Trabalho de Banco de Dados – Fauna Ameaçada

Igor de Jesus, Jhayson de Brito Jales, Pedro Arthur Santos Gama, Vitória de Souza Serafim

Instituto de Computação
Universidade Federal de Rio de Janeiro (UFRJ) – Rio de Janeiro, RJ – Brazil

{igorjs,jhaysonbj}@ic.ufrj.br, vitoriass@dcc.ufrj.br , pedrooarthursg@gmail.com

Abstract. *This Database work aims to build a web application that uses a modeled database to support the availability of data on the Brazilian fauna and in which category of threat of extinction it is found, as well as information on the units of existing conservation units and which units some species are associated with. The data comes from data.gov and the federal government's open data portal, being loaded, transformed to adapt to the modeling and loaded into the database.*

Resumo. *Este trabalho de Banco de Dados visa construir uma aplicação web que se utiliza de um banco de dados modelado para dar suporte a disponibilização de dados sobre a fauna brasileira e em qual categoria de ameaça de extinção ela se encontra, como também informações sobre as unidades de conservação existentes e a quais unidades algumas espécies estão associadas. Os dados provêm do dados.gov e do portal de dados abertos do governo federal, sendo carregados, transformados para se adaptar a modelagem e carregados no banco de dados.*

1. INFORMAÇÕES GERAIS

Utilizando dois datasets para as espécies ameaçadas, um do dados.gov e outro do portal de dados abertos, selecionamos algumas informações que consideramos relevantes para o escopo do nosso trabalho. Portanto as informações que selecionamos para a construção do banco de dados foram o nome científico da espécie, nome comum, estados a que pertence, unidades de conservação federal e estadual, grupo taxonômico, ordem, família e categoria de ameaça em 2021. Também utilizamos um terceiro dataset do portal de dados abertos sobre as unidades de conservação, onde havia mais informações sobre as unidades de conservação existentes no país. Desse dataset utilizamos as informações sobre a id da unidade de conservação, nome da unidade de conservação, esfera administrativa, ano de criação, unidades da federação que está a unidade de conservação e a quantidade de hectares de cada bioma que uma unidade de conservação possui.

2. MODELAGENS

2.1. MODELO ENTIDADE-RELACIONAMENTO (ER)

As entidades identificadas durante a modelagem foram:

- Espécie: com o objetivo de mapear o animal pelo nome científico, como o nome científico é único para cada espécie, utilizamos o nome científico como chave primária.
- Família: com o objetivo de mapear a família a qual cada espécie pertence. Como atributo colocamos nome_familia e id_familia
- Ordem: com o objetivo de mapear a ordem a qual cada família pertence. Como atributo colocamos nome_ordem e id_ordem
- Nome comum: com o objetivo de mapear os apelidos dados à espécie. Como atributos temos o id_nome_comum (chave primária) e nome_comum.
- Categoria de ameaça: com o objetivo de mapear os níveis de ameaça que determinada espécie possui, . Dentre as possibilidades de classificação, temos: Vulnerável (VU); Criticamente em Perigo provavelmente extinto (CR)(PEX); Extinta (EX); Criticamente Em Perigo (CR); Extinta na Natureza (EW); Regionalmente Extinta (RE); Subespécie que sai da Lista. Como atributos temos o id_categoria (chave primária) e a sigla da categoria.
- Grupo taxonômico: com o objetivo de mapear os grupos taxonômicos de cada espécie, possuindo dois atributos, id_grupo (chave primária) e nome_taxonomico. Dentre os grupos taxonômicos, temos: anfíbios; aves; invertebrados marinhos; invertebrados terrestres; invertebrados de agua doce; invertebrados de agua salgada ; Mamíferos; Mamíferos Aquáticos; Peixes Continentais; Peixes Marinhos; Répteis
- Estado: com o objetivo de mapear as unidades federativas do Brasil, incluindo o Distrito Federal. Como atributos temos a sigla (chave primária) e nome_estado.

- Unidade de conservação: com o objetivo de mapear todas as unidades de conservação do Brasil. Como atributos, mapeamos id_unidade (chave primária), esfera administrativa (federal, municipal ou estadual), ano de criação de criação e o nome da unidade.
- Bioma: com o objetivo de mapear todos os biomas do Brasil, sendo eles Amazônia, Caatinga, Cerrado, Mata Atlântica, Pampa, Pantanal, Marinho. Como atributos temos id_bioma(chave primária) e nome_bioma.

Os relacionamentos e cardinalidades identificadas durante a modelagem foram:

- Classificada: Indica que, cada espécie possui apenas 1 categoria de ameaça, mas cada categoria de ameaça classifica 1:n espécies.
- Apelidada por: Indica que, cada espécie pode possuir nenhum ou muitos apelidos. Já o apelido pode ser utilizado para apelidar 1 ou muitas espécies, isso por que o apelido é dado pela população e, de forma geral, se o animal não possuir características muito singulares, a população tende a dar um nome genérico que apelida mais de uma espécie.
- Pertence (grupo taxonômico e espécie) : Indica que, cada espécie pertence a apenas um grupo taxonômico, enquanto cada grupo taxonômico agrupa pelo menos 1 espécie, podendo agrupar várias, isso é devido a critérios biológicos.
- Pertence (espécie e família): Indica que, cada espécie pertence a apenas um uma família, enquanto cada família possui pelo menos 1 espécie, podendo possuir várias, isso é devido a critérios biológicos.
- Pertence (família e ordem): Indica que, cada família pertence a apenas um uma ordem, enquanto cada ordem possui pelo menos uma família, podendo possuir várias, isso é devido a critérios biológicos.
- Protege (Unidade de conservação e espécie): Indica que, cada espécie é protegida 0 ou ‘n’ unidades de conservação, além de refletir a realidade do nosso banco de dados, na prática essa cardinalidade também faz sentido, é de se esperar que nem todas as espécies ameaçadas estejam protegidas, pois, se fosse esse o caso, muito provavelmente essas espécies não estariam ameaçadas. A unidade de conservação protege 0 ou ‘n’ espécies, essa cardinalidade apenas faz sentido por que o dataset das unidades de conservação era mais atualizado se comparado ao dataset das espécies. Portanto, apesar de não refletir a realidade, no nosso banco de dados, possuímos unidades de conservação que não protegem nenhuma espécie, devido a assincronia dos datasets.
- É nativa de: Indica que, cada espécie é nativa de 0 ou ‘n’ estados, isso por que existem espécies com características mais gerais que não são nativas de nenhum estado, assim como existem espécies com características mais específicas que são nativas de mais de um estado. Cada estado do Brasil, incluindo o Distrito Federal, possui 1 ou ‘n’ espécies nativas, poderia ser o caso de um estado não possuir uma espécie nativa, mas analisando os dados, o grupo não identificou nenhum caso em que isso ocorresse.
- Pertence (unidade de conservação e estado): Cada unidade de conservação pode pertencer a 1 ou ‘n’ estados, isso por que existem unidades que estão localizadas

na divisão de algum estado, pertencendo a mais de um estado. Todos os estados do Brasil possuem pelo menos uma unidade de conservação, por isso o estado possui 1 ou ‘n’ unidades.

- Possui (unidade de conservação e bioma): Cada unidade de conservação possui 1 ou ‘n’ biomas, pela natureza dos dados analisados notamos que não há nenhuma unidade de conservação sem definição do bioma associado. Todo bioma é associado a pelo menos uma unidade de conservação, por isso o bioma possui 1 ou ‘n’ unidades de conservação. Além disso, do relacionamento entre unidade de conservação e bioma surge o hectare, cujo objetivo é indicar a quantidade de hectare que determinada unidade possui daquele bioma.

Vale destacar que, por transitividade, se cada espécie pertence a apenas uma família e, uma família pertence a apenas uma ordem, uma espécie pertence a apenas uma ordem. Com isso, concluímos que mapeamos implicitamente a relação que indica ordem de determinada espécie.

Além disso, como o dataset utilizado para armazenar as unidades de conservação é mais recente que o dataset que correlaciona a espécie com a unidade de conservação que a protege, o nosso banco de dados armazena unidades de conservação que teoricamente não possuem nenhuma espécie associada.

Outro fator importante é que, apesar de não refletir a realidade, no nosso banco de dados, possuímos unidades de conservação que não protegem nenhuma espécie, devido a assincronia dos datasets. Ainda que os dados de espécie e unidades de conservação fossem publicados no mesmo ano, é de se imaginar que mapear as unidades de conservação é um trabalho significativamente mais rápido do que mapear todas as espécies do Brasil. Então, ainda que os dados fossem publicados no mesmo ano, não necessariamente ambos foram coletados no ano de publicação. A respeito do ano de coleta dos dados, não possuímos informações nas páginas de referência dos datasets.

2.2. MODELO LÓGICO

O modelo lógico representa completamente o modelo ER. Nas relações entre entidades do tipo 1:n foram adicionadas chaves estrangeiras na entidade com cardinalidade máxima n que fazem referência a chave primária das entidades com cardinalidade máxima 1, isto é devido ao fato de que as entidades com cardinalidade máxima ‘n’ em relacionamentos 1:n estão relacionadas a, no máximo, uma instância da entidade de cardinalidade máxima 1.

Já para as relações do tipo n:n, foram criadas tabelas intermediárias que possuem chaves estrangeiras que fazem referência a cada uma das chaves primárias das entidades envolvidas no relacionamento. Então, para cada linha dessa tabela que representa o relacionamento teremos a informação de quais linhas de uma entidade estão relacionadas a quais linhas da outra entidade.

2.3. MODELO FÍSICO

O modelo físico é uma derivação direta do modelo lógico. Todos os modelos elaborados neste trabalho podem ser encontrados pelos link disponíveis nas referências.

3. TRATAMENTO DOS DADOS E SELEÇÃO DAS COLUNAS

A principal dificuldade enfrentada pelo grupo foi o tratamento de dados dos arquivos csv's. Alguns exemplos de dificuldades na manipulação de dados foram variações nos dados textuais, mas que se referiam a um mesmo ente do mundo real, abreviações de nomenclaturas das unidades de conservação, caracteres não identificados pelo encoding, entre outros.

Podemos separar em duas partes, os dados sobre as espécies e suas informações e os dados sobre as unidades de conservação.

Para as espécies havia dois arquivos csv´s, um proveniente do dados.gov e outro do portal de dados abertos, fizemos uma junção das duas tabelas, usando como critério o nome científico das espécies. Após isso, selecionamos os dados que desejávamos de cada uma das tabelas. Na tabela dos dados.gov, selecionamos as colunas que continham os dados sobre o Nome Científico, Nome Comum, Estado, Unidade de Conservação Federal e Unidade de Conservação Estadual. Já na tabela do portal de dados abertos, selecionamos as colunas Grupo Taxonômico, Ordem, Família, Sugestão de Categoria 2021. A partir disso, atribuímos id's as categorias de ameaça, grupos taxonômicos, nomes comuns, famílias e ordens, pois, eram informações fixas que se repetiam nas espécies. Dessa forma, era possível modelar em entidades separadas que manteriam essas informações a qual a espécie faria referência com uma chave estrangeira. Como a quantidade de estados do Brasil também é uma quantidade fixa, modelamos como uma entidade que armazena seu nome e é identificada por sua sigla, e da coluna de Estado das espécies que nos diz de quais estados elas pertencem criamos a relação 'eh_nativa_de' que podia ou não ter os estados a qual uma espécie pertencia. O que notamos foi que era comum espécies terem mais de um estado dos quais são nativas e as que não tinham dados relacionados em sua grande maioria são espécies marinhas e nossa hipótese é que espécies marinhas provavelmente não têm como ser atribuídas a localidades fixas.

Para as unidades de conservação utilizamos um terceiro arquivo csv do portal de dados abertos que continha bastante informações sobre cada unidade de conservação e selecionamos um subconjunto de colunas que continham informações de Id da Unidade, Nome da Unidade, Esfera Administrativa, Ano de Criação, Unidades da federação e um conjunto de colunas nomeadas com o nome de cada bioma brasileiro que representava a quantidade em hectares de cada bioma presente naquela determinada unidade. Como também o conjunto de biomas brasileiros é fixo e pequeno, modelamos como uma entidade que as unidades de conservação iriam se referenciar com a informação de quantos hectares possuem de cada bioma. E por fim, o maior desafio foi com base nas unidades de conservação que as espécies pertenciam, associá-las no relacionamento 'protege', onde conforme dado o nome científico da espécie informar a quais id's das

unidades de conservação que as protegem. Como eram de lugares diferentes (dados.gov e portal de dados abertos) havia diferenças na representação dos dados sobre as unidades de conservação. Por exemplo, na tabela das espécies era comum utilizarem abreviações como ESEC, PARNA, APA, dentre outras. Essas abreviações significam Estação Ecológica, Parque Natural e Área de Preservação Ambiental, respectivamente. Enquanto na tabela das unidades os nomes estavam escritos sem abreviações em sua grande maioria. Outro exemplo é o uso ou não de preposições que às vezes eram utilizadas ou não nos nomes das mesmas unidades, fazendo com que em alguns casos em uma tabela o nome da unidade tinha preposição no nome e na outra não tinha, ainda que estivessem se referindo a mesma unidade de conservação. Contornarmos tal problema retirando as abreviações e pontuações dos nomes das unidades na tabela das espécies e para todas que conseguimos posteriormente um resultado único ao pesquisar o nome da unidade na tabela das unidades de conservação associamos diretamente e as que não possuíam resultados únicos por ser um número consideravelmente pequeno selecionamos manualmente dentre os resultados observados pois era possível identificar somente analisando a qual unidade da lista de resultados estaria se referindo.

4. QUERIES CONSTRUÍDAS E IMPLEMENTADAS

Dentre as queries elaboradas, temos:

```
Unset
# QUERY 1 - Exibe o grupo taxonômico e a quantidade de espécies de cada
grupo taxonômico

SELECT grupo_taxonomico.nome_taxonomico, COUNT(especie.nome_cientifico)
AS total_especies
FROM grupo_taxonomico
JOIN especie ON grupo_taxonomico.id_grupo =
especie.fk_grupo_taxonomico_id_grupo
GROUP BY grupo_taxonomico.nome_taxonomico;
```

Utilizada no Dashboard de dados da Página.

```
Unset
# QUERY 2 - Exibe o grupo taxonômico, sua situação de risco e a
quantidade de espécies que estão simultaneamente nesse grupo taxonômico
e nessa categoria de risco

SELECT grupo_taxonomico.nome_taxonomico, categoria_de_ameaca.situacao,
COUNT(*) AS quantidade
FROM grupo_taxonomico
JOIN especie ON grupo_taxonomico.id_grupo =
especie.fk_grupo_taxonomico_id_grupo
JOIN categoria_de_ameaca ON especie.fk_categoria_de_ameaca_id_categoria
= categoria_de_ameaca.id_categoria
```

```
WHERE categoria_de_ameaca.situacao IN ('EX', 'EW', 'CR')
GROUP BY grupo_taxonomico.nome_taxonomico, categoria_de_ameaca.situacao;
```

Utilizada no Dashboard de dados da Página.

```
Unset
# QUERY 3 - Exibe os estados e a quantidade de animais nativos
associados a esse estado

SELECT estado.nome_estado,
COUNT(eh_nativa_de.fk_especie_nome_cientifico) AS total_animais_nativos
FROM estado
LEFT JOIN eh_nativa_de ON estado.sigla = eh_nativa_de.fk_estado_sigla
GROUP BY estado.sigla;
```

Utilizada no Dashboard de dados da Página e no Mapa onde era possível selecionar o dado por estado.

```
Unset
# QUERY 4 - Exibe os estados e a quantidade de unidades de conservação
associadas a esse estado

SELECT estado.nome_estado, COUNT(unidade_de_conservacao.id_unidade) AS
quantidade
FROM estado
INNER JOIN pertence ON estado.sigla = pertence.fk_estado_sigla
INNER JOIN unidade_de_conservacao ON
pertence.fk_unidade_de_conservacao_id_unidade =
unidade_de_conservacao.id_unidade
GROUP BY estado.nome_estado;
```

Utilizada no Dashboard de dados da Página e no mapa onde era possível selecionar o dado por estado.

```
Unset
# QUERY 5 - Exibe a família e a quantidade de espécies associadas essa
família, apenas quando essa quantidade superar 10 unidades

SELECT familia.nome_familia, (
       SELECT COUNT(*)
```

```
      FROM especie
      WHERE especie.fk_familia_id_familia = familia.id_familia
) AS quantidade_especies
FROM familia
HAVING quantidade_especies > 10;
```

Utilizada no Dashboard de dados da Página.

```
Unset
# QUERY 6 - Exibe o nome de cada bioma e a quantidade de unidades de
conservação associadas a cada bioma

SELECT bioma.nome_bioma, (
  SELECT COUNT(*)
  FROM possui
  INNER JOIN unidade_de_conservacao ON
possui.fk_unidade_de_conservacao_id_unidade =
unidade_de_conservacao.id_unidade
  WHERE possui.fk_bioma_id_bioma = bioma.id_bioma
) AS quantidade
FROM bioma;
```

Utilizada no Dashboard de dados da Página.

```
Unset
# QUERY 7 - Exibe a quantidade de animais em cada situação de risco

SELECT situacao, count(nome_cientifico) as total
FROM categoria_de_ameaca
JOIN especie on id_categoria=fk_categoria_de_ameaca_id_categoria GROUP
BY id_categoria ORDER BY total desc;
```

Utilizada no Dashboard de dados da Página.

```
Unset
# QUERY 8 - retorna a quantidade de espécies que não têm nome popular

SELECT count(nome_cientifico)
FROM especie
LEFT JOIN apelidada_por on nome_cientifico=fk_especie_nome_cientifico
WHERE fk_especie_nome_cientifico IS NULL;
```

Utilizada no Dashboard de dados da Página.

5. DESENVOLVIMENTO DA APLICAÇÃO WEB

Como parte da elaboração do trabalho foram desenvolvidas 3 páginas (Landing Page, Dashboard de dados e Mapa interativo) onde o principal objetivo era possibilitar a pesquisa dos dados que foram tratados e inseridos no nosso banco de dados. Devido ao curto prazo para o desenvolvimento das partes estéticas e construção do Backend da aplicação, foram decididas a utilização de apenas algumas queries onde fosse possível ver a funcionalidade do banco de dados, já que não seria viável chamadas mais rebuscadas que integrassem totalmente os usuários com os dados dispostos, visto que o tempo proposto de desenvolvimento era curto.

Então, para facilitar todo esse processo utilizamos as ferramentas nas quais o grupo tinha mais familiaridade e maior entendimento na hora de implementar. São essas: Typescript e Tailwind Css para o Front-End, além do Python e Flask para o Back-End. Essas escolhas foram feitas visando a simplicidade e melhor execução do projeto. Ainda assim, mesmo com o conforto do grupo com as stacks escolhidas, alguns impedimentos foram encontrados na codificação e nas buscas/requisições no backend, uma vez que somente com o conhecimento que tínhamos só era possível fazer buscas com queries fixas. Um exemplo disso é que não conseguimos buscar a quantidade de espécies de uma família específica apenas a quantidade de espécies de todas as famílias presentes no Brasil, deixando o Front recheado de filtros de dados - Uma observação é que provavelmente era possível fazer essas queries não serem tão estáticas mas não foi possível fazer uma pesquisa mais a funda para otimizar essas buscas.

Concluindo, embora o projeto atualmente seja simples, acreditamos que exista espaço para a implementação de funcionalidades mais ‘interativas’, assim como, a exibição de mais dados utilizando o mapa desenvolvido. Além disso, deixamos disponível para a visualização pública o repositório do projeto junto do banco de dados e da documentação necessária para servir a aplicação em Portal de Dados Animal Crossing .

6. ATRIBUIÇÕES DE CADA MEMBRO

Dentre as responsabilidades e obrigações de cada membro, subdividimos o grupo em duas principais obrigações: o subgrupo formado por Igor e Jhayson foi responsável pela modelagem até a integração do banco de dados com o back-end. O subgrupo formado por Vitória e Pedro foi responsável por integrar o back-end com o front-end, além do desenvolvimento da página web. De forma detalhada, temos:

Igor: Membro responsável pelo tratamento de dados; exploração dos datasets; modelagem ER; modelagem lógica; modelagem física; definição das cardinalidades dos relacionamentos; criação dos slides de apresentação; integrar o back-end com o banco de dados; desenvolvimento das queries utilizadas.

Jhayson: Membro responsável pela modelagem ER, modelagem lógica; modelagem física; definição das cardinalidades; criação da estrutura do banco de dados; inserção dos dados; auxiliou o Igor no tratamento de dados/ exploração dos datasets; criação dos

slides de apresentação; integração do back-end com o banco de dados; desenvolvimento das queries utilizadas.

Pedro : Membro responsável pelo desenvolvimento e estilização de parte da aplicação web , além da integração do back-end com o front, implementando as requisições das consultas com o Flask, e contribuição com parte da modelagem ER.

Vitória: Membro responsável pelo desenvolvimento da aplicação web nas partes de organização das páginas, criação e estilização dos componentes utilizados e integração das tabelas de dados fazendo as requisições no back através de rotas. Além da adaptação do mapa que originalmente era em html para a linguagem de Typescript e Tailwind Css. Também ficou responsável pela organização do ReadMe do projeto onde é detalhado as instruções para servir a aplicação junto com o banco de dados.

7. REFERÊNCIAS

1. Dataset das espécies da Fauna em Unidades de Conservação Federais https://dados.gov.br/dados/conjuntos-dados/especies-da-fauna-em-unidades-de-conservacao
2. Dataset das unidades de conservação https://dados.mma.gov.br/dataset/44b6dc8a-dc82-4a84-8d95-1b0da7c85dac
3. Dataset da fauna e as especies ameaçadas de extinção 2021 https://dados.mma.gov.br/dataset/especies-ameacadas
4. Dataset das unidades federativas do brasil https://github.com/leogermani/estados-e-municipios-ibge/blob/master/estados.csv
5. Link para as explicações conceitos biológicos utilizados (ordem e família) https://educacao.uol.com.br/disciplinas/biologia/taxonomia-como-funciona-o-sistema-de-classificacao-dos-seres-vivos.htm
6. Link para o projeto Back end e Front end no github https://github.com/VitoriaSerafim/portal-de-dados-mma/tree/main
7. Link para o modelo ER, modelo lógico e modelo físico utilizados https://github.com/Igor-dJS/Trabalho_BD/tree/main/modelos
8. Link para a tabela tratada, após a limpeza e seleção das colunas utilizadas https://github.com/Igor-dJS/Trabalho_BD/tree/main/dataset
9. Link para as tabelas utilizadas no banco de dados, com as devidas normalizações e chaves primárias, chaves estrangeiras e tabelas intermediárias em caso de relacionamento n:n

   https://github.com/Igor-dJS/Trabalho_BD/tree/main/tabelas
10. Link para o arquivo .sql com todos os dados inseridos no banco https://github.com/Igor-dJS/Trabalho_BD/blob/main/banco/trabalho.sql