

**FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA**

CARLOS CÉSAR DE OLIVEIRA FONSECA

**VISUALIZAÇÃO DE DADOS HIERÁRQUICOS EM COLEÇÕES
CIENTÍFICAS DE BIODIVERSIDADE: A CRIAÇÃO DE UMA
ÁRVORE TAXONÔMICA**

Rio de Janeiro
2023

CARLOS CÉSAR DE OLIVEIRA FONSECA

**VISUALIZAÇÃO DE DADOS HIERÁRQUICOS EM COLEÇÕES
CIENTÍFICAS DE BIODIVERSIDADE: A CRIAÇÃO DE UMA
ÁRVORE TAXONÔMICA**

Trabalho de conclusão de curso apresentada para a Escola de Matemática Aplicada (FGV/EMAp) como requisito para o grau de bacharel em Ciência de Dados e Inteligência Artificial.

Área de estudo: Visualização da Informação.

Orientador: Asla Medeiros e Sá

Rio de Janeiro

2023

Ficha catalográfica elaborada pela BMHS/FGV

Sobrenome, Nome

Visualização de dados Hierárquicos em Coleções Científicas de Biodiversidade: a criação de uma Árvore Taxonômica/ Carlos César de Oliveira Fonseca. – 2023.

51f.

Trabalho de Conclusão de Curso – Escola de Matemática Aplicada.

Advisor: Asla Medeiros e Sá.

Includes bibliography.

1. Matemática 2. Aplicada 2. na matemática I. Sobrenome professor, Nome professor II. Escola de Matemática Aplicada III. Visualização de dados Hierárquicos em Coleções Científicas de Biodiversidade

CARLOS CÉSAR DE OLIVEIRA FONSECA

**VISUALIZAÇÃO DE DADOS HIERÁRQUICOS EM COLEÇÕES
CIENTÍFICAS DE BIODIVERSIDADE: A CRIAÇÃO DE UMA
ÁRVORE TAXONÔMICA**

Trabalho de conclusão de curso apresentada para a Escola de Matemática Aplicada (FGV/EMAp) como requisito para o grau de bacharel em Ciência de Dados e Inteligência Artificial.

Área de estudo: Visualização da Informação.

E aprovado em 14/12/2023
Pela comissão organizadora

Asla Medeiros e Sá
Escola de Matemática Aplicada

Flávio Codeço Coelho
Escola de Matemática Aplicada

Cristiana Silveira Serejo
Museu Nacional - Universidade Federal do
Rio de Janeiro

Dedico este trabalho à distância que me separa daqueles que amo e a um futuro promissor que há de chegar.

Agradecimentos

Gostaria de agradecer a todos que proporcionaram o desenvolvimento desse trabalho, mas também desse curso.

Em especial:

- À professora Asla pela paciência, orientação e dedicação durante minha trajetória acadêmica.
- Aos curadores do Museu Nacional/UFRJ, que compartilharam suas bases de pesquisa e sempre estavam dispostos a contribuir com meu trabalho. Agradeço aos setores de:
 - Carcinologia (Cristiana Silveira Serejo)
 - Annelidologia (Joana Zanol e Camila Messias)
 - Herpetologia (Paulo Passos, Manoela Woitovicz Cardoso e Pedro Pinna)
 - Ornitologia (Renata Stopiglia)
- Ao CDMC e à EMAP por possibilitarem meus estudos na FGV, com toda sua excelência.
- Aos professores e aos colegas pela excelência do curso e pelos momentos de convivência.
- Aos amigos, tanto do Rio quanto de Patos, que me incentivaram e sempre estiveram juntos nos momentos de lazer ou de desespero.
- A toda minha família, aos meus pais, César e Júlia, à minha maninha, Ana Clara, aos meus tios, tias, primos, avós, que nunca deixaram de me apoiar e de incentivar a persistir nos meus sonhos, mesmo com a distância.
- E a todos os outros que, por momentos, não consigo me lembrar, mas sem sobra de dúvida são importantes para este momento.

Muito obrigado, e um futuro ainda melhor para todos nós.

*“Gráficos e Piadas tem uma semelhança.
Se tem que explicar muito é porque tem algo errado.”*

Minha Vó

*“Letras juntas formam uma palavra.
Palavras juntas formam uma frase. Frases juntas concatenam uma
ideia. A sua ideia são apenas palavras. Palavras são opiniões.”*

Paulo Autuori, 2023

Resumo

A pesquisa aborda a temática da visualização de dados hierárquicos, um tópico relevante, embora muitas vezes negligenciado devido à complexidade de sua implementação nos principais *frameworks* de visualização. O foco principal reside na aplicação dessa abordagem nas Coleções Científicas Zoológicas de Biodiversidade, priorizando a representação hierárquica natural presente na taxonomia dos animais. A implementação prática desse conceito é realizada por meio da linguagem de programação *Python*, onde os dados são manipulados, até que resultado seja listas de Nós e Arestas, uma forma eficaz de armazenar grafos. Essa manipulação simplificada facilita a plotagem dos dados utilizando bibliotecas de visualização da linguagem de programação escolhida. A conclusão do trabalho envolveu ajustes visuais essenciais para alcançar o objetivo estabelecido. Inclusive, a grande quantidade de dados se fez presente e foi necessário a utilização de abordagens específicas. Em suma, a pesquisa se concentra na aplicação de técnicas de Ciência de Dados para o desenvolvimento de uma visualização que auxilie curadores na exploração e apresentação de Registros Primários de Biodiversidade.

Palavras-chave: Visualização. Visualização de Árvore. Biodiversidade. Ciência de Dados

Abstract

The research addresses the theme of hierarchical data visualization, a relevant topic often overlooked due to the complexity of its implementation in major visualization frameworks. The primary focus lies in applying this approach to the Zoological Scientific Collections of Biodiversity, prioritizing the natural hierarchical representation present in the taxonomy of animals. The practical implementation of this concept is carried out through the Python programming language, where data is manipulated until the result is lists of Nodes and Edges, an effective way to store graphs. This simplified manipulation facilitates data plotting using visualization libraries of the chosen programming language. The conclusion of the work involved essential visual adjustments to achieve the established goal. Furthermore, the substantial amount of data was present, requiring the use of specific approaches. In summary, the research focuses on applying Data Science techniques to develop a visualization that assists curators in exploring and presenting Primary Biodiversity Records.

Keywords: Visualization. Tree Visualization. Biodiversity. Data Science.

Lista de ilustrações

Figura 1 – Servidor com a solução de sobreposição	17
Figura 2 – Imagem coleção	18
Figura 3 – Exemplares do acervo de répteis do setor de Herpetologia do MN/UFRJ	20
Figura 4 – Exemplares da coleção de Crustáceos	21
Figura 5 – Espécimes de anelídeos similares aos presentes na coleção de Annelidologia	21
Figura 6 – Exemplares de aves presentes na coleção de Ornitologia	22
Figura 7 – Representação da Taxonomia	27
Figura 8 – Representação de uma Árvore Filogenética	28
Figura 9 – Representação de uma Árvore em Teoria dos Grafos	28
Figura 10 – Exemplo de Grafo Gerado a partir das Listas de Nós e Arestas	34
Figura 11 – Representação Inicial da Coleção de Crustácea	35
Figura 12 – Grafo com filtro pela família	37
Figura 13 – Grafo com crescimento horizontal(troca dos eixos)	39
Figura 14 – Grafo com nível de registro	40
Figura 15 – Grafo com nível de registro e imagens na tooltip	41
Figura 16 – Árvores Taxonômicas	43
Figura 17 – Árvore taxonômica das Raposas	49
Figura 18 – Árvore das divisões regionais do Sudeste	51

Lista de tabelas

Tabela 1 – Resumo Taxonômico Quantitativo das Coleções	22
Tabela 2 – Exemplo de agrupamento dos dados	31
Tabela 3 – Exemplo da transformação para dados de grafo	32
Tabela 4 – Exemplo dos dados de localização	50

Sumário

1	INTRODUÇÃO	13
2	TRABALHOS ANTERIORES	15
2.1	Visualizações Coleções Científicas	15
2.2	Servidor	16
2.3	Iniciação Científica	16
3	DADOS	18
3.1	Estrutura dos Metadados	19
3.2	Coleções	20
3.3	Pré-processamento e tratamento de dados	22
4	ROTEIRO	24
4.1	Motivações	24
4.2	Circunstâncias	25
4.2.1	Dados Utilizados	25
4.2.2	Ferramenta	25
4.3	Expectativa	26
4.4	Fontes de Inspiração	27
5	EXECUÇÃO	29
5.1	Tratamento do dado até virar um grafo	29
5.1.1	Dados Faltantes	29
5.1.2	Agregação dos dados	31
5.1.3	Transformação em Grafos	31
5.2	Desenho do Grafo	33
5.2.1	Camadas do Grafo	33
5.2.2	Diferenças nos Comandos	34
5.2.3	Versão Inicial	35
5.2.4	Detalhes conforme a demanda	36
5.2.5	Inversão dos eixos	38
5.2.6	Representação dos Registros	38
5.2.6.1	Imagens nas Tooltips	40

6	RESULTADOS E CONCLUSÃO	42
6.1	Futuros passos	44
6.2	Considerações Finais	44
	Referências	45
	APÊNDICES	47
	APÊNDICE A – NOVA BASE	48
	APÊNDICE B – OUTRAS ÁRVORES	50

1 Introdução

Desde o desfecho da Segunda Guerra Mundial, o cenário global passou por transformações significativas em diversas esferas, desde o surgimento de novos países até as recentes tensões políticas e desafios sanitários globais. Essa trajetória tumultuada, marcada por guerras, epidemias e avanços tecnológicos, culminou em uma era em que a interconexão global é uma realidade. A crescente complexidade do mundo moderno foi acompanhada por avanços notáveis na tecnologia da informação, especialmente no campo da computação.

Com a ascensão de computadores mais poderosos e uma conectividade global sem precedentes, a cultura de armazenamento e análise de dados floresceu. No entanto, à medida que gigantescas bases de dados se tornaram parte integrante do cotidiano de instituições, surgiram desafios na interpretação efetiva dessas informações. A leitura e compreensão de conjuntos massivos de dados tabulares, frequentemente repletos de dezenas de metadados distintos, tornaram-se uma tarefa desafiadora, mesmo com ferramentas de filtragem e ordenação tradicionais.

Nesse contexto, a Visualização da Informação (InfoVis) emerge como uma área de pesquisa essencial e em expansão. Focada em auxiliar usuários na exploração, compreensão e análise visual de dados, a InfoVis oferece uma abordagem inovadora para revelar *insights* valiosos a partir de conjuntos complexos. Com aplicações já consolidadas em setores como esportes, negócios, dados textuais e audiovisuais, a InfoVis desempenha um papel fundamental em diversos campos.

Uma área específica em que a InfoVis se destaca é a biodiversidade. Instituições de todo o mundo, como herbários, museus e universidades, dedicam esforços consideráveis à digitalização de suas coleções biológicas. Esses esforços resultam em vastas coleções digitais, ricas em informações sobre a distribuição de espécies e os impactos das mudanças climáticas.

No contexto da biodiversidade, projetos como *AntMaps*([JANICKI et al., 2016](#)) e *Map of Life*([JETZ; MCPHERSON; GURALNICK, 2012](#)) representam exemplares da excelência alcançada em visualizações voltadas para a compreensão e disseminação de dados relacionados à diversidade biológica. A importância dessas visualizações transcende a mera apresentação de resultados, estendendo-se à exploração detalhada das informações disponíveis. Em uma extensa revisão de artigos científicos relacionados ao tema, Franklin Oliveira([OLIVEIRA, 2021](#)) identificou que diversas formas de visualizações, como aquelas de natureza geográfica, temporal e taxonômica, são frequentemente empregadas para comunicar resultados.

Contudo, o papel das visualizações no âmbito da biodiversidade vai além da

apresentação de dados, incluindo a importante função de facilitar a exploração dessas informações(FOX; HENDLER, 2011). Franklin Oliveira, em sua dissertação de mestrado(OLIVEIRA, 2021), delineou uma colaboração com o Museu Nacional/UFRJ, resultando na criação de uma galeria de visualizações projetadas para auxiliar pesquisadores na exploração detalhada de suas Coleções Científicas Zoológicas de Biodiversidade.

Essas visualizações desenvolvidas proporcionaram ferramentas eficazes para a exploração de bases de dados extensas. Entre as representações produzidas, destacam-se aquelas dedicadas às áreas de pesquisa geográfica, temporal e taxonômica, permitindo uma análise mais aprofundada de diferentes perspectivas sobre as coleções.

Mesmo existindo visualizações taxonômicas, elas apenas separam os espécimes de diferentes famílias em grupos distintos, ou seja, a representação taxonômica feita se limitava a separação de espécimes dentro de uma categoria taxonômica. Não existindo nenhuma representação que relacione diferentes grupos taxonômicos de um mesmo espécime nas visualizações feitas por Franklin(OLIVEIRA, 2021).

Apesar do avanço significativo, as visualizações taxonômicas ainda apresentam limitações notáveis. As representações existentes se restringem a separar espécimes em grupos distintos, geralmente delineados por categorias taxonômicas amplas, como famílias. No entanto, a lacuna persiste na ausência de visualizações que estabeleçam conexões significativas entre diferentes grupos taxonômicos pertencentes a um mesmo espécime. As criações não abordaram de maneira abrangente essa necessidade específica de relacionar diferentes categorias taxonômicas dentro de um único espécime, revelando um espaço valioso para inovação nesse campo.

Diante desse cenário, este trabalho propõe-se a sugerir uma abordagem diferenciada na criação de visualizações. O desafio aqui é desenvolver um gráfico que não apenas represente, mas também revele a hierarquia entre as categorias taxonômicas de cada espécie, conectando os diferentes grupos desde o reino até a espécie. A exploração dessa perspectiva busca contribuir para a compreensão e análise dos dados de biodiversidade, abrindo novas possibilidades de *insights* no campo de InfoVis aplicada à pesquisa em biodiversidade.

No decorrer dos próximos capítulos, apresentaremos uma (2) breve discussão sobre os trabalhos que deram origem a este, seguida de (3) uma descrição detalhada dos dados utilizados. Em seguida, abordaremos (4) a modelagem do problema e a definição de expectativas. Avançaremos para (5) a discussão sobre as criações e as decisões importantes tomadas durante o desenvolvimento. O capítulo (6) trará o resultado final, destacando os passos futuros e concluindo nosso estudo.

2 Trabalhos anteriores

Nesse capítulo propomos uma discussão sobre os trabalhos que antecederam à este. Três deles apresentaram papel crucial e são as bases da discussão: (1) Visualizações de Coleções Científicas, (2) Servidor e (3) Iniciação Científica.

2.1 Visualizações Coleções Científicas

Como já mencionado, este trabalho é desdobramento do projeto iniciado na dissertação de mestrado de Franklin Oliveira([OLIVEIRA, 2021](#)) em que é feita uma vasta pesquisa sobre Visualização da Informação e as suas aplicações ao contexto de biodiversidade. O autor levantou uma visão geral da utilização de Infovis em bases de Biodiversidade até o atual 'estado da arte'. Além disso, buscou artigos científicos relacionados a biodiversidade e quantificou a utilização de visualizações geográficas, temporais ou taxonômicas, de forma, a atestar a sua utilização de campo de biodiversidade.

Em um contexto geral de Infovis, Hedler([FOX; HENDLER, 2011](#)) diz que Visualização da informação é muito relevante para a apresentação de dados, mas também é importante para auxiliar pesquisadores a explorar seus dados, de forma complementar, Thomas e Mintz([THOMAS; MINTZ, 1999](#)) destacam a necessidade da manutenção de registros digitais fiéis ao original da coleção por parte dos curadores. Nesse cenário, Franklin Oliveira([OLIVEIRA, 2021](#)) propôs uma sequência de Visualizações para auxiliar curadores na exploração e verificação de suas coleções. Além de serem ferramentas no controle dos dados, essas visualizações podem ser utilizadas como apresentação dos dados, pois todas mostram os dados de forma clara e robusta.

O sucesso do trabalho e das Visualizações, sua adoção pelos pesquisadores e possíveis melhorias notadas graças ao uso, possibilitaram que outros trabalhos fossem desenvolvidos ([MEDEIROS E SÁ; OLIVEIRA; SILVEIRA SEREJO, 2021](#); [MEDEIROS E SÁ et al., 2022](#); [MESSIAS et al., 2023](#)). Esses projetos,juntamente com o trabalho primordial de Franklin Oliveira([OLIVEIRA, 2021](#)), forneceram uma forte base teórica e decisória para construção deste projeto.

Como mencionado anteriormente, o presente trabalho é uma continuação do projeto iniciado na dissertação de mestrado de Franklin Oliveira([OLIVEIRA, 2021](#)).

2.2 Servidor

Durante o mestrado de Franklin Oliveira([OLIVEIRA, 2021](#)), as visualizações se destacaram pela excelência, porém, enfrentou-se um desafio significativo: a defasagem entre os gráficos e as bases de dados à medida que eram atualizados. Essa defasagem decorria da abordagem assíncrona na geração dos gráficos, exigindo o envio das bases de dados pelos pesquisadores para receberem as visualizações correspondentes posteriormente.

Dado que o principal propósito desses gráficos é auxiliar os pesquisadores na manutenção da qualidade das bases, a esperada modificação dos dados resulta em gráficos desatualizados. Essa situação, por sua vez, conduz a uma potencial perda de relevância das visualizações ao longo do tempo, levando os pesquisadores a abandoná-las.

Para superar esse desafio, foi desenvolvido o servidor *VisZoo* ([MEDEIROS, 2023](#)). Ele foi criado para permitir que os pesquisadores atualizem suas visualizações de maneira ágil e dinâmica. Ao contrário da abordagem anterior, o *VisZoo* recebe as bases de dados e gera as visualizações automaticamente. Essa abordagem assegura que os pesquisadores tenham acesso a informações visualmente precisas e sempre atualizadas.

2.3 Iniciação Científica

O *VisZoo* foi desenvolvido como uma solução para possibilitar a produção dos gráficos pelos curadores sem o intermédio de especialistas em programação, contudo, a ferramenta web demanda de manutenção constante. Diante desse cenário, iniciei meu projeto de Iniciação Científica com o intuito de aprimorar o servidor e introduzir novas visualizações. Enquanto dedicava esforços à manutenção do servidor, deparei-me com dois desafios de visualização: (1) encontrar uma solução para a sobreposição de dados e (2) criar uma representação hierárquica dos dados taxonômicos.

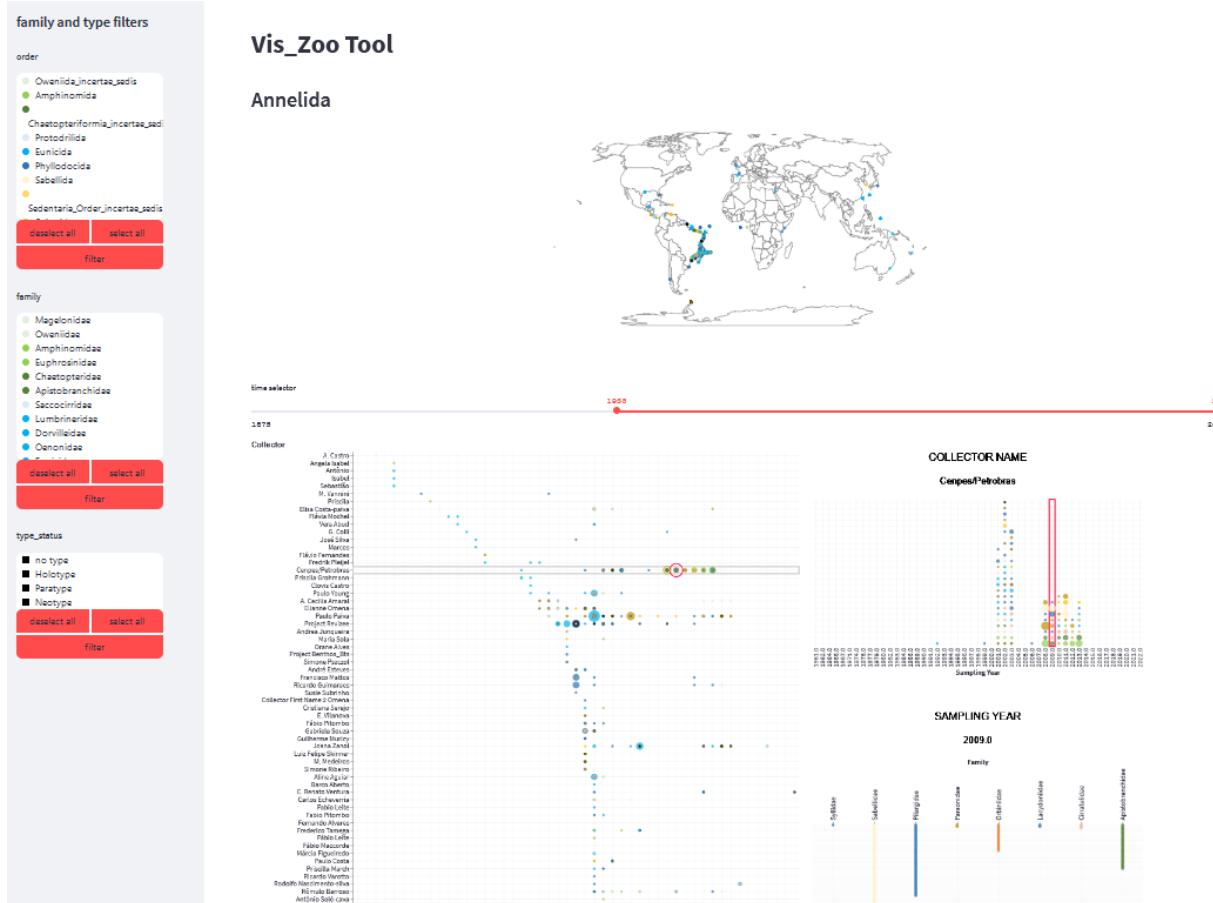
Durante a Iniciação Científica, abordamos questão de sobreposição de dados. A solução proposta foi baseada no mantra *Overview first, filter and zoom, details on demand*([SHNEIDERMAN, 1996](#)) e é composta pela visualização original, adicionada de dois *Dot plots*¹ auxiliares que inicialmente se encontram vazios, mas sempre que tem um item selecionado, é feito o desmembramento dos elementos contidos nesse item. Um exemplo dessa solução pode ser vista na Figura 1, onde a solução proposta para esse problema já está integrada ao servidor.

O segundo desafio, que tornou se o cerne deste trabalho e visa criar uma representação hierárquica dos dados. Esta abordagem ainda não implementada nessas bases de dados,

¹ *Dot Plot* é um gráfico de dispersão que com um eixo sendo uma variável, enquanto o outro eixo é a quantidade de elementos para cada um dos valores, parecido com um gráfico de barras mas sem agregação, veja <[https://en.wikipedia.org/wiki/Dot_plot_\(statistics\)](https://en.wikipedia.org/wiki/Dot_plot_(statistics))>

tem um futuro promissor na exploração e revisão de dados, também pode desempenhar um papel muito importante na apresentação das coleções.

Figura 1 – Servidor com a solução de sobreposição



Fonte: Elaboração própria com o uso do Viszoo.

3 Dados

O Museu Nacional/UFRJ representa um dos mais significativos centros de museologia do Brasil. Desde sua fundação em 1818, por meio de um decreto de Dom João VI, estabeleceu-se como a primeira instituição dedicada à museologia e pesquisa no país ([MUSEU NACIONAL, 2020](#)). Atualmente, é uma instituição autônoma integrante do Fórum de Ciência e Cultura da Universidade Federal do Rio de Janeiro (UFRJ). Com uma trajetória de mais de dois séculos, o Museu Nacional/UFRJ desempenhou um papel fundamental na coleta e registro de objetos de importância em diversas áreas das ciências naturais e antropológicas.

As diversas coleções do Museu Nacional/UFRJ incorporam uma extensa variedade de artefatos, peças, vestígios e fósseis, desempenhando um papel fundamental em diversas áreas do conhecimento, incluindo a biodiversidade e o estudo de vários grupos de seres vivos. Nesse contexto, a coleção científica zoológica emerge como uma das formas de pesquisa dedicada à biodiversidade, composta principalmente por espécimes de animais conservados. Vale ressaltar que o material de estudo primário não consiste em bases de dados tabelares ou em imagens, ao invés disso, reside nos próprios exemplares de espécies ou em lotes destes (veja um exemplo na Figura 2), coletados e preservados com cuidado pelos pesquisadores do Museu Nacional/UFRJ.

Para documentar e manter a coleção científica zoológica de biodiversidade, são utilizados os Registros Primários de Biodiversidade (PBR), compostos por planilhas que contêm metadados de cada exemplar. Devido à sua vasta diversidade, a coleção é

Figura 2 – Imagem coleção



Fonte: Acervo pessoal

subdividida em coleções menores, cada uma gerenciada por grupos de especialistas em áreas específicas de pesquisa. Estas subcoleções refletem categorias taxonômicas, como Classe ou Filo, resultando em coleções específicas, tais como Crustacea, Annelida, Répteis, Aves, entre outras. Parte desse acervo é acessível em bancos de dados online, como o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr), o Global Biodiversity Facility (GBIF) e o SpeciesLink. Essa abordagem permite uma gestão eficaz da diversidade biológica e a disseminação de conhecimento através de plataformas globais de biodiversidade.

Durante a condução da pesquisa, contamos com a valiosa contribuição dos Departamentos de Invertebrados e de Vertebrados, concentrando-nos especialmente nos setores de Annelidologia¹, Carcinologia², Herpetologia³ e Ornitologia⁴. Ao longo desse processo, esses setores compartilharam conosco seus Registros Primários de Biodiversidade (PBRs), acompanhados por imagens relevantes de espécimes das suas coleções.

3.1 Estrutura dos Metadados

Cada coleção (Crustacea, Annelida, Répteis, Aves, etc.) possui PBR que abrange todos os metadados relevantes para o grupo específico de espécimes. No entanto, é importante notar que nem todos os dados pertinentes a um grupo são aplicáveis a outros. Por exemplo, a coleção de Crustacea é uma coleção mega-diversa, apresentando divisões taxonômicas específicas para esse grupo, como a Subordem, que é um metadado relevante para essa categoria, mas não tem aplicabilidade para grupos como Annelida ou Aves. Outro exemplo é a consideração de dados geográficos, como profundidade e altitude, nos metadados. Enquanto dados geográficos como altitude são relevantes para regiões continentais, a profundidade é mais significativa para regiões oceânicas, demandando uma abordagem diferenciada desses metadados conforme a natureza das regiões coletadas.

Apesar dessas particularidades, é importante ressaltar que elas são específicas de cada coleção e não são amplamente presentes nas planilhas. De maneira geral, todas as planilhas compartilham uma majoritária interseção de metadados. A maioria dos dados é semelhante para todas as coleções. Em todas as Registros Primários de Biodiversidade(PBR) referentes às coleções, identificamos cinco categorias distintas de dados, sendo elas:

- **Dados Catalográficos:** Incluem informações relativas à catalogação, como o número de catálogo, a data de catalogação, a condição do objeto, entre outros.

¹ Annelidologia refere-se à área da biologia dedicada ao estudo dos anelídeos, um filo de animais invertebrados que inclui minhocas, sanguessugas e poliquetas.

² Carcinologia é a disciplina biológica que se dedica ao estudo dos crustáceos, abrangendo a diversidade, ecologia, morfologia e taxonomia desses animais, como caranguejos, camarões, lagostas e cracas.

³ Herpetologia foca no estudo dos répteis e anfíbios, abrangendo uma variedade de tópicos relacionados à biologia desses animais.

⁴ Ornitologia é o campo científico dedicado ao estudo das aves, investigando sua biologia, comportamento e ecologia para ampliar nosso entendimento e conservação dessas importantes espécies aladas.

- **Dados Taxonômicos:** Englobam dados relacionados à classificação taxonômica, como reino, filo, classe, ordem, família, gênero, espécie do espécime, nome do(s) pesquisador(es) responsável(is) pela identificação, data da identificação, e se o espécime é um tipo⁵.
- **Dados de Coleta:** Contêm informações sobre a coleta do espécime, como a data de coleta e o nome do(s) pesquisador(es) responsável(is) pela coleta.
- **Dados Geográficos:** Incluem informações sobre a localização em que o espécime foi encontrado, como latitude e longitude, profundidade ou altitude, cidade, estado, país, além de características específicas, como se o animal foi encontrado em água salgada ou doce, entre outros.
- **Dados Específicos:** Referem-se a informações específicas sobre o registro, como a solução em que o espécime está submetido para sua conservação, entre outros.

3.2 Coleções

Nesta seção, apresentamos breves descrições de cada uma das coleções, proporcionando uma visão sucinta dos animais que compõem cada uma delas.

Os répteis, cujo nome deriva do latim 'reptare' (rastejar), são notáveis pela sua adaptação a ambientes terrestres. Representando a classe Reptilia, esses animais de sangue frio e frequentemente escamados reproduzem-se por ovos. A Figura 3 ilustra alguns espécimes da coleção de répteis.

Figura 3 – Exemplares do acervo de répteis do setor de Herpetologia do MN/UFRJ



(a) *Salvator duseni*

(b) *Micrurus carvalhoi*

Fonte: <<http://www.herpetologiamuseunacional.com.br/galeria.html>> Acessado em 06/12/2023

Créditos: (a) Paula H. Valdujo (b) Thiago Silva-Soares

⁵ **Espécime tipo** é um exemplar ou conjunto de exemplares que serve como referência padrão para a descrição formal de uma nova espécie. O tipo é fundamental para garantir a consistência e a estabilidade na nomenclatura científica. Existem diferentes categorias de tipos, como holótipo, parátipo, e neótípico, cada um com uma função específica. Esses espécimes são guardados em instituições científicas e são referências oficiais para a identificação e estudo de uma espécie.

Os crustáceos, que incluem caranguejos, camarões e lagostas, fazem parte da classe Crustácea. O nome "crustáceo" tem origem no latim 'crustaceus', referindo-se à carapaça rígida que protege seus corpos. Esses artrópodes, muitas vezes encontrados em ambientes aquáticos, exibem uma notável variedade de formas e adaptações. A Figura 4 destaca espécimes da coleção de crustáceos.

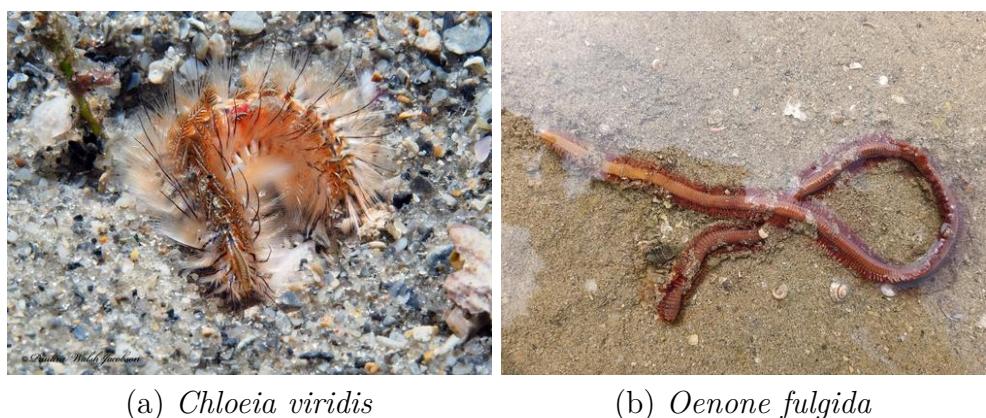
Figura 4 – Exemplares da coleção de Crustáceos



Fonte: Imagens cedidas pelo departamento de Carcinologia

Os anelídeos pertencem ao filo Annelida, e seu nome tem origem no grego 'anelos' (anéis). A segmentação distinta em anéis de seus corpos é uma característica marcante. A Figura 5 destaca alguns espécimes da coleção de anelídeos.

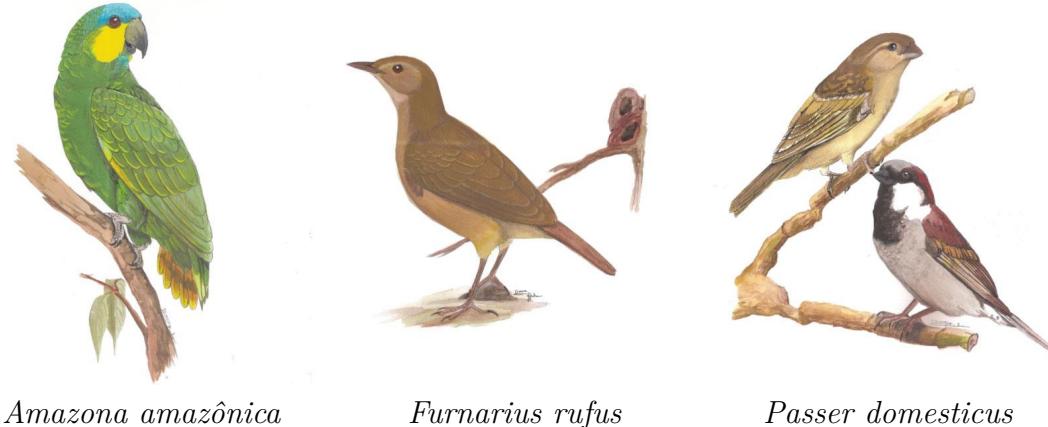
Figura 5 – Espécimes de anelídeos similares aos presentes na coleção de Annelidologia



Fonte: (a) <<https://inaturalist-open-data.s3.amazonaws.com/photos/167961510/medium.jpg>> Acessado em 06/12/2023
 (b) <<https://inaturalist-open-data.s3.amazonaws.com/photos/287249025/medium.jpeg>> Acessado em 06/12/2023

O termo 'ave', proveniente do latim 'avis', reflete a ancestralidade e singularidade desses seres. Desde majestosas águias até encantadores colibris, a coleção oferece um mergulho profundo na biologia, comportamento e ecologia dessas criaturas aéreas. A Figura 6 destaca exemplares da classe de aves.

Figura 6 – Exemplares de aves presentes na coleção de Ornitologia



Amazona amazônica *Furnarius rufus* *Passer domesticus*

Fonte: <<https://www.museunacional.ufrj.br/hortobotanico/Aves.html>> Acessado em 06/12/2023

Créditos: Diana Rocha Monteiro dos Santos

Dado um breve resumo de cada uma das coleções, a tabela a seguir contabiliza cada um dos grupos taxonômicos das coleções que tivemos acesso:

Tabela 1 – Resumo Taxonômico Quantitativo das Coleções

	Crustacea		Annelida		Repteis		Aves	
	Quantos	NaN	Quantos	NaN	Quantos	NaN	Quantos	NaN
Reino	1	9	1	0	1	12	1	6197
Filo	1	9	1	0	1	12	1	6197
Classe	10	604	3	0	2	12	1	6197
Ordem	43	1.493	13	233	6	32	24	6197
Infraordem	26	15.976	-	-	-	-	-	-
Subordem	34	6.376	-	-	6	36	-	-
Família	418	4.341	58	343	52	33	60	1
Gênero	1.308	10.433	304	2.413	278	63	583	2
Espécie	2.279	15.133	700	2463	893	514	1.348	81
Nº Registros	30.708	-	7.488	-	23.132	-	30.710	-

Fonte: Elaboração própria, baseado nas coleções

3.3 Pré-processamento e tratamento de dados

Nessa parte do trabalho, seguimos a abordagem proposta por Franklin Oliveira em sua dissertação de mestrado(OLIVEIRA, 2021) e, também, no trabalho de refatoração e transformação do código em funções para a utilização na versão inicial do Vis-Zoo(MEDEIROS, 2023). Durante minha Iniciação Científica e posteriormente, durante a realização deste trabalho, as funções sofreram ajustes e melhorias oportunas, que ajudaram no melhor aproveitamento dos dados, detalhadas no relatório da Iniciação Científica.

O processo inicia-se com a abertura dos PBRs do MN/UFRJ pelo script Python, selecionando a planilha que contém os metadados registrados e procedendo com seu tratamento. Durante essa etapa, são realizadas as seguintes operações: (1) concatenação de nomes originalmente armazenados como 'Primeiro Nome' e 'Sobrenome', (2) separação dos dados temporais em mês e ano em vez de uma data completa, e (3) transformação das latitudes e longitudes, passando de $[90^{\circ}S, 90^{\circ}N]$ para $[-90, 90]$ e de $[180^{\circ}W, 180^{\circ}E]$ para $[-180, 180]$.

Nesta etapa, há também uma seleção da classificação taxonômica mais atual. Isso ocorre devido à presença de mais de um conjunto de **Dados Taxonômicos** nas bases de dados. Geralmente, essas bases possuem todos os dados taxonômicos com o índice 1 e, em seguida, a repetição desses campos com o índice 2. Essa duplicidade reflete a dinâmica da classificação, sujeita a alterações ao longo do tempo. Por exemplo, se é identificado que uma espécie está com os dados imprecisos (como o caso do cachorro em 1993⁶), os dados taxonômicos são preenchidos de forma desatualizada e precisam ser atualizados, logo, a correção é implementada nos dados com o próximo índice disponível. Durante o processo de escolha do dado atualizado, foi incorporada uma verificação de 'preenchido' ou 'vazio' no índice mais alto, se os dados estiverem 'preenchidos', então os dados com índice 2 são utilizados; caso contrário, os dados com o índice 1 são mantidos, assegurando a retenção apenas do dado taxonômico mais atualizado.

Para finalizar o processo de pré-processamento, realiza-se uma filtragem dos dados, semelhante às abordagens anteriores. Nessa limpeza, são removidos os **Dados Específicos** e outros dados que, por enquanto, não são prioritários para análise. Essa etapa visa manter uma base de dados mais enxuta para evitar consumo excessivo de memória e facilitar procedimentos subsequentes.

⁶ Até o ano de 1993, o cachorro era considerado da espécie *Canis familiaris*, porém foi reclassificado como sendo uma subespécie do *Canis lupus*, sendo portanto: *Canis lupus familiaris*

4 Roteiro

Com a apresentação dos dados, é hora de abordarmos a criação da visualização. Desde meu programa de Iniciação Científica, a proposta de desenvolver uma visualização hierárquica de dados taxonômicos, utilizando as informações disponíveis, tem sido um desafio em aberto e agora é o objetivo deste trabalho.

Antes de entrarmos na renderização de nós e arestas, há uma etapa preliminar, considerada crucial pela literatura de visualização. Em *Storytelling com Dados*, Knaflic ([KNAFLIC, 2018](#)) destaca a importância de definir claramente alguns pontos de partida antes de iniciar a criação de conteúdo. Apesar do enfoque do livro ser em negócios, esses princípios são aplicáveis em diversos contextos. Os pontos de partida estão relacionados a quem/o que/como os dados devem comprovar uma informação.

Explorando ainda mais a literatura de visualização, em *Visualizing Data*, Kirk ([KIRK, 2023](#)) também aponta a necessidade de definir os pontos de partida. Além disso, destaca a importância de manter rigor nessa fase, pois atalhos podem acarretar em impactos negativos. Apesar das semelhanças, o processo proposto por Kirk, o '*Briefing*', abrange dimensões, como (1) Motivações (por que criar a visualização?), (2) Circunstâncias (Pessoas, Regras, Entregáveis, Ferramentas), (3) Expectativa de visualização (tipo da visualização) e (4) Aproveitamento de ideias (utilizar grandes visualizações como base para criar a sua).

Mesmo diante de divergências nas sugestões desses autores, ambos convergem para a importância de uma etapa introdutória que alinhe expectativas e oriente os objetivos do trabalho. Essa etapa é crucial, pois guia a criação da visualização, evitando que ela seja fruto do acaso ou criada por um evento totalmente aleatório. No entanto, roteiros ruins podem ser traçados; nesse caso, alterações e desvios de rota são válidos e necessários para alcançar um bom resultado final.

Baseados no modelo de Kirk ([KIRK, 2023](#)), mais completo e adaptado à nossa situação, neste capítulo, delinearemos o *Briefing* de como este trabalho será desenvolvido. No entanto, nem todos os passos de um *Briefing* padrão serão abordados, pois o escopo da visualização já está definido.

4.1 Motivações

No âmbito das motivações, convergem diferentes impulsos para a criação desta visualização. Os curadores possuem coleções científicas de biodiversidade e a representação por meio da taxonomia é uma maneira possível e ainda não explorada no âmbito do projeto

Viszoo. Portanto, têm a motivação de empregar essa abordagem visual. Desde o início, a presença de dados hierárquicos tem sido destacada, despertando o desejo de criar uma visualização distinta das que foram utilizadas e manipuladas durante a Iniciação Científica. Assim, com a união dessas motivações, emerge um objetivo.

4.2 Circunstâncias

Neste cenário, é fundamental ressaltar que a visualização terá dois tipos de usuários distintos: os curadores, que utilizarão as bases de dados como ferramenta de verificação, e os futuros interlocutores, que enxergarão as visualizações como apresentação das bases de dados. No entanto, concentraremos nossa atenção em um usuário que possuirá conhecimento suficiente para compreender o formato de árvore moldada pela taxonomia. Além disso, é relevante destacar que o grafo deve respeitar a taxonomia das espécies, e a manipulação de dados também deve seguir um padrão que não se afaste muito das normas científicas.

Nas duas subseções subsequentes, apresentaremos uma contextualização dos dados e da ferramenta, respectivamente.

4.2.1 Dados Utilizados

Dado que o objetivo é criar uma visualização hierárquica dos dados taxonômicos, os dados utilizados serão os próprios dados taxonômicos. Contudo, os dados catalográficos (de identificação) também foram preservados para uso, visto que sua inclusão tornará mais prática a identificação de possíveis incoerências.

4.2.2 Ferramenta

No contexto da visualização de grafos, uma área que frequentemente carece de bibliotecas especializadas em Python, a escolha de uma ferramenta torna-se ainda mais importante. VanderPlas et al. ([VANDERPLAS et al., 2018](#)) apresentaram o Altair([VEGA-ALTAIR, s.d.](#)) Python como uma biblioteca declarativa para visualização estatística, baseada na gramática de visualização interativa Vega-Lite. A escolha do Altair é ancorada na necessidade de representar interativamente as complexas relações interconectadas presentes nos dados.

É relevante destacar que algumas bibliotecas de visualização em Python, ao lidarem com a representação de grafos, frequentemente realizam transformações nos dados antes de representá-los graficamente. Posteriormente, elas recorrem a bibliotecas tradicionais para desenhar pontos e linhas. Esse processo é executado pelo NetworkX, que utiliza o Pyplot do Matplotlib para desenhar os grafos. O Altair, por sua vez, é capaz de lidar com essa complexidade de criar os gráficos, desde que os cálculos sejam feitos.

Um ponto crucial para a escolha do Altair é sua utilização na dissertação de mestrado de Franklin Oliveira ([OLIVEIRA, 2021](#)). A familiaridade prévia com a biblioteca e sua aplicação bem-sucedida em etapas anteriores, como na Iniciação Científica, fornecem uma base sólida para explorar suas capacidades na visualização de dados.

4.3 Expectativa

Pelo contato dos dados não ser recente e, principalmente, por já existir o desejo da representação hierárquica, o formato de representação já era de alguma forma pré-definido. De forma que inicialmente fosse uma ferramenta para verificação da base de dados, cumprindo a função exploratória, e num segundo momento fosse uma visualização capaz de apresentar a taxonomia da coleção, exercendo um papel mais explanatório. Além disso, a ideia era previamente conhecida e consistia de criar a visualização em grafo.

Os grafos, estruturas matemáticas com vértices e arestas, são amplamente usados na ciência da computação e teoria dos grafos, oferecendo uma abstração visual poderosa para análise e solução de problemas complexos, modelando relações entre entidades em diversas situações.

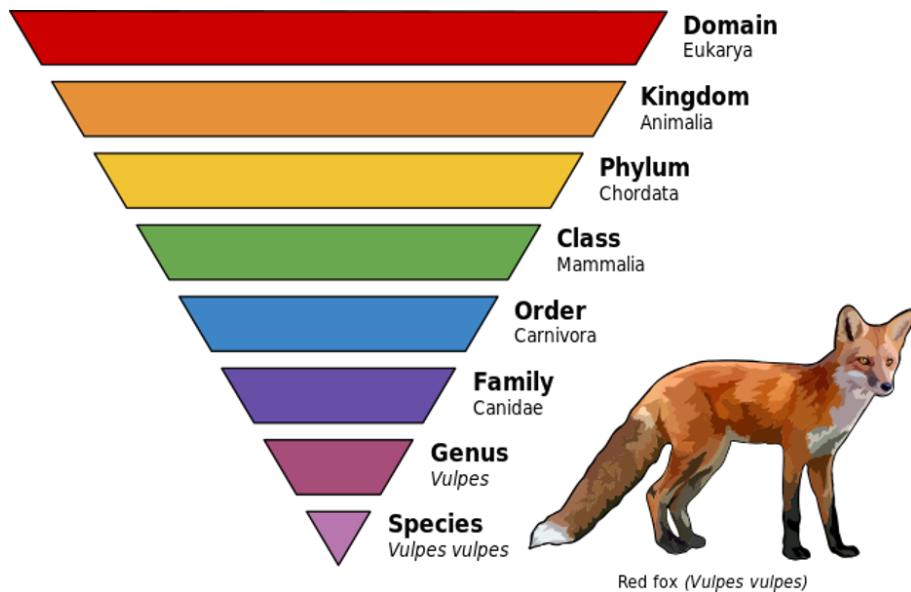
Representar os dados taxonômicos através de grafos seria definir cada um dos nomes de cada um dos grupos taxonômicos no grupo de vértices, enquanto uma aresta existiria apenas quando dois nós fizessem parte de um grupo taxonômicos distintos de um mesmo espécime, sendo que os dois grupos taxonômicos devem estar diretamente conectados na classificação (uma representação da taxonomia pode ser vista na Figura 7). Por exemplo, 'Gênero' conecta com 'Espécie' e 'Família', mas não se conecta com 'Filho', 'Classe' ou 'Ordem'. Apesar de estar parcialmente resolvido, ainda é necessário um estudo e um tratamento de como posicionar cada um dos nós, que será tratado no próximo capítulo.

Ao analisar o caso aprofundadamente, nota-se que todo e qualquer grafo, gerado sob tais premissas e com uma base de dados sem ruídos, na verdade, será uma árvore. Uma árvore em teoria dos grafos é um tipo especial de grafo acíclico, onde cada vértice está conectado a outro por uma única aresta, e não há ciclos. Essa estrutura hierárquica e sem *loops* torna as árvores uma representação eficaz para diversas aplicações, desde estruturas de dados até modelagem de relações em sistemas complexos.

A conclusão de que esses gráficos são árvores é obtida a partir do raciocínio de que elementos do grupo taxonômico mais amplo podem ter relação com vários elementos do grupo taxonômico menos amplo, porém, os elementos de um grupo menos amplo só podem ter conexão com apenas um elemento do grupo mais amplo. Novamente utilizando a taxonomia para exemplificar (veja Figura 7), uma 'Família' possui vários 'Gêneros', porém cada um dos vários 'Gêneros' possui apenas uma 'Família'.

Com a definição de que o trabalho será criar árvores, foi necessário definir qual o

Figura 7 – Representação da Taxonomia



Fonte: <https://en.wikipedia.org/wiki/Taxonomic_rank#/media/File:Taxonomic_Rank_Graph.svg> Acessado em 03/12/2023

seu formato. Neste contexto, o projeto '*Tree Vis*' (SCHULZ, 2011) facilita essa decisão, pois ele faz uma revisão atualizada da literatura de árvores utilizadas em artigos científicos e até mesmo fontes na *Internet* em geral e classifica e as exemplifica. Nesse contexto, escolhemos trabalhar com estruturas de árvore clássicas, com uma representação em 2 dimensões e sendo uma árvore explícita.

4.4 Fontes de Inspiração

Neste estágio, é relevante documentar as visualizações que desempenharam um papel significativo na concepção deste projeto. O projeto denominado '*Tree Vis*' (SCHULZ, 2011) e o '*Bio Vis Explorer*' (KERREN et al., 2017) são notáveis referências que proporcionaram *insights* valiosos.

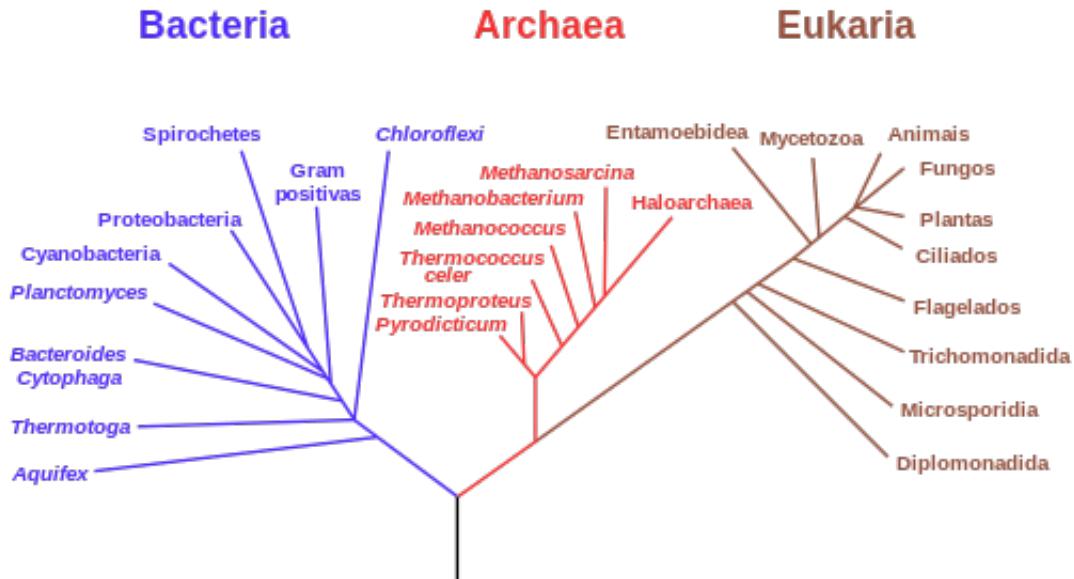
Destaca-se, especialmente, a influência das árvores filogenéticas, que desempenharam um papel crucial de referência visual no desenvolvimento deste trabalho. Um exemplo concreto dessa influência pode ser observado na representação gráfica apresentada na Figura 8. Essas representações hierárquicas são fundamentais para compreender as relações evolutivas entre diferentes espécies.

Além disso, a teoria dos grafos desempenha um papel significativo, sendo uma influência direta no design e na concepção do projeto. Os grafos, por si só, representam uma poderosa ferramenta, como evidenciado na Figura 9, onde a estrutura de nós e arestas é aplicada de maneira a elucidar conexões complexas entre entidades.

Essas inspirações provenientes de diversas fontes contribuíram para moldar a visão

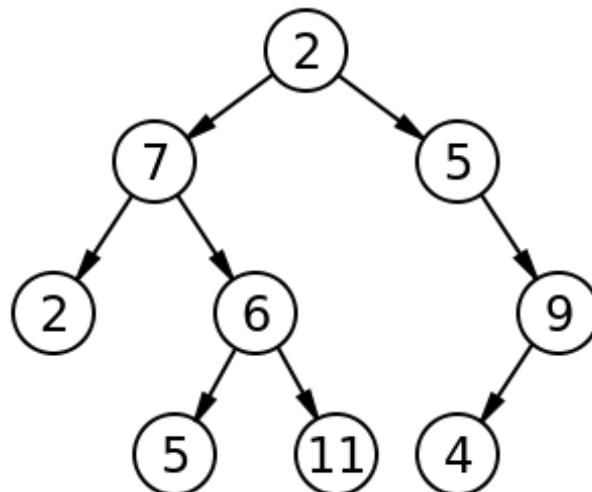
e abordagem adotada neste projeto, fornecendo um alicerce sólido para a criação de uma representação visual eficaz e informativa da taxonomia da coleção em questão.

Figura 8 – Representação de uma Árvore Filogenética



Fonte: <https://upload.wikimedia.org/wikipedia/commons/thumb/7/70/Phylogenetic_tree.svg/675px-Phylogenetic_tree.svg.png> Acessado em 04/12/2023

Figura 9 – Representação de uma Árvore em Teoria dos Grafos



Fonte: <https://upload.wikimedia.org/wikipedia/commons/f/f7/Binary_tree.svg> Acessado em 04/12/2023

5 Execução

Após a conclusão da fase de *Briefing*, delineando as estratégias e decisões-chave para a condução do projeto, este capítulo marca a transição para a execução efetiva do trabalho. Aqui, abordaremos a etapa prática de manipulação de dados e a criação de uma representação gráfica da taxonomia dos espécimes biológicos das Coleções Científicas do Museu Nacional/UFRJ.

A manipulação de dados será explorada em detalhes, enfrentando os desafios inerentes à natureza complexa dos conjuntos de dados biológicos. Em seguida, iniciaremos a construção do grafo taxonômico, utilizando a biblioteca Altair para a visualização informativa das relações taxonômicas.

Ao final deste capítulo, teremos superado os obstáculos práticos da manipulação de dados e criação de gráficos. Nesse processo, a representação visual resultante será não apenas uma ilustração, mas uma ferramenta para comunicar e compreender a riqueza taxonômica presente nos PBRs do MN/UFRJ.

5.1 Tratamento do dado até virar um grafo

Nesta seção, abordaremos a etapa fundamental de manipulação de dados no contexto da construção do grafo taxonômico com base nas coleções do Museu Nacional/UFRJ. A manipulação eficiente de dados é uma competência central na ciência de dados, possibilitando a extração de informações valiosas de conjuntos de dados complexos. Enfrentando o desafio inerente à incompletude dos dados biológicos, exploraremos estratégias práticas, como o preenchimento de lacunas. Adicionalmente, discutiremos transformações especiais necessárias para adequar os dados a formatos específicos, adequando os para etapas subsequentes na construção do grafo taxonômico. Este processo é crucial, uma vez que a manipulação eficaz de dados desempenha um papel vital na descoberta de informações.

5.1.1 Dados Faltantes

No que diz respeito à manipulação de dados, a ausência completa de dados faltantes é praticamente uma utopia. Em bases de dados reais, é praticamente impossível não haver dados incompletos. Portanto, uma etapa crucial é o tratamento desses dados, envolvendo processos como exclusão do registro com dados faltantes, aplicação de heurísticas para preenchimento dos dados ausentes ou consideração do dado faltante como tal. Cada técnica apresenta vantagens e desvantagens, sendo escolhida conforme a necessidade da modelagem e do problema a ser resolvido.

Nas coleções científicas de biodiversidade, a presença de dados faltantes é evidente, como ilustrado na Tabela 1. No mestrado de Franklin Oliveira ([OLIVEIRA, 2021](#)) e em trabalhos subsequentes, adotou-se a abordagem de registrar os dados faltantes como 'Não Identificado' em casos nominais e 'N/A' em dados temporais. Essa abordagem cumpre o objetivo desses trabalhos, auxiliando pesquisadores na detecção de possíveis incoerências em suas bases de dados.

Embora seja uma abordagem válida, ela apresenta complicações para o contexto desse trabalho, que podem ser evitadas com uma modelagem sucintamente diferente. Nessa perspectiva, propomos uma nova abordagem inspirada em uma prática usual na área de biologia. Quando a identificação completa da taxonomia de um espécime não é possível, adota-se a sigla '*sp*' para representar sua espécie. Por exemplo, se um animal pertence ao gênero '*Vulpes*', mas a espécie não pode ser identificada como raposa vermelha ('*Vulpes vulpes*') ou raposa das estepes ('*Vulpes corsac*'), ele é registrado como '*Vulpes sp*'.

Essa nova abordagem aplica o conceito da biologia discutido anteriormente e o estendendo para outros grupos taxonômicos, preenchendo '*Gen*' para gênero, '*Fam*' para família e utilizando siglas específicas para demais grupos. Importante salientar que, quando um espécime possui gênero indefinido, essa condição também é refletida no campo de espécie, sendo designado como '*Genus sp*'.¹ Cada categoria taxonômica possui um código representativo para o vazio, assegurando a representação de espécimes com múltiplos elementos não identificados em sua taxonomia.

Para prevenir conflitos durante a elaboração do grafo, especialmente ao lidar com dois espécimes de famílias distintas e gêneros não identificados^{2,3}, acrescenta-se o nome do grupo taxonômico de nível superior mais próximo e originalmente preenchido. Por exemplo, um espécime da ordem Primatas com família e gênero indefinidos teria o campo de família preenchido com '*Fam - Primates*', gênero com '*Gen - Primates*' e espécie com '*Gen sp - Primates*'. Essa estratégia se baseia na nomenclatura das espécies, que incorpora o nome do seu gênero.

Assim, a nova abordagem proposta incorpora a prática biológica de designar espécies não identificadas com a sigla '*sp*' e estende esse conceito para outros grupos taxonômicos. O preenchimento de elementos ausentes com códigos específicos, como '*Gen*' e '*Fam*', mantém a integridade da representação taxonômica, mesmo em situações em que diversos níveis da hierarquia não são identificados. Além disso, a inclusão do nome do grupo taxonômico de nível superior originalmente preenchido reduz possíveis ambiguidades, garantindo que a proposta não apenas lide eficazmente com dados faltantes, mas também preserve a precisão e clareza na representação taxonômica, minimizando interpretações equivocadas do grafo em construção.

¹ O nome das espécies é formado por *Genus species*

² os dois espécimes teriam o mesmo gênero ('*Gen*')

³ Pode ser quaisquer outros grupos taxonômicos, desde que o segundo está contido no primeiro

5.1.2 Agregação dos dados

Tabela 2 – Exemplo de agrupamento dos dados

(a) Formato original

Registro	Reino	Filo	Classe	Ordem	Família	Gênero	Espécie
EXP001	Animalia	Chordata	Mammalia	Carnivora	Canidae	Canis	Canis lupus
EXP002	Animalia	Chordata	Mammalia	Carnivora	Canidae	Vulpes	Vulpes vulpes
EXP003	Animalia	Chordata	Mammalia	Primates	Hominidae	Homo	Homo sapiens
EXP004	Animalia	Chordata	Mammalia	Carnivora	Canidae	Canis	Canis lupus
EXP005	Animalia	Chordata	Mammalia	Carnivora	Canidae	Vulpes	Vulpes vulpes
EXP006	Animalia	Chordata	Mammalia	Carnivora	Canidae	Canis	Canis lupus
EXP007	Animalia	Chordata	Mammalia	Primates	Hominidae	Homo	Homo sapiens
EXP008	Animalia	Chordata	Mammalia	Carnivora	Canidae	Canis	Canis lupus
EXP009	Animalia	Chordata	Mammalia	Carnivora	Canidae	Vulpes	Vulpes vulpes

(b) Formato após o agrupamento

Reino	Filo	Classe	Ordem	Família	Gênero	Espécie	Count
Animalia	Chordata	Mammalia	Carnivora	Canidae	Canis	Canis lupus	4
Animalia	Chordata	Mammalia	Carnivora	Canidae	Vulpes	Vulpes vulpes	3
Animalia	Chordata	Mammalia	Primates	Hominidae	Homo	Homo sapiens	2

Fonte: Elaboração própria

Ainda no âmbito da manipulação de dados, optamos pela agregação dos dados em todas as categorias taxonômicas. Essa escolha se justifica pelo fato de que muitos espécimes são da mesma espécie, consequentemente das mesmas categorias taxonômicas. De forma em que os dados, quando apresentados no grafo que estamos construindo, estarão se sobrepondo⁴. Mais adiante, apresentaremos uma estratégia específica para exibir informações no nível de registro e não perder nenhuma informação importante. O processo mencionado pode ser ilustrado na Tabela 2, onde (a) exemplifica dados semelhantes aos encontrados nos PBRs. Enquanto em (b), é apresentado como a tabela (a) seria reorganizada após esse processo.

É relevante salientar que a técnica de agrupamento adotada consiste na contagem da quantidade de registros e de espécimes tipo para cada combinação de categorias taxonômicas. Por exemplo, a contagem considera não apenas a família ou o gênero isoladamente, mas a combinação de todas as categorias taxonômicas relevantes. Reunindo apenas registros com a taxonomia exatamente igual e evitando um esforço de processamento desnecessário.

5.1.3 Transformação em Grafos

A transição dos dados originalmente tabulares para uma representação em grafo, como a estrutura de uma árvore taxonômica, envolve a escolha cuidadosa das estratégias

⁴ Essa sobreposição é esperada, pois os registros taxonômico de uma única espécie devem ser iguais, não fazendo sentido a repetição e o excesso no custo computacional

Tabela 3 – Exemplo da transformação para dados de grafo

Nós						Arestas		
Indice	Nome	Taxo	Eixo1	Eixo2	Count	Index	Nó1	Nó2
E1	Canis lupus	Espécie	7	1	4	a01	E1	G1
E2	Vulpes vulpes	Espécie	7	2	3	a02	E2	G2
E3	Homo sapiens	Espécie	7	3	2	a03	E3	G3
G1	Canis	Gênero	6	1	4	a04	G1	Fa1
G2	Vulpes	Gênero	6	2	3	a05	G2	Fa1
G3	Homo	Gênero	6	3	2	a06	G3	Fa2
Fa1	Canidae	Família	5	1.5	7	a07	Fa1	O1
Fa2	Hominidae	Família	5	3	2	a08	Fa2	O2
O1	Carnivora	Ordem	4	1.5	7	a09	O1	C1
O2	Primates	Ordem	4	3	2	a10	O2	C1
C1	Mammalia	Classe	3	2.25	9	a11	C1	Fi1
Fi1	Chordata	Filo	2	2.25	9	a12	Fi1	R1
R1	Animalia	Reino	1	2.25	9			

Fonte: Elaboração própria

Nota: Transformação dos dados da Tabela 2 em lista de nós e de vértices.

e técnicas adequadas. Existem várias maneiras de armazenar grafos na memória de um computador, cada uma com suas próprias características. As abordagens mais comuns incluem o uso de Matriz de Adjacência, Matriz de Incidência e Lista de Nós e Arestas. Diante da natureza tabular dos dados e da necessidade de incluir variáveis adicionais, optamos pela última abordagem, utilizando a flexibilidade da arquitetura de DataFrame do Pandas([TEAM, 2020](#)) para manipulação eficiente dos dados.

A construção das listas de nós e arestas começou com a adição de cada elemento de cada categoria taxonômica à lista de nós, de forma que cada registro fosse único. Isso impediu a duplicação desnecessária de elementos, fornecendo uma base sólida para a estrutura do grafo. Para a lista de arestas, todos os conjuntos de 2 elementos de grupos taxonômicos consecutivos em cada registro foram considerados. Essa representação de arestas permitiu capturar as relações entre os diferentes níveis taxonômicos de forma clara.

A escolha da disposição geométrica dos nós para a representação gráfica envolveu a definição de um eixo para cada categoria taxonômica. Assim, os valores do Eixo 1 foram atribuídos com base na hierárquica das categorias taxonômicas, proporcionando uma estrutura ordenada e intuitiva de que nós irmãos em uma geração do grafo seriam elementos de um mesmo grupo taxonômico. No entanto, o eixo 2 apresentou um desafio adicional, pois não havia uma ordenação natural como a taxonomia. Nesse contexto, optou-se por atribuir valores sequenciais aos elementos do grupo taxonômico mais subdividido(no caso geral Espécie) e calcular métricas, como a mediana dos valores nos nós filhos, para atribuir valor para os elementos dos outros grupos taxonômicos.

Ao discutir as estratégias para atribuir valores ao eixo 2, testou-se inicialmente a média dos valores do eixo 2 dos subgrupos conectados. Todavia, essa abordagem revelou-se sensível a *outliers*, podendo distorcer a estrutura do grafo. Como alternativa, a mediana foi

escolhida como base para esses cálculos, oferecendo uma medida mais resistente a extremos e, assim, preservando a integridade do grafo.

A Tabela 3 reflete o resultado final desse processo, proporcionando uma representação gráfica dos dados taxonômicos originalmente apresentados na Tabela 2. De forma a facilitar o processo de representação do grafo.

5.2 Desenho do Grafo

Esta seção explora o processo de criar uma representação gráfica da taxonomia dos espécimes biológicos das Coleções Científicas. Enfrentando a complexidade da plotagem de vértices, dividimos o processo em diversas subseções. Elas detalham a aplicação de técnicas como filtragem, inversão de eixos e integração de registros e imagens, fornecendo uma compreensão mais profunda da taxonomia e biodiversidade presentes nos PBRs do MN/UFRJ.

5.2.1 Camadas do Grafo

Na etapa de representação gráfica do grafo, a complexidade principal reside no posicionamento dos vértices e arestas, uma vez que os *frameworks* facilitam a transformação dos dados em elementos visuais. A geometria para posicionar os vértices já foi definida, assim, para criar o grafo, é necessário gerar dois gráficos interconectados: um gráfico de pontos para representar os nós e um gráfico de linhas para as arestas.

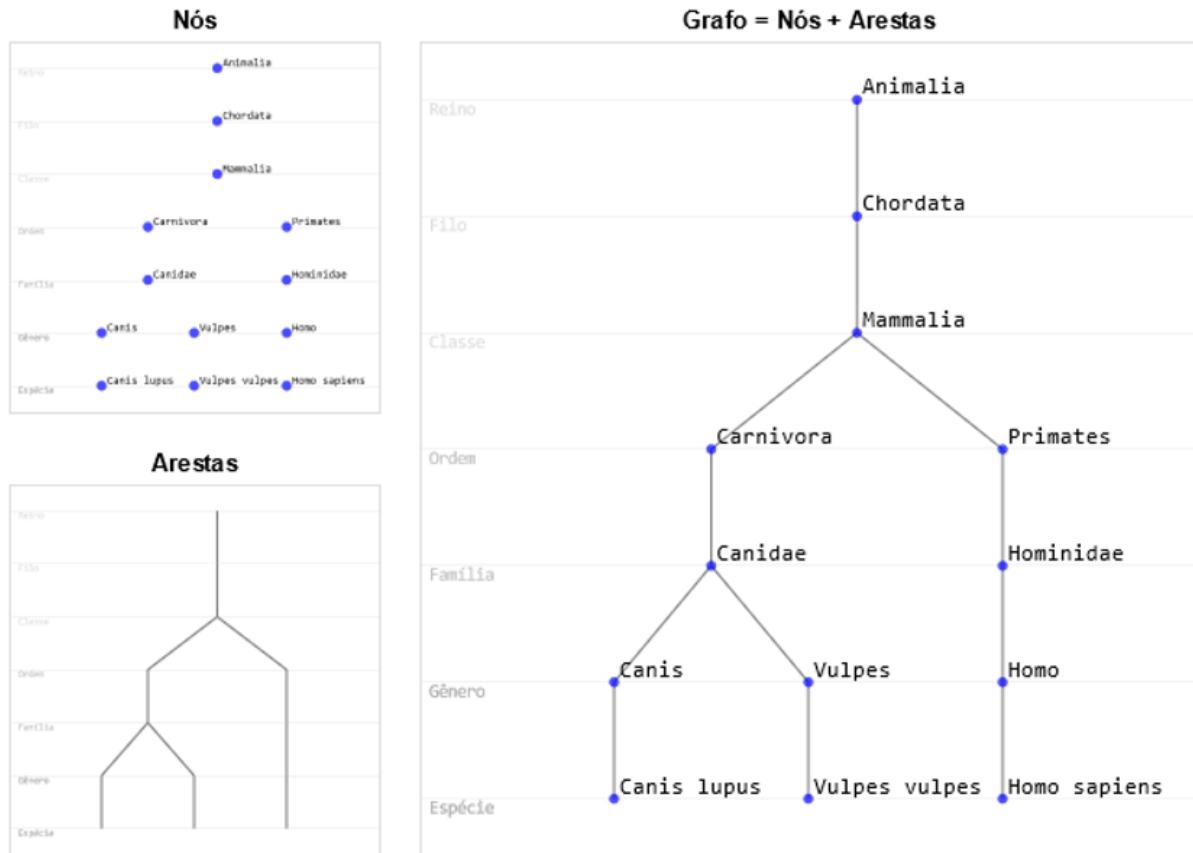
Com a geometria estabelecida, praticamente todas as bibliotecas padrão de visualização conseguem realizar a plotagem. Altair, a biblioteca escolhida desde o início, também consegue construir esses gráficos, por isso, se mostra capaz criar de grafos.

Para a criação do gráfico de nós, foi utilizado o comando *mark_point*, eficaz na geração de gráficos de dispersão. Nessa parte da visualização, o Eixo 1 foi atribuído ao eixo Y, e o Eixo 2 ao eixo X, determinando as posições dos pontos. Além disso, o nome, a categoria taxonômica, e os dados calculados na transformação de dados (quantidade de registros e quantidade de registros tipo) foram incluídos na *tooltip* de cada elemento.

No mesmo contexto, para gerar as arestas, empregou-se o comando *mark_rule*, que, embora comumente utilizado para *annotations* e marcações específicas nos gráficos, adapta-se facilmente aos nossos dados de arestas. Esse comando requer dois conjuntos de coordenadas, gerando uma linha que conecta essas duas coordenadas, exatamente a aplicação que queremos e os dados que temos.

Finalizando o processo, é necessário agrupar os gráficos em um único espaço e combiná-los. Na área de Visualização da Informação, esse conceito é conhecido como sobreposição de gráficos, ocorrendo quando dois gráficos compartilham o mesmo espaço,

Figura 10 – Exemplo de Grafo Gerado a partir das Listas de Nós e Arestas



Fonte: Elaboração própria.

Nota: A figura representa a formação do grafo a partir do exemplo da Tabela 3. Apresenta os nós e as arestas separadamente e a união dos dois, formando o grafo.

e os elementos de um se sobrepõem aos do outro. A ferramenta escolhida facilitou esse procedimento, sendo robusta e auxiliando na criação dessa técnica.

A Figura 10 representa o processo de criação do grafo gerado a partir das listas de nós e arestas apresentadas na Tabela 3. Na figura, estão evidenciadas a plotagem dos nós (um gráfico de dispersão), a plotagem das arestas (um gráfico de linhas) e a representação final mostra o grafo, que é a combinação desses dois elementos, plotado com sucesso.

5.2.2 Diferenças nos Comandos

Ao explorar o processo de criação da visualização, é importante destacar que o Altair possui funcionalidades distintas, cada uma com suas eficiências e aplicabilidades específicas. Essa diversidade de comandos proporciona uma experiência enriquecedora, incluindo equívocos e maus usos da ferramenta, que acabam contribuindo para o aprendizado ao longo do projeto.

No início, adotamos uma abordagem que, ao longo do desenvolvimento, revelou-se menos adequada para nosso propósito. Na tentativa inicial de traçar as arestas, utilizamos o comando '`mark_line`', utilizado para gerar linhas que conectam pontos. Inicialmente,

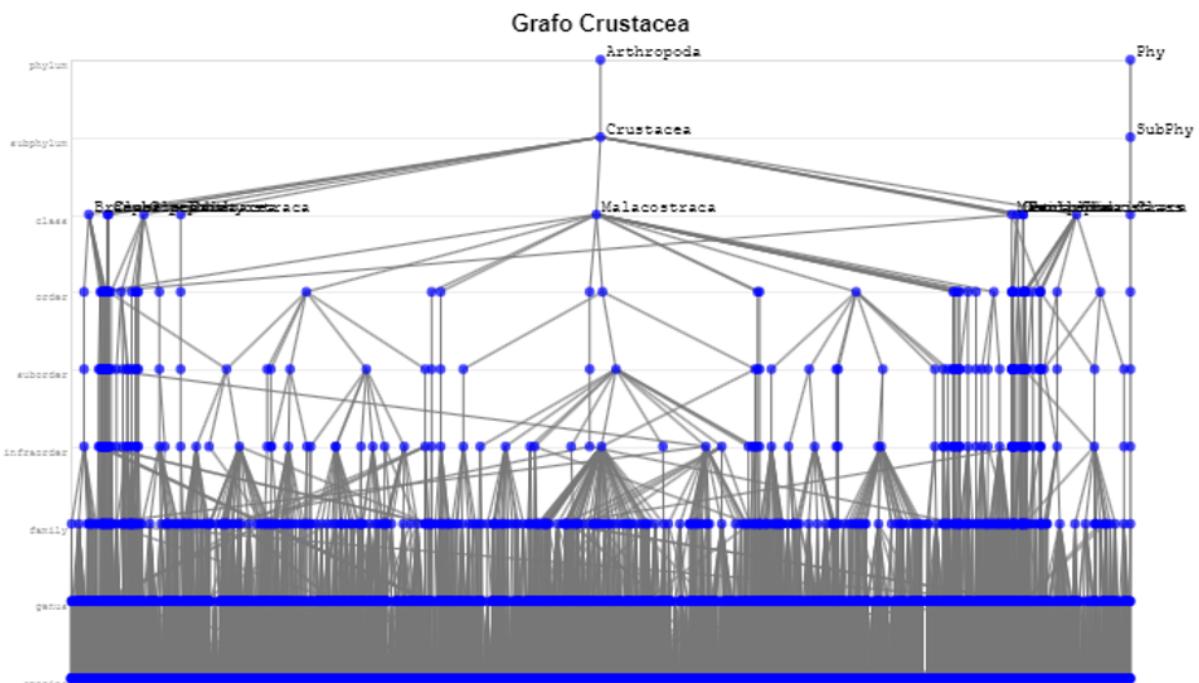
os dados de arestas foram formatados de maneira diferente, resultando em linhas que se estendiam desde o nível de Reino até o nível de Espécie. Em outras palavras, todos os grupos taxonômicos de um espécime eram conectados por uma única linha.

Visualmente, essa abordagem apresentava semelhanças com a versão final. Contudo, a execução desse comando da forma como o concebemos gerava uma única linha que conectava todas as categorias taxonômicas de um espécime. Isso poderia ser problemático, pois as linhas de dois espécimes compartilhando o mesmo gênero (nó da penúltima geração) se sobreponham em grande parte do trajeto.

Esse problema torna-se mais evidente ao lidar com conjuntos de dados extensos, ampliando significativamente a sobreposição e, consequentemente, aumentando o tempo necessário para criar as visualizações. Além disso, as visualizações tornavam-se mais pesadas, especialmente ao salvar em formato HTML, o padrão comum para visualizações interativas. Nesse formato, todos os caminhos são salvos, mesmo que grande parte deles compartilhem segmentos comuns, resultando em um custo considerável em termos de tamanho do arquivo.

5.2.3 Versão Inicial

Figura 11 – Representação Inicial da Coleção de Crustácea



Fonte: Elaboração própria, utilizando PBRs da Coleção de Crustácea.

Nota: A figura representa o grafo para a coleção de Crustácea. Ela apresenta a utilização da mesma estrutura do grafo de exemplo, mas pela quantidade substancialmente maior de elementos, ficou basicamente impossível ter alguma conclusão.

Retornando ao modelo descrito na Subseção 5.2.1, evidencia-se, pela Figura 10, que

o grafo foi criado com sucesso. No entanto, o escopo deste trabalho visa a aplicação desse modelo em Registros Primários de Biodiversidade (PBRs), que abrangem bases de dados substancialmente maiores do que o exemplo citado. Dessa forma, enfrentamos desafios significativos, os quais serão discutidos mais detalhadamente a seguir.

Na Figura 11, apresentamos a representação do grafo seguindo o mesmo modelo descrito na Subseção 5.2.1 para a coleção de Crustácea. Esse caso, no entanto, representa um extremo, pois a coleção de Crustácea é considerada uma coleção mega diversa, caracterizada por uma ampla variedade de espécies, gêneros, famílias, e assim por diante. Essa grande diversidade se torna evidente quando comparamos a coleção de Crustácea com outras coleções, como pode ser visto na Tabela 1.

Dada a magnitude da diversidade da coleção de Crustácea, foi necessário uma solução especial para essa coleção. Assim, em colaboração com os curadores da coleção, decidimos não criar um único grafo para a coleção como um todo. Optamos, em vez disso, por gerar vários grafos com base na ordem taxonômica, cada um contendo a árvore taxonômica completa presente nos PBRs de uma ordem específica. Assim, a coleção de Crustácea se encontra fragmentada em suas suas categorias taxonômicas de ordem.

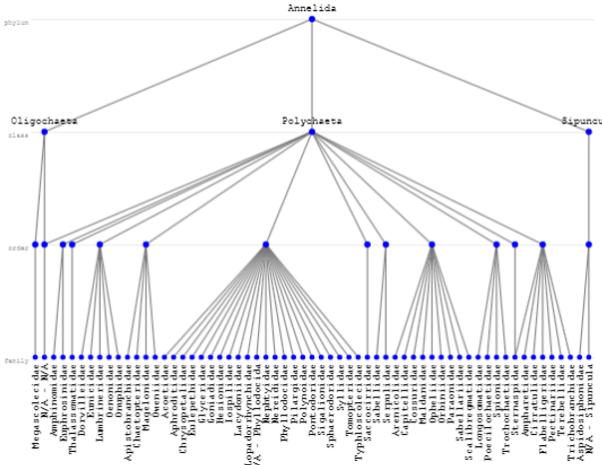
Embora a Figura 11 ilustre um cenário extremo devido à grande diversidade da coleção de Crustácea, outras coleções enfrentaram desafios semelhantes relacionados à quantidade de informação a ser apresentada. Por isso, na próxima seção, exploraremos uma solução geral para essa situação, que será eficaz para todos as coleções, inclusive as 'subcoleções' de Crustácea.

5.2.4 Detalhes conforme a demanda

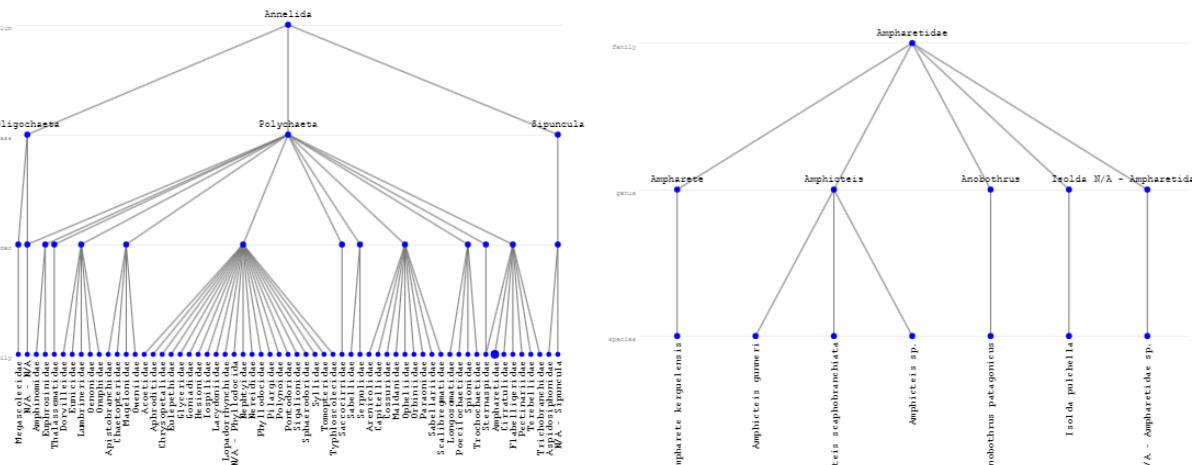
Diante do desafio decorrente da crescente complexidade das últimas gerações da árvore taxonômica, optamos por apresentar apenas uma porção dela como solução auxiliar para a coleção de Crustácea. Contudo, essa estratégia demanda a criação de múltiplas visualizações, o que acarreta prejuízos tanto na fase de produção quanto na experiência de utilização. Se mostrando uma solução ineficiente para aplicar para todas as coleções.

Neste contexto, fundamentamo-nos no princípio de Shneiderman ([SHNEIDERMAN, 1996](#)), conhecido como '*Overview first, filter and Zoom, then details on demand.*' Para aplicar uma nova abordagem, em que desenvolvemos uma visualização que abrange desde a primeira até a n-ésima geração da árvore taxonômica. Sendo que na n-ésima geração, incorporamos um filtro que, ao ser acionado por meio do clique em um elemento específico dessa geração, revela a sub-árvore que tem elemento clicado como raiz. Em outras palavras, proporciona uma visão detalhada da porção restante da árvore que se origina desse nó. Essa metodologia oferece uma visão geral inicial, permitindo que o usuário, de acordo com sua escolha, explore novos elementos e suas inter-relações.

Figura 12 – Grafo com filtro pela família

Grafo Annelida

(a) Estado inicial

Grafo Annelida

(b) Após selecionar a família Ampharetidae

Fonte: Elaboração própria, utilizando PBRs da Coleção de Annelida.

Nota: A figura representa o grafo para a coleção de Annelida. Ela apresenta a utilização do filtro pela família, ao selecionar uma família na primeira tela, a segunda tela abre o grafo referentes aquela família. Importante ressaltar que em (a) é o estado que o gráfico é inicialmente plotado, enquanto que em (b) é o resultado após selecionar uma família.

No que diz respeito às bases de biodiversidade, optamos por realizar esse corte na categoria taxonômica de família. Essa decisão foi fundamentada na análise dos dados e pela concordância dos curadores, indicando que seria um ponto de partida eficaz para a divisão entre a visão macro e micro das coleções. A Figura 12 apresenta o resultado desta nova abordagem aplicada à coleção de Annelida. Ao implementar esse método de filtragem na categoria taxonômica de família, observamos uma melhoria significativa na capacidade

do usuário de compreender a distribuição e interconexão das diferentes espécies. Esse refinamento proporciona uma transição suave entre uma visão geral abrangente e uma análise mais detalhada, permitindo uma exploração mais aprofundada da biodiversidade presente na coleção.

5.2.5 Inversão dos eixos

Após a elaboração dos dois grafos interconectados, tornou-se evidente a necessidade de uma reorganização mais cuidadosa para otimizar a experiência de visualização. A abordagem proposta envolve a disposição do primeiro grafo, que representa a visão geral, imediatamente seguido pelo segundo, de forma em que destaca a continuidade das categorias taxonômicas. Essa sequência cria um efeito contínuo, aproveitando o princípio psicológico da continuidade, estudado na psicologia de Gestalt([KNAFLIC, 2018](#)). Os conceitos derivados dessa área têm ramificações em disciplinas diversas, abrangendo não apenas a psicologia, mas também influenciando campos como Design, Arquitetura e a área de Infovis.

Dentro desse contexto, decidimos unir os dois gráficos em uma disposição sequencial, mantendo uma continuidade ao longo do eixo das categorias taxonômicas. Mesmo sendo grafos distintos, essa abordagem busca transmitir a percepção de continuidade e complementariedade entre as categorias taxonômicas presente no conjuntos de dados. Todavia, considerando a habitual orientação em paisagem de monitores e telas, que favorece uma maior largura em relação ao comprimento, optamos por alterar o eixo das categorias, originalmente no eixo Y, para o eixo X da visualização. Essa modificação visa oferecer uma representação visual mais clara e intuitiva do crescimento da árvore taxonômica.

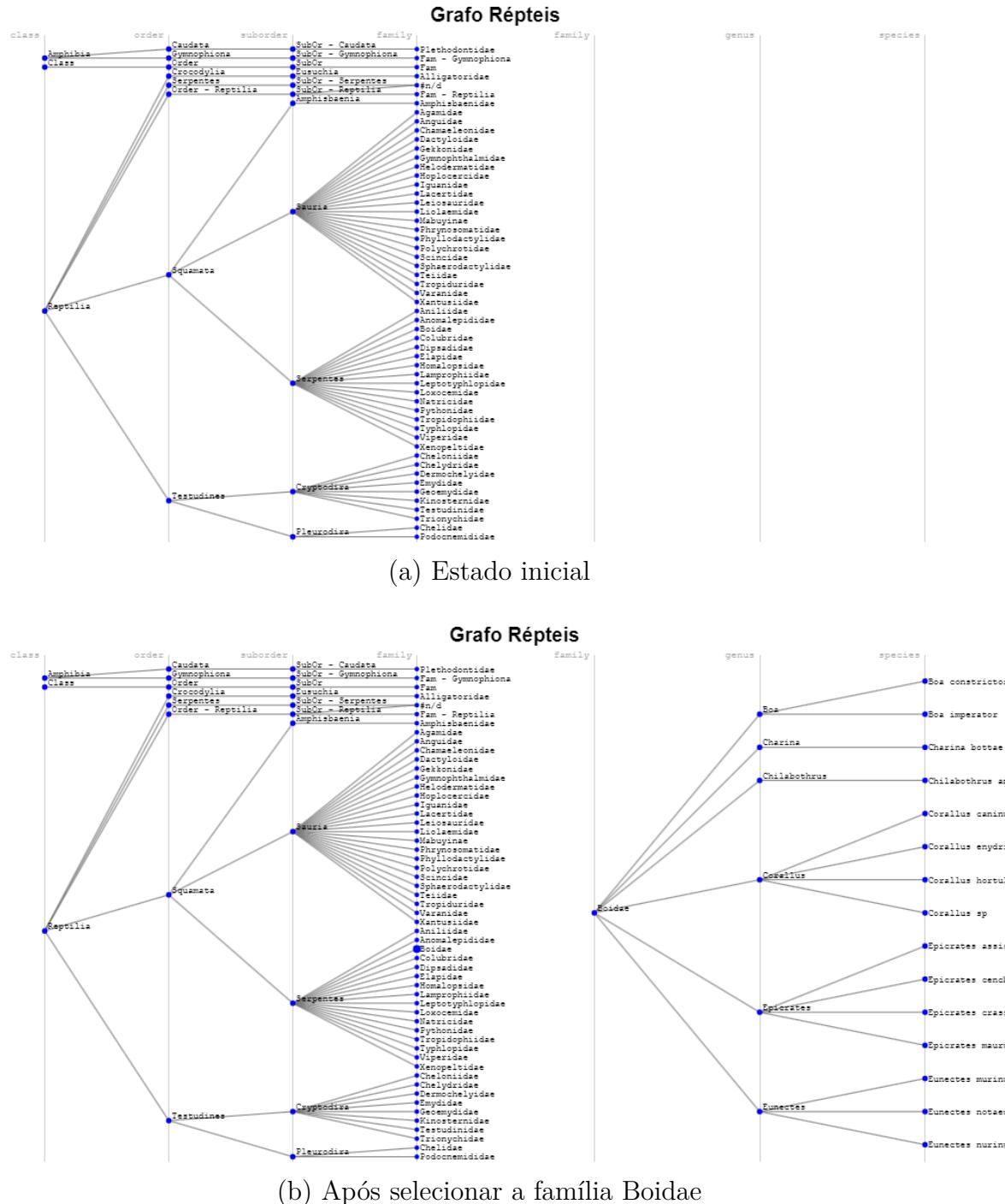
A Figura 13 apresenta o resultado dessa reorganização detalhada. Nela, é possível visualizar claramente o crescimento da árvore taxonômica e a interconexão dos diversos nós relacionados. Essa abordagem visa não apenas otimizar a compreensão da complexidade da biodiversidade, mas também proporcionar uma experiência visual mais informativa para o usuário.

5.2.6 Representação dos Registros

Desde o começo da pesquisa, os curadores expressaram a importância da relação entre registros e a árvore taxonômica. Por isso, um novo gráfico foi adicionado na sequência do grafo. Esse novo gráfico, tem uma abordagem semelhante à desenvolvida durante minha Iniciação Científica. Nesse contexto em que queríamos representar os dados que havíamos agrupado no início do trabalho, resolvemos adicionar um *Dot plot* contendo todos os registros associados à família selecionada pelo seletor do primeiro grafo.

Neste novo gráfico, *Dot Plot*, os registros referentes à família selecionada no seletor

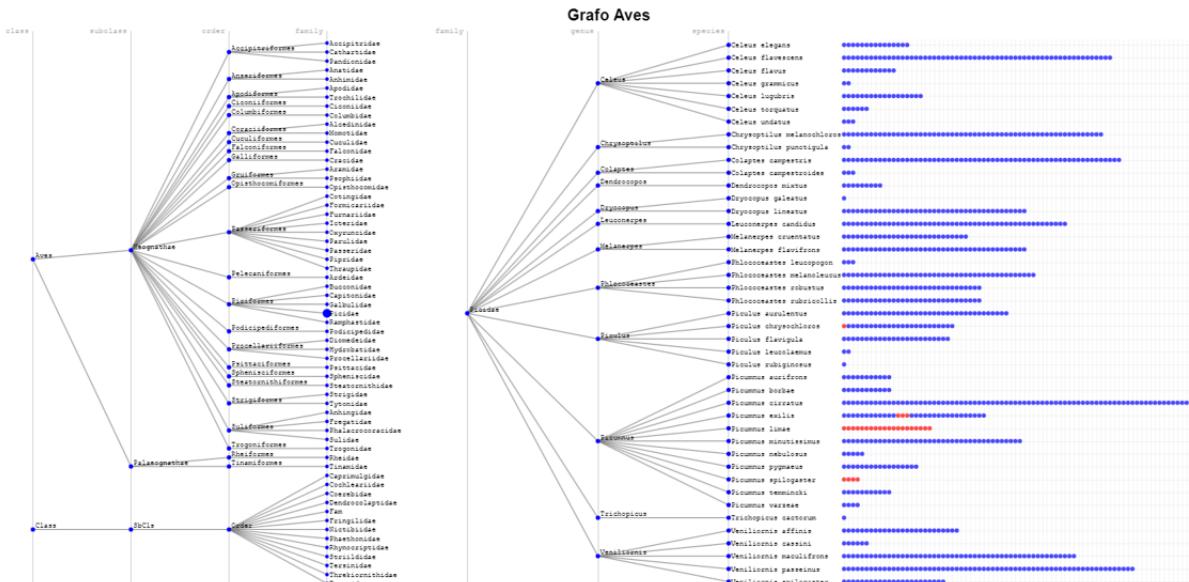
Figura 13 – Grafo com crescimento horizontal(troca dos eixos)



Fonte: Elaboração própria, utilizando PBRs da Coleção de Répteis.

Nota: A figura representa o grafo para a coleção de Répteis. Nela, apresenta a troca dos eixos. Importante ressaltar que em (a) é o estado que o gráfico é inicialmente plotado, enquanto que em (b) é o resultado após selecionar uma família.

Figura 14 – Grafo com nível de registro



Fonte: Elaboração própria, utilizando PBRs da Coleção de Aves.

Nota: A figura representa o grafo de Aves, com a utilização do *DotPlot* e representação a nível de registro, em que cada bolinha na terceira tela é registro na base de dados. Além disso, vale ressaltar a cor, se vermelho, é um registro tipo.

são dispostos em linhas conforme suas espécies, seguindo a mesma ordem estabelecida no grafo secundário. Isso proporciona uma percepção contínua e complementar entre as representações, questão já discutidas que tem os estudos da psicologia Gestalt(KNAFLIC, 2018) como fundamento.

Além disso, os curadores destacaram a relevância dos espécimes tipo e solicitaram uma forma de evidenciá-los. Assim, os pontos que indicam registros de espécimes tipo são diferenciados visualmente dos demais no *Dot Plot*. Essa escolha tem embasamento nos princípios da psicologia de Gestalt(KNAFLIC, 2018), especialmente nos conceitos de **Ponto Focal**, destacando elementos significativos, e **Similaridade**, ao agrupar os registros de espécimes tipo como parte de uma categoria distinta aos outros registros.

O resultado dessa nova abordagem pode ser visualizado na Figura 14, onde uma família foi selecionada, causando a abertura do segundo grafo e do *Dot plot*. Os pontos vermelhos no *Dot plot* indicam registros de espécimes tipo, proporcionando uma visão de cada registro, e evidenciando os espécimes tipo.

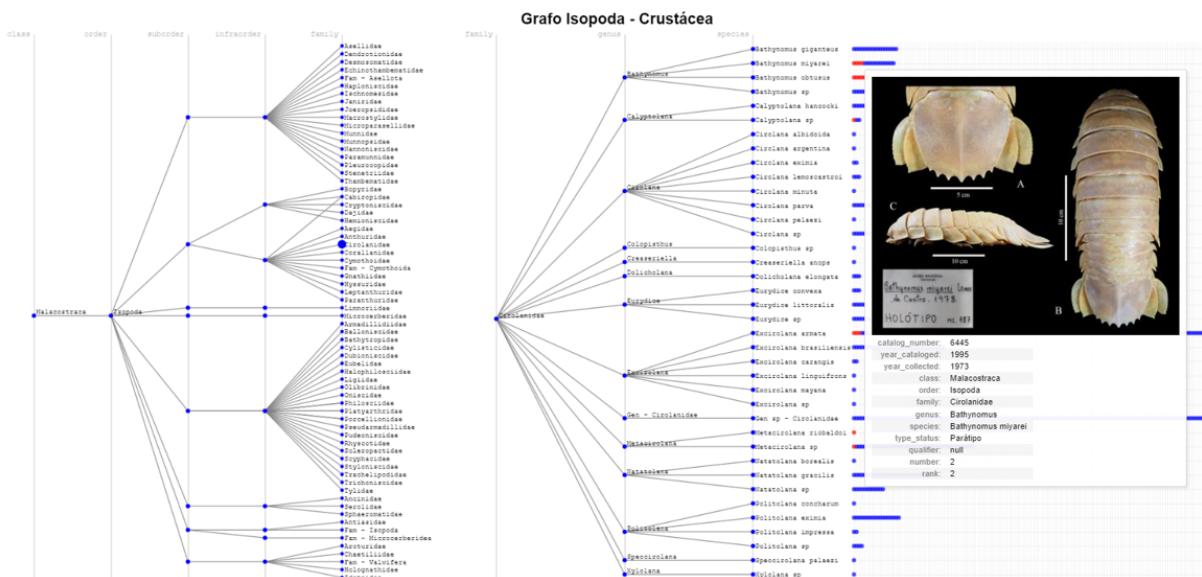
5.2.6.1 Imagens nas Tooltips

O Museu Nacional/UFRJ está empenhado na digitalização de suas coleções, incluindo a captura de imagens dos espécimes para criar uma base de dados fotográficos. Essa iniciativa, aliada ao interesse em trabalhar com imagens durante este projeto, resultou na integração de dados fotográficos na visualização.

Dado que o processo de fotografia está em andamento e nem todos os espécimes foram fotografados, decidimos focar em um conjunto de dados mais específico. Devido à importância dos espécimes tipo, muitos deles já foram fotografados, possibilitando a incorporação de imagens para esse grupo de registros.

Para enriquecer a experiência, optamos por adicionar imagens às *tooltips* do *Dot Plot* que contém os dados a nível de registro, isso acontece exclusivamente quando o registro corresponde a um espécime tipo. O resultado pode ser visto na Figura 15. Essa solução não apenas oferece informações visuais relevantes aos pesquisadores, mas também utiliza eficientemente as imagens disponíveis, aprimorando a compreensão sobre os espécimes tipo associados à árvore taxonômica.

Figura 15 – Grafo com nível de registro e imagens na tooltip



Fonte: Elaboração própria, utilizando PBRs e imagens da Coleção de Crustácea.

Nota: A figura representa o grafo para a classe de Isopoda da coleção de Crustácea, nela é possível notar os registros vermelhos que são os tipos e, com o mouse em um deles, a apresentação da *tooltip*, em que há um exemplo da utilização de imagem.

Encerramos esta seção com a visualização da árvore taxonômica, que se expande horizontalmente, proporcionando uma visão abrangente. A eficácia da filtragem por família foi notável, permitindo uma análise mais precisa. A conexão direta com os registros facilitou a identificação de cada elemento, e a inclusão de imagens enriqueceu ainda mais essa experiência visual. Contudo, devido à natureza confidencial dos dados utilizados, não é possível compartilhar publicamente essas visualizações específicas. No entanto, reproduzimos o mesmo processo utilizando uma base de dados aberta, apresentada no Apêndice A.

6 Resultados e Conclusão

Com o intuito de desenvolver uma visualização pouco difundida, estabelecemos uma parceria com o departamento de Invertebrados e Vertebrados do Museu Nacional/UFRJ. Nessa colaboração, compartilharam suas valiosas bases de dados, enquanto, em contrapartida, proporcionamos visualizações destinadas a auxiliá-los na exploração e apresentação de seus PBRs. Essa aliança remonta à dissertação de Franklin Oliveira([OLIVEIRA, 2021](#)), que serviu como alicerce teórico para esta pesquisa. No entanto, optamos por ir além, escolhendo desenvolver uma representação gráfica diferenciada – uma árvore taxonômica.

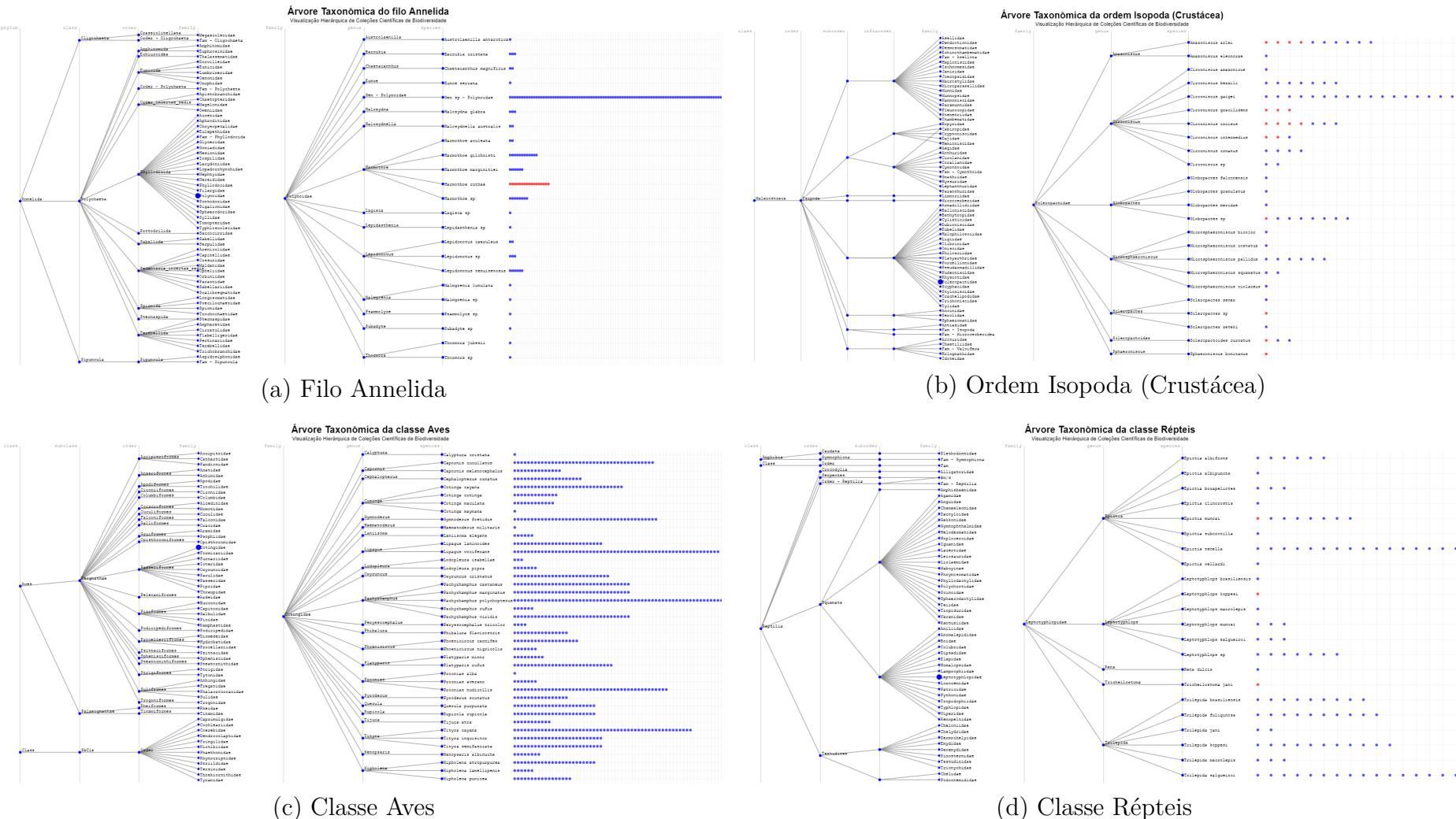
O desafio dessa abordagem residia na ausência de *frameworks* que oferecessem a capacidade desejada de interatividade na visualização de grafos. Dessa forma, embarcamos em um processo de aprendizado, convertendo dados tabulares em um formato acessível para criar os grafos. O posicionamento preciso de cada nó tornou-se a principal dificuldade, acompanhada por outros desafios secundários, como a gestão de dados faltantes e a eliminação de etapas redundantes.

Ao desenhar o grafo, observamos que bibliotecas padrões de visualização, como Altair, são capazes de representar grafos, contando que a posição dos nós esteja definida, resultado que já havíamos conseguido. Em sequência, foi necessário a implementação de aprimoramentos para melhor compreensão dos dados, incluindo a inversão dos eixos e uma representação mais granular.

Ao finalizar este trabalho, acreditamos ter alcançado com sucesso o objetivo de criar uma visualização capaz de revelar as diferentes categorias taxonômicas de um espécime. Além disso, proporcionamos aos curadores uma visão abrangente, identificando todos os registros de uma mesma espécie, a quantidade de registros em cada uma das nomenclaturas de categorias taxonômicas presente na base. Além disso, foi dado destaque aos espécimes tipo e indicado a quantidade existente em cada um dos grupos.

Vale destacar o desenvolvimento de um código abrangente, capaz de gerar grafos para todas as coleções disponíveis com pequenas alterações nos valores de entrada. Cada grafo utilizado neste trabalho representa uma coleção diferente, a versão final de cada um dos grafos das coleções disponíveis pode ser visto na Figura 16. Outro ponto importe é que o código, com ajustes pontuais, pode facilmente gerar árvores semelhantes a estas em outros contextos e com diferentes conjuntos de dados, como é feito no Apêndice B.

Figura 16 – Árvores Taxonômicas



Fonte: Elaboração própria, com base nas coleções
Nota: Grafo Final de cada uma das coleções que trabalhamos.

6.1 Futuros passos

Alcançamos o objetivo inicial neste trabalho. Contudo, é importante reconhecer que sempre há espaço para melhorias e evoluções. Neste contexto, duas possibilidades que já têm certo destaque para otimizar ainda mais nosso projeto: aprimoramentos na visualização, com foco na escrita dos nomes dos nós, e refinamentos na disposição geral do gráfico, concentrando-se na estratégia de posicionamento dos nós.

A inserção de texto em visualização é um desafio conhecido, por vezes exigindo intervenções manuais para ajustes precisos em cada elemento. Esse desafio é ampliado em um contexto de algoritmo genérico, como o que utilizamos, que emprega o mesmo *script* para gerar grafos para coleções distintas. Assim, reconhecemos a necessidade de estudos mais aprofundados e mais testes para alcançar melhorias substanciais nessa área.

Outro ponto destacável é a posição dos nós e a estrutura geral do grafo. Até o momento, nos limitamos à representação de árvores em que todos os caminhos partindo de um nó raiz, passa por um nó de cada geração e finaliza na última. No entanto, com estudos adicionais e modificações no código, acreditamos ser possível ampliar essa capacidade para desenhar outros tipos de grafos, proporcionando uma representação mais abrangente.

Após listar essas possíveis melhorias, é importante salientar que existam mais, ainda mais pelo contexto em que a vis foi feita. Como já sabemos, está visualização tem intuito de ajudar pesquisadores a explorar e a apresentar suas bases, assim, a utilização levará a novos desejos e necessidades, consequentemente, novas melhorias a serem feitas.

6.2 Considerações Finais

Ao concluir este trabalho, apresentamos uma árvore taxonômica que transcende a simples ilustração da hierarquia taxonômica de um grupo de animais. Este projeto não apenas oferece uma representação visual da biodiversidade, mas também busca inovação dentro dos padrões convencionais visando auxiliar o trabalho dos curadores.

Os resultados preliminares que foram compartilhados com os curadores das coleções científicas de biodiversidade foram recebidos com entusiasmo. Sua resposta positiva não apenas valida o trabalho realizado, mas também expressa o interesse em utilização da visualização como ferramenta. Isso reforça ainda mais a importância e o sucesso do projeto.

Em última análise, este projeto assumiu o papel de unir a Ciência de Dados com a biodiversidade, desempenhando a função de oferecer uma representação visual informativa. A utilização de recursos de visualização padrão e manipulação eficiente de dados permitiram o desenvolvimento um formato diferenciado, fugindo do convencional. Esta abordagem oferece uma compreensão mais rica da taxonomia, proporcionando uma experiência visual enriquecedora.

Referências

- FOX, Peter; HENDLER, James. Changing the Equation on Scientific Data Visualization. **Science (New York, N.Y.)**, v. 331, p. 705–8, fev. 2011. DOI: [10.1126/science.1197654](https://doi.org/10.1126/science.1197654).
- IBGE. **API de dados localidades do IBGE**. [S.l.]: IBGE (Instituto Brasileiro de Geografia e Estatística), 2017. Disponível em: <https://servicodados.ibge.gov.br/api/docs/localidades#api-_>.
- INATURALIST. [S.l.: s.n.]. Acessado em 30 de Novembro de 2023. Disponível em: <<https://www.inaturalist.org/>>.
- JANICKI, Julia et al. Visualizing and interacting with large-volume biodiversity data using client–server web mapping applications: The design and implementation of antmaps.Org. **Ecological Informatics**, v. 32, p. 185–193, mar. 2016. DOI: [10.1016/j.ecoinf.2016.02.006](https://doi.org/10.1016/j.ecoinf.2016.02.006).
- JETZ, Walter; MCPHERSON, Jana; GURALNICK, Robert. Jetz W, MacPherson J, and Guralnick RP. Integrating biodiversity distribution knowledge: toward a global map of life. Trends Ecol Evol. **Trends in ecology evolution**, v. 27, p. 151–9, mar. 2012. DOI: [10.1016/j.tree.2011.09.007](https://doi.org/10.1016/j.tree.2011.09.007).
- KERREN, Andreas et al. BioVis Explorer: A visual guide for biological data visualization techniques. **PLOS ONE**, Public Library of Science, v. 12, n. 11, p. 1–14, nov. 2017. DOI: [10.1371/journal.pone.0187341](https://doi.org/10.1371/journal.pone.0187341). Disponível em: <<https://doi.org/10.1371/journal.pone.0187341>>.
- KIRK, Andy. **Data visualisation: A handbook for data analysts, designers and communicators**. 4th. London: SAGE Publications, 2023.
- KNAFLIC, Cole Nussbaumer. **Storytelling com dados: um guia sobre visualização de dados para profissionais de negócios**. [S.l.]: Alta Books, 2018.
- MEDEIROS, Asla. **VisZoo**. [S.l.: s.n.], 2023. Disponível em: <https://github.com/aslamedeiros/Vis_Zoo---Fase-1>.
- MEDEIROS E SÁ, Asla; OLIVEIRA, Franklin; SILVEIRA SEREJO, Cristiana. Visualização de Informação como Ferramenta de Apoio à Curadoria de Dados em Coleções Biológicas. **Museologia amp; Interdisciplinaridade**, v. 10, Especial, p. 158–181, set. 2021. Disponível em: <<https://periodicos.unb.br/index.php/museologia/article/view/36709>>.

- MEDEIROS E SÁ, Asla et al. Visually Overviewing Biodiversity Open Data Digital Collections. English. In: PROCEEDINGS of the Symposium on Open Data and Knowledge for a Post-Pandemic Era ODAK22, UK (ODAK 2022). [S.l.]: Electronic Workshops in Computing, jul. 2022. (Open Data and Knowledge for a Post-Pandemic Era). Symposium on Open Data and Knowledge for a Post-Pandemic Era, ODAK 2022 ; Conference date: 30-06-2022 Through 01-07-2022. DOI: [10.14236/ewic/ODAK22.4](https://doi.org/10.14236/ewic/ODAK22.4). Disponível em: <<https://universityofbrighton.github.io/odak>>.
- MESSIAS, Camila Simões Martins de Aguiar et al. New perspectives of Annelida collection (National Museum/UFRJ) database: using data visualization to analyze and manage biological collections. **Ocean and Coastal Research**, 2023. Artigo aceito para publicação, ainda não publicado.
- MUSEU NACIONAL. **Museu Nacional: Panorama dos acervos: passado, presente e futuro**. Quinta da Boa Vista, São Cristóvão, Rio de Janeiro, RJ: Museu Nacional, Universidade Federal do Rio de Janeiro, 2020. v. 18. (Livro Digital). Disponível em: <https://www.museunacional.ufrj.br/destaques/panorama_de_acervos.html>.
- OLIVEIRA, Franklin Alves de. **Visualização de coleções científicas digitais de biodiversidade: um framework em Altair, Python**. 2021. Diss. (Mestrado) – Fundação Getulio Vargas, Rio de Janeiro.
- PYINATURALIST. [S.l.: s.n.], 2023. Disponível em: <<https://pyinaturalist.readthedocs.io/en/stable/index.html>>.
- SCHULZ, Hans-Jorg. Treevis.net: A Tree Visualization Reference. **IEEE Computer Graphics and Applications**, v. 31, n. 6, p. 11–15, nov. 2011. ISSN 1558-1756. DOI: [10.1109/MCG.2011.103](https://doi.org/10.1109/MCG.2011.103).
- SHNEIDERMAN, Ben. The eyes have it: a task by data type taxonomy for information visualizations. **Proceedings 1996 IEEE Symposium on Visual Languages**, p. 336–343, 1996.
- TEAM, The pandas development. **pandas-dev/pandas: Pandas**. [S.l.]: Zenodo, fev. 2020. DOI: [10.5281/zenodo.3509134](https://doi.org/10.5281/zenodo.3509134). Disponível em: <<https://doi.org/10.5281/zenodo.3509134>>.
- THOMAS, Selma; MINTZ, Ann. THE VIRTUAL AND THE REAL: MEDIA IN THE MUSEUM. **Curator: The Museum Journal**, v. 42, p. 55–58, 1999. Disponível em: <<https://api.semanticscholar.org/CorpusID:144208633>>.
- VANDERPLAS, Jacob et al. Altair: Interactive Statistical Visualizations for Python. **Journal of Open Source Software**, v. 3, p. 1057, dez. 2018. DOI: [10.21105/joss.01057](https://doi.org/10.21105/joss.01057).
- VEGA-ALTAIR. [S.l.: s.n.]. Acessado em 31 de Julho de 2023. Disponível em: <<https://altair-viz.github.io/index.html#>>.

Apêndices

APÊNDICE A – Nova base

Devido à natureza confidencial dos dados utilizados, optamos por não expor os resultados publicamente. Em vez disso, exploramos a ideia de coletar dados abertos de animais e reproduzir visualizações semelhantes às realizadas com os dados das Coleções Científicas do Museu Nacional/UFRJ, utilizando dados abertos coletados. Este apêndice descreve o processo de coleta de dados e apresenta os resultados obtidos.

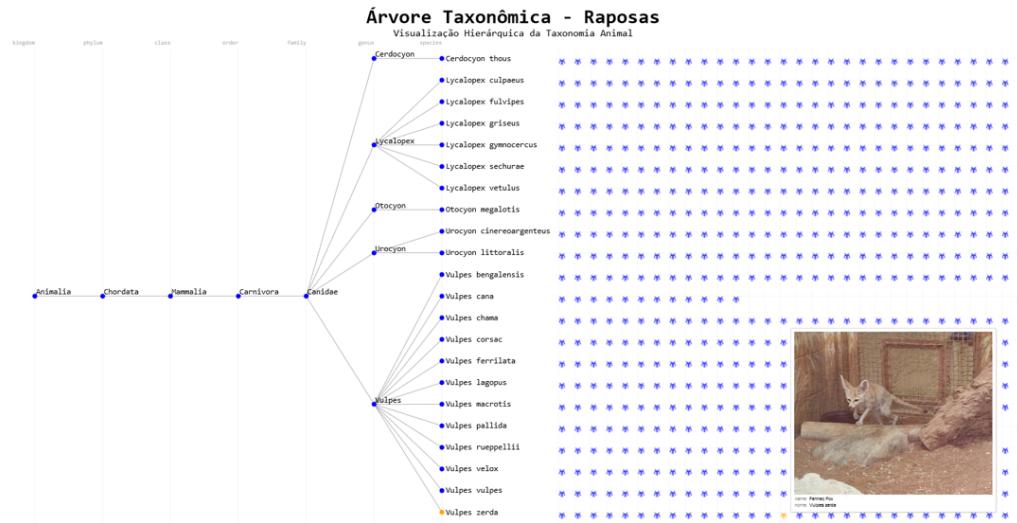
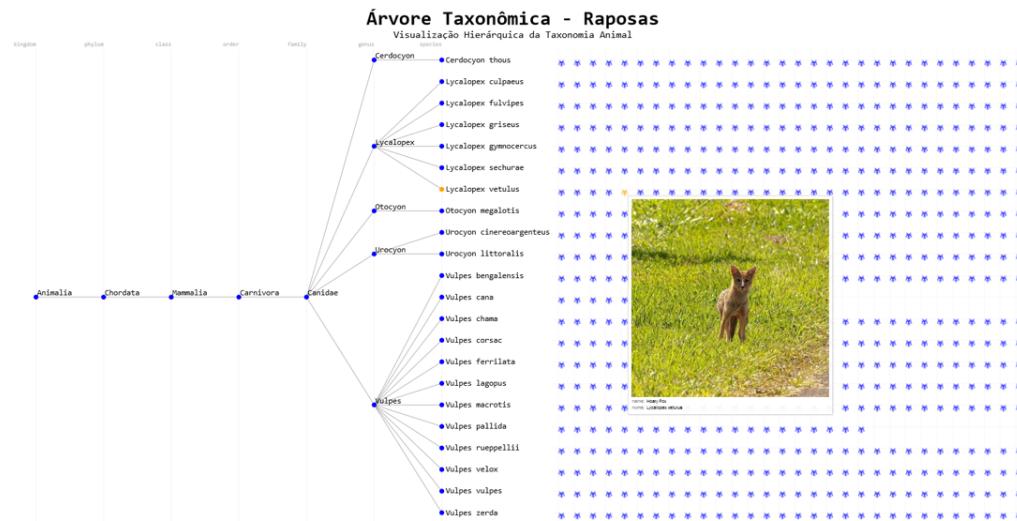
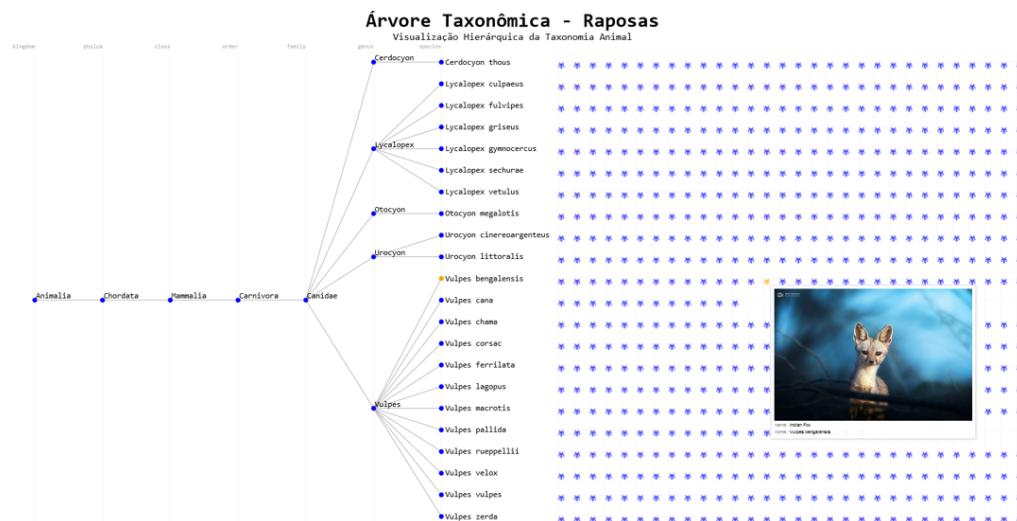
Para gerar as bases de dados, utilizamos a API do *iNaturalist*([INATURALIST, s.d.](#)), uma plataforma de dados abertos onde a comunidade compartilha fotos da fauna e flora globalmente. Inicialmente, acessamos esta plataforma na tentativa de encontrar imagens para utilizar nas visualizações do museu; no entanto, não prosseguimos com essa abordagem. A necessidade de criar uma nova base de dados nos levou a lembrar da API do *iNaturalist*, que não apenas fornece imagens, mas também dados taxonômicos dos registros. Embora não seja perfeita devido à sua natureza aberta, é útil no contexto de criar uma base de dados para exposição dos nossos resultados.

Para realizar a coleta de dados, utilizamos a biblioteca *pyinaturalist*([PYINATURALIST..., 2023](#)), que realiza buscas diretas na API do *iNaturalist*. A biblioteca requer uma lista de nomes científicos de espécies, os quais foram utilizados para fazer requisições individuais. Essa abordagem é semelhante à arquitetura usada na parte dos Registros Primários de Biodiversidade (PBRs) previamente empregados, e resultamos em um base similar a Tabela 2.

Portanto, com os dados integrados ao código previamente utilizado, foi estrategicamente fácil gerar as novas versões. Foram necessárias pequenas edições para adaptar a natureza um pouco diferente dos dados e para realizar melhorias eventuais nas visualizações. Assim, o *Dot Plot* que indica os dados relevantes do registro para as Coleções Científicas foram substituídos pela imagem da API.

A Figura 17 exibe o resultado da coleta de todas as espécies de raposas e a geração da árvore taxonômica correspondente. Para visualizar de maneira interativa, visite <<https://github.com/CarCesar/TCC-Arvore-Biodiversidade>> no *GitHub*, onde também está disponível o código utilizado.

Figura 17 – Árvore taxonômica das Raposas

(a) árvore taxonômica com imagem de uma *Vulpes zerda*(b) árvore taxonômica com imagem de uma *Lycalopex vetulus*(c) árvore taxonômica com imagem de uma *Vulpes bengalensis*

Fonte: Elaboração própria, utilizando dados coletados *iNaturalist*.

APÊNDICE B – Outras Árvores

Embora as árvores tenham sido projetadas para abordar a visualização da taxonomia, sua aplicação pode ser mais abrangente. É importante destacar que, para sua utilização em outros contextos, a estrutura de dados deve ser semelhante à encontrada na taxonomia, e mesmo assim, ainda haverá necessidade de algumas adaptações.

Sob certas condições, é possível criar outras árvores, desde que todos os caminhos partam do nó raiz, terminem apenas na última geração da árvore e percorram todas as gerações. A divisão das regiões do Brasil (Regiões, Estados, Mesorregiões, Microrregiões, Municípios) segue essas especificidades, possibilitando a criação da Árvore de Divisões do Brasil.

Para isso, utilizamos a API do IBGE([IBGE, 2017](#)) e coletamos os dados de localidade, especificamente da região Sudeste do Brasil. Após uma pequena manipulação nos dados, obtivemos uma tabela semelhante à apresentada na Tabela 2, mas com os dados de localização em vez dos dados de taxonomia, como podemos ver na Tabela 4:

Tabela 4 – Exemplo dos dados de localização

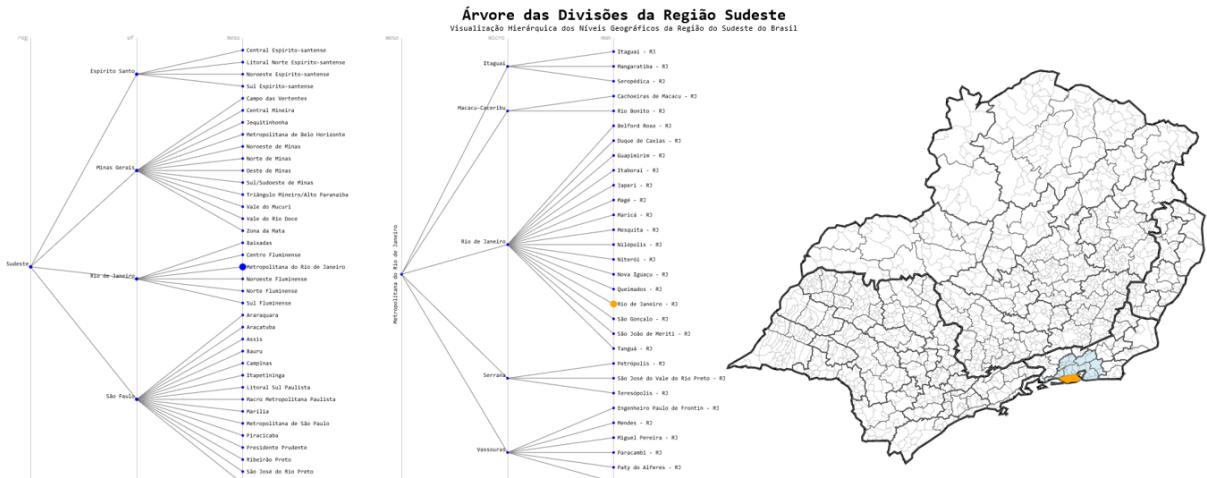
Região	UF	Mesorregiao	Microrregiao	Municipio
Sudeste	Rio de Janeiro	Metropolitana do Rio de Janeiro	Rio de Janeiro	Rio de Janeiro
Sudeste	Rio de Janeiro	Baixada	Lagos	Cabo Frio
Sudeste	São Paulo	Metropolitana São Paulo	São Paulo	São Paulo
Sudeste	São Paulo	Ribeirão Preto	Franca	Franca
Sudeste	Minas Gerais	Metropolitana Belo Horizonte	Belo Horizonte	Belo Horizonte
Sudeste	Minas Gerais	Triângulo Mineiro e Alto Paranaíba	Patos de Minas	Patos de Minas

Fonte: Elaboração própria, utilizando dados coletados IBGE.

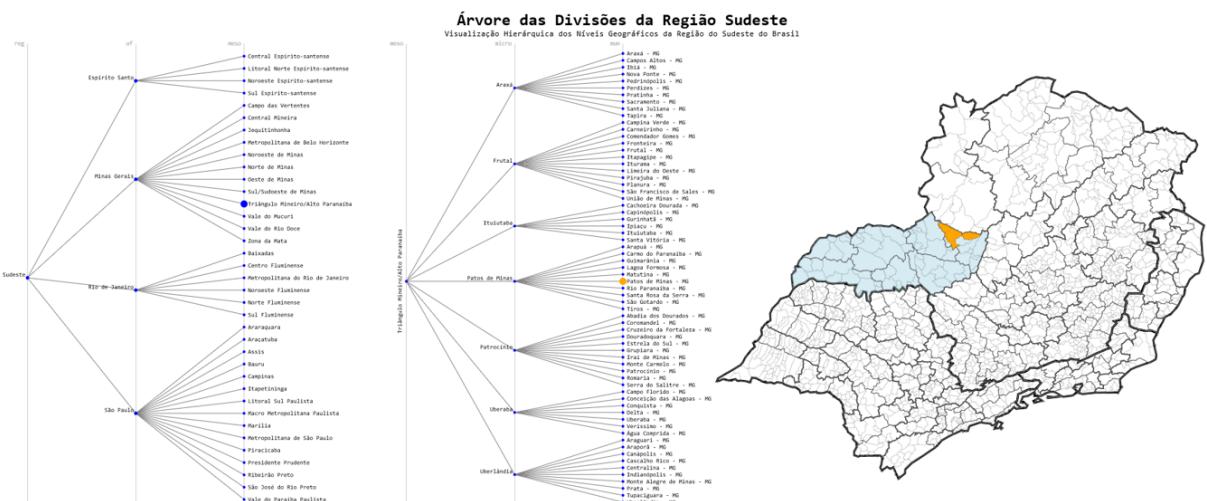
Com os dados estruturados, o processo, semelhante ao feito para a árvore taxonômica, foi realizado. Diferentemente do uso de um *DotPlot*, neste contexto, utilizamos um mapa para auxiliar na visualização, conforme mostrado na Figura 18.

O código e a visualização interativa estão disponíveis em <<https://github.com/CarCesar/TCC-Arvore-Biodiversidade>>.

Figura 18 – Árvore das divisões regionais do Sudeste



(a) árvore das divisões regionais que o município do Rio de Janeiro faz parte.



(b) árvore das divisões regionais que o município de Patos de Minas faz parte.

Fonte: Elaboração própria, utilizando dados coletados IBGE.