na figura, são contabilizados apenas espécimes da Ordem Decapoda, cujos registros foram revisados até a conclusão deste trabalho.

4.4.3 Cobertura temporal

Esta seção destina-se a apresentar o conjunto de gráficos criados com enfoque primordial na dimensão temporal, i.e., designados a responder à pergunta essencial "**Quando?**". A abordagem escolhida para representar esse tipo de informação busca evidenciar evoluções temporais tendo em vista a contribuição acumulada (ou total até certo período) de cada elemento para o crescimento da coleção em análise. Por esse motivo, o eixo Y codifica agregações quantitativas calculadas sobre os elementos selecionados. Para melhor compreender esse conceito, a figura 56 ilustra a quantidade acumulada de espécimes na coleção, coloridos por família (a mesma abordagem de mapeamento de cores usada nos gráficos exibidos até então), para a coleção de répteis do Museu Nacional (MNRJ).

Figura 56 – Crescimento acumulado da quantidade de espécimes por família na coleção de répteis (MNRJ).



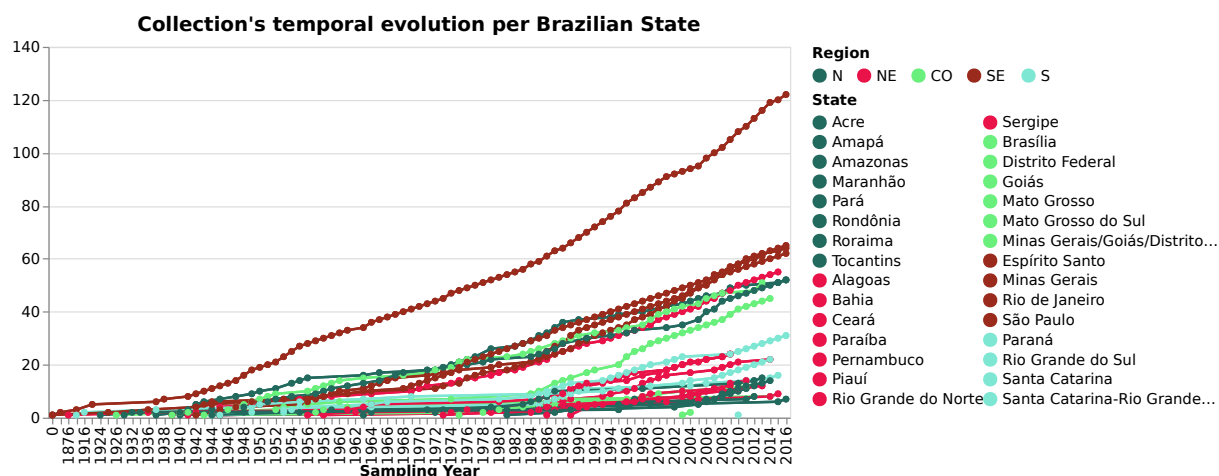
Fonte: Elaboração própria.

Nota: **Quando?** x **O quê?** x **Quanto?**: Quantidade acumulada de espécimes coletados por família. O gráfico exibido refere-se à coleção de répteis, do departamento de Herpetologia do Museu Nacional (MNRJ).

Além de "**Quando?**", também busca-se responder, com essa abordagem, *O quê foi coletado?* e *Quanto?*. No entanto, em relação aos demais gráficos, deve-se ressaltar os seguintes pontos em relação à essa proposta gráfica:

1. Assim como as demais visualizações apresentadas nesse capítulo, a análise visual desse gráfico é enriquecida pela presença de elementos de interatividade que permitem, por exemplo, a visualização da trajetória temporal de apenas uma família, ou a comparação direta de um subconjunto das categorias de cores;
2. No eixo Y, é possível visualizar o total de espécimes coletados até um dado ano. A quantidade de espécimes coletadas exclusivamente naquele ano é exibida na *tooltip*;
3. A distinção dos táxons é deixada completamente a cargo do mapeamento de cores. Isso implica que, representar outras informações que não as famílias acarreta na escolha de um novo mapa de cores, o que pode comprometer a consistência global do conjunto de visualizações propostas, conforme apontado por Qu e Hullman (2016). Por esse motivo, recomenda-se que essas duas propostas distintas não sejam apresentadas lado a lado quando houver mudanças no esquema de cores, por repousar sobre o leitor uma maior carga cognitiva.

Figura 58 – Crescimento acumulado da quantidade de espécimes por Estados brasileiros na coleção de répteis (MNRJ).



Fonte: Elaboração própria.

Nota: **Quando?** x **Onde?** x **Quanto?**: Quantidade acumulada de espécimes coletados por Estados brasileiros. Há mais de um seletor disponível, a nível de região e unidades da federação brasileira. Ao clicar em uma região, apenas as curvas referentes aos Estados daquela região são exibidas. Ao clicar em um Estado, especificamente, é dada maior ênfase à sua respectiva curva, fazendo uso do paradigma frente e fundo. O gráfico exibido refere-se à coleção de répteis, do departamento de Herpetologia do Museu Nacional (MNRJ).

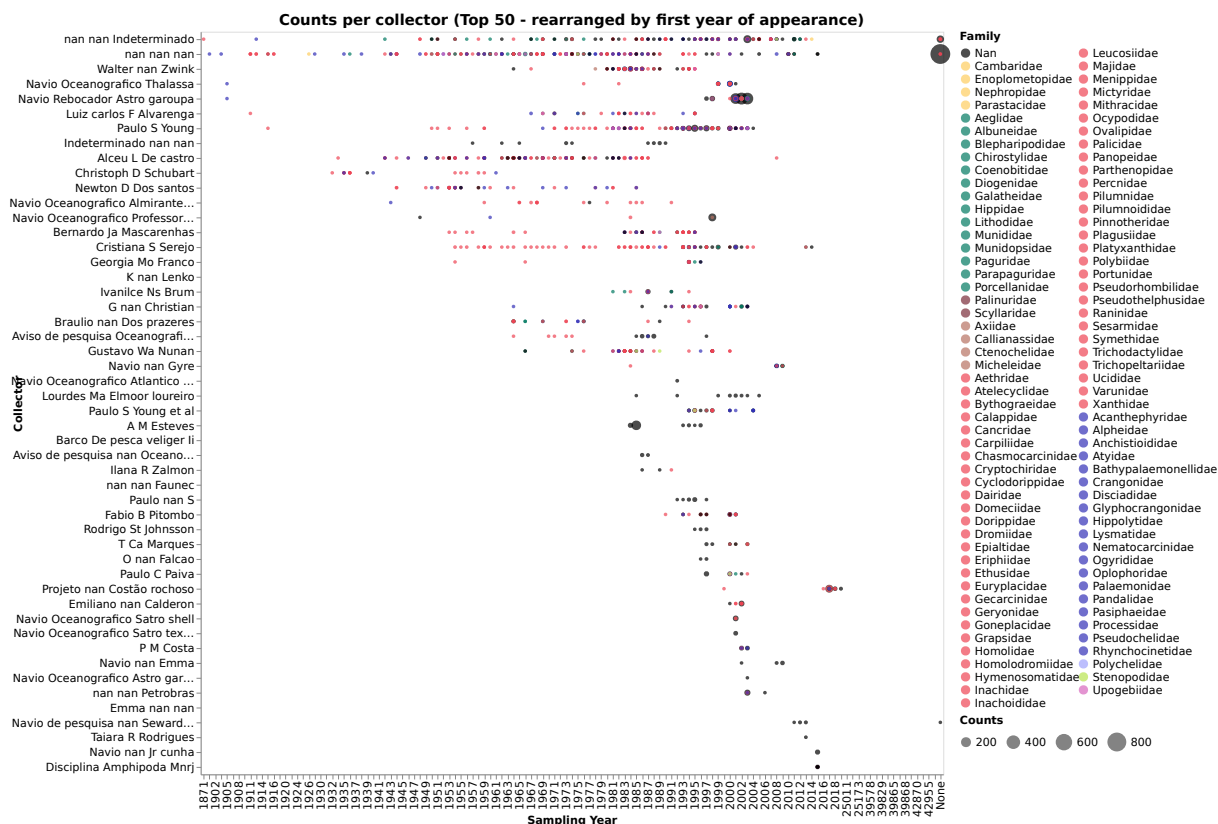
4.4.4 Contribuição de indivíduos para as coleções

Por fim, o foco no **"Quem?"** configura uma abordagem inédita explorada neste trabalho. Nesse contexto, o principal interesse está em visualizar a contribuição dos indivíduos responsáveis pelas diferentes etapas do ciclo de registro de um novo espécime na coleção. Especificamente, é explorada a contribuição dos responsáveis pela coleta e identificação de cada exemplar.

Não obstante, a figura 59 revisita a proposta gráfica matricial para expor a quantidade de espécimes de diferentes famílias coletados por cada indivíduo responsável registrado na base de dados de crustáceos (MNRJ), identificados pelo seu nome completo³⁴, codificado no eixo Y. Por vezes, o campo associado ao nome dos responsáveis pela coleta dos espécimes contém o nome da expedição que viabilizou a captura do exemplar em questão. Ainda, os nomes são apresentados segundo o primeiro ano em que o referido coletor (ou expedição) apresentou seu primeiro registro na base de dados, i.e., coletores mais antigos são expostos no topo, enquanto mais abaixo localizam-se àqueles que iniciaram suas contribuições em períodos mais recentes.

³⁴ Primeiro nome, nome do meio e último nome, quando disponíveis. Para as bases de dados de répteis e poliquetas, apenas o primeiro e o último nome eram apresentados.

Figura 59 – Quantidade de espécimes capturados pelos 50 coletores mais expressivos, distinguidos por família, para a coleção de crustáceos (MNRJ).



Fonte: Elaboração própria.

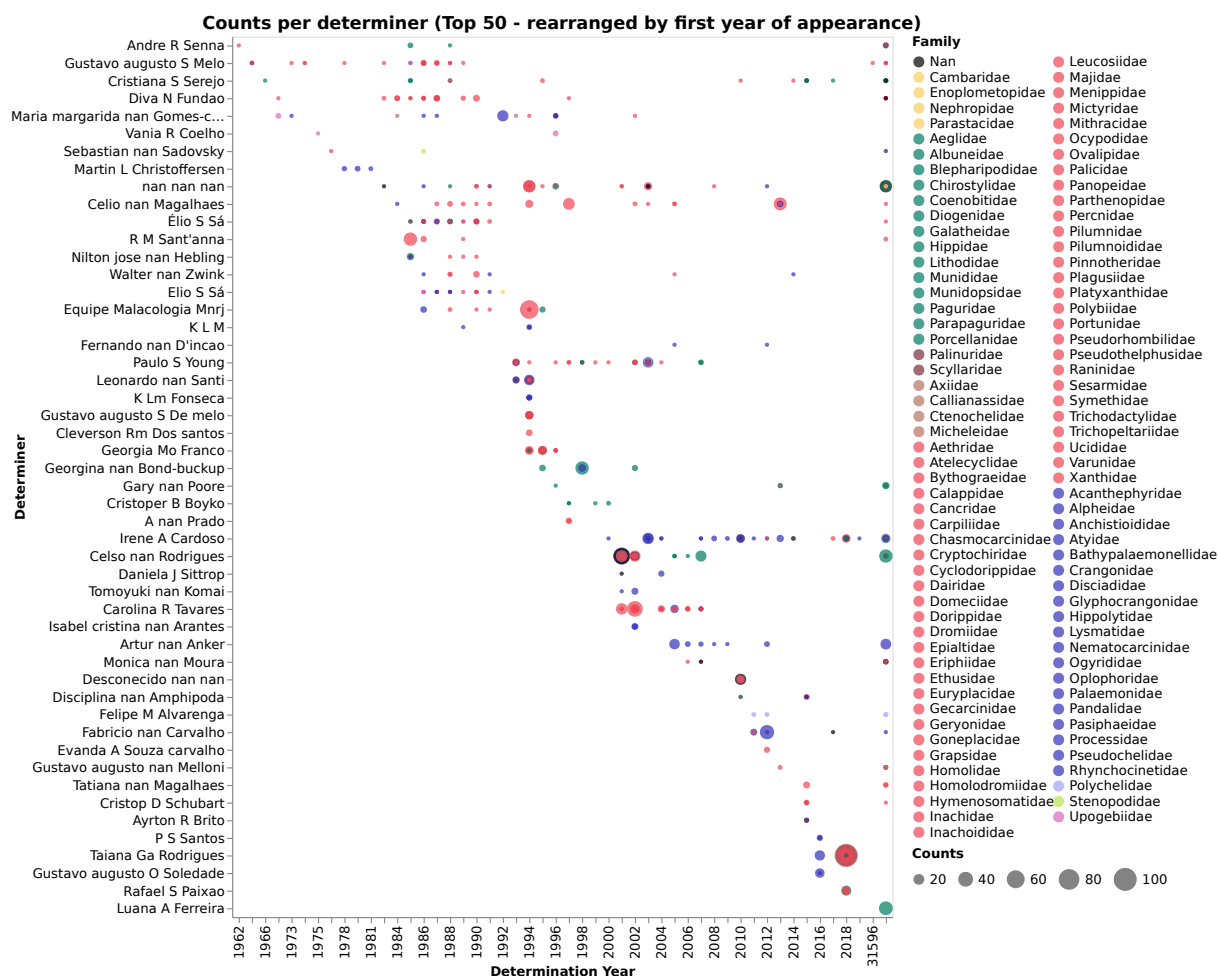
Nota: **Quem?** x **Quando?** x **O quê?** x **Quanto?**: Quantidade de espécimes coletados a cada ano por coletor, colorido pelo nível taxonômico de famílias. Devido à grande quantidade de coletores presente na base de dados, são apresentados apenas os 50 mais expressivos (i.e., com maiores volumes associados de coleta). O gráfico exibido refere-se à coleção de carcinos (crustáceos), do departamento de Carcinologia do Museu Nacional (MNRJ). Na figura, são contabilizados apenas espécimes da Ordem Decapoda, cujos registros foram revisados até a conclusão deste trabalho.

As vantagens e desvantagens dessa abordagem já foram discutidas previamente, nas subseções 4.4.1, 4.4.2 e 4.4.3. Aqui cabe apenas o comentário pertinente de que, por meio da seleção de diferentes categorias da legenda de cores, pode-se observar a especialidade de cada indivíduo, ou da expedição, no que se refere às famílias mais observadas em sua linha. Ainda, é importante citar a dificuldade em associar um indivíduo à todas as suas contribuições de fato, uma vez que seu nome, em alguns casos, é reportado de formas diferentes, por exemplo, usando abreviações ou simplesmente devido a erros de digitação.

A mesma discussão se aplica ao caso de determinadores³⁵, cujo gráfico é exposto a seguir, na figura 60.

³⁵ Pesquisadores e curadores responsáveis pela identificação ou re-classificação taxonômica dos exemplares da coleção.

Figura 60 – Quantidade de espécimes identificados ou reclassificados pelos 50 determinadores mais expressivos, distinguidos por família, para a coleção de crustáceos (MNRJ).

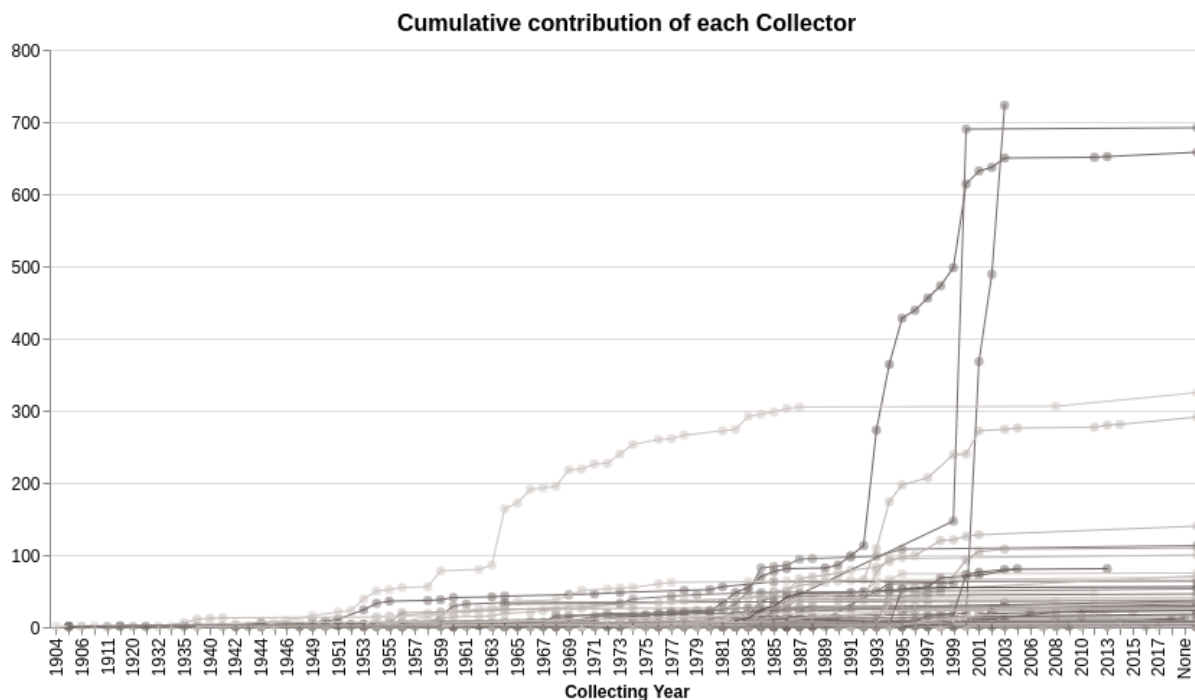


Fonte: Elaboração própria.

Nota: **Quem?** x **Quando?** x **O quê?** x **Quanto?**: Quantidade de espécimes identificados ou reclassificados a cada ano por determinador, colorido pelo nível taxonômico de famílias. Devido à grande quantidade de determinadores presente na base de dados, são apresentados apenas os 50 mais expressivos (i.e., com maiores volumes associados de classificação). O gráfico exibido refere-se à coleção de carcinos (crustáceos), do departamento de Carcinologia do Museu Nacional (MNRJ). Na figura, são contabilizados apenas espécimes da Ordem Decapoda, cujos registros foram revisados até a conclusão deste trabalho.

As abordagens gráficas trazidas nas figuras 59 e 60 expõem a contribuição pontual de cada indivíduo para a composição da coleção em ênfase. Para sanar a curiosidade de se observar a contribuição total (acumulada) de cada indivíduo para a coleção, foram criadas as representações nas figuras 61 e 62.

Figura 61 – Contribuição acumulada de coletores para a coleção de crustáceos (MNRJ).



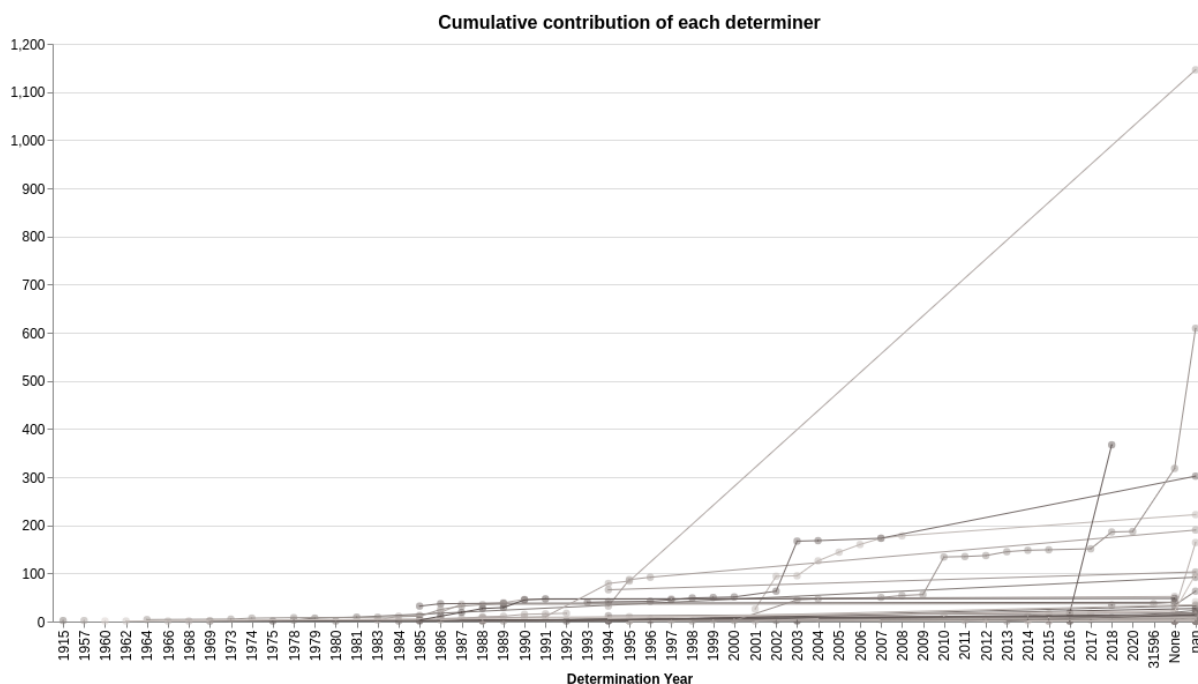
Fonte: Elaboração própria.

Nota: **Quem?** x **Quando?** x **Quanto?**: Quantidade acumulada de espécimes coletados por cada indivíduo. Informações específicas à cada coletor (**Quem?**) são exibidas ao aproximar o mouse da sua respectiva curva, enfatizando sua trajetória pelo paradigma de frente e fundo. Deve-se atentar para a presença de anos inconsistentes (localizados ao final do eixo X, mais à direita), o que pode comprometer a interpretação desse gráfico. O gráfico exibido refere-se à coleção crustáceos (carcinós), do departamento de Carcinologia do Museu Nacional (MNRJ). Foram considerados apenas registros da ordem Decapoda, revisados pela equipe responsável até a finalização deste estudo.

Deve-se apontar, no entanto, que a escolha de um mapeamento de cores, nesses casos, revelou-se bastante desafiadora e foge ao escopo abordado na seção 4.2.1. Por esse motivo, optou-se por representar todos os indivíduos em cores neutras (tons de cinza) e permitir a exibição de informações específicas por meio da interatividade. Ao aproximar o mouse de uma das curvas, toda a trajetória é realçada empregando-se o paradigma de frente e fundo, como definido em (KIRK, 2016), e informações pontuais são exibidas na *tooltip* de cada ponto³⁶.

³⁶ Por uma limitação da ferramenta Altair, não conseguiu-se implementar o reconhecimento da proximidade do mouse à toda a curva, mas sim em relação aos pontos (vértices) os quais ela liga.

Figura 62 – Contribuição acumulada de determinadores para a coleção de crustáceos (MNRJ).



Fonte: Elaboração própria.

Nota: **Quem?** x **Quando?** x **Quanto?**: Quantidade acumulada de espécimes identificados ou reclassificados por cada determinador. Informações específicas à cada determinador (**Quem?**) são exibidas ao aproximar o mouse da sua respectiva curva, enfatizando sua trajetória pelo paradigma de frente e fundo. Deve-se atentar para a presença de anos inconsistentes (localizados ao final do eixo X, mais à direita), o que pode comprometer a interpretação desse gráfico. O gráfico exibido refere-se à coleção de crustáceos (carcinos), do departamento de Carcinologia do Museu Nacional (MNRJ). Foram considerados apenas registros da ordem Decapoda, revisados pela equipe responsável até a finalização deste estudo.

5 Análise de resultados

O capítulo 4 apresentou todos os elementos necessários à construção de um *framework* visual para auxiliar pesquisadores e curadores em suas tarefas diárias no que se refere à verificar, identificar e corrigir possíveis inconsistências em suas coleções científicas biológicas digitais. Mais especificamente, a seção 4.4 apresentou um conjunto de gráficos proposto nesse estudo para servir como tal ferramenta de visualização. Isso posto, este capítulo apresenta, na seção 5.1, exemplos de usabilidade com o intuito de expor como as soluções gráficas propostas cumpriram o seu propósito em evidenciar deficiências qualitativas dos dados, considerando sua aplicação em três coleções biológicas digitais cedidas pela parceria com o Museu Nacional (MNRJ): crustáceos, répteis e poliquetas¹. A seção 5.2 finaliza esse capítulo comentando a respeito de novas formas de exploração visual disponíveis para usufruto dos pesquisadores por meio de elementos de interatividade, i.e., como a aplicação de filtros e seleções pode destacar padrões e criar novos gráficos a partir das soluções já expostas.

5.1 Visualização como ferramenta de apoio aos curadores

Conforme citado previamente, o principal objetivo do presente trabalho é propor um *framework* constituído pela aplicação de um conjunto de técnicas, recomendações e boas práticas da área de InfoVis com o objetivo de guiar a construção de propostas gráficas capazes de evidenciar aspectos associados à qualidade dos dados em coleções científicas biológicas digitais. O público-alvo desta aplicação são especialistas de domínio, leia-se pesquisadores e curadores de tais bases de dados.

Neste sentido, este estudo contou com a colaboração de três grupos de especialistas ligados ao Museu Nacional (MNRJ)² em todas as etapas do seu desenvolvimento. Tal colaboração envolveu a concessão do acesso à três coleções de dados e a condução de reuniões periódicas nas quais novas propostas visuais eram apresentadas às equipes, coletadas sugestões para melhorias e construção de novos gráficos, além de receber *feedback* direto das equipes a respeito do quanto os recursos visuais lhes foram úteis como ferramentas de auxílio para melhoria de seus conjuntos de dados.

A seguir, são apresentados três exemplos de como os recursos visuais podem auxiliar na detecção de inconsistências nos dados. Assim, a figura 63 ilustra como se dá a detecção de possíveis erros analisando-se as anotações dos eixos X e Y.

¹ Pertencentes aos departamentos de Carcinologia, Herpetologia e Polychaeta, respectivamente.

² Integrantes responsáveis pela manutenção diária das coleções pertencentes aos setores de Carcinologia, Herpetologia e Polychaeta.

é o valor que o intérprete Python³ associa à células vazias na planilha de dados. Ao clicar nesse seletor, então, é possível exibir todos os pontos correspondentes à registros cujas famílias não foram devidamente anotadas.

Fugindo às formas mais óbvias de se identificar possíveis erros - por meio de anotações gráficas - o pesquisador precisa estar atento à identificação de padrões anômalos nos dados que, apesar de exigirem maior nível de atenção para sua detecção, podem ser realçados visualmente. Um exemplo de um padrão desse tipo - que demanda investigação mais profunda para se verificar se os registros estão, de fato, corretos - é exposto na figura 64. Nessa figura, em particular, são visualizadas as contribuições pontuais de cada determinador em identificar espécimes a cada ano. Uma observação importante, nesse caso, é que o eixo X representa o ano de determinação dos espécimes, possibilitando empregar o raciocínio que virá a seguir.

Na figura 64, ao representar os dados dessa forma, cada linha pode ser interpretada como a exposição de todas as contribuições feitas ao longo da carreira profissional de cada pesquisador. Conforme destacado nessa mesma figura, padrões anômalos podem surgir à medida que são indicados períodos de carreira muito extensos, por exemplo, superiores à 100 anos de dedicação. Ainda, a partir de conhecimentos prévios a respeito do trabalho de cada pesquisador, especialistas podem prontamente apontar períodos inconsistentes por recaírem fora do seu período de contribuição. Nesse exemplo, a equipe de especialistas do departamento de Polychaeta do Museu Nacional (MNRJ) apontou as contribuições referentes ao ano de 1905 como passíveis de maiores investigações por meio da análise desta visualização⁴.

Deve-se ainda destacar que, a aplicação de critérios relevantes para a reorganização das informações em eixos nominais, como sugerido por Bertin (1983), pode facilitar a evidência de padrões nos dados passíveis de maior suspeita, uma vez empregado esse tipo de proposta de visualização matricial.

³ Esse padrão acontece pois trabalhou-se com as bibliotecas *Pandas* e *Numpy* para manuseio de dados com Python.

⁴ O ano 0 foi identificado prontamente por estar anotado no eixo X.

Figura 64 – Identificando padrões anômalos nos dados - exemplo para a coleção de poliquetas (MNRJ).

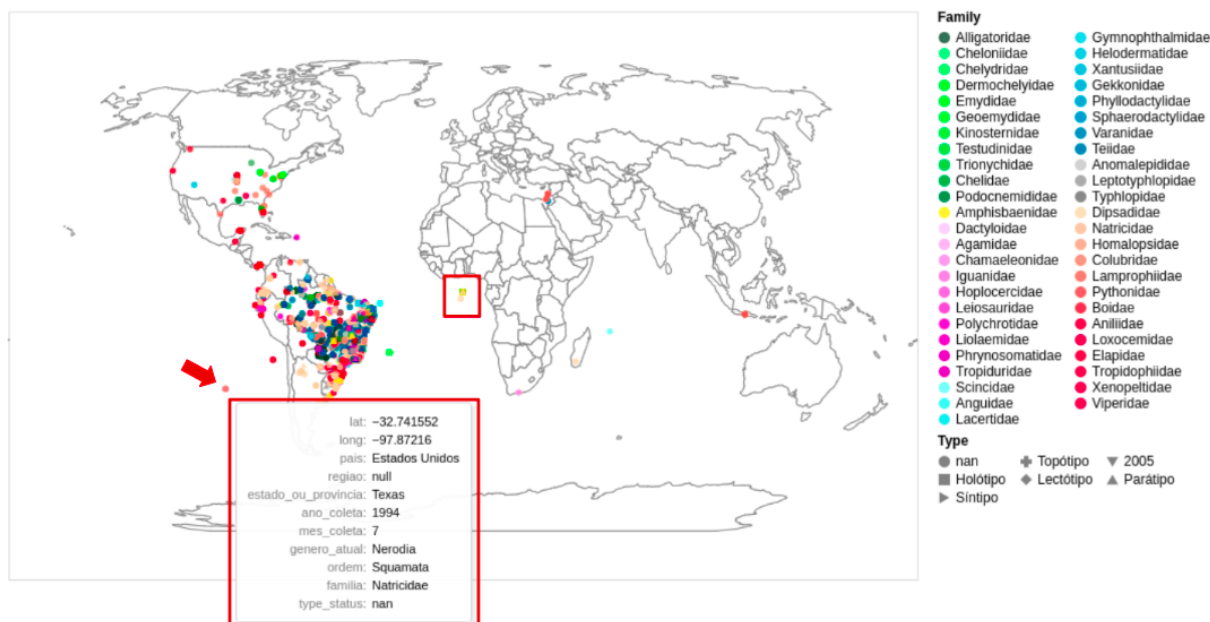


Fonte: Elaboração própria.

Nota: **Quem?** x **Quando?** x **O quê?** x **Quanto?**: Quantidade de espécimes identificados ou reclassificados a cada ano por determinador, colorido pelo nível taxonômico de famílias. Devido à grande quantidade de determinadores presente na base dedados, são apresentados apenas os 50 mais expressivos (i.e., com maiores volumes associados de classificação). O gráfico exibido refere-se à coleção de poliquetas, do departamento de Polychaeta do Museu Nacional (MNRJ).

Por fim, uma vez identificados padrões visuais de interesse, ou um conjunto de pontos para investigação mais detalhada, informações específicas a cada ponto podem ser obtidas por meio do elemento interativo *tooltip*, conforme exibido na figura 65. Aqui, fica claro a concordância com o mantra "*overview first, zoom and filter, then details on demand*" (SHNEIDERMAN, 2003). Nesse exemplo, por se tratar da coleção de répteis - que são animais predominantemente terrestres - suspeita-se de pontos marcados sobre o oceano, demandando uma investigação mais detalhada a fim de verificar, por exemplo, se tal registro aponta para uma ilha. Para tal, as informações detalhadas na *tooltip* se mostram de fundamental importância.

Figura 65 – Identificando inconsistências por meio da *tooltip* - exemplo para a coleção de répteis (MNRJ).



Fonte: Elaboração própria.

Nota: **O quê?** x **Onde?**: Exemplo de incongruência confirmada pelo uso das informações específicas apontadas na *tooltip*. O gráfico representa as coordenadas de latitude e longitude do ponto de coleta de cada um dos espécimes presentes na coleção de répteis do departamento de Herpetologia do Museu Nacional (MNRJ). Mapeamento de cores por famílias e forma (*shape*) designado para material-tipo.

Na figura 65, é fácil ver que, o espécime teoricamente localizado no Oceano Pacífico, na verdade, está demarcado como coletado nos Estados Unidos. Ainda, outro conjunto de pontos pode ser de particular interesse por constituir, exatamente, zero grau de latitude e longitude, isto é, registros que não apresentam coordenadas especificadas ou que as demarcam como nulas.

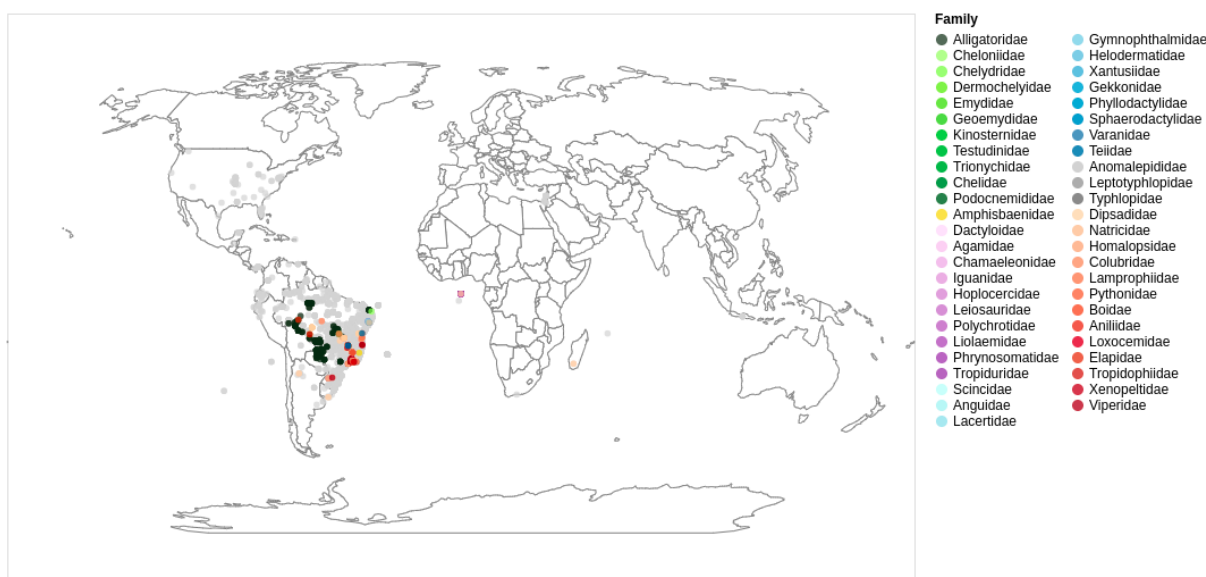
Em suma, os exemplos apresentados nessa seção - nas figuras 63, 64 e 65 - têm o propósito de ilustrar possíveis casos de uso em que o conjunto de visualizações proposto pode apontar registros que necessitam de maior nível de verificação por parte dos curadores das coleções científicas digitais de biodiversidade.

5.2 Criação de novas visualizações a partir da interatividade

Esta breve seção tem como objetivo primordial ilustrar como o conjunto de visualizações, por meio de filtros e seleções, podem realçar determinados padrões de interesse para exploração em publicações das coleções de dados. Em outras palavras, é ilustrado o processo de criação de novos gráficos a partir do uso de elementos interativos presentes nas visualizações aqui propostas.

Não obstante, a figura 66 mostra como padrões de dados podem surgir a partir de mudanças em pequenas configurações das visualizações aqui propostas. Especificamente, pela aplicação de um pequeno filtro adicional para o ano de coleta, pode-se observar o padrão destacado para a família *Alligatoridae* que, por sua vez, está associado à realização de uma expedição cujo foco foi capturar espécimes desta família, no ano de 1987.

Figura 66 – Localização geográfica dos exemplares coletados em 1987 para a coleção de répteis (MNRJ).

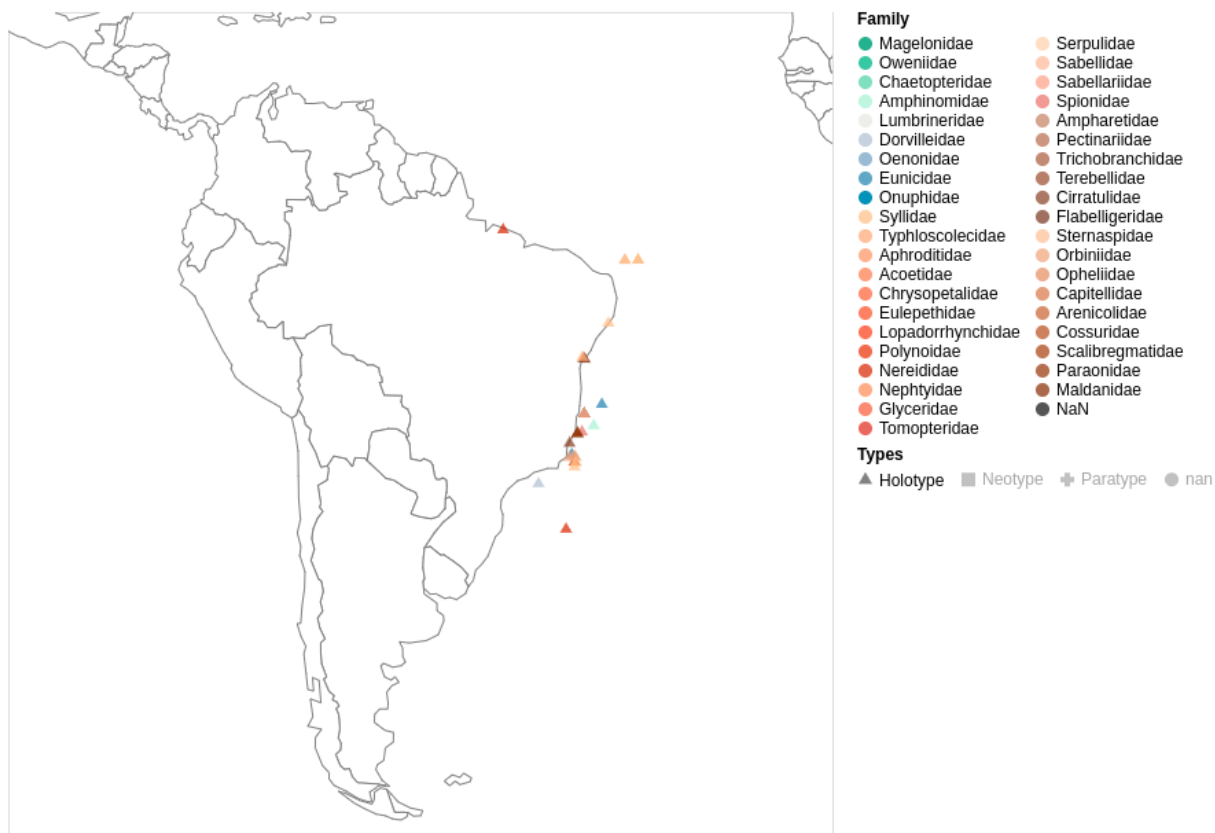


Fonte: Elaboração própria.

Nota: **Onde? x O quê?:** O gráfico representa as coordenadas de latitude e longitude do ponto de coleta de cada um dos espécimes, para o ano de 1987, presentes na coleção de répteis do departamento de Herpetologia do Museu Nacional (MNRJ). Mapeamento de cores por famílias.

Por fim, devido à raridade da presença de materiais-tipo em coleções científicas de biodiversidade, em especial para o tipo *holótipo* (que é o espécime diretamente responsável pela nomenclatura do táxon em questão), pode ser de interesse particular dos pesquisadores evidenciar a riqueza de holótipos em suas coleções por meio das propostas visuais aqui empregadas. Nesse sentido, as figuras 67 e 68 mostram como a aplicação de um filtro para espécimes desse tipo pode destacar suas presenças geográficas e sua distribuição em relação a níveis taxonômicos, respectivamente.

Figura 67 – Localização geográfica dos holótipos coletados na América Latina para a coleção de poliquetas (MNRJ).



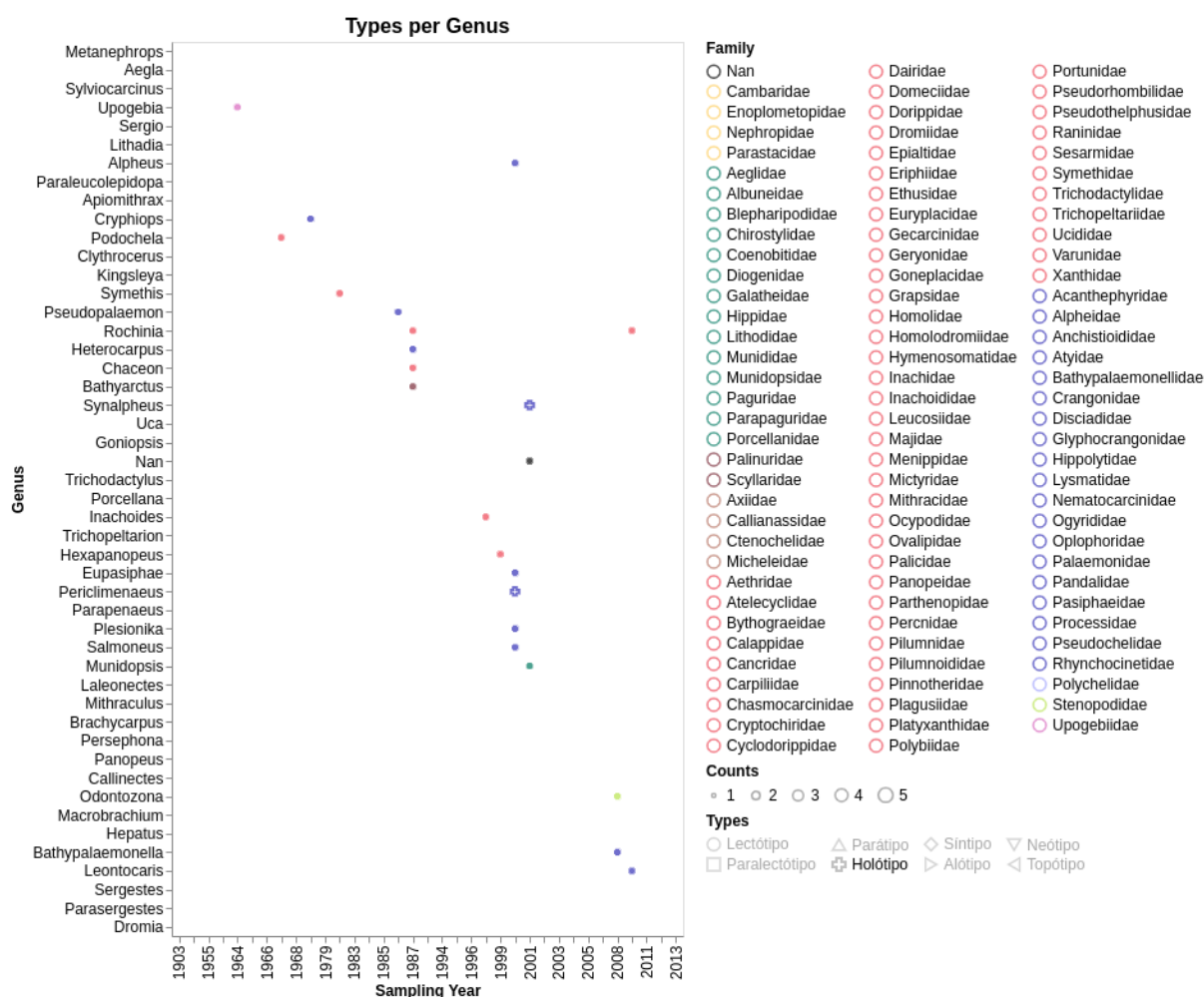
Fonte: Elaboração própria.

Nota: **Onde? x O quê?**: O gráfico representa as coordenadas de latitude e longitude do ponto de coleta de cada um dos espécimes do tipo holótipo, para a América Latina, presentes na coleção de poliquetas do departamento de Polychaeta do Museu Nacional (MNRJ). Mapeamento de cores por famílias e forma (*shape*) designado para material-tipo.

Os exemplos apresentados nessa seção - nas figuras 66, 67 e 68 - ilustram como o uso dos recursos de interatividade são encorajados para permitir novos níveis de exploração das bases de dados, podendo também servir para a criação de novas visualizações que oferecem maior ênfase a padrões desejados pelos pesquisadores. Tais visualizações podem ser empregadas em futuras publicações das coleções de dados em *data papers* de biodiversidade⁵, fornecendo diferentes perspectivas visuais em contraste com os recursos gráficos usualmente empregados em artigos dessa classe.

⁵ Documentos revisados pela comunidade científica que descrevem um conjunto particular de dados de biodiversidade.

Figura 68 – Distribuição dos holótipos por gênero na coleção de crustáceos (MNRJ).



Fonte: Elaboração própria.

Nota: **O quê? x Quando? x Quanto?**: Distribuição de holótipos coletados a cada ano para diferentes Gêneros, colorido pelo nível taxonômico de famílias. O gráfico exibido refere-se à coleção de carcinos (crustáceos), do departamento de Carcinologia do Museu Nacional (MNRJ). Na figura, são contabilizados apenas espécimes da Ordem Decapoda, cujos registros foram revisados até a conclusão deste trabalho.

6 Conclusão

Grandes instituições de história natural - Museus, Herbários, Universidades, Institutos de Pesquisa, dentre outras - têm um papel crucial uma vez que foram as maiores responsáveis documentar e armazenar dados de diferentes espécies levantados por pesquisadores taxonomistas ao longo do tempo. Registros primários de biodiversidade (do inglês *Primary Biodiversity Records* - PBR) contém informações como data e local de coleta, identificação taxonômica, o taxonomista responsável pela classificação daquele exemplar, dentre outras.

Ainda, tais coleções de dados apresentam registros precisos que podem ser usados para reconstruir o histórico de diferentes espécies em uma extensão que vai além do tempo de carreira de um único pesquisador (DREW, 2011). Várias décadas de esforço em catalogar e armazenar estas informações deram origem a grandes coleções de dados que dão suporte a inúmeras pesquisas em diferentes áreas do conhecimento biológico. Iniciativas de digitalização, ainda em curso, objetivaram a disponibilização dessas coleções científicas em ambientes *online*.

No entanto, garantir a qualidade desses registros é uma preocupação presente na literatura, discutida desde o trabalho de Thomas e Mintz (1998). Fatores como o grande volume de dados e a interdependência entre múltiplas variáveis dificultam determinar até que ponto esses dados estão completos, corretos e se, de fato, fornecem uma boa cobertura geográfica, temporal e taxonômica das espécies em questão. Garantir a acurácia no registro de espécimes é uma tarefa que inicia-se no campo de coleta e perpetua-se até o momento em que aquele novo registro é gravado, sob a responsabilidade dos curadores, especialistas e equipes técnicas envolvidos. Porém, por motivos que vão desde a grande extensão temporal desses registros até dificuldades de acesso ao local de coleta, raridade de certas espécies e questões orçamentárias referentes à grandes expedições, a existência de uma ferramenta capaz de verificar e disponibilizar os dados existentes de forma a possibilitar a detecção de erros e incongruências é de importância direta.

Nesse sentido, o presente trabalho buscou aplicar princípios e técnicas advindos da literatura de Visualização da Informação ao contexto de coleções biológicas digitalizadas, visando fornecer um conjunto de técnicas e visualizações capazes de expor tais coleções de forma tal que a qualidade dos registros seja facilmente acessada pelo público especializado. À luz dos trabalhos de Broeck et al. (2005) e Liu et al. (2018), tais elementos foram agrupados em um *framework* de visualização para dados biológicos do tipo PBR, voltado para auxiliar na melhoria de qualidade dessas informações sob a supervisão de *experts*, em etapa prévia à sua publicação. Adicionalmente, é importante citar que o desenvolvimento

do *framework* foi executado em parceria com equipes de especialistas responsáveis pela manutenção de três coleções de dados (crustáceos, répteis e poliquetas), cujo acesso foi concedido por meio de uma parceria com o Museu Nacional¹ (MNRJ). Diversos ajustes nas bases foram feitos ao longo do processo, sempre buscando integrar as ferramentas visuais ao dia a dia dos curadores.

Entende-se, com esse estudo, que o potencial de utilização de técnicas de visualização de dados para o apoio à curadoria e verificação de registros em coleções científicas digitais de biodiversidade é confirmado na prática. Essas atividades já são parte do cotidiano de especialistas, que verificam suas bases de dados por meio de ferramentas tipicamente baseadas em interface textual (como OpenRefine²). Vale destacar que as ferramentas de representação visual aqui propostas não buscam competir com as demais soluções usadas com esse propósito, mas sim complementá-las ao adicionar a capacidade de revelar padrões e anomalias nos dados visualmente, apelando para diferentes aspectos da cognição humana.

Uma vez que possíveis incongruências sejam resolvidas, as visualizações propostas neste estudo podem servir ao propósito de apresentar a coleção à comunidade científica, em publicações do tipo *data papers*, destacando suas riquezas em termos de cobertura geográfica, temporal e entre diferentes níveis taxonômicos. Adicionalmente, o conjunto de gráficos elaborado pode ser exposto em ambientes web, contendo elementos dinâmicos e interativos capazes de elevar o grau de interação do usuário com a coleção, além de facilitar a geração de diferentes visualizações estáticas que podem ser facilmente publicadas em veículos de circulação impressa, como artigos científicos, buscando melhor atender ao objetivo de destacar padrões nos conjuntos de dados que sejam de interesse particular dos pesquisadores.

O presente trabalho contribuiu para a literatura de biodiversidade ao apontar um conjunto de técnicas e propostas gráficas sugeridas por estudos de visualização da informação que, portanto, têm o potencial de representar esse tipo de dado geoespacial de maneira compreensível à especialistas, além de possibilitar o uso de recursos visuais para identificar e corrigir possíveis inconsistências nos registros de localização nas coleções por seus mantenedores. Para a literatura de InfoVis, as contribuições deste estudo podem ser condensadas na reunião e aplicação de diferentes conceitos e técnicas visando identificar e sanar incongruências nos conjuntos de dados, além do contexto no qual estes foram empregados.

Adicionalmente, destacar a contribuição de cada pesquisador e coletor para a constituição da coleção ao longo do tempo foi uma abordagem inédita explorada neste estudo. Tais representações são comuns em outros contextos - como o de esportes³, por

¹ Mais especificamente, com seus departamentos de Carcinologia, Herpetologia e Polychaeta.

² <<https://openrefine.org/>>

³ Vide visualização intitulada "*The many types of sumo careers*", em Perin et al. (2018).

exemplo - mas não são conhecidas aplicações deste tipo de representação ao contexto de dados biológicos.

Finalmente, vale ressaltar que, apesar de o escopo deste trabalho estar delimitado à apresentação de bases de dados de biodiversidade, alguns de seus produtos podem ser facilmente ajustados para auxiliar no processo de comunicação e análise de outras bases de dados biológicas. Ainda, as técnicas aqui empregadas podem ser aplicadas em coleções científicas de diversas naturezas, não se restringindo apenas a dados de biodiversidade. No entanto, esse exercício é deixado a cargo de trabalhos futuros. Adicionalmente, todos os elementos utilizados se enquadram no paradigma WIMP tradicional - no qual se espera que os usuários interajam com as visualizações utilizando mecanismos tradicionais, como *mouse* e teclado - ficando como sugestão para trabalhos futuros o desenvolvimento de uma ferramenta, com base nas recomendações aqui propostas, cuja utilização seja otimizada para interfaces sensíveis ao toque, como em celulares e *tablets*.

Referências

- ADDIS, M. J.; MARTINEZ, K.; LEWIS, P.; STEVENSON, J.; GIORGINI, F. New ways to search, navigate and use multimedia museum collections over the web. 2005.
- ALONSO, P.; IRIONDO, J. M. Urjc gb dataset: Community-based seed bank of mediterranean high-mountain and semi-arid plant species at universidad rey juan carlos (spain). *PhytoKeys*, Pensoft Publishers, n. 35, p. 57, 2014.
- ALTAIR: Declarative Visualization in Python. 2016. <<https://altair-viz.github.io/>>. Acessado em: 30/01/2021.
- ANSCOMBE, F. J. Graphs in statistical analysis. *The american statistician*, Taylor & Francis Group, v. 27, n. 1, p. 17–21, 1973.
- AntMaps (aplicação web). 2016. <<https://antmaps.org/>>. Acessado em: 03/02/2021.
- AquaMaps. 2019. <<https://www.aquamaps.org/>>. Acessado em: 04/02/2021.
- ARTS, K.; WAL, R. van der; ADAMS, W. M. Digital technology and the conservation of nature. *Ambio*, Springer, v. 44, n. 4, p. 661–673, 2015.
- AUER, T.; MACEACHREN, A. M.; MCCABE, C.; PEZANOWSKI, S.; STRYKER, M. Herbariaviz: A web-based client–server interface for mapping and exploring flora observation data. *Ecological Informatics*, Elsevier, v. 6, n. 2, p. 93–110, 2011.
- BAKIEV, A. et al. Reptile occurrences data in the volga river basin (russia). *Biodiversity Data Journal*, Pensoft Publishers, v. 8, 2020.
- BARBET-MASSIN, M.; ROME, Q.; VILLEMANT, C.; COURCHAMP, F. Can species distribution models really predict the expansion of invasive species? *PloS one*, Public Library of Science San Francisco, CA USA, v. 13, n. 3, p. e0193085, 2018.
- BARVE, V.; OTEGUI, J. bdvis: visualizing biodiversity data in r. *Bioinformatics*, Oxford University Press, v. 32, n. 19, p. 3049–3050, 2016.
- BDJ | Biodiversity Data Journal. 2021. <<https://bdj.pensoft.net/>>. Acessado em: 03/01/2021.
- BEKRI, N. E.; PEINSIPP-BYMA, E. Assuring data quality by placing the user in the loop. In: IEEE. *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*. [S.l.], 2016. p. 468–471.
- BERTIFIER. Revisiting Bertin’s Matrices: New interactions for crafting tabular visualizations. 2016. <<https://aviz.fr/bertifier>>. Acessado em: 30/01/2021.
- BERTIN, J. *Semiology of graphics; diagrams networks maps*. [S.l.], 1983.
- BOAKES, E. H. et al. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol*, Public Library of Science, v. 8, n. 6, p. e1000385, 2010.

- BOSTOCK, M.; HEER, J.; OGIEVETSKY, V. D3. js. *Data Driven Documents*, v. 3, n. 5, 2019.
- BROECK, J. Van den; CUNNINGHAM, S. A.; EECKELS, R.; HERBST, K. Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med*, Public Library of Science, v. 2, n. 10, p. e267, 2005.
- CAMACHO, A. I.; DORDA, B. A.; REY, I. Iberian peninsula and balearic island bathynellacea (crustacea, syncarida) database. *ZooKeys*, Pensoft Publishers, n. 386, p. 1, 2014.
- CECCARELLI, S. et al. Analysis of chagas disease vectors occurrence data: the argentinean triatomine species database. *Biodiversity data journal*, Pensoft Publishers, v. 8, 2020.
- CHEN, T.; LU, A.; HU, S.-M. Visual storylines: Semantic visualization of movie sequence. *Computers & Graphics*, Elsevier, v. 36, n. 4, p. 241–249, 2012.
- CLEVELAND, W. S.; MCGILL, R. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, Taylor & Francis Group, v. 79, n. 387, p. 531–554, 1984.
- COLOR Crafter. 2020. <<http://cu-visualab.org/ColorCrafter/>>. Acessado em: 10/01/2021.
- COLORGORICAL. 2020. <<http://vrl.cs.brown.edu/color>>. Acessado em: 10/01/2021.
- CONABIO | Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. 2020. <<https://www.gob.mx/conabio>>. Acessado em: 03/09/2020.
- COOPER, A.; REIMANN, R. et al. *About face 2.0: The essentials of interaction design*. [S.l.]: Wiley Indianapolis, 2003. v. 17.
- CUI, W. et al. Textflow: Towards better understanding of evolving topics in text. *IEEE transactions on visualization and computer graphics*, IEEE, v. 17, n. 12, p. 2412–2421, 2011.
- DILLEN, M. et al. A benchmark dataset of herbarium specimen images with label data. *Biodiversity data journal*, Pensoft Publishers, n. 7, 2019.
- DOMADIYA, N. H.; RAO, U. P. Hiding sensitive association rules to maintain privacy and data quality in database. In: IEEE. *2013 3rd IEEE International Advance Computing Conference (IACC)*. [S.l.], 2013. p. 1306–1310.
- DREW, J. The role of natural history institutions and bioinformatics in conservation biology. *Conservation Biology*, JSTOR, v. 25, n. 6, p. 1250–1252, 2011.
- ESCRIBANO, N.; GALICIA, D.; ARIÑO, A. H.; ESCALA, C. Long-term data set of small mammals from owl pellets in the atlantic-mediterranean transition area. *Scientific data*, Nature Publishing Group, v. 3, n. 1, p. 1–8, 2016.
- ESPINOSA, M.; LÓPEZ, J. Herbarium of vascular plants collection of the university of extremadura (spain). *PhytoKeys*, Pensoft Publishers, n. 25, p. 1, 2013.

- ESTRADA-PÉÑA, A.; FUENTE, J. de L. Species interactions in occurrence data for a community of tick-transmitted pathogens. *Scientific data*, Nature Publishing Group, v. 3, n. 1, p. 1–13, 2016.
- FAN, W.; GEERTS, F. Foundations of data quality management. *Synthesis Lectures on Data Management*, Morgan & Claypool Publishers, v. 4, n. 5, p. 1–217, 2012.
- FEKETE, J.-D. The infovis toolkit. In: IEEE. *IEEE Symposium on Information Visualization*. [S.l.], 2004. p. 167–174.
- FLEMONS, P.; GURALNICK, R.; KRIEGER, J.; RANIPETA, A.; NEUFELD, D. A web-based gis tool for exploring the world's biodiversity: The global biodiversity information facility mapping and analysis portal application (gbif-mapa). *Ecological informatics*, Elsevier, v. 2, n. 1, p. 49–60, 2007.
- FOLEY, J. D. et al. *Computer graphics: principles and practice*. [S.l.]: Addison-Wesley Professional, 1996. v. 12110.
- FOSTER, Z. S.; CHAMBERLAIN, S.; GRÜNWALD, N. J. Taxa: an r package implementing data standards and methods for taxonomic data. *F1000Research*, Faculty of 1000 Ltd, v. 7, 2018.
- FOSTER, Z. S.; SHARPTON, T. J.; GRÜNWALD, N. J. Metacoder: An r package for visualization and manipulation of community taxonomic diversity data. *PLoS computational biology*, Public Library of Science, v. 13, n. 2, p. e1005404, 2017.
- FOX, P.; HENDLER, J. Changing the equation on scientific data visualization. *Science*, American Association for the Advancement of Science, v. 331, n. 6018, p. 705–708, 2011.
- FRIENDLY, M. A brief history of data visualization. In: CHEN, C.; HÄRDLE, W.; UNWIN, A. (Ed.). *Handbook of Computational Statistics: Data Visualization*. Heidelberg: Springer-Verlag, 2006. III, p. ???–??? (In press).
- GARCÍA-SÁNCHEZ, J.; CABEZUDO, B. Herbarium of the university of malaga (spain): Vascular plants collection. *PhytoKeys*, Pensoft Publishers, n. 26, p. 7, 2013.
- GBIF | Global Biodiversity Information Facility. Free and open access to biodiversity data. 2020. <<https://www.gbif.org/>>. Acessado em: 28/08/2020.
- GEISLER, G. Making information more accessible: A survey of information visualization applications and techniques. URL: <http://www.ils.unc.edu/~geisg/info/infovis/paper.html>, Citeseer, 1998.
- GRAHAM, M.; KENNEDY, J. Vesper: Visualising species archives. *Ecological informatics*, Elsevier, v. 24, p. 132–147, 2014.
- GRAMAZIO, C. C.; LAIDLAW, D. H.; SCHLOSS, K. B. Colorgorical: creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2017.
- GSCHWANDTNER, T. et al. Timecleanser: A visual analytics approach for data cleansing of time-oriented data. In: *Proceedings of the 14th international conference on knowledge technologies and data-driven business*. [S.l.: s.n.], 2014. p. 1–8.

- GURALNICK, R.; HILL, A. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics*, Oxford University Press, v. 25, n. 4, p. 421–428, 2009.
- HEER, J.; CARD, S. K.; LANDAY, J. A. Prefuse: a toolkit for interactive information visualization. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. [S.l.: s.n.], 2005. p. 421–430.
- HOLMES, M. W. et al. Natural history collections as windows on evolutionary processes. *Molecular Ecology*, Wiley Online Library, v. 25, n. 4, p. 864–881, 2016.
- HOWE, D. et al. The future of biocuration. *Nature*, Nature Publishing Group, v. 455, n. 7209, p. 47–50, 2008.
- IABIN | Inter-American Biodiversity Information Network. 2020. <<https://www.oas.org/en/sedi/dsd/iabin/about.asp>>. Acessado em: 03/09/2020.
- JACOBS, B.; BORONYAK, L.; MITCHELL, P.; VANDENBERG, M.; BATTEN, B. Towards a climate change adaptation strategy for national parks: adaptive management pathways under dynamic risk. *Environmental science & policy*, Elsevier, v. 89, p. 206–215, 2018.
- JANICKI, J.; NARULA, N.; ZIEGLER, M.; GUÉNARD, B.; ECONOMO, E. P. Visualizing and interacting with large-volume biodiversity data using client–server web-mapping applications: The design and implementation of antmaps. org. *Ecological Informatics*, Elsevier, v. 32, p. 185–193, 2016.
- JETZ, W.; MCPHERSON, J. M.; GURALNICK, R. P. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in ecology & evolution*, Elsevier, v. 27, n. 3, p. 151–159, 2012.
- KECK, M.; GROH, R.; VOSOUGH, Z. A didactic methodology for crafting information visualizations. In: IEEE. *2020 IEEE Visualization Conference (VIS)*. [S.l.], 2020. p. 186–190.
- KEIM, D. A. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, IEEE, v. 8, n. 1, p. 1–8, 2002.
- KELLING, S. et al. Data-intensive science: a new paradigm for biodiversity studies. *BioScience*, American Institute of Biological Sciences Circulation, AIBS, 1313 Dolley . . . , v. 59, n. 7, p. 613–620, 2009.
- KIRK, A. *Data visualisation: A handbook for data driven design*. [S.l.]: Sage, 2016.
- KO, S.; MACIEJEWSKI, R.; JANG, Y.; EBERT, D. S. Marketanalyzer: an interactive visual analytics system for analyzing competitive advantage using point of sale data. In: WILEY ONLINE LIBRARY. *Computer Graphics Forum*. [S.l.], 2012. v. 31, n. 3pt3, p. 1245–1254.
- LISTER, A. M.; GROUP, C. C. R. et al. Natural history collections as sources of long-term datasets. *Trends in ecology & evolution*, Elsevier, v. 26, n. 4, p. 153–154, 2011.
- LIU, S. et al. Steering data quality with visual analytics: The complexity challenge. *Visual Informatics*, Elsevier, v. 2, n. 4, p. 191–197, 2018.

- LIU, S.; CAO, N.; LV, H. Interactive visual analysis of the nsf funding information. In: IEEE. *2008 IEEE Pacific Visualization Symposium*. [S.l.], 2008. p. 183–190.
- LIU, S.; CHEN, C.; LU, Y.; OUYANG, F.; WANG, B. An interactive method to improve crowdsourced annotations. *IEEE transactions on visualization and computer graphics*, IEEE, v. 25, n. 1, p. 235–245, 2018.
- LIU, S.; CUI, W.; WU, Y.; LIU, M. A survey on information visualization: recent advances and challenges. *The Visual Computer*, Springer, v. 30, n. 12, p. 1373–1393, 2014.
- MABROUKI, Y.; TAYBI, A.; CHAVANON, G.; BERRAHOU, A.; MILLÁN, A. Distribution of aquatic beetles from the east of morocco (coleoptera, polyphaga). *Arxius de Miscel·lània Zoològica*, p. 185–211, 2018.
- MACKINLAY, J. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, Acm New York, NY, USA, v. 5, n. 2, p. 110–141, 1986.
- Map of Life (aplicação web). 2016. <<https://www.mol.org/>>. Acessado em: 04/02/2021.
- Marshall III, W. F. et al. Detection of borrelia burgdorferi dna in museum specimens of peromyscus. *Journal of Infectious Diseases*, The University of Chicago Press, v. 170, n. 4, p. 1027–1032, 1994.
- MARX, V. The big challenges of big data. *Nature*, Nature Publishing Group, v. 498, n. 7453, p. 255–260, 2013.
- MAVRAKI, D. et al. Rescuing biogeographic legacy data: The "thor" expedition, a historical oceanographic expedition to the mediterranean sea. *Biodiversity data journal*, Pensoft Publishers, n. 4, 2016.
- MCCURDY, N.; GERDES, J.; MEYER, M. A framework for externalizing implicit error using visualization. *IEEE transactions on visualization and computer graphics*, IEEE, v. 25, n. 1, p. 925–935, 2018.
- MERINO-SÁINZ, I.; ANADÓN, A.; TORRALBA-BURRIAL, A. Harvestmen of the bos arthropod collection of the university of oviedo (spain)(arachnida, opiliones). *ZooKeys*, Pensoft Publishers, n. 341, p. 21, 2013.
- MODI, C. N.; RAO, U. P.; PATEL, D. R. Maintaining privacy and data quality in privacy preserving association rule mining. In: IEEE. *2010 Second International conference on Computing, Communication and Networking Technologies*. [S.l.], 2010. p. 1–6.
- MODICA, L.; LANUZA, P.; GARCÍA-CASTRILLO, G. Surrounded by microplastic, since when? testing the feasibility of exploring past levels of plastic microfibre pollution using natural history museum collections. *Marine Pollution Bulletin*, Elsevier, v. 151, p. 110846, 2020.
- MONFILS, A. K. et al. Natural history collections: teaching about biodiversity across time, space, and digital platforms. *Southeastern Naturalist*, BioOne, v. 16, n. sp10, p. 47–57, 2017.
- Movebank. 2021. <<https://www.movebank.org/>>. Acessado em: 04/02/2021.

MULLER, S. Intérêt des herbiers pour la connaissance des dynamiques spatio-temporelles des invasions biologiques. *Revue d'écologie*, Société nationale de protection de la nature et d'acclimatation de France . . . , 2015.

Museu Nacional. *Museu Nacional: Panorama dos acervos: passado, presente e futuro*. Universidade Federal do Rio de Janeiro - Quinta da Boa Vista, São Cristóvão, Rio de Janeiro, RJ: Série Livro Digital - Disponível em <http://www.museunacional.ufrj.br/destaques/panorama_de_acervos.html>, 2020. v. 18. 120 p.

NOCAJ, A.; BRANDES, U. Organizing search results with a reference map. *IEEE Transactions on Visualization and Computer Graphics*, IEEE, v. 18, n. 12, p. 2546–2555, 2012.

NORMAN, D. *The design of everyday things: Revised and expanded edition*. [S.l.]: Basic books, 2013.

NUALART, N. et al. Dataset of herbarium specimens of threatened vascular plants in catalonia. *PhytoKeys*, Pensoft Publishers, n. 77, p. 41, 2017.

OBIS | Ocean Biodiversity Information System. 2020. <<https://obis.org/>>. Acessado em: 28/08/2020.

OLIPHANT, T. E. Python for scientific computing. *Computing in Science & Engineering*, IEEE, v. 9, n. 3, p. 10–20, 2007.

ONDOV, B. D.; BERGMAN, N. H.; PHILLIPPY, A. M. Krona: interactive metagenomic visualization in a web browser. *Encyclopedia of Metagenomics*. Springer US, p. 339–346, 2015.

ORTIZ-YUSTY, C. E.; DAZA, J. M.; PAEZ, V. P.; BOCK, B. C. The collection of the herpetological museum of the university of antioquia (northwestern colombia). *Biodiversity data journal*, Pensoft Publishers, n. 3, 2015.

OTEGUI, J.; ARIÑO, A. H. Biddsat: visualizing the content of biodiversity data publishers in the global biodiversity information facility network. *Bioinformatics*, Oxford University Press, v. 28, n. 16, p. 2207–2208, 2012.

PAINI, D. R. et al. Global threat to agriculture from invasive species. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 113, n. 27, p. 7575–7579, 2016.

PANDO, F.; DUEÑAS, M.; LADO, C.; TELLERIA, M. T. The flora mycologica iberica project fungi occurrence dataset. *MycoKeys*, Pensoft Publishers, v. 15, p. 59, 2016.

PÉREZ-LUQUE, A. J.; BAREA-AZCÓN, J. M.; ÁLVAREZ-RUIZ, L.; BONET-GARCÍA, F. J.; ZAMORA, R. Dataset of passerine bird communities in a mediterranean high mountain (sierra nevada, spain). *ZooKeys*, Pensoft Publishers, n. 552, p. 137, 2016.

PÉREZ-LUQUE, A. J. et al. Sinfonevada: Dataset of floristic diversity in sierra nevada forests (se spain). *PhytoKeys*, Pensoft Publishers, n. 35, p. 1, 2014.

PERIN, C.; DRAGICEVIC, P.; FEKETE, J.-D. Revisiting bertin matrices: New interactions for crafting tabular visualizations. *IEEE transactions on visualization and computer graphics*, IEEE, v. 20, n. 12, p. 2082–2091, 2014.

- PERIN, C. et al. State of the art of sports data visualization. In: WILEY ONLINE LIBRARY. *Computer Graphics Forum*. [S.l.], 2018. v. 37, n. 3, p. 663–686.
- PERSING, D. H. et al. Detection of borrelia burgdorferi dna in museum specimens of ixodes dammini ticks. *Science*, American Association for the Advancement of Science, v. 249, n. 4975, p. 1420–1423, 1990.
- PINTO, C. M. et al. Using museum collections to detect pathogens. 2010.
- PRETORIUS, A. J.; BRAY, M.-A.; CARPENTER, A. E.; RUDDLE, R. A. Visualization of parameter space for image analysis. *IEEE Transactions on Visualization and Computer Graphics*, IEEE, v. 17, n. 12, p. 2402–2411, 2011.
- PRUDENTE, A. L. da C. et al. Dataset from the snakes (serpentes, reptiles) collection of the museu paraense emílio goeldi, pará, brazil. *Biodiversity Data Journal*, Pensoft Publishers, v. 7, 2019.
- QU, Z.; HULLMAN, J. Evaluating visualization sets: Trade-offs between local effectiveness and global consistency. In: *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*. [S.l.: s.n.], 2016. p. 44–52.
- RADJI, R. et al. Togo national herbarium database. *PhytoKeys*, Pensoft Publishers, n. 109, p. 1, 2018.
- READY, J. et al. Predicting the distributions of marine organisms at the global scale. *Ecological Modelling*, Elsevier, v. 221, n. 3, p. 467–478, 2010.
- Rebioma. 2021. <<http://www.rebioma.net/>>. Acessado em: 04/02/2021.
- REICHMAN, O. J.; JONES, M. B.; SCHILDHAUER, M. P. Challenges and opportunities of open data in ecology. *Science*, American Association for the Advancement of Science, v. 331, n. 6018, p. 703–705, 2011.
- ROBBIRT, K. M.; DAVY, A. J.; HUTCHINGS, M. J.; ROBERTS, D. L. Validation of biological collections as a source of phenological data for use in climate change studies: a case study with the orchid ophrys sphegodes. *Journal of Ecology*, Wiley Online Library, v. 99, n. 1, p. 235–241, 2011.
- ROSSUM, G. V. *Python*. 1991. <<https://www.python.org/>>. Acessado em 30/01/2021.
- SANTOS, B. F.; AGUIAR, A. P.; TEDESCO, A. M.; FONTENELLE, J. C. Long-term seasonal dominance of the wasp trihopsis polita townes (hymenoptera, ichneumonidae) in the brazilian atlantic forest. *Biodiversity Data Journal*, Pensoft Publishers, n. 5, 2017.
- SARNAT, E. Piakey: Identification guide to ants of the pacific island. *University of California, Davis*, 2008.
- SATYANARAYAN, A.; MORITZ, D.; WONGSUPHASAWAT, K.; HEER, J. Vega-lite: A grammar of interactive graphics. *IEEE transactions on visualization and computer graphics*, IEEE, v. 23, n. 1, p. 341–350, 2016.
- SCHLOSS, K. B.; PALMER, S. E. Aesthetic response to color combinations: preference, harmony, and similarity. *Attention, Perception, & Psychophysics*, Springer, v. 73, n. 2, p. 551–571, 2011.

- SHIRAVI, H.; SHIRAVI, A.; GHORBANI, A. A. A survey of visualization systems for network security. *IEEE Transactions on visualization and computer graphics*, IEEE, v. 18, n. 8, p. 1313–1329, 2011.
- SHNEIDERMAN, B. The eyes have it: A task by data type taxonomy for information visualizations. In: *The craft of information visualization*. [S.l.]: Elsevier, 2003. p. 364–371.
- SISTEMA de Informação sobre a Biodiversidade Brasileira. 2020. <<https://www.sibbr.gov.br/>>. Acessado em: 14/02/2020.
- SMART, S.; WU, K.; SZAFIR, D. A. Color crafting: Automating the construction of designer quality color ramps. *IEEE transactions on visualization and computer graphics*, IEEE, v. 26, n. 1, p. 1215–1225, 2019.
- SONG, H.; SZAFIR, D. A. Where's my data? evaluating visualizations with missing data. *IEEE transactions on visualization and computer graphics*, IEEE, v. 25, n. 1, p. 914–924, 2018.
- SRIVASTAVA, V. et al. Assessing niche shifts and conservatism by comparing the native and post-invasion niches of major forest invasive species. *Insects*, Multidisciplinary Digital Publishing Institute, v. 11, n. 8, p. 479, 2020.
- SUAREZ, A. V.; TSUTSUI, N. D. The value of museum collections for research and society. *BioScience*, American Institute of Biological Sciences, v. 54, n. 1, p. 66–74, 2004.
- TARLI, V. D.; GRANDCOLAS, P.; PELLENS, R. The informative value of museum collections for ecology and conservation: A comparison with target sampling in the brazilian atlantic forest. *PloS one*, Public Library of Science San Francisco, CA USA, v. 13, n. 11, p. e0205710, 2018.
- THOMAS, S.; MINTZ, A. *Virtual and the Real: Media in the Museum*. [S.l.]: American Association of Museums, 1998.
- TSIKY, R. Rebioma data portal, tool for conservation planning in madagascar. *Biodiversity Information Science and Standards*, Pensoft Publishers, 2018.
- UETZ, P.; STYLIANOU, A. The original descriptions of reptiles and their subspecies. *Zootaxa*, v. 4375, n. 2, p. 257–264, 2018.
- VANDERPLAS, J. et al. Altair: interactive statistical visualizations for python. *Journal of open source software*, v. 3, n. 32, p. 1057, 2018.
- VEGA-LITE: A Grammar of Interactive Graphics. 2016. <<https://vega.github.io/vega-lite/>>. Acessado em: 01/02/2021.
- WANG, R.; PEREZ-RIVEROL, Y.; HERMJAKOB, H.; VIZCAÍNO, J. A. Open source libraries and frameworks for biological data visualisation: A guide for developers. *Proteomics*, Wiley Online Library, v. 15, n. 8, p. 1356–1374, 2015.
- WICKHAM, H. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 19, n. 1, p. 3–28, 2010.
- WILKINSON, L. The grammar of graphics. In: *Statistics and Computing*. [S.l.]: Springer, 1999.

- WINDHAGER, F. et al. Visualization of cultural heritage collection data: State of the art and future challenges. *IEEE transactions on visualization and computer graphics*, IEEE, v. 25, n. 6, p. 2311–2330, 2018.
- WU, Y.-H.; CHIANG, C.-M.; CHEN, A. L. Hiding sensitive association rules with limited side effects. *IEEE Transactions on Knowledge and Data engineering*, IEEE, v. 19, n. 1, p. 29–42, 2006.
- YI, J. S.; KANG, Y. ah; STASKO, J.; JACKO, J. A. Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics*, IEEE, v. 13, n. 6, p. 1224–1231, 2007.
- ZHELEZNOVA, G.; SHUBINA, T.; DEGTEVA, S.; CHADIN, I.; RUBTSOV, M. Moss occurrences in yugyd va national park, subpolar and northern urals, european north-east russia. *Biodiversity Data Journal*, Pensoft Publishers, v. 7, 2019.
- ZHELEZNOVA, G.; SHUBINA, T.; RUBTSOV, M.; LITVINENKO, G.; CHADIN, I. Bryophytes occurrences dataset based on syko herbarium moss collection. *Biodiversity Data Journal*, Pensoft Publishers, v. 8, 2020.
- ZUDALOVA, E.; ADRIAANSEN, T.; VANLIERE, R. Trends in interactive visualization: State-of-the-art survey. *Advanced Information and Knowledge Processing*, Springer, 2009.

APÊNDICE A – Ordenamento de Cores

Esse capítulo tem o simples objetivo de ilustrar como foi feito o ordenamento de cores no exemplo apresentado na seção 4.2 do capítulo 4, ilustrado na figura 41. Ao final, será apresentado o código Python usado para reproduzir o exemplo. Partindo do pressuposto que se deseja construir um mapeamento de cores para $N > 9$ elementos, esse exemplo pode ser sintetizado nos seguintes passos:

1. É inserida uma sugestão de cor centroide¹ (de interesse do usuário) na ferramenta Color Crafter[®] (SMART; WU; SZAFIR, 2019).
2. Com base em uma abordagem algorítmica construída sobre 220 esquemas de cores criados diretamente por *designers*, o Color Crafter[®] sugere algumas paletas de cores sequenciais (gradações) discretas com base em configurações pré-carregadas de seus parâmetros de matiz (*hue*), luminância e saturação.
3. A partir de fatores que vão desde interesses e gostos pessoais, até características dos objetos que se quer representar, o usuário seleciona uma das paletas de cores sugeridas pela ferramenta. No entanto, o Color Crafter[®] sugere apenas 9 códigos de cor para cada *seed* fornecida.
4. Para criar novas cores que habitam a mesma vizinhança induzida pela cor centroide dada como *input*, o usuário pode fazer pequenas alterações nos parâmetros da ferramenta a fim de criar variações perceptualmente distintas mas, ao mesmo tempo, semelhantes às cores da paleta anterior.

Vale destacar que, caso se esteja trabalhando com múltiplas cores centroides, são preferidas alterações no parâmetro de luminância (deslocamentos horizontais na curva C-L) pois, assim, há menos chances de que o espaço de cores definido pelas demais centroides seja invadido pelos elementos da nova paleta, dessa forma, garantindo que a paleta resultante de cada centroide sejam perceptualmente distintas. Caso se esteja trabalhando com cores muito saturadas, ajustes na saturação podem ajudar a obtenção de resultados satisfatórios. Em contrapartida, ajustes de matiz são fortemente desencorajados nesse caso para se preservar níveis de distância perceptual.

5. Uma vez selecionados dois conjuntos de cores distintos, porém situados na mesma vizinhança da cor centroide, pode-se agregá-los para formar uma lista com 18 entradas de códigos hexadecimais.

¹ Um código de cor hexadecimal.

6. Por fim, para melhorar a apresentação do esquema de cores estendido ou atribuir significado relevante à diferentes intensidades, recomenda-se ordenar suas entradas segundo seu valor de luminância.

É importante pontuar que foram testadas três maneiras de se calcular a luminância de uma cor a partir de seu código RGB²:

- a) $Y = 0.2126 \cdot R + 0.7152 \cdot G + 0.0722 \cdot B$, usada em espaços colorimétricos que usam primárias ITU-R BT.709;
- b) $Y = 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$, definida como o brilho percebido de uma cor (vide <<https://www.w3.org/TR/AERT/#color-contrast>> para mais detalhes);
- c) $Y = \sqrt{0.299 \cdot R^2 + 0.587 \cdot G^2 + 0.114 \cdot B^2}$, versão normalizada para o brilho percebido de uma cor.

Esta última medida, de fato, foi a que apresentou os resultados mais satisfatórios ao longo deste estudo.

Não obstante, é apresentado o código Python usado para combinar duas paletas de cores fornecidas pelo Color Crafter[®] e ordenar sua versão estendida a partir do cálculo da luminância.

```

1  ## Codigo em Python 3
2  # imports
3  import numpy as np
4
5  # funcao para converter de hexadecimal para RGB
6  def get_rgb(HEX):
7      HEX = HEX.lstrip("#")    # remove # de codigos hexadecimais no formato Web
8      r, g, b = (int(HEX[i:i+2], 16) / 255.0 for i in range(0,5,2))
9      return r, g, b
10
11 # funcoes para calcular luminancia
12 def calc_luminancia1(HEX):
13     '''
14     Calcula luminancia como Y = 0.2126*R + 0.7152*G + 0.0722*B (proposta 1).
15     '''
16     r, g, b = get_rgb(HEX)
17     return 0.2126*r + 0.7152*g + 0.0722*b
18
19 def calc_luminancia2(HEX):
20     '''
21     Calcula luminancia como Y = 0.299*R + 0.587*G + 0.114*B (proposta 2).
22     '''
23     r, g, b = get_rgb(HEX)
24     return 0.299*r + 0.587*g + 0.114*b

```

² Para tal, é necessário converter todos os códigos hexadecimais para o padrão RGB.

```

25
26 def calc_luminancia3(HEX):
27     '''
28     Calcula luminancia como  $Y = \sqrt{0.299*R + 0.587*G + 0.114*B}$  (proposta 3).
29     '''
30     r, g, b = get_rgb(HEX)
31     return np.sqrt(0.299*r + 0.587*g + 0.114*b)
32
33
34 # inserindo esquemas de cores escolhidos no exemplo do capitulo 4
35 cores = ['#ffd9d3', '#ffc0b9', '#ffa8a1', '#ff908b', '#ff7777', '#fb5e65', '#
    ee4454', '#e12445', '#d20037']
36 cores_alteradas = ['#ff9a9b', '#ff8182', '#ff686d', '#ee4d58', '#d52f44', '#
    bc0030', '#a3001e', '#8a000c', '#710000']
37
38 # estendendo paleta de cores
39 cores.extend(cores_alteradas)
40
41 # ordenando de acordo com a luminancia
42 cores.sort(key=calc_luminancia3, reverse=True) # basta substituir a key por uma
    das funcoes definidas

```

Ainda, é apresentado um código simples para visualizar o esquema de cores gerado com a biblioteca Altair.

```

1 ##Codigo em Python 3
2 # imports
3 import numpy as np
4 import pandas as pd
5 import altair as alt # (tutorial de instalacao em https://altair-viz.github.io/
    getting_started/installation.html)
6
7 # objeto cores (definido anteriormente) contem todas as cores da paleta estendida
8 cores
9
10 # criando um pandas DataFrame com todas as cores e uma coluna inteira de 1's (
    altura do quadrado)
11 df = pd.DataFrame([cores, [int(i) for i in np.repeat(1, len(cores))]], index=['
    cores', 'ones']).transpose()
12
13 # visualizando o esquema de cores (semelhante ao exposto no exemplo)
14 alt.Chart(df, height=40, width=800).mark_rect().encode(
15     x = alt.X('cores:N', title='', axis= None),
16     y = alt.Y('ones:Q', title='', axis= None),
17     color= alt.Color('cores:N', scale= alt.Scale(range=cores), legend=None)
18 )

```