MUSEOLOGIA & INTERDISCIPLINARIDADE Vol. 10, nº especial, set. de 2021

DOSSIÊ

Visualização de Metadados como Ferramenta de Apoio à Curadoria Digital de Coleções Científicas Biológicas

Metadata Visualization as a Support Tool for Curatorship of Biodiversity Scientific Digital Collections

Asla Medeiros e Sá[1]
Franklin Oliveira [2]
Cristiana Silveira Serejo [3]

DOI 10.26512/museologia.v10i19.36709

Resumo

Coleções científicas biológicas são a base primária para o descobrimento e entendimento da biodiversidade mundial, sendo geridas por instituições como museus, herbários e instituições científicas que são fiéis depositárias de guarda de acervos. São constituídas de espécimes coletados em campo, acrescidos de informações relacionados ao material de estudo, tais como, coordenadas geográficas, altitude, profundidade, tipo de habitat, data de coleta e coletor, e que uma vez recebendo um número de registro, são incorporados à coleção. Tais informações são os metadados relativos aos espécimes da coleção, estes, juntamente com dados da digitalização do acervo como fotos, tomografias computadorizadas, registros sonoros etc., compõem a coleção digital. Garantir a qualidade e usabilidade desses metadados é uma tarefa iminente dos curadores, sendo fundamental para a longevidade e relevância do acervo e está atrelada diretamente ao que chamamos de curadoria digital dos acervos. Nesse sentido, o presente trabalho se propõe a aplicar de forma inédita os princípios e técnicas de visualização da informação com base nos bancos de dados de coleções científicas biológicas, ao fornecer um conjunto de representações visuais cuja finalidade primária é a de promover a verificação da qualidade dos metadados associados pelos especialistas. O processo de manuseio de dados foi realizado empregando-se a linguagem de programação Python 3 e a biblioteca de visualização Altair, onde foram gerados *scripts* de trabalho e que podem servir de base para uso em outras coleções com a mesma tipologia.

Palavras-chave

Visualização de Informação. Coleções científicas biológicas. Qualidade de registros. Representação visual de dados.

Abstract

Biological scientific collections are the primary basis for discovering and understanding the world's biodiversity, being managed by institutions such as museums, herbaria and scientific institutions that are faithful custodians of collections. They consist of specimens collected in the field, plus information related to the study material, such as geographic coordinates, altitude, depth, habitat type, date of collection and collector, and that, once receiving a registration number, are incorporated into the collection. Such information is the metadata related to the specimens in the collection, these, together with data from the digitization of the collection such as photos, topographies, sound recordings, etc., make up the digital collection. Ensuring the quality and usability of these metadata is an imminent task for curators, being fundamental to the longevity and relevance of the collection, and it is directly linked to what we call digital curation of the collections. In this sense, the present work proposes to apply, as an original contribution, the principles and techniques of information visualization based on biological scientific collections databases, by providing a set of visual representations whose primary purpose is to promote quality verification metadata associated by the experts. The data handling process was carried out using Python 3 and the Altair visualization library, the proposed scripts can serve as a basis for use in other collections with the same typology.

Keywords

Information Visualization. Scientific Collections. Biodiversity Collections. Data Quality. Data Visual Representation.

1 FGV/EMAp

2 FGV/EMAp

3 MN/UFRJ

Introdução

Nas últimas décadas, o advento da Internet e o aumento do poder computacional revolucionaram a forma de criar, armazenar e recuperar dados. Nesse contexto, técnicas para representar dados de forma a facilitar sua manipulação e compreensão são de extrema importância. Um caso particular, objeto de estudo do presente artigo, é o caso das coleções científicas sob gestão e curadoria de instituições GLAM[4]. Coleções científicas biológicas são a base primária para o descobrimento e entendimento da biodiversidade mundial, constituídas, portanto, de informações dos espécimes relacionados ao seu habitat. O material é catalogado, incluindo informações básicas de coordenadas geográficas, altitude, profundidade, tipo de habitat, data de coleta e coletor, e que uma vez recebendo um número de registro, é incorporado à coleção. Sobre a qualificação do material, esta vai estar condicionada à disponibilidade de especialistas trabalharem com o acervo e promoverem, em última instância, a publicação do mesmo.

No contexto desse artigo vamos focar no tratamento dos metadados digitais gerados nas coleções científicas. Para a compreensão do que são acervos digitais temos duas categorias: os acervos digitalizados e os nato digitais. Os acervos digitalizados possuem uma base física, um espécime no caso de coleções biológicas, que passam pelo processo de digitalização (fotos, tomografias computadorizadas). Por sua vez, os nato digitais não têm uma fonte física, já nascendo no formato digital. Isso se aplica a muitos materiais contemporâneos por excelência, como fotos, vídeos e gravações sonoras em formato digital (IBRAM, 2020).

Com a crescente geração de metadados das coleções se faz necessária uma constante curadoria digital, que tem como objetivo reduzir as ameaças ao valor informacional e garantir a manutenção de características como qualidade, confiabilidade, integridade e usabilidade dos dados. Entende-se que a manutenção de tais características está apoiada no tratamento proveniente da atribuição de metadados (Triques, 2020). A aplicação de princípios e técnicas de visualização da informação junto aos bancos de dados de coleções científicas biológicas irão justamente trabalhar para a manutenção, limpeza e integridade dos dados, funcionando como ferramenta importante na curadoria digital.

A discussão sobre quais metadados utilizar para catalogação dos itens e como preenchê-los é um campo de estudo que depende do tipo de dado e é muito vasto. A questão se torna mais crítica na era digital, dando origem à discussão sobre padrões de metadados, que visam a integração e compartilhamento dos mesmos na Internet e permite que buscas por itens possam ser executadas por algoritmos remotamente e não apenas através de consultas individualizadas e locais. O paradigma de *Linked Open Data* (Dados Abertos Interligados) mudou significativamente a forma como a informação é disponibilizada na internet; é uma iniciativa que preconiza a adoção sistemática de padrões de metadados pelas diferentes instituições que aderirem ao modelo, enriquecendo semanticamente os dados disponibilizados e permitindo buscas mais eficazes e interligadas entre instituições.

Uma coleção pode, então, ser entendida como um repositório de espécimes/objetos atrelados a seus metadados a partir do qual novos conhecimentos podem ser gerados. Esse ciclo de geração de conhecimento se torna possível uma vez que passamos a ter a capacidade de explorar os dados e

4 GLAM é acrônimo do inglês Galleries, Libraries, Archives & Museums: Galerias, Bibliotecas, Arquivos e Museus.

analisar padrões que estejam presentes no coletivo de itens. O cenário digital potencializa a capacidade de gestão e análise de dados, possibilitando a utilização de ferramentas e técnicas disponíveis também em outros domínios de conhecimento, como é o caso da visualização de informação, por exemplo. Em seu artigo Navarrete (2017) discute a digitalização de coleções de herança cultural dentro das instituições GLAM como um indicador de inovação de conhecimento e conclui que existe uma relação causal entre inovação organizacional e a adoção de fluxos de trabalho digitais. Tal característica se reflete na atenção para o reuso criativo de coleções digitais e na presença de corpo técnico especializado para avançar nas estratégias encontradas nas instituições com um maior compartilhamento das coleções digitalizadas.

Após décadas de investimentos em digitalização, ainda em curso, diversas coleções científicas podem, atualmente, ser acessadas via web. As interfaces que permitem a exploração dos dados são heterogêneas e, em muitos casos, pouco amigáveis. Exemplos mais recentes de interfaces implementam paradigmas de busca que vão além da representação usual baseada em busca textual ou exibição matricial dos itens da coleção. Discutiremos mais a fundo este tópico na Seção 2 deste artigo.

Dentre as coleções científicas digitalizadas, destacam-se as coleções biológicas. Nesse cenário, grandes instituições têm sua importância revelada dado que foram responsáveis por coletar e documentar dados de biodiversidade ao longo do tempo. Essa documentação baseia-se em registros primários de biodiversidade, os PBR[5], que contém metadados tais como a data e o local de coleta, tipo de habitat, coletor da amostra e identificação taxonômica.

Iniciativas de digitalização das coleções de registro de biodiversidade foram realizadas nas últimas décadas por instituições de pesquisa como museus, herbários, universidades etc., dando origem à um vasto banco de dados de coleções digitais. Uma iniciativa nacional e com grande êxito foi o Projeto Reflora[6], liderado por pesquisadores do Jardim Botânico do Rio de Janeiro e que vem realizando o resgate de imagens dos espécimes da flora brasileira e das informações a eles associadas, depositados nos herbários nacionais e estrangeiros para a construção do Herbário Virtual Reflora.

Um grande diferencial que caracteriza a comunidade de dados de biodiversidade é a convergência em torno do padrão de metadados *Darwin Core* (DwC) que é um conjunto de normas definido pelo *Biodiversity Information Standards* ou *Taxonomic Databases Working Group* (TDWG). O DwC é baseado principalmente em dados taxonômicos e em sua ocorrência na natureza, documentada por meio de observações de espécimes e amostras. A convergência em torno de um padrão permitiu a criação de agregadores de banco de dados de coleções bem como a ampliação do acesso às mesmas em toda parte do mundo. Iniciativas pioneiras como o *Global Biodiversity Information Facility* (GBIF), maior plataforma de dados abertos em biodiversidade, *Inter-American Biodiversity Information Network* (IABIN) e *Comisión Nacional para el Conocimiento y Uso de la Biodiversidad* (CONABIO) juntas viabilizam acesso a bilhões de registros de coleções de biodiversidade. Tais bases dão suporte a estudos que buscam, por exemplo, compreender padrões de distribuição e variação temporal, mudanças climáticas, interações entre diferentes espécies, presença ou ausência de espécies em um dado bioma etc. e estão disponibilizados em plataformas de dados abertos.

---

5 PBR é acrônimo no inglês para *Primary Biodiversity Record*: Registro primário de biodiversidade.

6 http://reflora.jbrj.gov.br/reflora

Para a comunidade científica, a qualidade dos registros que constam nos bancos de dados é de suma importância para garantir a precisão dos estudos desenvolvidos. 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. Muitas são as ferramentas disponíveis para produção de visualizações de coleções de dados biológicos. Wang (2015) faz uma boa síntese de ferramentas *open source* 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 informações efetivas que permitam a melhoria da qualidade dos registros analisados.

No presente artigo, discutimos um conjunto de técnicas e visualizações providas de recursos dinâmicos e interativos que têm o potencial de facilitar a identificação de possíveis inconsistências na base de dados. Além disso, permite ao usuário especialista obter uma visão geral dos dados e que possibilita a interpretação de padrões visuais globais. Vale ressaltar que as ferramentas de visualização desenvolvidas neste trabalho têm o propósito de instrumentar o especialista na fase de preparação para publicação da base e não para usá-las diretamente como interface com o público em geral, apesar de poderem ser ajustadas para essa finalidade.

Conforme Keim (2002) destaca, para que a exploração de dados seja efetiva, é importante que o aspecto humano seja integrado ao processo, combinando sua flexibilidade, criatividade e conhecimento à grande capacidade de armazenamento e poder de processamento dos computadores. Neste cenário, 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 detecção de padrões globais presentes na base de informação bem como a identificação de anomalias (Shiravi, 2011). Diversas são as aplicações em diversos contextos amplamente comentados em (Liu, 2014).

O *framework* apresentado neste trabalho foi desenvolvido no contexto de uma parceria entre a FGV/EMAp e Museu Nacional/Universidade Federal do Rio de Janeiro (MNRJ), mais especificamente, com curadores de algumas coleções biológicas. Parte das coleções biológicas (Zoologia e Botânica) do MN pode ser acessada em bancos de dados abertos que estão disponíveis no Sistema de Informações sobre Biodiversidade Brasileira[7] (SIB-Br), lançado em 2014 e que é um nó do *Global Biodiversity Facility*[8] (GBIF). Outra iniciativa importante é a base SpeciesLink[9] que conta com cerca de 15 milhões de registros de coleções online, incluindo dados do Herbário do MN.

Atualmente, o fluxo de trabalho nos metadados de coleções biológicas do MNRJ está sendo aprimorado e as ferramentas de visualização aqui propostas fazem parte dessa discussão. Os exemplos explorados dizem respeito à Coleção de Carcinologia[10] (MNRJ – CARCINO), valendo ressaltar que, por se tratar de metadado padronizado, a mesma discussão se aplica a quaisquer das bases de coleções biológicas que compartilhem tal padrão (*Darwin Core* - DwC).

7 https://ipt.sibbr.gov.br/mnrj/

8 https://www.gbif.org/publisher/4205110f-3f0f-40d8-bd0f-2fa71bc827b5

9 http://www.splink.org.br/

10 https://ipt.sibbr.gov.br/mnrj/resource?r=mnrj_carcinologia

No desenvolvimento do *framework* foi adotado o ecossistema da linguagem de programação *Python*, além de 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 às coleções de dados de natureza semelhante.

O presente artigo se desenvolve a partir da seguinte estrutura: na Seção 2 falamos sobre os conceitos que estruturam o raciocínio de desenvolvimento de representações visuais para dados, em particular, para coleções típicas de instituições GLAM. Na Seção 3 buscaremos expor os conceitos relacionados ao processo de verificação de qualidade de dados de coleções científicas. Na Seção 4 é apresentada a aplicação de visualização de metadados em coleções de dados de biodiversidade. Na Seção 5, apresentaremos a base de dados MNRJ - CARCINO e, na Seção 6, passaremos ao estudo de caso do uso de ferramentas de visualização de metadados como apoio à curadoria de dados. Por fim, tecemos conclusões e apontamos trabalhos futuros.

Sistematização das Representações Visuais de Dados e Coleções

Ao desenvolver soluções de representação visual de dados 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, 2003). Diferentes tipos de dados têm diferentes características e padrões de interesse que requerem um conjunto de ferramentas especializadas para visualizá-los (LIU, 2014).

Para o contexto de coleções digitais de herança cultural, Windhager, (2018) publicou um panorama do estado da arte no uso de ferramentas de visualização aplicadas à coleções de herança cultural e propõe uma categorização de sistemas de visualização, levando em conta os seguintes critérios: (1) características dos dados, (2) quem são os usuários, (3) quais são as tarefas a serem resolvidas através da representação gráfica, (4) qual a granularidade de apresentação dos dados e (5) qual a codificação visual adotada. Detalharemos estes critérios um pouco mais a seguir, chamando atenção para o fato de que, apesar da classificação ter sido realizada com base em coleções de herança cultural, a mesma também se aplica às coleções científicas biológicas. Nos restringiremos à discussão sobre recursos para visualização de metadados de dados cujo paradigma de registro é centrado no objeto, que é o caso que corresponde diretamente às coleções de dados de biodiversidade (PBR). Quanto aos usuários, a concepção de ferramentas de visualização de dados deve ser levada em conta, tipicamente dividindo-os em duas classes: (a) usuários especialistas, e (b) usuários casuais ou leigos. Caso o usuário não seja bem especificado, corre-se o risco de desenvolver um produto que não atende bem a nenhum dos dois públicos!

As tarefas a serem realizadas através da visualização também são de extrema importância para o desenvolvimento de uma solução efetiva. São classificadas em duas principais classes: *tarefas elementares*, relativas a itens da coleção ou busca por itens; e *tarefas sinópticas*, dizem respeito à busca de padrões presentes no conjunto como um todo. Tarefas sinópticas podem ser compreendidas como atividades analíticas que dão suporte ao usuário para uma compreensão global da coleção.

O sucesso de uma representação visual para um conjunto de dados é, em grande parte, associado à sua capacidade de facilitar a identificação de padrões presentes nos dados; e, em alguns casos, destacar a ausência de padrão

significativo também pode ser relevante. Nesse quesito, a clássica obra de Bertin (1983) traz uma rica discussão acerca de como técnicas tão simples quanto reordenar um eixo de variáveis categóricas podem revelar padrões visuais 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. O método de Bertin pode ser explorado na ferramenta Bertifier[11], implementada para web, e seus conhecimentos atemporais são empregados sempre que possível no contexto do presente trabalho.

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 ao processo, combinando sua capacidade cognitiva e conhecimento de domínio ao grande poder de processamento dos computadores. El Bekri (2016) também chama a atenção para importância do conceito de *user-in-the-loop* e destaca que, em muitos casos, o conhecimento de domínio de usuários especialistas é determinante para a boa 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.

De forma geral, sistemas e *frameworks* de visualização seguem o *pipeline* descrito por Liu (2014) que consiste de cinco etapas:

• **Transformação e análise de dados**: esse módulo geralmente consiste na produção de um novo conjunto de variáveis e/ou de registros a partir do conjunto de dados original, utilizando, quando necessário, técnicas de redução de dimensionalidade, suavização, redução de ruído, análise de conglomerados e interpolação a depender dos dados em questão.

• **Filtragem**: porções de dados com interesse específico são selecionadas a partir do *output* do módulo anterior para que sejam visualizados.

• **Mapeamento**: nesta etapa, dados pré-selecionadas são mapeados em primitivas geométricas (ex.: pontos, linhas, barras) e atributos (ex.: cor, posição, forma, tamanho).

• **Renderização**: nessa etapa, dados geométricos são transformados em imagens. Em outras palavras, as primitivas gráficas especificadas na etapa de mapeamento são processadas de modo a produzir uma imagem na tela respeitando as especificações tais como altura, largura, quantidade de pixels, etc.

• **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 resultados corretos e eficazes, são necessários diversos tipos de conhecimento específico. Por exemplo, a seleção e transformação inicial dos dados exige conhecimento de domínio para que, ao final do processo, as informações exibidas sejam fidedignas à base original. Sem esse tipo de especialidade, distorções podem ser criadas e levar a conclusões errôneas ou equivocadas. A ferramenta desen-

11 https://aviz.fr/bertifier

volvida neste trabalho contou com a colaboração de especialistas de domínio para o manuseio e filtragem de dados bem como na avaliação de utilidade das representações visuais propostas.

Para além da escolha da representação visual para os dados, uma camada de fundamental importância no paradigma *human-in-the-loop* diz respeito à interatividade implementada na visualização que pode permitir uma dinâmica de exploração dos dados favorável à checagem de informações por um especialista. Como paradigma de implementação de interatividade Shneiderman (2003) propõe o seguinte mantra, largamente usado em implementação de interfaces de visualização: "*overview first, zoom and filter, then details on demand*", em livre tradução poderia ser dito da seguinte forma: "Visão geral primeiro, zoom e filtragem na sequência, por fim, detalhes sob demanda".

Para coleções digitais Windhager (2018) apresenta um paradigma atualizado, alinhado com o mantra de Shneiderman, que é o conceito de granularidade visual. A granularidade visual se refere ao nível de agregação em que os dados estão sendo apresentados. A figura 1 ilustra esse conceito.

Fig. 1. Níveis de granularidade em coleções de objetos.

Fonte: elaboração dos autores

A apresentação de um único objeto (*single-object preview*) é um nível em que se representa um item individual da coleção, com acesso aos dados do registro. 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). Exemplos que implementam a transição entre distintos níveis de granularidade com mestria podem ser explorados na página do projeto Virkus Viewer[12]. Neste ponto sugerimos que o leitor explore alguns dos exemplos e ganhe familiaridade com os conceitos de visualização de coleções discutidos até agora.

Chama atenção o fato de que a grande maioria das soluções propostas para visualização de coleções pressupõem que os metadados tenham sido preparados previamente e estejam prontos para publicação. No entanto, o ferramental proposto na literatura de Visualização de Informação sugere um grande potencial de auxiliar no processo de preparação dos dados, a partir da utilização de representações gráficas desenhadas para exploração e análise dessas informações. No presente trabalho buscamos preencher essa lacuna ao propor uma solução na forma de *framework*, desenhada a partir de demandas de especialistas de domínio, com o objetivo de facilitar as tarefas de triagem, diagnóstico e

12 https://vikusviewer.fh-potsdam.de/

correção ao explorar visualmente as bases de dados em diferentes níveis de granularidade.

A presente aplicação inclui a visualização de dados de biodiversidade estruturadas no padrão *Darwin Core*, e que visa auxiliar usuários especialistas na tarefa da verificação da qualidade dos registros da base. Durante o desenvolvimento do conjunto de visualizações buscou-se adotar representações voltadas para a demanda do grupo de usuários principal, i.e., curadores especialistas e técnicos em coleções biológicas. Tal abordagem permitiu um processo de avaliação iterativo em todas as etapas de desenvolvimento (NORMAN, 2013) e com foco em dar apoio à curadoria de dados de coleções científicas do MN.

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. O emprego desses elementos visa auxiliar especialistas de domínio a realizar tarefas mais complexas como identificar possíveis anomalias em seus conjuntos de metadados e informações auxiliares que apontem para o registro exato na planilha de dados, por exemplo, exibir o número de catálogo do referido exemplar em uma *tooltip*, para que a correção seja facilmente efetuada, seguindo o paradigma do mantra de Shneiderman (2003).

Visualização de metadados como ferramenta de apoio a curadoria digital

Passaremos agora à discussão sobre o uso de modelos de visualização para assegurar a qualidade dos registros em um *dataset* e a usabilidade das informações também discutida em Liu (2018), Song (2018) e McCurdy (2018).

Nas etapas de coleta e pré-processamento de dados, podem ocorrer inconsistências como duplicações, imprecisões, perda de registros ou, até mesmo, aplicação de transformações irreversíveis nas informações originais. Artigos que abordam a preparação dos metadados costumam abordar o tema de forma muito técnica e frequentemente pouco acessível, distanciando o curador especialista da área que é objeto da coleção dessa discussão.

Ferramentas de preparação dos dados são muitas e vêm sendo cada vez mais usadas. No entanto, checagens de maior complexidade sobre os dados e quais tarefas de caráter sinóptico demandam se tornam desafiadoras. Para além disso, a interface dessas ferramentas é, usualmente, baseada em busca textual, ou seja, pressupõe que o usuário já sabe o que está procurando limitando a capacidade de serem detectados erros desconhecidos.

Em muitas aplicações, o processo de limpeza dos dados não pode ser completamente automatizado devido à ambiguidade de alguns erros (LIU, 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 é o caminho (EL BEKRI, 2016). Como já mencionado anteriormente, o conceito de *human-in-the-loop* tem 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.

A visualização de dados, através de gráficos simples, pode auxiliar enormemente essa tarefa de checagem, funcionando como um guia na tarefa de preparação e normalização da base. O uso desse tipo de recurso visual tem o potencial de facilitar, por exemplo, a detecção de outliers ao reforçar aspectos da cognição humana e acessar o conhecimento prévio de analistas.

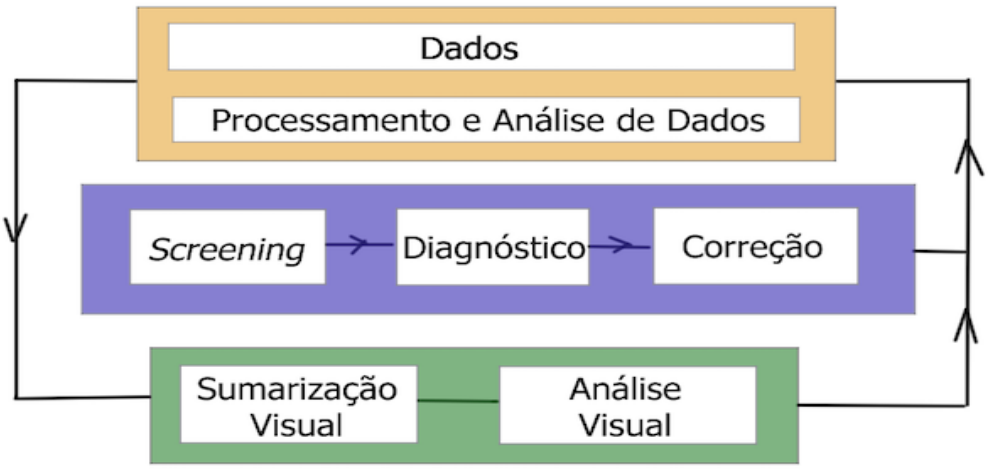
ISSN 2238-5436

Nesse sentido, Van den Broeck (2005) propõe um *framework* baseado em três etapas: Triagem[13]-> Diagnóstico-> Correção. Na etapa de triagem, o autor 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 devem ser restringidos exclusivamente à métodos estatísticos. Muitos *outliers* são detectados devido à experiência prévia do investigador, estudos-piloto, evidências na literatura ou, apenas, bom senso (VAN DEN BROECK, 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 positivo. 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 (2018) estende o *framework* proposto por Van den Broeck (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 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 2 ilustra as camadas que compõem o *framework* desenvolvido em Liu (2018): (1) manuseio de dados; (2) visualização conectada à camada (3) interação.

Fig. 2. Camadas do framework de análise visual proposto por Liu (2018): (1) manuseio de dados – em laranja; (2) visualização – em verde e conectada à camada (3) interação – em azul.



Fonte: elaboração dos autores

Cada um desses módulos foi, respectivamente, desenhado para: (1) extração de *insights*; (2) representação intuitiva e interpretação de dados; (3) facilitar exploração e análise segundo o conceito de *human-in-the-loop* (LIU, 2018) seguindo as seguintes etapas:

13 Livre tradução do termo original *Screening*.

• A partir dos dados coletados como dados de entrada, é realizada uma etapa de pré-processamento com o objetivo de revelar padrões, descobrir *outliers* ou, até mesmo, recuperar dados ausentes.

• Com base nos resultados da etapa anterior, é acionada uma etapa de interação incorporando os módulos originalmente propostos em Van den Broeck (2005), com a adição de representações visuais dos dados:

• Triagem: nesse módulo, o usuário é capaz de captar uma visão geral da coleção de dados, medidas estatísticas e determinados padrões por meio de gráficos intuitivos.

• Diagnóstico: por intermédio das visualizações, o usuário é capaz de identificar potenciais inconsistências e dados faltantes, duplicações, etc.

• Correção: Finalmente, a correção dos dados pode ser efetuava por métodos interativos.

Nesse ponto, torna-se claro o potencial de sistemas de visualização no suporte à interpretação de dados e tomada de decisão (LIU, 2018) e essa discussão pode ser estendida facilmente ao contexto de coleções de biodiversidade.

Visualizando Metadados de Coleções Biológicas

Coleções biológicas são constituídas por registros do tipo *Primary Biodiversity Records* (PBR), que consistem em metadados associados ao evento de coleta dos espécimes, identificadas pelo local e data de coleta, o coletor e a classificação taxonômica, que é uma classificação com estrutura hierárquica feita por um taxonomista identificado na base como determinador; além de eventuais campos adicionais que podem identificar o bioma ou outras particularidades. Consumidores dessas informações são, geralmente, especialistas de domínio e a exploração desses dados está presente em seu cotidiano, seja para a realização de estudos científicos ou para a manutenção e curadoria do material depositado nessas coleções. Para assegurar a precisão e a veracidade dos estudos conduzidos, é de fundamental importância que essas informações tenham alto nível de qualidade. A importância dessas coleções é amplamente reconhecida na literatura (SUAREZ, 2004) e muitas são suas aplicações nos mais diversos contextos, em especial, para avanços científicos das ciências biológicas.

Grandes instituições de história natural (museus, herbários, universidades, institutos de pesquisa, etc.) têm um papel crucial na manutenção e ampliação destes registros, uma vez que são responsáveis pela coleta e documentação de dados dos diferentes espécimes ao longo do tempo e espaço. 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. Contudo, para as instituições responsáveis por ampliar, manter e realizar a curadoria desses conjuntos de dados, com verificação de seus atributos qualitativos, nem sempre é uma tarefa trivial. Fatores como escala de grandes proporções e interdependência entre variáveis são desafiadores para a tarefa de correção e completude dos registros por parte dos curadores e técnicos responsáveis.

A partir das iniciativas de digitalização de coleções biológicas, milhões de registros de biodiversidade tiveram acesso facilitado em plataformas *online*

como resultado de esforços da digitalização efetuados por grandes instituições como museus, herbários, institutos de pesquisa e universidades. Artigos como Arts (2015), Marx (2013) e Reichman (2011), 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, 2009).

Para guiar a elaboração de representações visuais dos dados para a finalidade aqui proposta, se faz útil a adoção de uma metodologia de organização do processo criativo, capaz também de revelar diferentes possibilidades de exploração visual. Tal metodologia se baseia em responder perguntas fundamentais sobre os dados e tomá-las como base para criar os gráficos. Algumas perguntas se fazem relevantes em geral, são:

• O quê? Busca responder o que é alvo de estudo nos registros. No presente caso, refere-se ao conjunto de informações associadas à identificação taxonômica dos espécimes tombados. Por exemplo, acessando os campos de sua classificação taxonômica (Reino, Filo, Ordem, Gênero, Espécie).

• Onde? Indaga, literalmente, a localização geográfica em que um dado exemplar foi coletado. Normalmente, essa informação é gravada em coordenadas de latitude e longitude a partir de leitura via GPS diretamente no campo de coleta. A partir dessas coordenadas, consegue-se recuperar outras informações como País, Estado e cidade mais próximos. Contudo, devido à grande extensão temporal de coleções científicas, técnicas de georreferenciamento não estavam disponíveis quando alguns espécimes foram coletados. Nesses casos, aplica-se o processo inverso, isto é, estima-se as coordenadas de latitude e longitude a partir de informações como província e município de coleta, o que necessariamente guarda uma imprecisão inerente ao método.

• Quando? Esta pergunta diz respeito às datas de coleta e identificação taxonômica dos espécimes. Em casos de registros históricos em que uma data de identificação não tenha sido registrada na base, especialistas de domínio podem recorrer a datas de publicação disponíveis em artigos científicos referentes a exemplares presentes em suas coleções.

• Quem? Essa pergunta diz respeito às pessoas envolvidas nos processos de coleta, identificação e revisão dos registros digitais da coleção (coletores, taxonomistas e curadores). Como é de praxe, muitos dos curadores são docentes efetivos e especialistas em determinados grupos taxonômicos das respectivas coleções. Tal correlação pode ser evidenciada no banco de dados onde se vê o ciclo temporal da pesquisa e dados como ampliação, identificação e permuta de material realizada pelos curadores. Além disso, a duração da carreira destes especialistas pode coincidir com o período dedicado ao estudo de certos grupos, refletindo um padrão a ser explorado nas bases de dados. Estes dados estão altamente relacionados à memória institucional como mantenedora da coleção.

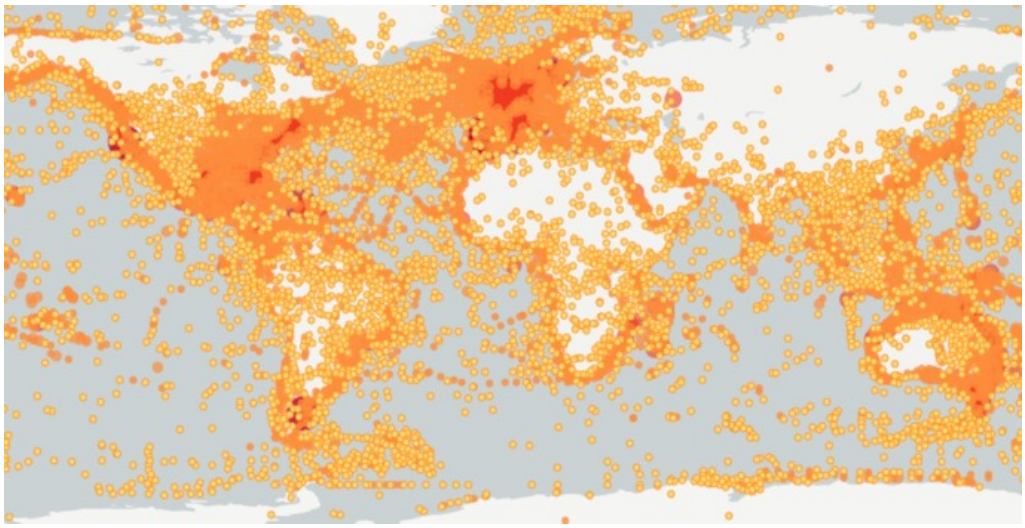
• Quanto? Essa pergunta diz respeito ao volume de espécimes coletados com certa característica de interesse. Sua resposta está associada a agregações quantitativas, como contagens, soma, média, dentre outras, para um conjunto específico.

Note que cada uma das perguntas essenciais pode ser associada a um ou mais campos das coleções de dados. Por exemplo, a pergunta "O quê?" está

diretamente associada aos campos que contém os diferentes níveis de classificação taxonômica (Reino, Ordem, Família, Gênero, Espécie, etc.). Por sua vez, a pergunta "Onde?" pode ser respondida por meio de coordenadas de latitude e longitude, como ilustra o exemplo da figura 3, que cumpre bem seu objetivo de mostrar a abrangência espacial da ordem Decapoda nas diferentes coleções disponibilizadas através da plataforma GBIF. O exemplo ilustra o potencial desse tipo de recurso visual para destacar e revelar padrões relevantes ao mesmo tempo em que apresenta a totalidade dos registros presentes nas bases de dados.

Estas perguntas podem guiar a estratégia de abordagem visual dos dados e a implementação de propostas de visualização, elevando seu grau de eficácia para a verificação de dados relacionados aos campos sujeitos à verificação. Ainda, soluções visuais, por *default*, têm a capacidade de apresentar mapas mentais de informações (LIU, 2014) e codificar múltiplas variáveis das bases de dados em diferentes dimensões, nos eixos X e Y, e em elementos estéticos da marca gráfica, tais como como cor, tamanho e forma. Desse modo, um único gráfico pode permitir a usuários explorar diferentes ângulos de visão das bases de dados a partir de combinações das perguntas essenciais. As questões essenciais podem ser combinadas para criar diferentes representações gráficas.

Fig. 3: Onde? Distribuição espacial de registros da ordem Decapoda disponíveis na plataforma GBIF (total de 2.339.670 de ocorrências no conjunto das bases).



Fonte: elaboração dos autores utilizando a plataforma de acesso livre GIBF (https://www.gbif.org/)

De forma similar ao exemplo de distribuição geográfica, gráficos de diferentes naturezas ilustram, de forma mais ou menos agregada, a informação contida nas bases de dados. Usualmente, gráficos que ilustram a evolução temporal das coleções 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. Estas decisões de agregação dos dados dialogam com o conceito de granularidade já mencionado.

Preparando os dados da Coleção de Carcinologia - MNRJ

As coleções biológicas são a base primária para as pesquisas da biodiversidade no mundo. O material acumulado a partir de expedições de coleta, permuta com outras instituições e doações, forma uma memória única da ocorrên-

cia, abundância e distribuição das espécies no espaço e no tempo, fornecendo um importante contexto histórico ao acervo para o entendimento da evolução da vida na terra. Em um contexto acadêmico as coleções científicas biológicas vêm sendo utilizadas em trabalhos de sistemática, biogeografia, geocronologia, estudos climáticos, evolução do sistema solar, evolução humana, entre outros. Em um contexto aplicado, tais coleções servem para subsidiar áreas estratégicas do governo como gestão, manejo e conservação ambiental, patrimonial e cultural. São também um recurso vital para responder às grandes questões de hoje, incluindo a manutenção da segurança alimentar, soluções para os cuidados de saúde, conservação da biodiversidade e as alterações climáticas.

Fundado em 1808, o Museu Nacional (MNRJ) é considerado a primeira instituição museal do Brasil. Suas coleções científicas vêm se formando desde meados do século XIX e são a base para as atividades de pesquisas, educação e formação de alunos do museu, que integra a Universidade Federal do Rio de Janeiro desde 1946. O Museu Nacional possui 15 coleções biológicas (não fósseis) e que estão distribuídas por quatro departamentos: Botânica (DB), Entomologia (DE), Invertebrados (DI) e Vertebrados (DV) (Serejo, 2020). A Coleção de Carcinologia (MNRJ – CARCINO) é uma das maiores da América Latina, com cerca de 30.000 registros, e está alocada no DI com outras sete coleções (Aracnologia, Celenterologia, Echinodermatologia, Malacologia, Polychaeta, Porifera e Inver Outros). Parte das coleções do DI encontrava-se no palácio de São Cristovão, que sofreu um grande incêndio em setembro de 2018 causando duras perdas ao acervo biológico do MNRJ. Apesar da perda física de muitos lotes/espécimes, os dados digitais permanecem em custódia dos curadores, que é o testemunho da coleção em si e reflete o resultado de investimento nessa área (digitalização dos dados) nos últimos 30 anos. Além da digitalização dos dados, se faz necessário a implementação de programas de gerenciamento de coleções (SPECIFY[14], JABOT[15], etc.) que vai proporcionar ao curador ferramentas de curadoria e gestão de dados. E como reflexo de uma boa gestão teremos um banco de dados limpo, atualizado e de preferência acessível ao público. Nesse sentido, as técnicas de visualização da informação atreladas aos bancos de dados de coleções científicas biológicas é um grande avanço na facilitação da qualidade e posterior divulgação dos mesmos para a comunidade em geral.

Fig. 4: Exemplar tipo da Coleção de Carcinologia/MNRJ coletado por Carlos Moreira em 1903. Lagostim *Metanephrops rubellus* (Moreira, 1903). Ordem Decapoda: Infraordem Astacidea: Família Nephropidae.

Foto: Cristiana S. Serejo

14 https://www.specifysoftware.org/

15 http://jabot.jbrj.gov.br/v3/consulta.php

O acervo utilizado no presente trabalho faz parte da Coleção de Carcinologia do Museu Nacional/MNRJ que compreende registros de animais invertebrados pertencentes ao filo Arthropoda – subfilo Crustacea. A título de exemplo, utilizamos os registros da ordem Decapoda para mostrar os resultados de visualização do presente trabalho. Esse grupo é bastante conhecido do grande público uma vez que inclui espécies economicamente importantes como os caranguejos, lagostas e camarões e é bem representado na coleção. A figura 4 ilustra um exemplar tipo, que são marcados com fitas vermelhas, de um lagostim da coleção de Carcinologia/MNRJ. O material tipo é considerado o testemunho que está perpetuamente associado ao nome de cada espécie de animais e plantas por ocasião da publicação original e que possui alto valor científico.

Algumas estatísticas descritivas da coleção de Carcinologia/MNRJ restrita a ordem Decapoda são: trata-se de um subconjunto que totaliza 8323 registros; cuja cobertura taxonômica compreende 108 famílias, 350 gêneros e 592 espécies; cobertura geográfica compreendendo 5 continentes, 23 países e 27 estados brasileiros; e cobertura temporal compreendendo um intervalo para ano de coleta entre 1871 e 2020, e para ano de determinação entre 1905 e 2020.

O processo de manuseio de dados foi realizado empregando-se a linguagem de programação Python 3. Em linha com as observações dadas por Oliphant (2007), essa escolha está fundamentada nos seguintes fatores:

• Python é uma linguagem amplamente adotada pela comunidade científica por ter alta flexibilidade, com sintaxe limpa e facilmente adaptável para lidar com bases de dados de diferentes tipos e tamanhos;

• 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;

• É defensável como uma linguagem de programação de baixo custo de entrada para usuários não programadores.

Essas são características desejáveis que visam prover uma solução sustentável para as equipes de curadoria do Museu Nacional/MNRJ. Vale observar que, ao trabalhar com coleções históricas, devemos ter em mente que muitos dos registros 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. Mesmo quando a digitação dos registros, em tempos mais recentes, é feita diretamente em ambiente computacional, caso não seja adotado um dicionário de vocabulário controlado, o mesmo tipo de erro de digitação ou variações de grafia de nomes próprios, são muito prováveis de acontecer. Nesse cenário, exemplos de erros comuns podem ser apontados tais como: (a) Confusão de caracteres com grafia semelhante. (b) Troca de um caractere por outro cuja tecla situa-se próxima à que se deseja pressionar.

Com o intuito de corrigir erros ocasionais de digitação a base de dados passou por uma etapa de pré-processamento seguindo as seguintes etapas:

• Colunas textuais, i.e., que contém nomes de pessoas, de lugares e classificação taxonômica, foram controladas para a presença de caracteres indevidos como espaços, 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 foram tratados para certificar-se de que todos os seus registros estavam representados na mesma unidade de medida. Foram retirados caracteres não-númericos (i.e., espaços, letras e símbolos). Por fim, tomou-se o cuidado de representar o separador decimal de maneira consistente.

• Coordenadas de latitude e longitude foram sujeitas ao seguinte processo:

• Foram eliminados quaisquer caracteres de caráter não-numérico, e todos os valores foram representados com o mesmo separador de casas decimais;

• Coordenadas representadas em formatos diferentes (i.e., DMS e UTM), foram convertidas para a convenção decimal;

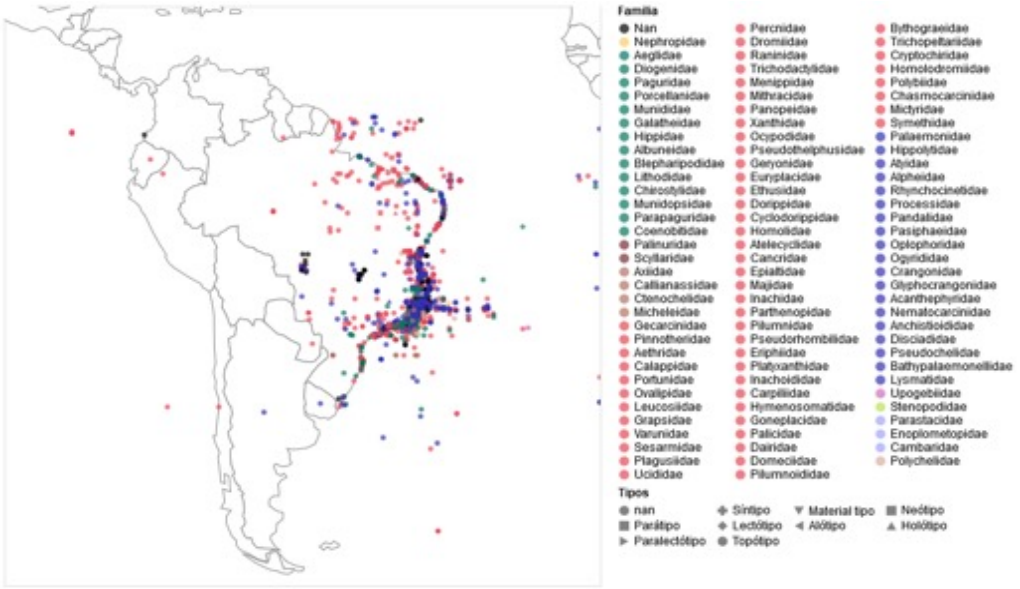
• Uma vez que todos os pares de coordenadas estavam representados no mesmo formato, foi verificada a presença de pontos que extrapolam os valores extremos (i.e., valores fora do intervalo [- 90o, +90o] para latitude e [- 80o, +180o] para longitude).

Apesar de os dados terem sido submetidos a uma etapa de pré-tratamento, 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). Finalmente, uma vez controlada a presença de pequenos erros de digitação e diferenças de formatação, a coleção está pronta para a etapa de codificação e exposição visual.

Visualizando a Coleção de Carcinologia/MNRJ

Seguindo a metodologia que busca responder às perguntas básicas a serem respondidas consultando os dados, as representações visuais produzidas buscam exibir as principais características dos dados, isto é, **cobertura geográfica**, **taxonômica** e **temporal**, além de um conjunto de visualizações que têm enfoque nos indivíduos responsáveis pelo ciclo de pesquisa dos espécimes, como **coletores** e **determinadores**. Aqui apresentaremos alguns gráficos selecionados, como ilustra a figura 5, e que se revelaram especialmente úteis para a conferência de informações da base. O conjunto de gráficos resultante é adequado para integração em ambiente web, para construção de *dashboards* ou para ilustrações de produções acadêmicas.

Fig.5: O quê? x Onde?: Distribuição geográfica da ordem Decapoda indicada pelas famílias representadas na coleção de Carcinologia/MNRJ, incluindo as informações dos tipos. As cores representam as diferentes infraordens vide fig. 6.



Fonte: elaboração dos autores

Para a implementação das visualizações, é necessário o uso de uma ferramenta de linguagem declarativa de gráficos. Nesse sentido, seguimos alinhados com Wang (2015) que buscam guiar o público interessado (biólogos e desenvolvedores) à visualização de dados biológicos, apresentando um conjunto de bibliotecas e instrumentos *open source*. Neste trabalho, dentre opções como ggplot2[16], D3.js[17], a dupla Vega e Vega-Lite[18] e sua prima-irmã Altair[19] optou-se pelo uso da ferramenta Altair, desenvolvida por VanderPlas (2018), por possibilitar um ambiente de visualização integrado com os recursos de tratamento e análise da base de dados em linguagem de programação Python, além de produzir gráficos finais que podem ser exibidos diretamente em *browsers* (em formato .html ou .svg). Dessa forma, é possível aliar os benefícios de uma linguagem de programação flexível, estruturas de gramática interativa de gráficos e a praticidade de uma exibição final em *browsers*. É importante salientar que essa escolha também leva em conta a manutenção e continuidade a longo prazo dos recursos aqui criados para as equipes técnicas e de curadoria do Museu Nacional/ MNRJ, que estão envolvidos em projetos que vão fomentar tal iniciativa.

16 https://ggplot2.tidyverse.org/

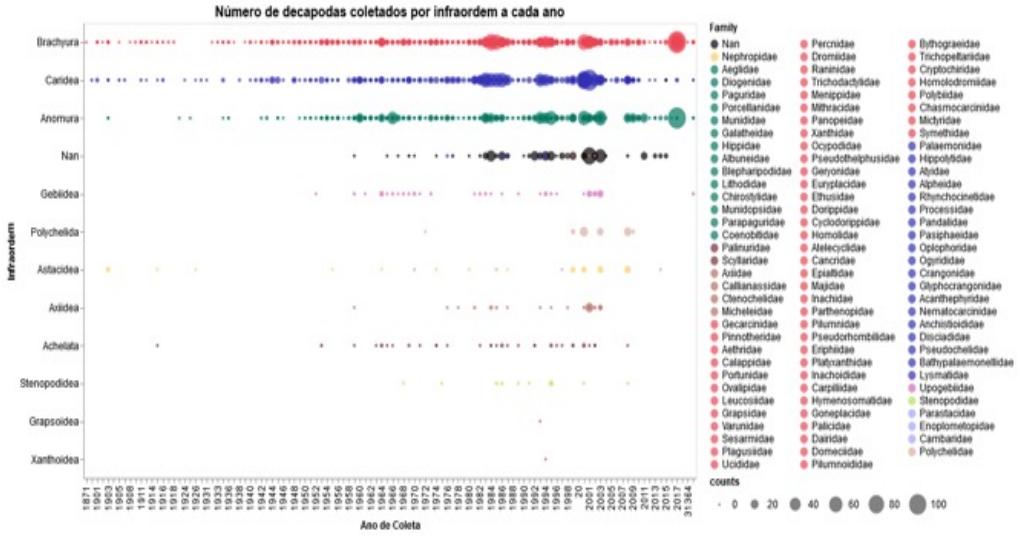
17 https://d3js.org/

18 https://vega.github.io/

19 https://altair-viz.github.io/#

ISSN 2238-5436

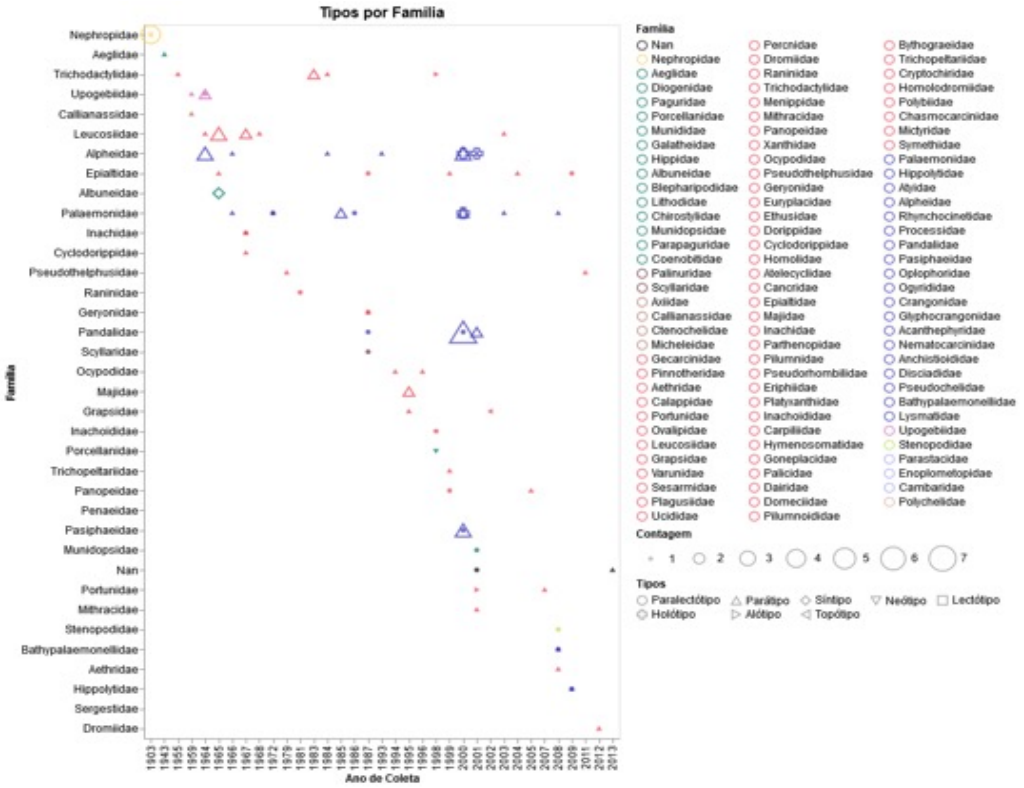
Fig.6: O quê? x Quando? Contagem de espécies por infraordem (nível da classificação taxonômica relevante para a ordem Decapoda) que ocorre na base.



Fonte: elaboração dos autores

Para produzir um conjunto sistemático de representações visuais da base de dados, foram propostas representações gráficas levando em consideração os níveis de granularidade definidos em Windhager (2018). Foram propostas representações de nível global (*multi-object previews*) que provêm a apresentação de registros agrupados para os diferentes níveis da classificação taxonômica além de buscar também apresentar o nível mais alto aonde todos os registos são exibidos no mesmo gráfico (*collection overview*). As figuras 6 e 7 exemplificam um par de representações gráficas que ilustram uma visão global da coleção sob distintos aspectos, no caso da figura 6 o enfoque é dado em um nível da classificação taxonômica (infraordem), enquanto na figura 7 o enfoque está no material tipo presente na coleção. Em todos os casos, o nível mais fino da granularidade, que diz respeito à informação individual de um registro, foi implementado através de recurso de interatividade.

Fig.7: O quê? x Quando?: Contagem de espécies tipo por família que ocorre na base. A presença de espécies tipo é de grande relevância para uma coleção de biodiversidade.



Fonte: elaboração dos autores

Para a concepção das funcionalidades de interatividade, buscou-se permitir o tráfego em profundidade entre diferentes níveis taxonômicos observados em gráficos distintos e, ao mesmo tempo, prover navegação horizontal através da aplicação de filtros e seleção de diferentes objetos. Isto posto, em cada visualização proposta neste estudo, foram empregadas as seguintes funcionalidades para:

• Imersão vertical: implementada por meio do recurso gráfico chamado de tooltip: uma moldura ou card flutuante que é exibido na tela quando se passa o mouse sobre um elemento da interface, contendo 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, através da informação exibida na tooltip, possibilitar a identificação exata do registro a ser verificado na base, ou para guiar a aplicação de filtros diretamente nas planilhas de dados, tornando mais eficaz a conferência da base.

• Navegação horizontal: esse tipo de interação está associado ao objeto da legenda de cores, permitindo a aplicação de filtros referentes às categorias codificadas 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.

ISSN 2238-5436

Os canais de codificação visual do Vega-Lite, e consequentemente Altair[20], apresentada em Satyanarayan (2016), são: X e Y (posição), cor (color), tamanho (size) e forma (shape). Cleveland (1984) sugere um ordenamento destes canais de codificação segundo sua efetividade em apresentar dados de acordo com seu tipo (nominal, ordinal e quantitativo). Buscou-se associar os canais visuais às variáveis da base de dados, sempre auxiliados pelos especialistas de domínio, buscando potencializar a revelação de padrões que pudessem responder ao conjunto de perguntas essenciais selecionado. Na figura 7 temos um exemplo em que o eixo X codifica o tempo, o eixo Y elenca as famílias para as quais ocorrem espécies tipo na base. A paleta de cores utilizada é a paleta de cores definida para as famílias utilizada consistentemente em todos os gráficos. Por fim, a forma diferencia os diferentes tipos, que são um conceito muito específico de domínio.

De modo geral, esse processo de associação entre variável visual e variável da base pode ser analisado sob critérios objetivos levando-se em consideração a pergunta a ser respondida pela visualização. As seguintes situações exemplo visam esclarecer o conceito:

a) Se a intenção é visualizar a cobertura espacial de espécies coletados presentes na base, vide figura 5, os campos correspondentes às coordenadas de latitude e longitude são priorizados para serem associados às coordenadas X e Y.

b) Caso deseje-se representar a contribuição de cada coletor ao longo do tempo, vide figura 8, devemos priorizar associar os campos que contém o nome dos responsáveis pela coleta dos espécimes e suas respectivas datas de coleta às coordenadas X e Y.

Esse tipo de fundamentação teórica é de crucial importância para se criar propostas de visualização eficazes. Tratando, ainda, dos demais canais de codificação, a cor recebeu atenção especial, sendo solicitado aos especialistas de domínio que agrupassem as famílias em animais que guardassem similaridades entre si, no presente caso, os tons de vermelho dizem respeito a distintas espécies de caranguejos, os azuis/roxos são espécies de camarões, os verdes são espécies de lagostins, e as demais cores são animais com menor diversidade de famílias. Esse agrupamento por famílias similares, produzida pelos especialistas, foi crucial para a efetividade do uso do canal de cor nas representações gráficas propostas. No conjunto de visualizações proposto neste estudo, buscou-se empregar as recomendações dadas por Qu (2016), especialmente mantendo-se fixa a paleta de cores usada no conjunto dos gráficos. O uso consistente da paleta de cores garante maior fluidez na leitura entre gráficos, além de evitar possíveis confusões aos especialistas no momento de executar as tarefas de manutenção e curadoria das coleções.

20 O Vega-Lite e Altair são ferramentas correspondentes entre si, diferindo na sintaxe mas não nos recursos implementados. O Altair utiliza sintaxe Python e trata-se de uma "tradução" automatizada da biblioteca Vega-Lite. Desse modo, uma vez que a biblioteca Veja-Lite é atualizada, o Altair é gerado novamente produzindo também uma atualização compatível com a nova versão.

Fig.8: Quem? x O quê? x Quando?: Contabilização de contribuição para a coleção dos 50 coletores que mais contribuíram para a coleção da ordem Decapoda.

Fonte: elaboração dos autores

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

• O canal visual de tamanho da marca gráfica tamanho (size) foi reservado para codificar dados quantitativos, nesse caso, quando foi empregado algum tipo de agregação estatística como contagem, soma, média, etc. Representar dados ordinais dessa forma é desencorajado por solicitar uma maior carga cognitiva do usuário.

• Informação temporal, quando retratada, foi codificada sempre no eixo X. A ordem intuitiva dessa informação sempre foi respeitada nas visualizações propostas.

• O canal de cor foi reservado para uso com variáveis nominais, atendendo a necessidade reportada por especialistas em representar a taxonomia dos espécimes visualizados nos diferentes gráficos.

Passamos agora a tecer comentários sobre como as representações visuais propostas para os dados foram capazes de potencializar a identificação de possíveis inconsistências nas bases de dados e, sempre que possível, 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. No caso da figura 8, podemos identificar potenciais correções a serem realizadas na grafia dos coletores. São observadas também algumas ocorrências de anos na linha do tempo que indicam uma necessidade de revisão no campo ano para o registro correspondente. Chamamos atenção para o fato de que, em mais de uma ocasião, os especialistas da coleção identificaram registros suspeitos muito complexos de serem detectados ao reconhecerem o nome de um pesquisador ao qual estão familiarizados

ISSN 2238-5436

apresentar uma ocorrência de coleta em um tempo incompatível com o tempo de atuação desse pesquisador. A origem do erro no registro digital é difícil de ser traçada, no entanto, a correção do erro pode ser feita a partir de consulta ao acervo ou a publicações acadêmicas relacionadas. A interatividade dos gráficos, permitindo a identificação do registro suspeito é uma funcionalidade imprescindível para a eficácia da tarefa de verificação da base.

Conclusões e Trabalhos Futuros

No presente trabalho buscamos aplicar princípios e técnicas advindos da literatura de Visualização da Informação aplicados aos metadados das coleções biológicas do Museu Nacional. Tais visualizações irão permitir explorar esses metadados de forma a potencializar a detecção de inconsistências na base de dados e contribuir diretamente para a atividade de curadoria digital dos acervos. Os códigos bem como exemplos que complementam a discussão aqui apresentada podem ser acessados em Oliveira (2021).

O *framework* desenvolvido no presente estudo se destinou à exibição e tratamento de coleções de dados do tipo PBR (*Primary Biodiversity Records*) em etapa previa à sua publicação. Garantir a qualidade desses registros é uma responsabilidade iminente dos curadores, especialistas e equipe técnica envolvidos, desse modo, o desenvolvimento do *framework* foi feito em parceria com os especialistas responsáveis pela qualidade dos registros e diversos ajustes nas bases foram feitas ao longo do processo, sempre sob supervisão direta dos curadores. As representações gráficas propostas contêm elementos dinâmicos e interativos capazes de elevar o grau de interação do usuário com a coleção.

Vale observar que, as mesmas representações visuais podem ser efetivamente utilizadas para a ilustrar as bases em *data papers*[21], i.e., documentos revisados pela comunidade científica que descrevem um conjunto particular de dados de biodiversidade. Além disso, 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 dos dados, no entanto, esse exercício é deixado a cargo de trabalhos futuros.

Entendemos, 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 biológicas se confirma na prática. Destacamos que essas atividades já são parte do dia a dia dos especialistas, que se valem de ferramentas tipicamente baseadas em interface textual (por exemplo, *OpenRefine*[22]), para a realização de tarefas de verificação das bases. No entanto, as ferramentas de representação visual complementam essas ferramentas de interface textual e se destacam pela capacidade de revelar padrões e anomalias nos dados de forma eficaz.

Por fim, uma contribuição original do presente trabalho no contexto de coleções biológicas é a proposta de representar visualmente a contribuição de cada pesquisador e coletor para a constituição da coleção ao longo do tempo. Tais representações são comuns em outros contextos, como o de esportes, por exemplo, mas não é de nosso conhecimento a utilização deste tipo de representação no contexto de coleções biológicas.

---

21 Para mais informações a respeito de *data papers*, vide \url{https://www.gbif.es/en/datos-biodiversidad/participa-en-gbif-es/data-papers/

22 https://openrefine.org/

Como trabalhos futuros, destaca-se a possibilidade de integrar as ferramentas de representação visual ao fluxo de trabalho dos especialistas, bem como a realização de pequenos ajustes nas visualizações propostas para a finalidade de divulgação de coleções para o público em geral.

Referências

ARTS, Koen; VAN DER WAL, René e ADAMS, William M. Digital technology and the conservation of nature. *Ambio*. vol. 44, no. 4, pp. 661- 673, Springer, 2015.

BERTIN, Jacques. *Semiology of graphics*: diagrams networks maps. ESRI Press, 1983.

COOPER, Alan; REIMANN, Robert et alli. *About face 2.0*. The essentials of interaction design. Vol. 17, Wiley Indianapolis, 2003.

CLEVELAND, William S e McGILL, Robert. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*. vol. 79, no. 387, pp. 531-554, Taylor & Francis Group, 1984.

EL BEKRI, Nadia e PEINSIPP-BYMA, Elisabeth. Assuring Data Quality by Placing the User in the Loop; 2016 *International Conference on Computational Science and Computational Intelligence* (CSCI)}, pp 468-471, IEEE, 2016.

FEKETE, J-D. The infovis toolkit. IEEE *Symposium on Information Visualization*, pp. 167-174, IEEE, 2004.

GURALNICK, Robert e HILL, Andrew. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics*. vol. 25, no. 4, pp. 421- 428, Oxford University Press, 2009.

HEER, Jeffrey; CARD, Stuart K e LANDAY, James A; Prefuse: a toolkit for interactive information visualization. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 421-430, 2005.

IBRAM. *Acervos digitais nos museus*: manual para realização de projetos, pp. 1–140, 2020.

KEIM, Daniel A. Information visualization and visual data mining, *IEEE transactions on Visualization and Computer Graphics*. vol. 8 no 1, pp 1-8, IEEE, 2002.

LIU, Shixia; CUI, Weiwei; WU, Yingcai e LIU, Mengchen. A survey on information visualization: recent advances and challenges. *The Visual Computer*, vol. 30, no. 12, pp 1373-1393, Springer, 2014.

LIU, Shixia; ANDRIENKO, Gennady; WU, Yingcai; CAO, Nan; JIANG, Liu; SHI, Conglei; WANG, Yu-Shuen e HONG, Seokhee. Steering data quality with visual analytics: The complexity challenge. *Visual Informatics*, vol. 2, no. 4, pp. 191—197, Elsevier, 2018.

MARX, Vivien. The big challenges of big data. *Nature*, vol. 498, no. 7453, pp. 255-260, Nature Publishing Group, 2013.

McCURDY, Nina; GERDES, Julie e MEYER, Miriah. A framework for externalizing implicit error using visualization. *IEEE transactions on visualization and computer graphics*. vol. 25, no. 1, pp. 925-935, IEEE, 2018.

NAVARRETE, Trilce e BOROWIECKI, Karol J. "Digitization of heritage collections as indicator of innovation". *Economics of Innovation and New Technology*. vol. 26, no 3, pp 227 – 246, 2017.

NORMAN, Don. *The design of everyday things*: Revised and expanded edition; Basic books, 2013.

OLIPHANT, Travis E. Python for scientific computing. *Computing in Science & Engineering*. vol. 9, no. 3, pp. 10-20, IEEE, 2007.

OLIVEIRA, Franklin A. *Visualização de coleções científicas digitais de biodiversidade:* um framework em Altair, Python. Dissertação de Mestrado em Modelagem Matemática, FGV\EMAp, https://bibliotecadigital.fgv.br/dspace/handle/10438/30711

QU, Zening e HULLMAN, Jessica. Evaluating visualization sets: Trade-offs between local effectiveness and global consistency. *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization,* pp. 44—52, 2016.

REICHMAN, O James; JONES, Matthew B. e SCHILDHAUER, Mark P. Challenges and opportunities of open data in ecology. *Science*, vol. 331, no. 6018, pp. 703-705, American Association for the Advancement of Science, 2011.

SATYANARAYAN, Arvind; MORITZ, Dominik; WONGSUPHASAWAT, Kanit e HEER, Jeffrey. Vega-lite: A grammar of interactive graphics. *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 341-350, IEEE, 2016.

SEREJO, C.S. *Panorama dos Acervos*: Passado, Presente e Futuro. Museu Nacional, Ed. Museu Nacional - Série Livros 18, 120 p, 2020.

SHIRAVI, Hadi; SHIRAVI, Ali e GHORBANI, Ali A. A survey of visualization systems for network security. *IEEE Transactions on visualization and computer graphics*, vol. 18, no. 8, pp. 1313-1329, IEEE, 2011.

SHNEIDERMAN, Ben. The eyes have it: A task by data type taxonomy for information visualizations. In: *The craft of information visualization*. 364--371, Elsevier, 2003.

SILVA, D.L.; CORRÊA, P.L.P.; JUAREZ, K.M.; FONSECA, R.L. *Diretrizes para a Integração de Dados de Biodiversidade*. Brasília: MMA, 100 p., 2015.

SONG, Hayeong e SZAFIR, Danielle Albers. Where's my data? evaluating visualizations with missing data. *IEEE transactions on visualization and computer graphics*. vol. 25, no. 1, pp. 914-924, IEEE, 2018.

SUAREZ, Andrew V e TSUTSUI, Neil D. The value of museum collections for research and society. *BioScience*. vol. 54, no. 1, pp. 66-74. American Institute of Biological Sciences, 2004.

THOMAS, Selma e MINTZ, Ann. *Virtual and the Real*: Media in the Museum. American Association of Museums, 1998.

TRIQUES. *A dimensão relacional entre curadoria digital e metadados*. Universidade Federal de São Carlos, SP. Centro de Educação e Ciências humanas - Programa de Pós-Graduação em Ciência da Informação. Tese de Doutorado, pp 128. 2020.

VAN DEN BROECK, Jan; CUNNINGHAM, Solveig Argeseanu; EECKELS, Roger e HERBST, Kobus. Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med*, vol. 2, no. 10, pp e267, Public Library of Science, 2005.

VANDERPLAS, Jacob; GRANGER, Brian E; HEER, Jeffrey; MORITZ, Dominik; WONGSUPHASAWAT, Kanit; SATYANARAYAN, Arvind; LEES, Eitan; TIMOFEEV, Ilia; WELSH, Ben e SIEVERT, Scott. Altair: interactive statistical visualizations for Python. *Journal of open source software*. vol. 3, no. 32, pp. 1057. 2018.

WANG, Rui; PEREZ-RIVEROL, Yasset; HERMJAKOB, Henning e VIZCAÍNO, Juan Antonio. Open source libraries and frameworks for biological data visualisation: A guide for developers. *Proteomics*, vol. 15, no. 8, pp 1356-1374, Wiley Online Library, 2015.

WINDHAGER, Florian. Visualization of cultural heritage collection data: State of the art and future challenges; WINDHAGER, Florian; FEDERICO, Paolo; SCHREDER, Gunther; GLINKA, Katrin; DORK, Marian; MIKSCH, Silvia e MAYR, Eva. *IEEE transactions on visualization and computer graphics*. vol 25, no. 6, pp 2311-2330, IEEE, 2018.

ISSN 2238-5436