SADD - Sistema de Armazenamento e Disponibilização de Dados Oceanográficos*

Almir Monteiro[1], Cláudio Alves[1], Janice Trotte[3], Fernando Monteiro[3], Eduardo Bezerra[1,2], Pedro Henrique Gonzalez[1]

[1]Programa de Engenharia de Sistemas e Computação
Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro, RJ - Brasil

{amonteirojr,claudioalves,pegonzalez}@cos.ufrj.br

[2]Escola de Informática e Computação
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - Cefet/RJ

ebezerra@cefet-rj.br

[3]Instituto Nacional de Pesquisas Oceânicas

{janice.trotte,fernando.monteiro}@inpo.org.br

Abstract. *Oceanographic data play a critical role in advancing climate research, safeguarding marine ecosystems, and mitigating extreme environmental events. However, their effective use is often hampered by restricted access and the fragmentation of datasets across multiple repositories. We introduce the Data Availability and Distribution System (SADD) — a collaborative platform developed for the National Institute for Oceanic Research (INPO), designed to centralize the sharing, access, and analysis of Brazilian oceanographic data. SADD enables researchers to contribute and retrieve datasets in a streamlined, cooperative environment, thereby fostering scientific progress in the field. Moreover, the platform integrates an online analysis environment that supports the processing of both hosted datasets and user-uploaded files.*

Resumo. *Dados oceanográficos desempenham um papel fundamental no avanço da pesquisa climática, na proteção de ecossistemas marinhos e na mitigação de eventos ambientais extremos. No entanto, seu uso efetivo é frequentemente dificultado por restrições de acesso e pela fragmentação dos conjuntos de dados em múltiplos repositórios. Apresentamos o SADD (Sistema de Armazenamento e Disponibilização de Dados) — uma plataforma colaborativa desenvolvida para o Instituto Nacional de Pesquisas Oceânicas (INPO), para centralizar o compartilhamento, o acesso e a análise de dados oceanográficos brasileiros. O SADD permite que pesquisadores contribuam com dados e os acessem de forma integrada. Além disso, a plataforma incorpora um ambiente online de análise de dados que viabiliza o processamento tanto de conjuntos de dados hospedados quanto de arquivos fornecidos pelos próprios usuários.*

1. Introdução

Importante fonte de conhecimento para o entendimento do clima e a proteção de ecossistemas, dados oceanográficos desempenham papel crucial em tarefas como o

*Link para o vídeo de demonstração: https://youtu.be/R26xDyuN3mc

monitoramento da saúde dos oceanos, a segurança marítima e a previsão de eventos climáticos extremos [Xie et al. 2019, Singh et al. 2024, Capotondi and et al. 2024]. No Brasil, diversas instituições e órgãos governamentais realizam coletas sistemáticas de dados oceanográficos. Dois exemplos dessas iniciativas são o Portal SIMCosta[1] e o Banco Nacional de Dados Oceanográficos[2]. Contudo, essas iniciativas ocorrem de forma descentralizada e muitas vezes desconectada, resultando na dispersão das informações em múltiplos repositórios, frequentemente sem um processo padronizado de catalogação [Tanhua et al. 2019]. Nesse cenário, se torna importante a criação de um sistema unificado de armazenamento e compartilhamento como uma proposta estratégica para integrar essas diferentes fontes, promovendo a interoperabilidade dos dados e ampliando sua acessibilidade.

Com o intuito de atuar como repositório de conhecimento da oceanografia brasileira, o SADD (Sistema de Armazenamento e Disponibilização de Dados) foi desenvolvido para o Instituto Nacional de Pesquisas Oceânicas (INPO)[3]. Neste sistema, pesquisadores poderão contribuir com dados provenientes de seus estudos, assim como terão a possibilidade de acessar dados submetidos por outros pesquisadores da área. Desta forma, o acesso aos dados passa a ser centralizado, simplificando o acesso a conhecimento prévio, impulsionando novos estudos e descobertas. Tal iniciativa representa um passo importante para a democratização do acesso a informações oceanográficas de alto valor científico e para o fortalecimento da pesquisa nacional na área. Ao passo que possibilita a consulta e o *download* das bases de dados disponíveis, o SADD também disponibiliza um ambiente completo de análise de dados, permitindo que estudos e análises sejam realizados de forma *online*. Desta forma, o pesquisador pode trabalhar tanto com as fontes de dados existentes na plataforma quanto realizar o *upload* de seus próprios arquivos para subsidiar seus estudos.

Este artigo apresenta detalhes de implementação e do funcionamento do SADD, com as demais seções organizadas conforme segue. Na Seção 2, apresentamos aspectos das arquiteturas física e lógica do sistema. Na Seção 3, introduzimos as principais funcionalidades do SADD. Na Seção 4, as considerações finais e indicações de trabalhos futuros são apresentadas.

2. Arquitetura do Sistema

O SADD foi concebido como um portal *Web*, e foi desenvolvido em Python. Utilizamos o *framework Flask*[4] para o desenvolvimento do frontend. O backend é composto por um sistema de gerenciamento de banco de dados, um data lake e um servidor de notebooks Jupyter. Em relação ao gerenciador de banco de dados, foi utilizado o *PostgreSQL*. Porque as fontes de dados submetidas para a plataforma sejam georreferenciadas, a extensão *PostGIS* foi ativada em seu gerenciador. Esta extensão permite tanto que dados geograficamente referenciados sejam salvos diretamente no banco de dados quanto possibilita a criação de consultas geoespaciais para cruzamento e recuperação desses dados.

Considerando a vasta diversidade de tipos e tamanhos de arquivos que o SADD

[1] `https://simcosta.furg.br/home`
[2] `https://www.marinha.mil.br/chm/bndo2`
[3] `https://inpo.org.br`
[4] `https://flask.palletsprojects.com/`

deve receber, optamos pela utilização de um *datalake* [Zagan and Danubianu 2020] para armazená-los. Esta estratégia permite armazenar os arquivos recebidos em seu formato original, sem a necessidade de algum procedimento de adequação a um formato específico. Também torna-se uma abordagem vantajosa por lidar bem com grandes volumes de dados. Para tanto, optamos pela utilização do *MinIO*[5], ferramenta que possibilita o armazenamento e posterior acesso aos arquivos submetidos pelos usuários.

Para a funcionalidade de análise de dados, optamos por utilizar o *JupyterHub*[6]. Esta ferramenta permite servir uma plataforma de análise de dados completa para os usuários, de forma online, sem a necessidade de que qualquer configuração ou instalação adicional seja realizada. Esta abordagem é justificada pela utilização rotineira de *notebooks Jupyter* na área de análise de dados [Thomas and et al 2016]. Além das funcionalidades comuns a esse tipo de tecnologia, o ambiente de análise de dados do SADD é disponibilizado de forma hermética a cada usuário, garantindo segurança e privacidade aos pesquisadores. Ainda, este ambiente provê acesso a modelos de Inteligência Artificial (IA) generativa para apoiar nas tarefas de análise. A Figura 1 apresenta como o SADD está estruturado: os módulos do sistema fazem acesso a um *backend* composto por diversos servidores.

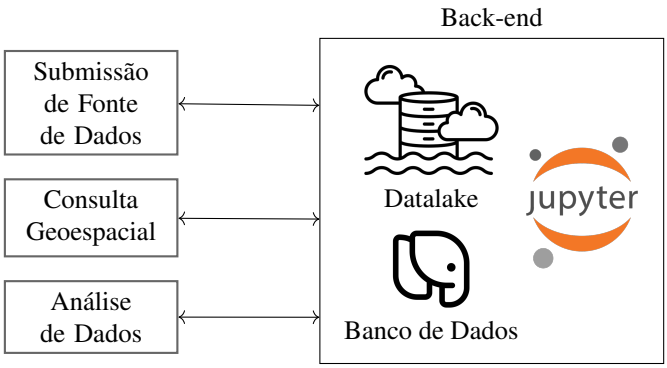


Figura 1. Estrutura do SADD. Um backend (formado por servidor de banco de dados, servidor de notebooks e um data lake) provê serviços para os módulos do sistema.

Em termos de organização lógica, o SADD está dividido em três módulos: *registro de fonte de dados*, *ferramenta de consulta* e *análise de dados*. A seguir, o propósito de cada módulo é apresentado.

O módulo de registro de fonte de dados é responsável por receber, processar e armazenar as fontes de dados submetidas pelos pesquisadores. Por meio de um formulário, dados descritivos sobre a fonte de dados são informados pelo usuário, validados no momento da submissão e salvos no banco de dados. De forma similar, os arquivos que compõem a fonte de dados são processados. Contudo, como esses arquivos podem ser de diferentes formatos, o SADD os armazena em um *datalake*, depositando no registro do banco de dados apenas o seu endereço de acesso para posterior recuperação do arquivo.

A ferramenta de consulta é responsável por apresentar ao usuário as fontes de

[5]https://min.io/
[6]https://jupyter.org/hub

dados disponíveis no sistema, permitindo a verificação de seu conteúdo e o seu *download*. Por meio de uma consulta realizada através da demarcação de uma área geográfica de interesse, o usuário do sistema recebe uma lista das fontes de dados que possuam ao menos uma interseção com a região geográfica demarcada.

O módulo de análise de dados é responsável por prover um ambiente completo de análise de dados para o usuário. Nele, é possível acessar as fontes de dados disponíveis no SADD, assim como realizar o *upload* de arquivos complementares que componham o estudo em questão. Além disso, este módulo possui uma IA generativa, disponível para auxiliar nas tarefas de análise, na forma de um co-piloto.

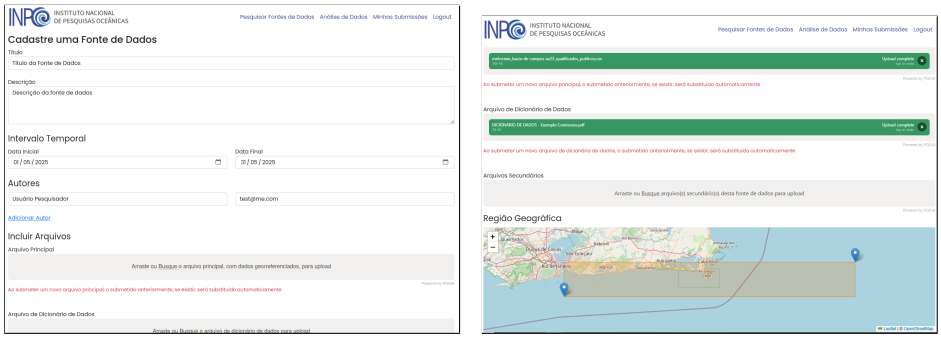
3. Funcionalidades

Nesta seção são apresentadas as principais funcionalidades do sistema, organizadas por módulo: submissão de fonte de dados (Seção 3.1), consulta geoespacial de fontes (Seção 3.2), e ambiente de análise de dados (Seção 3.3).

3.1. Submissão de Fonte de Dados

Qualquer usuário autenticado pode realizar a submissão de uma fonte de dados que deseja disponibilizar no portal. Embora o formulário de submissão de fonte contenha vários passos, este cadastro pode ser realizado aos poucos, com as informações parciais informadas salvas para posterior complemento. O processo de submissão envolve fornecer os arquivos da fonte de dados, além de dados dos autores.

Neste formulário, além dos dados cadastrais básicos sobre a fonte de dados que está sendo submetida, o pesquisador também deve enviar os dados georreferenciados e um dicionário de dados, que detalha os tipos e formas dos dados submetidos, visando facilitar a compreensão daqueles que farão uso destes dados posteriormente. Ao submeter um arquivo que possui informações de localização geográfica, o sistema automaticamente identifica e delimita a região compreendida, exibindo-a em um mapa, para a verificação do pesquisador. Caso o formato do arquivo não permita esta identificação automática, o SADD solicita ao usuário que informe quais conjuntos de dados se referem à latitude e longitude de cada registro. Confirmada esta informação, o sistema prossegue, calculando e exibindo os limites, como no cenário anterior. A Figura 2 exibe o formulário de submissão, com um arquivo de fonte de dados carregado e seus limites geográficos apresentados no mapa.



(a) Inclusão de dados gerais na submissão de uma fonte de dados.

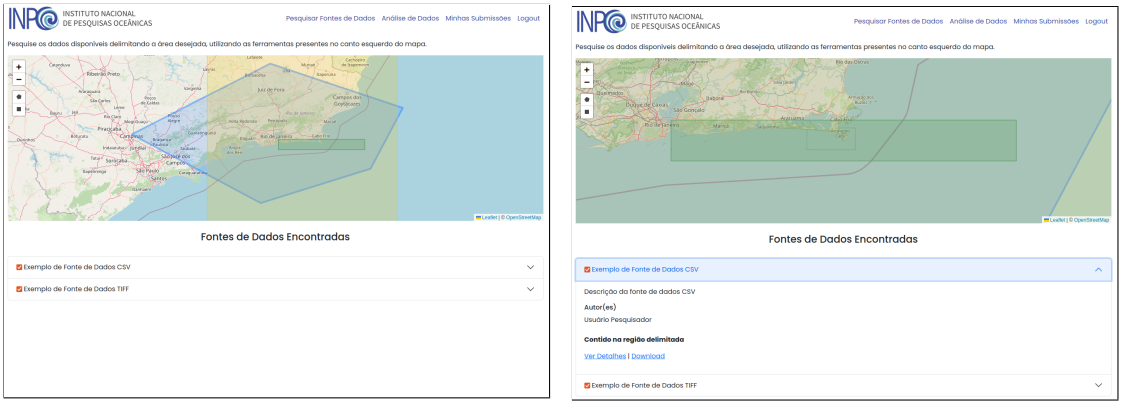
(b) Delimitação automática da área geográfica correspondente.

Figura 2. Funcionalidade de submissão de fonte de dados oceanográficos.

A versão atual do SADD provê suporte aos seguintes formatos de arquivos, definidos após consultas a especialistas da área: CSV, GeoJSON, NetCDF, HBR e TIFF. Mais formatos estão previstos para receber suporte posteriormente.

3.2. Consultas geoespaciais

Um usuário (autenticado ou não) pode pesquisar as fontes disponíveis no SADD. Por meio de um mapa apresentado na tela, o usuário marca qual a região geográfica de interesse. Como resposta, o sistema apresenta todas as fontes de dados disponíveis que possuam interseção com a área delimitada. A partir daí, o usuário pode verificar os detalhes daquela fonte de dados, como seus autores, sua breve descrição, o dicionário de dados e estatísticas tais como contagem de registros e médias. Caso esteja logado no SADD, o pesquisador também poderá realizar o download das bases disponíveis. A Figura 3 mostra a tela de consulta, com o mapa delimitado e a lista de resultados.



(a) Área delimitada durante a consulta. **(b) Detalhe de uma das fontes retornadas.**

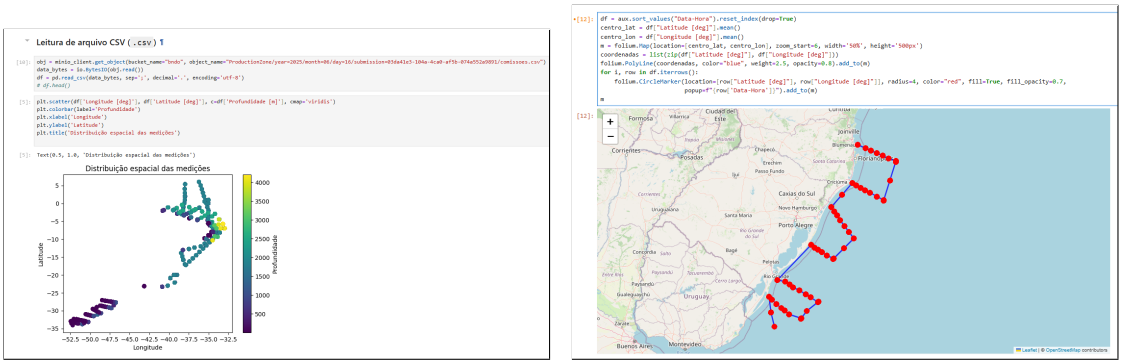
Figura 3. Consulta de dados no SADD com resultados exibidos conforme a área geográfica selecionada pelo usuário.

3.3. Análise de Dados

Esse módulo foi projetado para ampliar a autonomia dos pesquisadores, ao permitir a análise de dados diretamente na plataforma. Desta forma, o sistema se apresenta também como um espaço de exploração científica, e não apenas um repositório de dados. Esse módulo permite a análise tanto com as fontes de dados disponíveis no SADD quanto com arquivos enviados pelo próprio pesquisador, por meio de ferramenta de *upload*. Logo, o pesquisador pode realizar estudos sem estar preso a uma máquina ou localidade específica, necessitando apenas de acesso à internet.

O ambiente de análise de dados é isolado por usuário, garantindo a segurança e a privacidade das análises realizadas. Este módulo é baseado em *notebooks Jupyter*, onde códigos podem ser desenvolvidos e executados de forma interativa. O suporte à execução de códigos diretamente sobre os dados armazenados na plataforma reduz a barreira de entrada para pesquisadores com menos familiaridade com infraestrutura computacional, ao mesmo tempo que proporciona flexibilidade para usuários avançados automatizarem fluxos e integrarem modelos preditivos ou de aprendizado de máquina. Essa estrutura robusta, combinada à possibilidade de importar dados próprios, torna o SADD um laboratório digital de pesquisa aplicada em oceanografia, onde tarefas diversas podem ser rea-

lizadas, como limpeza de dados, apresentações gráficas e uso de ferramentas estatísticas, dentre outras. A Figura 4 exibe dois exemplos de gráficos gerados a partir do acesso programático a dados no *data lake*.



(a) Batimetria de uma variável oceanográfica (cor representa a profundidade de medição).

(b) Uma das rotas da comissão coletada pelo Navio de Pesquisa Hidroceanográfico "Vital de Oliveira".

Figura 4. Módulo de análise de dados do SADD exibindo notebook Jupyter.

4. Considerações Finais

Este trabalho apresentou o SADD[7], um sistema cujo objetivo é simplificar a disponibilização e disseminação de bases de dados oceanográficas, facilitando o acesso a pesquisadores e outros interessados nesta área de estudo. O SADD permite tanto a pesquisa de bases existentes quanto a submissão de novas bases, o que contribui para o crescimento constante do volume de dados disponibilizados nesta importante área de estudo. Como trabalhos futuros, consideramos aumentar o número de formatos de arquivo suportados pelo sistema, além de prover melhorias no módulo de análise de dados, como a possibilidade de que o próprio usuário defina qual modelo de IA generativa deseja utilizar.

Referências

Capotondi, A. and et al. (2024). A global overview of marine heatwaves in a changing climate. *Communications Earth & Environment*, 5(1):701.

Singh, R. P., Singh, B., and Popli, N. K. (2024). Temporal analysis of oceanographic data: Insights into environmental variability and trends. In *2024 IEEE Pacific Rim Conf. on Communications, Computers and Signal Processing (PACRIM)*, pages 1–9.

Tanhua, T., Pouliquen, S., Hausman, J., O'brien, K., Bricher, P., De Bruin, T., Buck, J. J., Burger, E. F., Carval, T., Casey, K. S., et al. (2019). Ocean fair data services. *Frontiers in Marine Science*, 6:440.

Thomas, K. and et al (2016). *Jupyter Notebooks - a publishing format for reproducible computational workflows*. IOS Press.

Xie, C., Li, M., Wang, H., and Dong, J. (2019). A survey on visual analysis of ocean data. *Visual Informatics*, 3(3):113–128.

Zagan, E. and Danubianu, M. (2020). Data lake approaches: A survey. In *2020 International Conf. on Development and Application Systems (DAS)*, page 189–193. IEEE.

[7]https://github.com/AILAB-CEFET-RJ/inpo