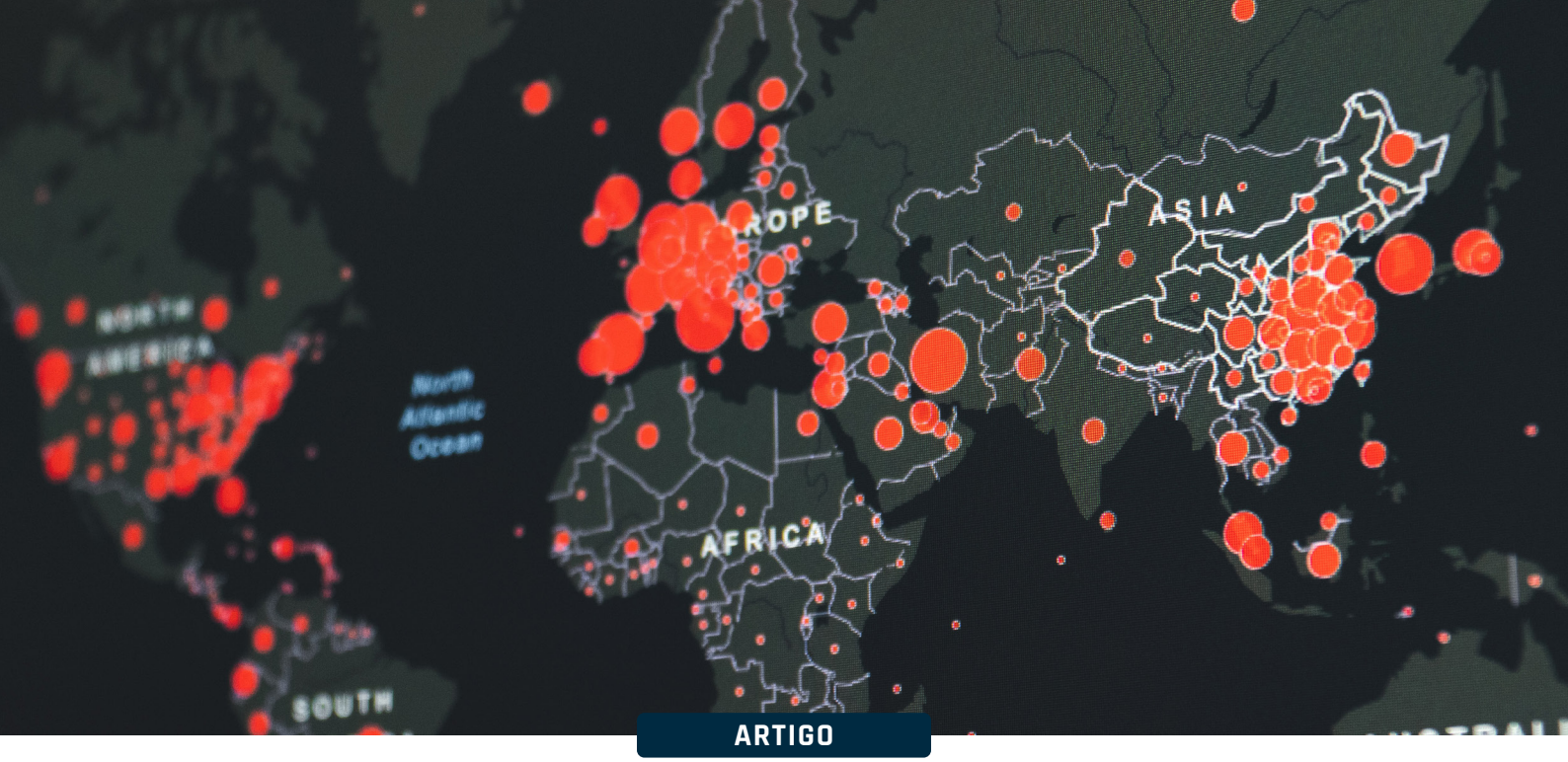

ARTIGO

COVID-19 DATASHARING/BR: UMA INFRAESTRUTURA SUSTENTÁVEL PARA DADOS DE PESQUISA ABERTOS

POR

Fátima L. S. Nunes, João Eduardo Ferreira

fatima.nunes@usp.br, jef@ime.usp.br

Alguns momentos da história exigem ações efetivas e rápidas. Assim foi e continua sendo a pandemia da COVID-19 que, a essas alturas, dispensa apresentação. Logo que os casos começaram a se multiplicar, cientistas do Brasil todo direcionaram suas pesquisas para responder às inúmeras perguntas elaboradas por eles próprios e pela sociedade em geral. Na atualidade nunca exigiu-se e cobrou-se tanto da ciência brasileira e mundial. A Direção Científica da Fapesp (Fundação de Amparo à Pesquisa do Estado de São Paulo) assumiu seu papel de liderança e rapidamente aglutinou colaboradores - instituições de saúde que possuíam dados sobre COVID-19 - dispostos a disponibilizar seus dados sob o paradigma de ciência aberta.

Embora disponibilizar arquivos de dados pudesse ser uma ação sem grandes complicações, a iniciativa estabeleceu alguns princípios e necessidades que exigiam uma plataforma que garantisse segurança, sustentabilidade, obediência a princípios legais e éticos, identificação dos usuários, além de disponibilidade no conhecido modo "24x7". Somado aos itens anteriores, o projeto apresentava o pré-requisito de disponibilização imediata. A Universidade de São Paulo (USP), mais especificamente a Superintendência de Tecnologia da Infor-

mação (STI), aceitou o desafio perante o convite realizado pela Fapesp e se juntou às demais instituições para contribuir com o projeto. A USP já possuía experiência em disponibilizar dados abertos por meio de seu repositório de dados científicos, disponível a todos os docentes e pesquisadores da Universidade [1]. Por isso, em poucas semanas, a equipe da USP duplicou e adaptou o sistema corporativo do repositório para as necessidades do projeto em questão, dando vida ao repositório COVID-19 DataSharing/BR [2].

O repositório está hospedado e utiliza o arcabouço de *hardware* denominado interNuvem USP [3]. O interNuvem USP é um conjunto integrado de servidores, dispositivos de armazenamento e rede de dados que estão sendo disponibilizados, para fins de pesquisa, para a comunidade USP e para pesquisadores externos, por meio de interface Web. Consiste em um arcabouço expansível e sustentável, que garante ao repositório a alocação de recursos computacionais compatível com seu crescimento.

A STI-USP destinou parte deste arcabouço para uso do repositório COVID-19 DataSharing/BR, reservando recursos para instalação de servidor web seguro, servidor de banco de dados e codificação do sistema propriamente dito. Além da estrutura de *backup* que garante a segurança dos dados, as permissões de acesso à infraestrutura são efetuadas por meio de sistema corporativo da USP, o que garante que somente usuários autorizados por meio de senha consigam incluir e alterar os dados existentes (Figura 1).

A infraestrutura de *software* que compõe o repositório COVID-19 DataSharing/BR é composta de três camadas: plataforma de análise e de inserção de dados, plataforma para navegação e plataforma para visibilidade dos dados.

Conforme exemplificado na Figura 1, **a plataforma de análise e de inserção de dados** consiste em um sistema desenvolvido com as características dos sistemas corporativos da USP que grava os metadados (dados sobre o conjunto de dados que serão disponibilizados para busca e navegação) e os dados propriamente ditos. Em uma fase paralela ao desenvolvimento da plataforma, uma equipe técnica composta por docentes e profissionais de Tecnologia da Informação das instituições envolvidas discutiu e estabeleceu regras para padronizar os dados a serem disponibilizados. Além da definição do formato dos dados e dos dicionários de dados, regras específicas de anonimização na área de saúde foram investigadas e definidas, visando garantir a não identificação de indivíduos [4]. Um sistema de validação automática foi desenvolvido para verificar o cumprimento das regras pelas instituições parceiras. Também foi disponibilizado um modelo relacional de banco de dados e respectivos programas na linguagem Python para utilização deste modelo [5]. Após a validação, o conjunto de dados é inserido no sistema por meio da interface apresentada na Figura 1.

FIG. 01 | **INTERFACES DO REPOSITÓRIO COVID-19 DATASHARING/BR PARA INCLUSÃO DE DADOS (ESQUERDA) E NAVEGAÇÃO (DIREITA)**

Na Figura 1 também é apresentada a interface da **plataforma para navegação,** construída com base na plataforma *open source dSpace* [6]. A escolha desta plataforma considerou fatores como customização, validação, flexibilidade na indexação e no acesso aos dados. Como pode ser observado na Figura 1 e no próprio repositório [2], é possível a consulta considerando os metadados fornecidos por meio da plataforma de inserção de dados. Além disso, a plataforma disponibiliza estatísticas de uso e de acesso rápido a dados recentes. Para cumprir integralmente as regras definidas, foram adicionadas funcionalidades para monitoramento e armazenamento dos dados de acesso, assim como disponibilização de termos de concordância em relação à responsabilidade ética e legal sobre o uso dos dados.

A terceira camada (**plataforma para visibilidade dos dados**) consistiu em incorporar a plataforma construída na rede de repositórios de dados abertos de pesquisa de São Paulo. O repositório foi incluído em uma plataforma previamente construída pela USP, denominada Metabuscador [7]. As principais funções do metabuscador são agregar metadados de repositórios provenientes de diversas instituições, disponibilizar tais metadados para consulta e direcionar o usuário, quando solicitado, para obtenção dos dados diretamente na sua origem.

A viabilização da infraestrutura de *hardware* e *software* em tempo recorde, somente foi possível porque a USP também disponibilizou recursos humanos para implementação e manutenção do projeto. Para garantir o pleno funcionamento do repositório, o volume de dados é constantemente monitorado para assegurar a alocação de recursos adicionais quando necessário. Assim, analistas experientes em redes de computadores são responsáveis por alocar e por configurar tais recursos. Da mesma forma, uma equipe experiente em bancos de dados garante o acompanhamento das necessidades de instalação, de configuração, de segurança e de uso de servidores nesta área. As equipes responsáveis pela implementação das

diferentes plataformas garantem o atendimento aos requisitos funcionais do projeto. Somada a essas atribuições, tem-se uma equipe de gerenciamento técnico e político do projeto.

Agradecimentos

Os autores agradecem à Fapesp e aos integrantes das diversas equipes que viabilizaram o projeto COVID-19 DataSharing/BR.

Referências

1. Repositório de dados científicos da USP. Disponível em: https://uspdigital.usp.br/repositorio. Acesso: outubro, 2021.
2.FAPESP. FAPESP COVID-19 Data Sharing/BR, Available from https://repositoriodatasharingfapesp.uspdigital.usp.br/. Acesso: outubro, 2021.
3.InterNuvem. Disponível em: http://cetisp.sti.usp.br/competencias/internuvem. Acesso: outubro, 2021.
4.Mello, L et al. Opening Brazilian COVID-19 patient data to support world research on pandemics. Disponível em https://doi.org/10.5281/zenodo.3966427. Acesso: outubro, 2021.
5.Carlotti, D; Ferreira, J. E.; Nunes, F. L. S. Relational data model and programs to use and view data from the FAPESP COVID-19 Data Sharing/BR repository. Disponível em: http://repositorio.uspdigital.usp.br/handle/item/243. Acesso: outubro, 2021.
6.dSpace. Disponível em: https://www.dspace.com. Acesso: outubro, 2021.
7.Metabuscador de dados de pesquisa. Disponível em: https://metabuscador.uspdigital.usp.br. Acesso: outubro, 2021.

FÁTIMA L. S. NUNES é Professora Titular da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (USP). Atua na área de Processamento Gráfico, Bancos de Dados e Sistemas de Informação. Atualmente é Diretora do Centro de Tecnologia da Informação de São Paulo, pertencente à Superintendência de Tecnologia da Informação da USP e coordena o projeto FAPESP COVID-19 Data Sharing/BR.

JOÃO EDUARDO FERREIRA é Professor Titular do Instituto de Matemática e Estatística da Universidade de São Paulo (USP). Atua na área de Bancos de Dados e Sistemas de Informação. Atualmente é Superintendente de Tecnologia da Informação da USP e foi responsável pela definição da arquitetura e infraestrutura do projeto FAPESP COVID-19 Data Sharing/BR.