Fonte: Freepik.com. Reprodução

A publicação de observações e resultados das atividades científicas é um dos focos da Gestão de Dados Científicos.

Gestão de dados científicos

por Pedro Luiz Pizzigatti Corrêa, Alan James Peixoto Calheiro e Luciana Varanda Rizzo

Introdução

A preservação e a reutilização de dados científicos não foram estabelecidas individualmente por pesquisadores, mas sim através de uma iniciativa global, viabilizada por instituições de pesquisa e governos em vários países.[1] A natureza do processo científico atual tem sofrido importantes alterações que precisam ser reconhecidas. Para entender melhor essas questões, são necessários estudos que enfatizem como essas metodologias estão sendo empregadas e quais as soluções que vêm sendo utilizadas.

A publicação de observações e resultados das atividades científicas voltadas para análise, uso e reutilização de dados é um dos focos da Ciência Aberta[2] ou, mais especificamente, da Gestão de Dados Científicos, aqui definida como Ciência dos Dados Abertos. De acordo com Tenopir *et al.*:[3]

> *[...] O tema do compartilhamento de dados é uma parte importante do debate acadêmico moderno. O movimento de acesso aberto focado em disponibilizar livremente artigos da pesquisa publicadas, cresceu para englobar os dados associados à pesquisa.*

Além dos aspectos apontados anteriormente, outra questão importante é que a Ciência Aberta apoia fortemente o reaproveitamento de metodologias científicas e dados primários. Para Gray:[4]

> *"[...] a possibilidade de tratar os dados, em vez de somente colecioná-los, permite*

que os esforços concentrem-se na análise dos dados existentes. Os dados podem então ser compartilhados, reduzindo custos e permitindo avanços rápidos e eficazes na ciência".

Os autores citados anteriormente concordam que a Ciência dos Dados Abertos não é necessariamente uma nova disciplina científica, mas sim uma proposta que trata da criação de infraestruturas computacionais aprimoradas, que promovem soluções adequadas para apoiar a Ciência, permitindo um "padrão-ouro" na ciência moderna que é alcançar replicabilidade total de experimentos científicos ou, quando não for possível, alcançar a reprodutibilidade.[5] Outra característica importante é que a Ciência dos Dados traz um novo paradigma, chamado Quarto Paradigma. De acordo com Tenopir *et al*.:[3]

[...] Essa nova era tem sido chamada de "O Quarto Paradigma": descoberta científica baseada no uso intensivo em dados" na qual toda a literatura científica é digital, todos os dados científicos são digitais e interoperáveis". Os dados digitais não são apenas os resultados da pesquisa, mas fornecem insumos para novas hipóteses, permitindo novos insights científicos e impulsionando a inovação.

Tenopir *et al*.[3] complementa:

"[...] A ideia do Quarto Paradigma,[4] baseia-se no uso intensivo de dados entre domínios distintos, sendo o caminho do futuro para as redes distribuídas e colaborativas de pesquisadores, trabalhando juntas para enfrentar problemas da pesquisa científica".

Dessa forma, o contexto brasileiro atual ainda se apresenta num estágio preliminar, embora a realidade já esteja mudando em muitas instituições. Nos últimos anos, as instituições brasileiras de pesquisa iniciaram ações para abordar alguns dos desafios no contexto contemporâneo de Ciência dos Dados Abertos, considerando principalmente a criação de repositórios de dados.

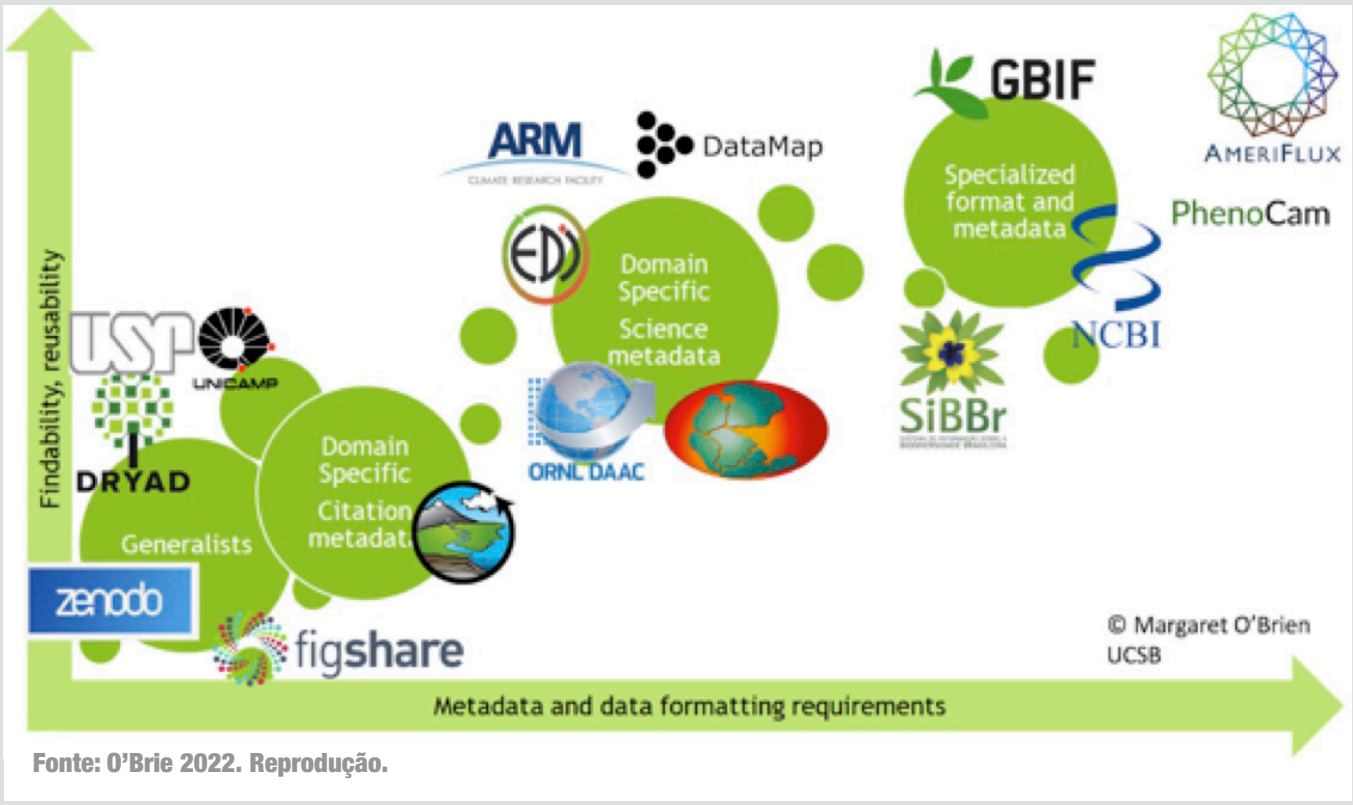
Repositório de Dados

Os repositórios de dados são responsáveis por manter, por um longo período, dados, *softwares* e outros objetos digitais necessários para permitir a replicabilidade ou reprodutibilidade[6] de um experimento científico. Pode-se classificar em repositórios generalistas ou repositórios voltados para domínios específicos da ciência.[7, 8]

Os repositórios generalistas manipulam dados de várias disciplinas e aceitam vários tipos de dados. Esses repositórios são comumente utilizados em situações em que os dados não podem ser publicados em um repositório de domínio ou disciplina da ciência específico. No Brasil os repositórios generalistas são disponibilizados por instituições de pesquisa, universidades, ou mesmo por agências financiadoras de pesquisa (Repositórios Institucionais).

Um repositório especializado em um domínio ou disciplina específica aumentará significativamente a probabilidade de que os dados e *softwares* depositados atendam aos princípios FAIR,[9] sendo: localizáveis (*Findable*), ou seja, facilmente encontrados por humanos e máquinas por meio de metadados e identificadores persistentes; acessíveis (*Accessible*), garantindo que possam ser recuperados, preferencialmente de forma automatizada, mesmo a longo prazo; interoperáveis (*Interoperable*), permitindo integração com outros dados e ferramentas, usando formatos e padrões amplamente aceitos; e reutilizáveis (*Reusable*), com metadados ricos e bem documentados, facilitando seu uso em diferentes contextos. Esses repositórios são geralmente mantidos por redes de pesquisa, programas científicos ou organizações científicas, governamentais ou acadêmicas.

A Figura 1, reproduzida do artigo de O'Brie,[10] cita vários repositórios de dados, permitindo observar que repositórios especializados, confiáveis e aceitos pela comunidade de uma determinada disciplina permitem que os dados possam ser mais facilmente encontrados e reutilizados.



Fonte: O'Brie 2022. Reprodução.

Figura 1. Exemplos de repositórios de dados científicos.

Serviços dos repositórios

Os repositórios de dados disponibilizam serviços necessários para apoiar a gestão dos objetos digitais (e.g. conjuntos de dados, metadados, códigos computacionais, modelos, resultados de análises, documentação, publicações científicas e identificadores persistentes) gerados durante o processo científico. Os serviços básicos estão relacionados com:

a) Armazenamento confiável e seguro dos objetos digitais e seus descritores (metadados);
b) Publicação dos *datasets* para garantir a sua proveniência, através da atribuição de identificadores únicos e persistentes, como o Digital Object Identifier (DOI), sob responsabilidade do DataCite que realiza o gerenciamento dos DOIs;
c) Busca, seleção e *download* dos objetos digitais.

Os serviços especializados estão relacionados às funcionalidades específicas dos repositórios, citadas a seguir:

a) Curadoria e armazenamento confiável e seguro dos objetos digitais e seus descritores (metadados), mantendo o controle de versões e a sua imutabilidade;
b) Gestão de identificadores persistentes (DOIs) em diferentes estados (por exemplo: o embargo temporário de *dataset* associado a uma publicação durante a sua submissão);
c) Integração com outros repositórios e agregadores como Google Dataset Search, DataOne e agregadores de publicações e *datasets* (Scholix);
d) Visualização dos objetos digitais e infraestrutura computacional para apoiar análises;
e) Formação, capacitação e treinamento de novos curadores e gestores de dados que desenvolvem atividades junto aos grupos de pesquisa associados.

Além dos serviços mencionados, os repositórios diferenciam pelo comprometimento e aderência a padrões vigentes, tais como os princípios FAIR, a certificação com os requisitos que refletem as principais características de repositórios de dados confiáveis CoreTrustSeal, governança de dados indígenas CARE e *framework* genérico para discussão e a implementação de melhores práticas em preservação digital TRUST.

O uso de repositórios de dados brasileiros tem crescido gradualmente, refletindo um esforço nacional por maior adesão aos princípios da Ciência Aberta, que vem sendo aplicada como política editorial das principais revistas científicas, das instituições de pesquisa e das agências financiadoras brasileiras. Um levantamento do DataCite[11] indica o Brasil entre os principais países latino-americanos na publicação de *datasets*. Até 2024, repositórios de dados brasileiros publicaram aproximadamente 210.000 DOIs. Porém, considerando o número de repositórios científicos registrados no Re3data, o Brasil atualmente tem somente 23 repositórios, enquanto os EUA, que tem um papel de liderança mundial na produção científica, têm 1.190 repositórios de dados científicos.

Iniciativa de Repositório de Dados da Amazônia – DataMap/Amazon

A Amazônia é crucial para o equilíbrio climático e a preservação da biodiversidade, tornando essencial a criação de repositórios de dados dedicados. Esses repositórios podem assegurar a coleta, gestão e acessibilidade de informações ambientais e climáticas, promovendo a transparência e a colaboração científica. A integração de dados sobre observações visando quantificar o balanço de gases de efeito estufa e mudanças climáticas na Amazônia apresenta uma oportunidade única de agregar

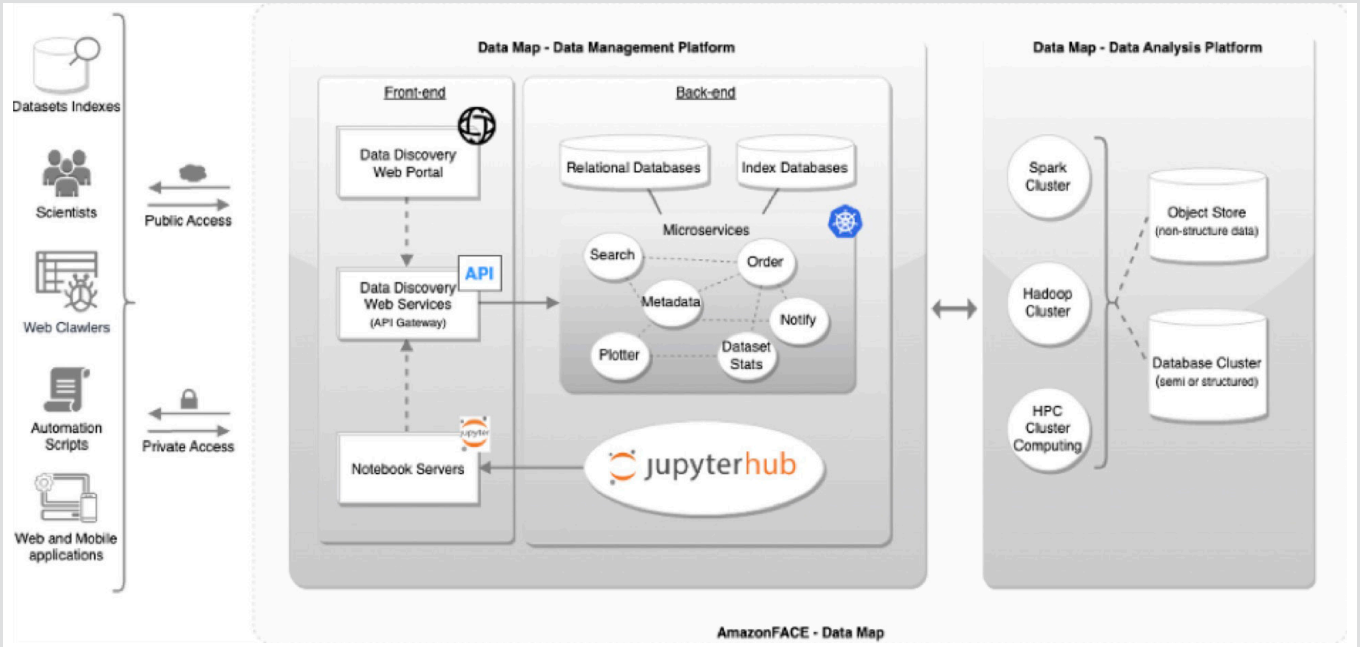
> "A gestão de dados científicos é imprescindível para a construção de um ambiente de ciência aberta, transparente e colaborativa, que seja capaz de responder aos desafios globais."

dados num repositório científico para melhor entender o funcionamento dos fenômenos ambientais na Amazônia, ao mesmo tempo, em que permite a síntese da enorme quantidade de dados gerados nesta região. Com essa visão, concebeu-se o DataMap/Amazon, iniciativa relevante para a ciência brasileira e latino-americana que carecem de ferramentas computacionais que integrem e auxiliem na síntese necessária do conhecimento de uma vasta quantidade de informações coletadas por pesquisadores de diferentes instituições como o Instituto Nacional de Pesquisas da Amazônia (Inpa), Instituto Nacional do Espaço (Inpe), Universidade de São Paulo (USP) a Universidade Estadual de Campinas (Unicamp), dentre outras. O DataMap/Amazon é uma iniciativa do Centro de Estudos Amazônia Sustentável (CEAS/USP) cujo objetivo é promover a produção e disseminação da ciência para o desenvolvimento sustentável da Amazônia.

O DataMap/Amazon vem consolidando, preliminarmente, dados coletados por projetos de pesquisa envolvendo pesquisadores do Inpe, USP e Unicamp, utilizando a expertise e as contribuições tecnológicas de iniciativas internacionais consolidadas, como o *Atmospheric Radiation Measurement* (ARM) do Departamento de Energia (DOE) dos EUA, especializado em pesquisas envolvendo a coleta e síntese de dados climáticos e atmosféricos coletados *in situ*. O ARM colabora com o DataMap/Amazon como referência do estado da arte em tecnologias computacionais para armazenamento e recursos analíticos para processar grandes volumes de dados atmosféricos específicos da Amazônia, ao mesmo tempo, em que adere aos principais padrões de ciência aberta. Assim, esta plataforma computacional visa apoiar a gestão de dados científicos na Amazônia, envolvendo principalmente programas de pesquisa de coleta de dados de longo prazo, como os citados a seguir:

a) O Large-Scale Biosphere-Atmosphere Experiment in Amazonia (LBA) é uma iniciativa brasileira para gerar dados científicos para entender melhor o funcionamento da Amazônia.
b) O Free-Air CO2 Enrichment Experiment in the Amazon (AmazonFACE) coletará dados florestais *in situ* (observações de longo prazo).

Ambas iniciativas demandam curadoria e publicação de dados, considerando os princípios e responsabilidades de repositórios voltados para a ciência aberta, além da necessidade de formação, capacitação e treinamento dos



Fonte: DataMap/Amazon. Reprodução.

Figura 2. Visão geral da arquitetura da plataforma DataMap/Amazon.

gestores de dados envolvidos nessas iniciativas.

O DataMap/Amazon disponibiliza uma plataforma para visualizar, descobrir, catalogar, publicar e processar *datasets* e *softwares*. Com o DataMap/Amazon, pesquisadores e cientistas de dados podem explorar conjuntos de dados usando ferramentas de visualização e pesquisa, promovendo a pesquisa colaborativa e aderindo aos princípios da ciência aberta. A arquitetura computacional do DataMap utiliza tecnologias baseadas em serviços, através de uma infraestrutura híbrida, utilizando tanto nuvem computacional como servidores locais, segundo a Figura 2.

O repositório DataMap/Amazon estabeleceu um processo de curadoria e de qualidade dos dados para os *datasets* coletados previamente pelos programas de pesquisa citados. Foram desenvolvidas ferramentas de *softwares*, documentos e treinamentos que apoiam o levantamento dos metadados, a validação dos *datasets* e a publicação no DataMap/Amazon.

Em 2024, a iniciativa DataMap/Amazon foi premiada pelo DataCite e Fundação Zuckerberg com recursos que permitiram o desenvolvimento e a disponibilização do repositório de dados atmosféricos científicos da Amazônia. Este repositório foi lançado em outubro de 2024, durante o VII Workshop on Data Science. Atualmente, conta com aproximadamente 150 *datasets* que passaram por um processo de curadoria e estão publicados no repositório.

> "Essas plataformas não apenas garantem a integridade dos conjuntos de dados coletados, mas também promovem a sua reutilização e a colaboração entre diferentes áreas de pesquisa."

Conclusão

A gestão de dados científicos é imprescindível para a construção de um ambiente de ciência aberta, transparente e colaborativa, que consiga responder aos desafios globais. Iniciativas como aquelas associadas ao repositório DataMap/Amazon ilustram o impacto positivo que podem ter na preservação, curadoria e disseminação de dados. Essas plataformas não apenas garantem a integridade dos conjuntos de dados coletados, mas também promovem a sua reutilização e a colaboração entre

diferentes áreas de pesquisa, ampliando, assim, as possibilidades de inovação científica. No caso da Amazônia, em específico, o armazenamento e a gestão eficaz de dados são de extrema importância para compreender fenômenos ambientais e climáticos, assim, subsidiando políticas públicas para o desenvolvimento sustentável da região e sua preservação.

No Brasil, os avanços rumo a essas práticas de ciência aberta ainda enfrentam desafios, como a necessidade de uma maior capacitação de profissionais na área, a cultura do tratamento de dados e o fortalecimento de infraestrutura tecnológica, que requerem recursos constantes. Porém, cabe ressaltar que iniciativas como o DataMap/Amazon demonstram que o país caminha na direção correta, desenvolvendo ferramentas, alinhadas aos padrões FAIR, e incorporando as melhores práticas globais possíveis. Contudo, é importante salientar que, é necessário o apoio contínuo de instituições de fomento à pesquisa, órgãos governamentais, empresas e organizações internacionais, de forma que o Brasil possa consolidar sua posição como um ator relevante no cenário da ciência aberta na América Latina, contribuindo para avanços científicos e o bem-estar da sociedade.

Pedro Luiz Pizzigatti Corrêa é professor do Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica da Universidade de São Paulo. Tem experiência na área de Ciência da Computação, com ênfase em Banco de Dados Distribuídos.

Alan James Peixoto Calheiros é tecnologista do INPE e membro do grupo do Laboratório Associado de Computação e Matemática Aplicada (LabAC/CoCTE/INPE) e do permanente do programa de pós-graduação do INPE em Computação Aplicada (CAP).

Luciana Varanda Rizzo é docente do Instituto de Física da Universidade de São Paulo (USP) e integra o Laboratório de Física Atmosférica (LFA).

Agradecimentos

Os autores agradecem ao DataCite Global Access Fund (Chan Zuckerberg Initiative) pelo apoio recebido para o desenvolvimento e disponibilização do repositório DataMap/Amazon (https://doi.org/10.14454/edht-vs98).

REFERÊNCIAS

[1] Serwadda, David, Paul Ndebele, M. Kate Grabowski, Francis Bajunirwe, e Rhoda K. Wanyenze. 2018. "*Open data sharing and the Global South—Who benefits?*" Science 359 (6376): 642–43. Disponível em: https://doi.org/10.1126/science.aap8395.

[2] UNESCO. 2021. UNESCO Recommendation on Open Science. UNESCO. Disponível em: https://doi.org/10.54677/MNMH8546.

[3] Tenopir, Carol, Allard suzie, e Mike Frame. 2015. "*Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide*", 2015. Disponível em: https://doi.org/10.1371/journal.pone.0134826.

[4] Gray, Jim. 2009. "Jim Gray on eScience: a transformed scientific method". In: *The Fourth Paradigm*: Data-intensive scientific discovery. Microsoft Research.

[5] Peng, Roger D. 2011. "*Reproducible Research in Computational Science*". Science 334 (6060): 1226–27. Disponível em: https://doi.org/10.1126/science.1213847.

[6] National Academies of Sciences, Engineering, and Medicine. 2019. "*Reproducibility and Replicability in Science*". Washington, DC: The National Academies Press. Disponível em: https://doi.org/10.17226/25303.

[7] NATURE. 2025. "*Data Repository Guidance*". Disponível em: https://www.nature.com/sdata/policies/repositories.

[8] AGU. 2021. "*Domain-Discipline Repositories Useful to AGU Journals*". Disponível em: https://data.agu.org/resources/useful-domain-repositories.

[9] Wilkinson, Mark D., Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, *et al.* 2016. "*The FAIR Guiding Principles for Scientific Data Management and Stewardship*". Scientific Data 3 (março):160018. Disponível em: https://doi.org/10.1038/sdata.2016.18.

[10] O'Brie, Margareth. 2022. "*Spectrum of Data Repositories*". Apresentado em VII Workshop on Data Science - Approaches of open science and synthesis techniques, São Paulo, outubro 5. Disponível em: http://wds.poli.usp.br/wds6/

[11] Garduño-Magaña, A. (2024). *Infrastructure and Awareness Landscape Analysis in Latin America (1.0)*. Zenodo. Disponível em: https://doi.org/10.5281/zenodo.14010858