Revista Eletrônica de Iniciação Científica em Computação, 2025, 23:1
ISSN: 3085-8461 • doi: 10.5753/reic.2025.5984

Artigo de Pesquisa/Research Paper

Portal de Gestão e Qualidade de Dados em Big Data

Big Data Management and Data Quality Portal

Flávio Midea [Universidade de São Paulo |*flavio.midea@usp.br*]
Felipe Almeida [Universidade de São Paulo |*fvalencia@usp.br*]
Pedro Corrêa [Universidade de São Paulo |*pedro.correa@usp.br*]
Alan Calheiros [Instituto Nacional de Pesquisas Espaciais |*alan.calheiros@inpe.br*]

Escola Politécnica, Universidade de São Paulo, Av. Prof. Luciano Gualberto, 380, Butantã, São Paulo, SP, 05508-010, Brasil.

Resumo. O crescente volume de dados demanda maior controle e monitoramento das informações geradas. Nesse contexto, ferramentas de gerenciamento de dados tornam-se essenciais para atender às necessidades de grandes instituições no Brasil. Órgãos como o ICMBio e o INMET já utilizam soluções desse tipo, mas o INPE ainda carece de um portal específico que atenda a essas demandas. Este artigo apresenta o desenvolvimento de uma ferramenta baseada nos princípios FAIR, voltada à inserção e indexação padronizada de dados em um banco de dados não relacional. Os resultados demonstram que a ferramenta atende adequadamente aos princípios FAIR.

Abstract. The growing volume of data requires increased control and monitoring of the information produced. In this context, data management tools have become essential in meeting the needs of major institutions in Brazil. Agencies such as ICMBio and INMET already employ such solutions; however, INPE still lacks a dedicated portal to address these demands. This paper presents the development of a tool based on the FAIR principles, designed for standardized data entry and indexing in a non-relational database. The results demonstrate that the tool effectively meets the FAIR principles.

Palavras-chave: Big data, qualidade de dados, portal de dados

Keywords: Big data, data quality, data portal

Recebido/Received: 01 May 2025 • **Aceito/Accepted:** 02 October 2025 • **Publicado/Published:** 17 October 2025

1 Introdução

Gerir dados é a base para qualquer área de atuação, além de que, organizar as informações e o conhecimento são ações que provêm de séculos. Nos dias atuais, possuir informações com alta qualidade sobre seus usuários, produtos, serviços e operações podem resultar em melhores decisões. Porém, produzir e administrar esses dados é um processo complicado, ainda mais em contextos de grande volume, sendo necessários padrões que sigam conceitos regidos para a adequação destas informações, junto a um sistema capaz de controlar este fluxo [Barbosa *et al*., 2021; Hazen *et al*., 2014].

O Brasil hoje carece de portais capazes de fornecer um alto volume de dados oriundos de diferentes projetos de pesquisa, sendo um dos órgãos afetados com essa necessidade o Instituto Nacional de Pesquisas Espaciais (INPE). Instituições como o INPE trabalham com diversas fontes e tipos de dados aeroespaciais que são estudados constantemente. Esses estudos buscam avaliar como essas informações influenciam tanto na qualidade de vida da população, quanto no desenvolvimento de novas tecnologias para estimular o desenvolvimento do País.

Diante do elevado volume de dados processados pelo INPE, torna-se evidente a necessidade de monitorar e gerenciar cuidadosamente a entrada dessas informações no sistema, a fim de assegurar sua qualidade e conformidade com os padrões exigidos para a pesquisa científica. Por se tratarem de dados aeroespaciais coletados ao longo de extensos períodos de tempo — muitas vezes durante anos — e com altas frequências de amostragem (como uma medição por segundo), esses dados acumulam um volume significativo de informações.

Fatores como esses mostram a necessidade de uma gestão ativa que vai além de apenas armazenar essas informações, requerendo um sistema confiável para a publicação de dados, baseado em uma robusta infraestrutura digital. Assim, o potencial de reuso das informações para pesquisas está fortemente relacionado à adoção de melhores práticas de gestão, buscando a estruturação de interoperabilidade adequada, reusabilidade dos metadados de qualidade e na acessibilidade dos dados [Henning *et al*., 2018; Sales and Sayão, 2018].

Um padrão importante utilizado como base são os princípios FAIR, abrangentes e que buscam garantir que a pesquisa produzida a partir dos dados obtidos possa ser sempre validada de maneira independente por outros pesquisadores. Os princípios FAIR são acrônimo para **F**indable (localizável), **A**ccessible (acessível), **I**nteroperable (interoperável) e **R**eusable (reutilizável). Esses princípios surgiram com o propósito de melhorar a infraestrutura e a prática de compartilhamento de dados. O detalhamento de cada princípio pode ser visto através do domínio GoFAIR [1] [Wilkinson *et al*., 2016].

Iniciativas como GoFAIR buscam em âmbito internacional desenvolver uma infraestrutura global para o compartilhamento e interoperabilidade de ciência de dados, visando a execução dos princípios FAIR em todos os âmbitos da informação cientifica. Expandir e adotar essas novas praticas ocasiona em diversos fatores como: eficiência, visibilidade, transparência, confiabilidade e controle das pesquisas. Por

[1] https://go-fair.org/fair-principles/

conta disso, fazer parte desse esforço global em busca do reuso e compartilhamento de dados é estimular o progresso da ciência e impulsionar a criação de novas oportunidades de pesquisa [Henning *et al*., 2019].

O objetivo deste trabalho é desenvolver uma ferramenta que atenda os princípios FAIR e que seja capaz de auxiliar na necessidade do INPE de gerir um grande volume de dados. Por mais que não existam menções oficiais ao volume de dados total armazenado pelo INPE, estimasse que ele seja da ordem de centenas de terabytes, composto principalmente por dados de sensoriamento remoto [INPE, 2018]. O projeto visa também auxiliar no *workflow* dos pesquisadores deste instituto, com uma plataforma intuitiva e de rápida aprendizagem.

2 Portais de dados

Esta seção foi organizada em duas subseções. A primeira subseção explora portais de dados relevantes nacionais e internacionais, fazendo uma análise de cumprimento dos princípios FAIR por esses portais. A segunda subseção apresenta o projeto DataMap, onde esse trabalho está inserido.

2.1 Soluções Existentes

Um portal de dados tem como objetivo garantir os princípios da boa governança de dados. Assegurar a qualidade e a veracidade das informações é de grande importância, visto que serão de âmbito científico. Prezar por essa qualidade é direcionar as decisões referentes aos dados de forma correta, sem que haja enganos em relação a eles. A gestão de dados se mostra importante para isso, tanto a curadoria quanto o armazenamento influenciam diretamente na qualidade [Teixeira and Santos, 2019].

Uma grande inspiração para o projeto é o portal do *Atmospheric Radiation Measurement* (ARM). O ARM é uma organização vinculada ao Departamento de Energia (DoE) dos Estados Unidos voltada para pesquisas atmosféricas que reúne dados de diversos centros de pesquisa ao redor do mundo [Peppler *et al*., 2016; Palanisamy, 2016]. Seu portal de dados é gerenciado por uma equipe específica, localizada no Laboratório Nacional de Oak Ridge.

No cenário nacional, existem diversos portais de dados gerenciados por organizações públicas e privadas. Os portais aqui apresentados foram selecionados com base em sua relevância e cobertura de um domínio, como saúde e educação.

Entre as instituições nacionais, destaca-se primeiramente o ICMBio (Instituto Chico Mendes de Conservação da Biodiversidade), cujo portal de dados armazena e disponibiliza informações voltadas à proteção do patrimônio natural, contribuindo para o desenvolvimento socioambiental. Esse portal reúne conjuntos de dados relacionados à fauna e flora brasileiras ICMBio [2010]. Por sua vez, o INMET (Instituto Nacional de Meteorologia) mantém um portal dedicado ao monitoramento, análise e previsão do tempo e do clima em todo o território nacional, além de apoiar pesquisas aplicadas que fornecem informações essenciais sobre secas, enchentes e desastres naturais.

No que diz respeito a dados de saúde, destaca-se o DATASUS, responsável por coletar, processar e armazenar informações de saúde provenientes do Sistema Único de Saúde (SUS). Além disso, o IBGE (Instituto Brasileiro de Geografia e Estatística) atua na produção de estatísticas sociais, demográficas e econômicas, sendo também o órgão responsável pela realização dos censos nacionais.

A Tabela 1 lista os portais citados anteriormente e quais princípios FAIR as estruturas dos portais cumprem. O critério F1, por exemplo, verifica se os dados e metadados associados possuem um identificador único e persistente, enquanto o critério A1 analisa se os dados e metadados associados são obtidos por meio de um protocolo padronizado. Uma lista completa destes princípios juntamente com sua documentação e exemplos de casos de sucesso está disponível no domínio GoFAIR.

Percebe-se que, em grande parte, todos são acháveis e reutilizáveis segundo os princípios. Isso se dá por conta da facilidade em encontrar os dados brutos, pois são listados e separados por categorias, além de possuírem tabelas e gráficos que facilitam a sua exploração. Em contrapartida, encontrar os metadados das informações é possível em somente um portal. Fazer o *download* dos metadados não é uma função presente em todos os portais, o que compromete diretamente a acessibilidade. Essa falta se dá pela não utilização de formatos de arquivo que armazenem essas propriedades. Por exemplo, em portais como INMET e DATASUS é utilizado um formato de texto comum (*Comma-Separated Values* - CSV) quando os dados são transferidos. Um formato que possui a capacidade de armazenar esses metadados é o NetCDF (*Network Common Data Form*) [Rew and Davis, 1990], sendo aceito na ferramenta projetada neste trabalho. No INMET, por exemplo, único portal onde é possível encontrar os metadados, eles não são listados juntos aos dados, sendo possível somente obter informações da estação que coletou os dados.

Em suma, os portais analisados atingem parcialmente os princípios FAIR, sendo evidente a carência dos princípios de acessibilidade e interoperabilidade. Nenhum deles possui estrutura para vincular um DOI (*Digital Object Identifier*) aos dados, possuindo como identificador único os seus indexamentos nos sites dos portais. E, conforme discutido, a falta de metadados compromete a interoperabilidade dos dados brutos.

2.2 Projeto DataMap

DataMap é um projeto interdisciplinar, cujo propósito é revolucionar o ecossistema de dados nacionais, por meio do desenvolvimento de um portal de dados inovador. Seu time é composto por professores e pesquisadores de instituições nacionais como USP, UNICAMP e INPE e de instituições internacionais como o DataCite. O caráter inovador deste projeto está relacionado não somente aos desafios intrínsecos de um ambiente *Big Data*, mas também em fornecer um ambiente completo para armazenamento e manipulação de dados, segundo boas práticas internacionais. Atualmente uma versão alfa do portal está disponível em: `https://datamap.pcs.usp.br/`, sendo necessário que os usuários sejam autenticados via ORCID. Por mais que existam iniciativas semelhantes a nível internacional, como o portal de dados do ARM já citado, pelo conhecimento dos autores essa iniciativa se destaca no cenário nacional, ao propôr uma solução que adota boas práticas da comunidade e adequá-la ao cenário brasileiro.

O portal de dados aqui apresentado está inserido no escopo do projeto DataMap. Seu propósito é implementar em

Portal	Tipo	F1	F2	F3	F4	A1	A2	I1	I2	I3	R1
ARM[2]	Pesquisas atmosféricas	X	X	X	X	X	X	X	X	X	X
ICMBio[3]	Meio ambiente	X			X			X			X
INMET[4]	Meteorologia	X			X	X		X		X	X
DATASUS[5]	Saúde				X			X			X
IBGE[6]	Estatísticas	X		X	X		X	X		X	X

Tabela 1. Portais mais conhecidos

escala reduzida, com vertente de prototipação, um ambiente projetado com base em diretrizes internacionais, como os princípios FAIR, e que implementa técnicas para garantir uma boa gestão e qualidade de dados. Na subseção 5.7 é apresentada uma discussão sobre a integração futura entre o trabalho aqui apresentado e o DataMap, com enfoque nos desafios da manutenção do portal em um cenário real.

3 Dados Atmosféricos

Os dados atmosféricos são fundamentais para estudo, previsão e modelagem climática, e no monitoramento ambiental. O foco desse trabalho são os dados aeroespaciais, em especial os dados atmosféricos, que comumente são coletados em campanhas. Uma campanha refere-se ao processo de coleta de dados por diferentes instrumentos em uma determinada região geográfica, com o propósito de analisar fenômenos atmosféricos. Cada campanha reúne diferentes instrumentos que atuam juntos para coletar dados geralmente agrupados em torres ou estações de medição. Esses dados, além de densos e volumosos, apresentam um grande valor científico e estratégico.

Um desafio recorrente nas campanhas é manter a precisão de metadados dos equipamentos utilizados visto que cada instrumento cumpre um papel na geração de dados brutos da campanha. Assim, certos equipamentos possuem seu próprio DOI vinculado ao DOI da campanha. Isso garante uma maior confiabilidade e rastreabilidade, mas também demanda ferramentas que organizem e unifiquem tais informações para facilitar o reuso.

Um exemplo de campanha foi a *Green Ocean Amazon* (GOAmazon) Martin *et al*. [2016], que ocorreu no estado do Amazonas nos anos de 2014 e 2015, onde foram coletados dados inéditos na região. Essa campanha permitiu o estudo de aerossóis e suas iterações para formação de nuvens e o tempo de vida delas. Essas observações forneceram dados que até hoje são importantes fontes de estudo para diversos projetos, como o deste trabalho que usou os dados dessa campanha como teste para as funcionalidades do portal [Macedo and Fisch, 2018].

Um recurso geral de Ciência dos Dados que é utilizado neste trabalho no contexto dos dados atmosféricos são os *Data Quality Reports* (DQRs). A função de um DQR é indicar uma ocorrência que gera um impacto na qualidade dos dados armazenados. Por exemplo, a interrupção na coleta de dados em intervalos de tempo pode gerar lacunas no conjunto de dados. Um *Data Quality Report* pode informar ao usuário do portal de dados sobre essa interrupção. Outra situação que pode ser destacada por um DQR é a existência de dados incorretos, informando o usuário sobre impactos na confiabilidade dos conjuntos de dados. O uso de DQRs no portal aproxima-se diretamente dos princípios FAIR, ao tornar as informações mais transparentes, acessíveis e reutilizáveis, fortalecendo a confiança no processo de análise científica.

4 Desenvolvimento

Nesta seção, será apresentado o desenvolvimento de cada componente do portal, como o banco de dados escolhido, as medidas para assegurar os princípios FAIR e as bibliotecas na linguagem Python utilizadas na aplicação.

4.1 Banco de dados

É fundamental que a ferramenta possua um banco de dados confiável e que cumpra com as necessidades dos dados inseridos. Sabendo disso, foi selecionado um formato não relacional para armazenar os dados. Este modelo possui um alto nível de escalabilidade, para trabalhar com grandes volumes de dados, e alta disponibilidade visando oferecer o menor tempo de resposta aos seus usuários de Oliveira [2014]. Além disso, para cumprir com a demanda esperada do portal, é necessário flexibilidade para manipular dados de diversos esquemas (formatos variáveis).

Assim, foi escolhido o MongoDB, pois esta ferramenta já é conhecida por processar grande volumes de dados, além de possuir uma escalabilidade voltada para a análise de dados e *Big Data*. Porém, o grande diferencial desse sistema é a sua flexibilidade, diferente de outros bancos de dados, o MongoDB segue um modelo não relacional para gerir esses dados, armazenando as informações no formato BSON (*Binary JSON*), um derivado do JSON (*JavaScript Object Notation*), assim permite que a estrutura não siga esquemas fixos, sendo facilmente implementada em Python [Hows *et al*., 2019]. O formato JSON é amplamente utilizado para troca de dados por ser simples, leve e legível por humanos, como visto no Listing 1. Seguindo uma estrutura de chave-valor, onde cada elemento é rotulado por uma chave, facilitando a sua indexação [Smith, 2015]. Baseando-se nesse formato, surgiu o BSON, desenvolvido pelo MongoDB, que mantém a estrutura do JSON, mas em formato binário. Essa transformação permite maior eficiência no armazenamento e transmissão dos dados, melhorando o seu desempenho em sistemas com armazenamento de grandes volumes de dados.

Listing 1: Exemplo JSON do GOAmazon.

```
{
  "base_time": {
    "$date": "2014-08-10T00:00:00.000Z"
  },
  "timestamp": {
    "$date": "2014-08-10T00:04:00.000Z"
  },
  "atmos_pressure": 100.5,
  "qc_atmos_pressure": 0,
```

```
  "temp_mean": 26.450000762939453,
  "qc_temp_mean": 0,
  "temp_std": 0.009999999776482582,
  "rh_mean": 97,
  "qc_rh_mean": 0,
  "rh_std": 0.05400000140070915
  ....
}
```

No MongoDB as informações são divididas em coleções, ou tabelas quando comparado com outros modelos, para o portal, além da coleção com os dados brutos das campanhas existirão também outros três tipos de coleção, voltados para guardar os metadados das campanhas inseridas no sistema. Essas coleções estarão presentes no banco de dados exclusivo de cada usuário do sistema.

Através do diagrama de banco de dados (Figura 1) são observadas as três coleções de metadados, com suas variáveis e respectivos tipos. A coleção *Campaign* guarda as informações das campanhas, armazenando os ids de campanha e do usuário que a criou, além do nome, descrição e data na qual os dados foram adquiridos. Em seguida temos a coleção *Dataquality*, responsável por guardar o *data quality report* da campanha, na variável *data* (dicionário em JSON), que consegue guardar as datas e as notas referentes à análise de qualidade feita. Por fim temos a coleção *Headers*, que guarda os parâmetros como nome da coluna, mínimo e máximo, unidade de medida e entre outros, assim como na coleção anterior essas informações são armazenadas na variável *header*.

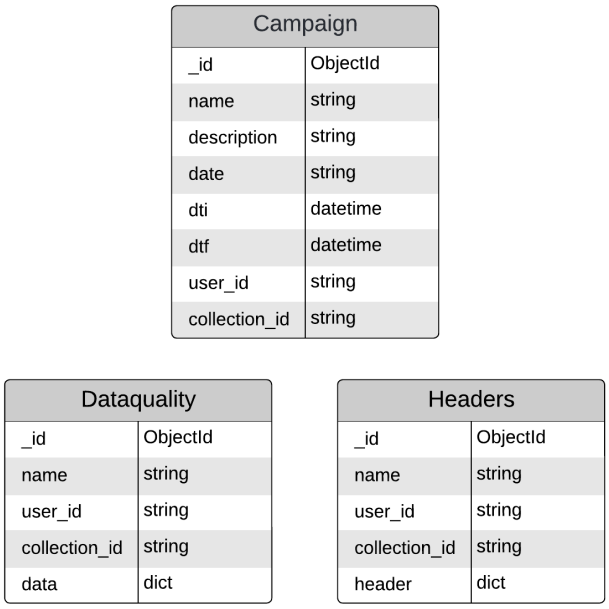

Figura 1. Diagrama das coleções de metadados

4.2 Princípios FAIR

O portal foi desenvolvido tendo como objetivo seguir os princípios FAIR. Desta forma, nesta seção será explicado como os recursos implementados no portal atingem estes princípios.

4.2.1 *Acessível*

A aplicação deve ser acessível para humanos e máquinas, para isso foi utilizado o MongoDB. Esse sistema de gerenciamento nos permite administrar os dados mediante a uma autenticação com login e senha, além guardar os metadados das informações mesmo que as mesmas já não estejam mais disponíveis. Por fim, ele é de código aberto, tornando-o universalmente implementável.

4.2.2 *Localizável*

A localização de dados é feita de duas formas distintas. Primeiramente deve-se localizar o *collection_id* em que ela está inserida, ele possui um nome único, composto pelo nome da campanha, definida através do usuário que a criou, seguido por um identificador único universal (UUID v4) aleatório e separados por um underline. Esse tipo de identificação permite que exista mais de uma campanha com o mesmo nome para o mesmo usuário e entre outros usuários. Após encontrar a campanha desejada, as informações referente a ela pode ser localizada individualmente por um id do tipo *ObjectId* próprio do MongoDB. Com esses dois métodos é possível localizar qualquer informação dentro de cada coleção de metadados, como data quality, parâmetros das variáveis e descrição, e assim entregá-las para o usuário modificar ou visualizar.

4.2.3 *Interoperável*

Os dados são documentados no banco de dados em formato JSON, sendo esse formato um vocabulário que já segue os princípios, além de usar uma linguagem formal para as variáveis. Por último, o *collection_id* é uma referência que está presente e é qualificada para todos os dados, atuando como principal localizador de cada campanha.

4.2.4 *Reutilizável*

A reutilização dos dados é proporcionada pela coleção de parâmetros, onde nela é possível descrever o contexto sob qual esses dados foram gerado, valores mínimos e máximos, nome por extenso de cada parâmetro, unidade de medida e entre outros. Além disso, é possível inserir informações específicas da estação de medição, como o local da medição, e colocar o identificador DOI usado globalmente.

4.3 Bibliotecas Python

O código do portal foi escrito na linguagem Python, utilizando bibliotecas de manipulação e análise de dados. A biblioteca *Streamlit* foi utilizada pelo seu enfoque no projeto de aplicações que manipulem dados. Essa biblioteca é um *framework* de aplicações web rápida e flexível, com o foco de ajudar os desenvolvedores a reduzir o tempo de desenvolvimento desses aplicativos. Nesse trabalho o *Streamlit* foi usado para fazer a prototipagem do portal, com proposito de realizar a interação do usuário com o sistema [Richards, 2021].

A biblioteca *Pandas* [Wes McKinney, 2010] possui ampla utilização em projetos de Ciência de Dados em Python. Neste trabalho ela foi responsável por processar os arquivos de entrada e gerar a estrutura de dados utilizada no código [Brunner and Kim, 2016]. Também visando suporte à manipulação de dados, foi utilizado o pacote *Xarray*. Esse pacote se diferencia no modo em que os dados são processados, se baseando na estrutura de matrizes multidimensionais do NetCDF [Hoyer and Hamman, 2017].

Para realizar operações matemáticas, foi utilizado o *Numpy* [Harris *et al.*, 2020], um pacote com diversas funções amplamente usadas em ciência e engenharia. Na geração

de gráficos, implementou-se duas bibliotecas, o *Matplotlib* [Hunter, 2007] em conjunto com o *Seaborn* [Waskom, 2021]. Com o intuito de fazer a comunicação entre o código e o servidor foram usados dois pacotes. O *PyMongo*, uma biblioteca mantida pela própria MongoDB a fim de ser o *framework* entre a linguagem Python e os bancos de dados [O'higgins, 2011]. Seguindo essa linha, a biblioteca *Beanie* tem a mesma função do *PyMongo*, porém opera de forma assíncrona.

5 Resultados e Discussões

Nesta seção serão apresentados os resultados obtidos através do desenvolvimento, demonstrando o cenário que representa um pesquisador adicionando os dados de uma campanha no portal. Atualmente parte do processo deve ser feito manualmente pelo pesquisador, porém, versões futuras terão maior grau de automatização.

Esse processo de inserção de dados no portal aborda todas as seções existentes, passando pelas funcionalidades presentes, as opções de modificação e por fim a visualização. O funcionamento do portal é dividido em quatro blocos, como apresentado na Figura 2, sendo eles:

1. Fonte de dados: É o bloco de entrada, composto pelos dados oriundos de estações de medição.
2. Processamento: Neste bloco os dados são tratados e separados nas campanhas. Além disso, são criados os metadados de *data quality* e de parâmetros (*Headers*) do conjunto de dados.
3. Armazenamento: Após o processamento é feito o *upload* das coleções no banco de dados, onde os dados são colocados em uma coleção única, enquanto os metadados são inseridos nas coleções de cada usuário.
4. Visualização: Este bloco é responsável por resgatar os dados das coleções e os metadados referentes. É possível gerar gráficos, fazer o download e editar os metadados, como parâmetros e *data quality* de cada campanha.

O portal é dividido em áreas que englobam os blocos citados anteriormente. A página inicial é a de login, local onde o usuário vai realizar a autenticação para o acesso à aplicação. Após efetuado o login, o usuário é direcionado para a página com as opções de gerenciamento, dividida por abas. É possível realizar a criação de uma nova campanha, editar os metadados, fazer o *upload* de novos dados e visualizar os dados.

5.1 Criação

A primeira aba é responsável por gerar a campanha e inserir as informações como nome, período e descrição, no sistema. O usuário deve preencher conforme a necessidade, definindo um nome e o período correto em que os dados foram coletados. Opcionalmente é possível colocar uma descrição, inserindo informações como posição geográfica, DOI, instrumento usado e outros elementos que o usuário achar pertinente. Após a confirmação do usuário, a aplicação gera um *collection_id*, utilizando o nome seguido de um identificador aleatório e com as informações preenchidas, insere esses metadados na coleção de campanhas.

5.2 Edição

Esta aba permite ao usuário modificar metadados da campanha, além da possibilidade de excluí-la. Dentro dela existem dois tipos de opções: a primeira é direcionada para a edição de cabeçalho, local onde estão os parâmetros dos dados, com a possibilidade de modificar e atualizar essas informações. A segunda, possui a opção de deletar a campanha, necessitando de uma confirmação que é feita ao digitar o nome da campanha a ser excluída antes de ser atualizada no sistema.

O funcionamento consiste em: fazer um requerimento dos metadados presentes no servidor, mostrá-los aos usuários e, por fim, permitir a modificação deles como visto na Figura 3. Após as alterações, o programa faz o *upload* dessas novas informações no banco de dados e substitui os metadados dos antigos. Paralelamente, a aba de deletar, após a validação da ação, realiza a exclusão dos dados brutos das campanhas.

5.3 Administrar

É nesta aba que os dados serão inseridos, além de possuir a área de *data quality report*. Os dados podem ser inseridos através de quatro extensões de arquivos: *csv*, *tsv* (*Tab-Separated Values*), *cdf* e *nc*, sendo as duas últimas extensões de arquivos NetCDF. Após realizar o *upload* do arquivo com os dados, a aplicação apresentará uma pré-visualização das informações, contendo as quatro primeiras linhas e as quatro últimas linhas que serão adicionadas. Em seguida, é gerado na área de *data quality report* (Figura 4) uma análise de quais dados estão faltando e oferecendo três opções para o usuário inserir no *report*, são elas: um dado suspeito, uma nota sobre esses dados e os dados incorretos. As opções de *report* estão disponíveis para cada uma das medidas inseridas da campanha. O *upload* de dados opera da seguinte forma sequencial:

1. Os arquivos de entrada são convertidos para um dataframe da biblioteca Pandas ou pela biblioteca do xarray, se for um arquivo do tipo NetCDF.
2. Caso o arquivo possua tipo NetCDF, os parâmetros presentes neste tipo de dado são separados para serem carregados na coleção de metadados (*Headers*).
3. O dataframe é analisado e gera o *data quality report* de dados faltantes.
4. A aplicação mostra para o usuário a prévia dos dados e permite a edição do *data quality report* para inserir mais informações.
5. Aguarda alguma mudança do *data quality report* pelo usuário.
6. Após a confirmação do usuário a ferramenta através de bibliotecas como o Pymongo e o Beanie realizam o *upload* no banco de dados desse dataframe em uma coleção própria e em seguida dos metadados nas coleções referentes.

5.4 Visualização

A última aba é voltada para a visualização dos dados. Esta área dispõe de uma visualização das linhas de dados, valores que estão faltando, estatísticas como média, mínimo, máximo e por último uma visualização por gráficos, que são de três tipos, dispersão, boxplot e histograma. O usuário tem a opção de escolher quais colunas serão utilizadas no gráfico. Além disso, é possível definir o intervalo de tempo apresentado pelo gráfico. A responsável pela elaboração do gráfico é a

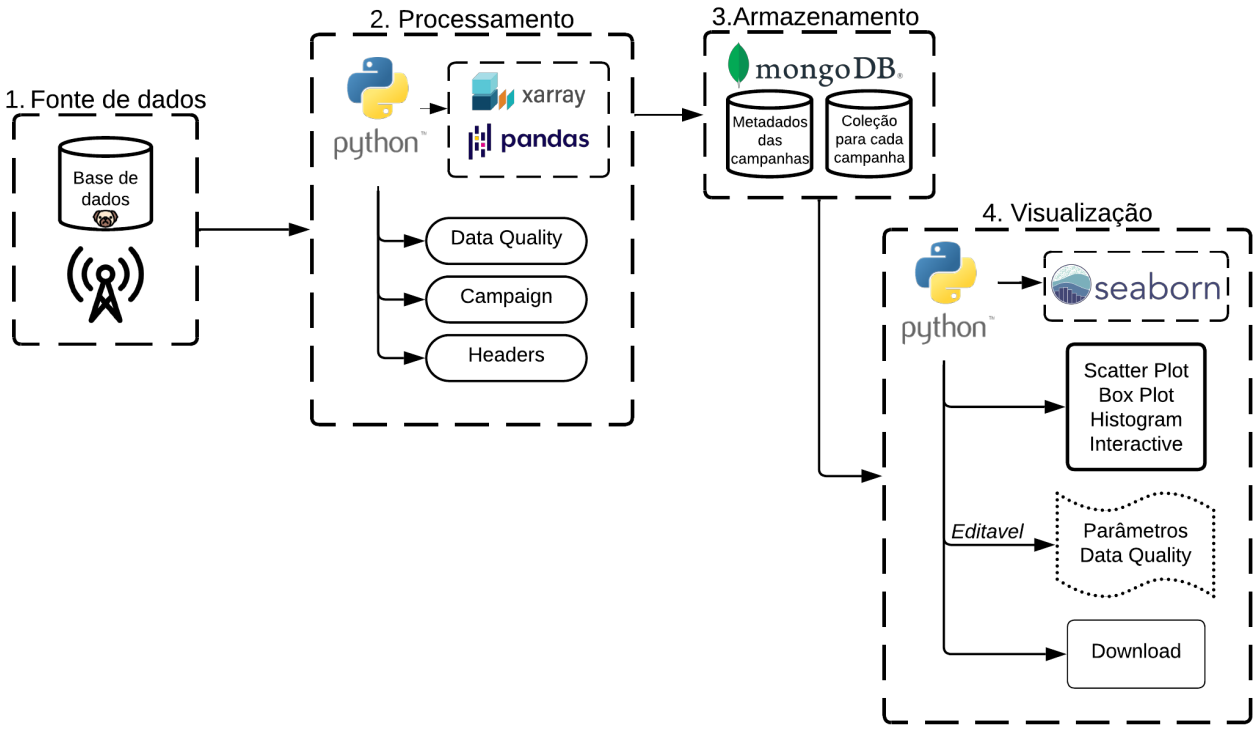

Figura 2. Divisão de blocos da aplicação

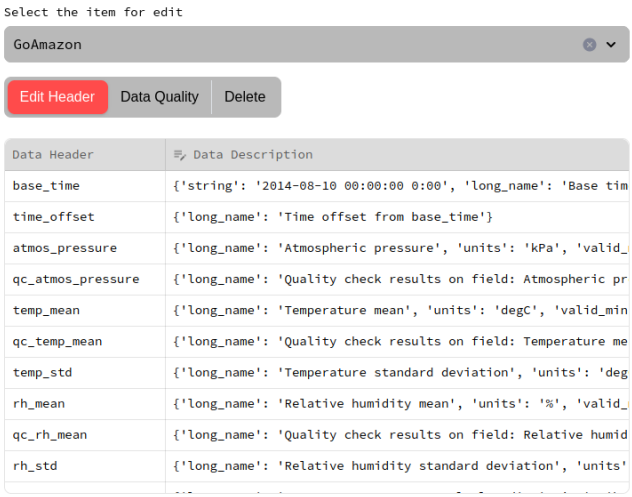

Figura 3. Aba de edição dos metadados

biblioteca Seaborn, utilizando os dados enviados pelo sistema após a requisição.

Como exemplo da capacidade do portal, foi realizado a inserção de uma semana de dados do GOAmazon [Martin *et al.*, 2016], possuindo mais de 11.500 linhas de informação. A Figura 5 mostra o gráfico de dispersão para a média de temperatura durante esses sete dias. Nela é possível analisar uma discrepância nos dias 15 e 16, quando comparados aos outros dias, algo que pode ser analisado na própria ferramenta alterando a faixa de visualização do gráfico. Outro tipo de visualização é a linha do tempo do *data quality report* (Figura 6) que fornece uma visualização rápida e simplificada da qualidade geral dos dados, quais deles estão incorretos, quais estão faltando e eventuais anotações.

Figura 4. Data quality report

5.5 Desempenho

Os dados do GOAmazon [Martin *et al.*, 2016] foram usados também para medir o desempenho do projeto. Foram adicionados dados coletados durante um mês, a fim de medir os consumos de memória que o MongoDB teria para armazenar essas informações. Ao final do teste, 31 arquivos que juntos ocupam 9 MB de espaço resultaram em um espaço alocado nos servidores de 8.2 MB de memória compactada. A importação desses dados no portal apresentou uma redução de 9% do tamanho original deles, e o tempo para inserção dessas informações, cerca de 44 mil linhas, foi de 3 minutos.

Princípio	Como foi atingido
F1	Identificador através do *collection_id*
F2	Com as coleções *Campaign*, *Headers* e *Dataquality*
F3	É possível encontrar cada campanha baseada no seu *collection_id*
F4	Cada linha possui um valor ObjectID único
A1	Autenticação via banco de dados
A2	Exclusão somente o valor bruto dos dados, não seus metadados
I1	Utiliza linguagem formal para variáveis
I2	Documentos são baseados no formato JSON
I3	Cada dado possui sua própria referência
R1	Possui descrição especifica onde pode ser inserido um DOI

Tabela 2. Resultados atingidos pela ferramenta

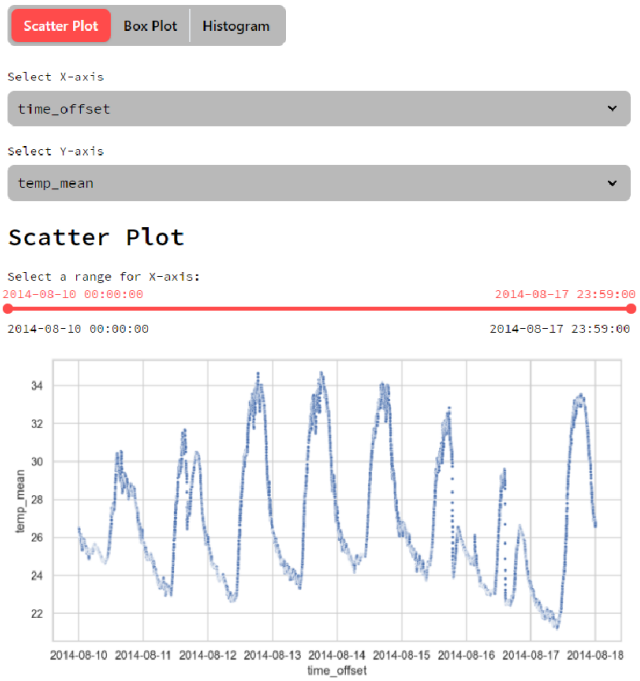

Figura 5. Gráfico de dispersão

5.6 Princípios FAIR atingidos

Com estes resultados é possível sumarizar como cada princípio foi atingido nas áreas da ferramenta. A Tabela 2 mostra como cada um dos quatro pontos destes princípios foram atingidos.

5.7 Integração com DataMap

Conforme apresentado, o DataMap já está disponível para uso em uma versão alfa. Além disso, o portal aqui desenvolvido foi projetado em caráter de protótipo, sendo que os testes realizados tiveram como enfoque a adequação do portal aos requisitos funcionais estabelecidos. A futura integração desse portal com o DataMap será realizada considerando uma migração das funções e estruturas aqui utilizadas para o DataMap. Com isso, serão realizados testes de escalabilidade e disponibilidade, utilizando maiores volumes de dados e acessos simultâneos por múltiplos usuários.

Entende-se que a manutenção de um portal de dados em um cenário real de aplicação requer um time especializado para conseguir assegurar não somente o seu correto funcionamento, mas também sua constante inovação. Por meio do monitoramento do fluxo de usuários e de dados será possível uma alocação dinâmica de recursos computacionais

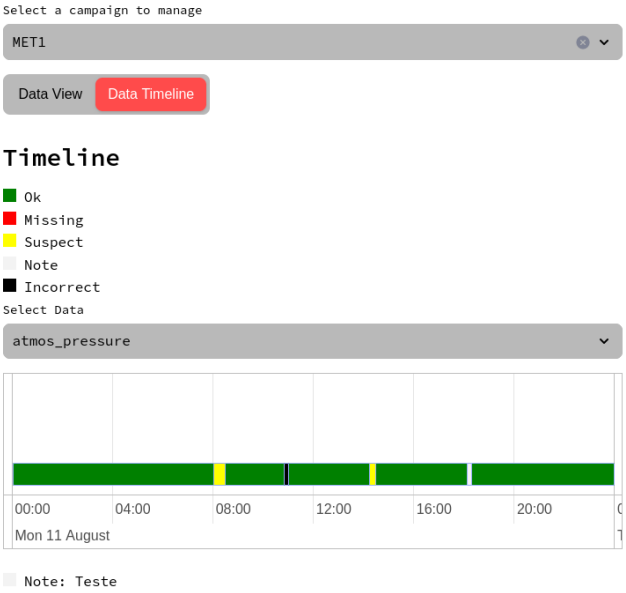

Figura 6. Timeline do data quality report

disponíveis, garantindo a disponibilidade do portal.

6 Conclusão

Considerando o escopo do desenvolvimento de um portal para gestão e qualidade de dados em *Big Data* todos os objetivos propostos foram atingidos. Diante dos resultados, foi possível concluir que o portal é capaz de auxiliar pesquisadores a gerir uma grande volume de dados. Seguindo os princípios da boa governança para esses dados, baseando-se em sistemas já conhecidos e usados. Além disso, a ferramenta possui uma interface simplificada, sendo intuitiva, prática e acessível. Mesmo sendo voltada para pesquisadores, não possui uma curva de aprendizagem longa, utilizando conceitos já conhecidos em portais de dados meteorológicos. No entanto, a ferramenta apresentada volta-se para os pesquisadores do INPE. Ademais, a aplicação é feita em Python e possui código aberto, permitindo que pesquisadores externos modifiquem a aplicação, dentro de seu próprio sistema interno, conforme suas necessidades.

Por mais que tenha sido discutida a integração do portal de dados aqui projetado com o DataMap, é importante destacar que este portal de dados continuará sendo incrementado para prototipação. Futuramente, espera-se implementar no projeto mais ferramentas de auxílio ao *workflow* do usuário, como integração com Jupyter Notebook, inclusão de uma área

pública para disponibilizar os dados desses estudos para o público e anexar o código no repositório oficial de terceiros do Python (PyPI).

Declarações complementares

Financiamento

Flavio Midea gostaria de agradecer ao INPE pela bolsa de Iniciação Científica fornecida (Processo 122503/2024-4) que viabilizou o desenvolvimento deste trabalho. Felipe Almeida gostaria de agradecer ao CNPq (Processo 140253/2021-1).

Contribuições dos autores

Conceptualização: FM, FA, PC e AC. Recursos: FM e FA. Software: FM. Supervisão: PC e AC. Validação: FA, PC e AC. Administração do Projeto: PC e AC. Escrita do Manuscrito: FM e FA. Revisão e Edição do Manuscrito: FM, FA, PC e AC. FM é o principal responsável pela realização deste trabalho. Todos os autores leram e aprovaram o manuscrito final.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

Disponibilidade de dados e materiais

Os conjuntos de softwares gerados durante o estudo atual estão disponíveis em `https://github.com/flagar0/DataQualityDataMap`

Referências

Barbosa, L., Shayer Lyra, R., *et al*. (2021). Governança de dados-2021. *Escola Nacional de Administração Pública*. Disponível em: http://repositorio.enap.gov.br/handle/1/7092.

Brunner, R. J. and Kim, E. J. (2016). Teaching data science. *Procedia Computer Science*, 80:1947–1956. DOI: 10.1016/j.procs.2016.05.513.

de Oliveira, S. S. (2014). Bancos de dados não-relacionais: um novo paradigma para armazenamento de dados em sistemas de ensino colaborativo. *Revista da Escola de Administração Pública do Amapá*, 2(1):184–194. Disponível em: https://www2.unifap.br/oliveira/files/2016/02/35-124-1-PB.pdf.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362. DOI: 10.1038/s41586-020-2649-2.

Hazen, B. T., Boone, C. A., Ezell, J. D., and Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, 154:72–80. DOI: 10.1016/j.ijpe.2014.04.018.

Henning, P., Ribeiro, C. J. S., Sales, L., Moreira, J., and da Silva Santos, L. O. B. (2018). Desmistificando os princípios fair: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos dados fair. *Tendências da Pesquisa Brasileira em Ciência da Informação*, 11(1). Disponível em: https://revistas.ancib.org/tpbci/article/view/458.

Henning, P. C., Ribeiro, C. J. S., Santos, L. O. B., and dos Santos, P. X. (2019). Go fair e os princípios fair: o que representam para a expansão dos dados de pesquisa no âmbito da ciência aberta. *Em Questão*, 25(2):389–412. DOI: 10.19132/1808-5245252.389-412.

Hows, D., Membrey, P., and Plugge, E. (2019). *Introdução ao MongoDB*. Novatec Editora.

Hoyer, S. and Hamman, J. (2017). xarray: Nd labeled arrays and datasets in python. *Journal of open research software*, 5(1):10–10. DOI: 10.5334/jors.148.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95. DOI: 10.1109/MCSE.2007.55.

ICMBio, L. (2010). Instituto chico mendes de conservação da biodiversidade. Disponível em: https://www.gov.br/icmbio/.

INPE (2018). Inpe amplia monitoramento da amazônia com dados da nova geração de satélites ambientais. Disponível em: http://www.inpe.br/noticias/noticia.php?Cod_Noticia=4829. Acesso em: 01/10/2025.

Macedo, A. d. S. and Fisch, G. (2018). Variabilidade temporal da radiação solar durante o experimento goamazon 2014/15. *Revista Brasileira de Meteorologia*, 33(2):353–365. DOI: 10.1590/0102-7786332017.

Martin, S. T., Artaxo, P., Machado, L. A. T., Manzi, A. O., Souza, R. A. F. d., Schumacher, C., Wang, J., Andreae, M. O., Barbosa, H., Fan, J., *et al*. (2016). Introduction: observations and modeling of the green ocean amazon (goamazon2014/5). *Atmospheric Chemistry and Physics*, 16(8):4785–4797. DOI: 10.5194/acp-16-4785-2016.

O'higgins, N. (2011). *MongoDB and Python: Patterns and processes for the popular document-oriented database*. "O'Reilly Media, Inc.".

Palanisamy, G. (2016). Arm data file standards version 1.2. Technical report, DOE Office of Science Atmospheric Radiation Measurement (ARM) Program.

Peppler, R., Kehoe, K., Monroe, J., Theisen, A., and Moore, S. (2016). The arm data quality program. *Meteorological Monographs*, 57:12.1–12.14. DOI: 10.1175/AMSMONOGRAPHS-D-15-0039.1.

Rew, R. and Davis, G. (1990). Netcdf: an interface for scientific data access. *IEEE computer graphics and applications*, 10(4):76–82. DOI: 10.1109/38.56302.

Richards, T. (2021). *Getting Started with Streamlit for Data Science*. Packt Publishing. Disponível em: https://ieeexplore.ieee.org/abstract/document/10163434.

Sales, L. F. and Sayão, L. F. (2018). A ciência invisível: revelando os dados da cauda longa da pesquisa. In *XIX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (XIX ENANCIB)*.

Smith, B. (2015). *Beginning JSON*. Apress.

Teixeira, M. P. and Santos, G. C. (2019). Gestão de dados científicos para pesquisadores. *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*, 17:e019035. DOI: 10.20396/rdbci.v17i0.8657527.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021. DOI: 10.21105/joss.03021.

Wes McKinney (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61. DOI: 10.25080/Majora-92bf1922-00a.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., *et al*. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9. DOI: 10.1038/sdata.2016.18.