dezembro/2021 • n. 46

# COMPUTAÇÃO BRASIL

REVISTA DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO

# O Papel da Computação na Ciência Aberta

SBC
Sociedade Brasileira de Computação

# EDITORIAL

**RAIMUNDO JOSÉ DE ARAÚJO MACÊDO**
*Presidente da Sociedade Brasileira de Computação (SBC)*

Passados quase dois anos do início da pandemia do coronavírus, estamos finalmente entrando em uma fase de esperança pelo retorno à normalidade sanitária, ainda que a transição precise ser feita com todo o cuidado para evitar novas ondas da doença.

A SBC se mobilizou em práticas remotas para o trabalho administrativo e de eventos, apoiando e orientando nossa comunidade. Criamos estratégias para garantir a sustentabilidade de nossos serviços e atuamos na defesa de instituições científicas e de educação nacionais. Também nos engajamos em muitas outras frentes, com novos grupos de trabalho já apresentados em nosso relatório anual.

Esses meses de grande luto trouxeram também importantes aprendizados, desde o valor da ciência ante a desafios sanitários sem precedentes, até as potencialidades e as dificuldades do trabalho remoto e seus efeitos no mundo da educação, da indústria e do comércio. Ficou ainda mais evidente a relevância da computação e de suas tecnologias como potencializadoras do desenvolvimento social e econômico, e seu papel no enfrentamento de desafios planetários, como a produção de vacinas e as mudanças climáticas.

A ciência não se faz de forma isolada dos problemas que afligem a sociedade. Pelo contrário, a ciência é instrumento fundamental para a superação desses desafios, emergentes ou de longo prazo. Em particular, a computação e suas tecnologias, quando bem aplicadas, representam oportunidade de geração de riquezas e trabalho.

Para melhor aproveitamento dos esforços da pesquisa em computação, é imperativa a busca de uma agenda nacional que oriente desenvolvimentos e promova o crescimento industrial do setor, aliada à busca por melhores práticas para a construção e difusão do conhecimento científico.

Muito recentemente, no mês de novembro de 2021, a UNESCO, Organização das Nações Unidas para a Educação, a Ciência e a Cultura, em sua 41.ª Conferência Geral, adotou, por decisão unânime, sua Recomendação para Ciência Aberta (UNESCO Recommendation on Open Science), que tem o potencial de tornar o processo científico mais transparente e democrático - ajudando na articulação entre ciência, tecnologia e inovação, em diálogo amplo com a sociedade.

Em sintonia com os propósitos da Ciência Aberta, em 2020 propusemos e aprovamos no conselho da SBC, uma diretriz para que todas nossas publicações estejam disponíveis no modo aberto e na SOL, nosso repositório aberto de publicações. No mesmo sentido, criamos um grupo de trabalho para promover a disseminação de *hardware* aberto, nos manifestamos pela participação da SBC na especificação e implementação de iniciativas de Ciência Aberta no Brasil e dedicamos este número especial da *Computação Brasil* ao tema da Ciência Aberta. Agradecemos ao trabalho do editor, professor Alirio Sá, da editora convidada deste número, professora Claudia Bauzer Medeiros, e de quem se dispôs a produzir os artigos publicados nesta edição.

Seguiremos para 2022 com as atividades rotineiras, e novos desafios: ajustes ao nosso sistema de submissão de artigos, atualização de nosso estatuto e código de ética, proposições de utilização dos meios digitais para enfrentamento de mudanças climáticas, orientações sobre ensino de computação no nível básico, campanha para redução de evasão de estudantes de computação, diretrizes curriculares para cursos de graduação emergentes, campanhas por mais recursos para ciência e tecnologia, entre outros.

No retorno ao presencial possível que se avizinha, precisamos repensar eventos aproveitando a experiência dos tempos remotos, e fazer novas campanhas para aglutinar membros e instituições que tiveram dificuldades nesse período.

Agradecemos o empenho e o engajamento dos membros de nossa sociedade, da diretoria e do conselho, das comissões especiais e grupos de interesse, das secretarias e das escolas regionais, de quem organiza eventos, das representações institucionais e estudantis, assim como de nossa equipe técnico-administrativa.

A união tem sido o ingrediente essencial do sucesso e prestígio da SBC. Continuemos unidos em busca da computação como instrumento de desenvolvimento social, econômico e humano, inclusivo e sustentável.

**dezembro**/2021 • **n**. 46

# COMPUTAÇÃO BRASIL

REVISTA DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO

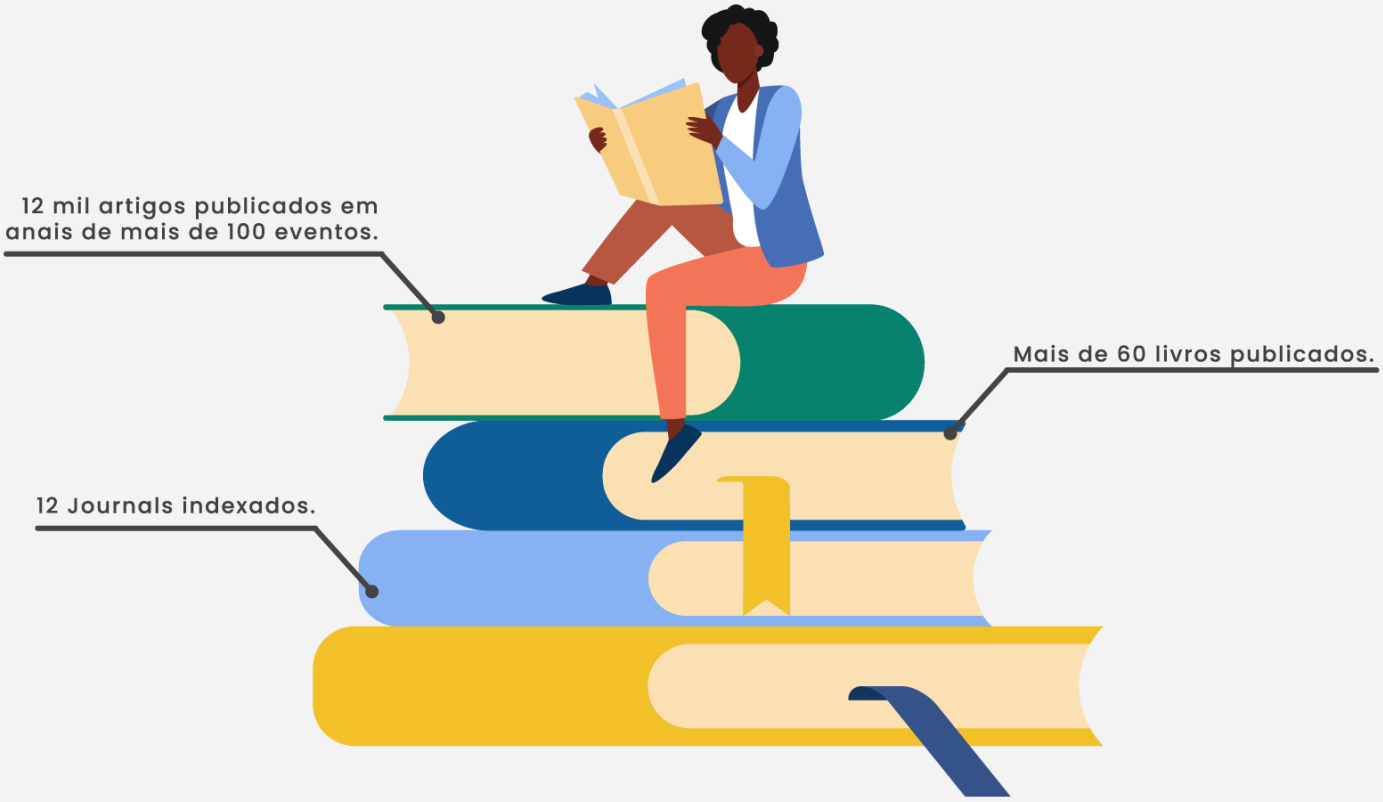
Caixa Postal 15012
CEP: 91.501-970 – Porto Alegre/RS
Av. Bento Gonçalves, 9.500 - Setor 4 – Prédio 43412 – Sala 219
Bairro Agronomia - CEP: 91.509-900 - Porto Alegre/RS
Fone: (51) 3308.6835 | Fax: (51) 3308.7142
marketing@sbc.org.br | sbc.org.br



**Diagramação:** Priscila Krüger | priscilahbk@gmail.com | 84 99112-7473
**Revisão** | Andrea Linhares
**Imagens ilustrativas:** Unsplash.com

# ACESSE A SBC OPENLIB

A SBC OpenLib (SOL) reúne a produção científica de eventos, journals e livros na área de Computação e afins. O acervo é alimentado diariamente e o conteúdo possui visibilidade internacional.




**O acesso à biblioteca é aberto e gratuito.**
Para mais informações sobre como publicar na SOL, entre em contato conosco em publicacoes@sbc.org.br.

Acesse sol.sbc.org.br

COMPUTAÇÃO BRASIL

# ÍNDICE

## O Papel da Computação na Ciência Aberta

Computação Brasil | Dezembro 2021



> Até o fim de 2021, a UNESCO irá votar um documento que recomenda fortemente a todos seus países-membros que adotem práticas de Ciência Aberta
>
> *-Claudia Bauzer Medeiros*
>
> *p. 06*

COMPUTAÇÃO ABERTA

# APRESENTAÇÃO

**POR**
Claudia Bauzer Medeiros
cmbm@ic.unicamp.br

Até o fim de 2021, a UNESCO irá votar um documento que recomenda fortemente a todos seus países-membros que adotem práticas de Ciência Aberta [1]. Este documento levou três anos para ser redigido, com reuniões regionais sucessivas em todo o mundo. Sua introdução destaca que "facilita o compartilhamento de conhecimento científico, dados e informação" visando "decisões baseadas em conhecimento". Estas duas expressões indicam objetivos e a importância da Ciência Aberta para acelerar descobertas científicas, produzir conhecimento, e beneficiar a ciência, a tecnologia, a inovação e a sociedade como um todo.

Há várias definições de Ciência Aberta, inclusive há quem diga que ela surgiu no século 17 com as primeiras publicações científicas. Para a maioria das pessoas, o movimento nasceu graças à Internet e à

pesquisa e ao desenvolvimento em tecnologias da informação e comunicação, pois o "compartilhamento do conhecimento, dados e informação" é, principalmente, o digital.

Este número especial da Computação Brasil cobre alguns dos principais aspectos computacionais associados à Ciência Aberta - dentre outros, dados, software, hardware e publicações. O conjunto de artigos, escritos por especialistas, está organizado de forma a apresentar as bases, seguidas por um exemplo da aplicação de várias dessas bases.

Escrevi o primeiro artigo dando uma visão geral de Ciência Aberta e me concentrei nos aspectos de dados abertos. Ele é seguido pelo artigo de von Flach e Kon, sobre software aberto, e as barreiras para sua adoção. Em Ciência Aberta, resultados de pesquisa devem ser disponibilizados de forma a serem encontrados, acessíveis, interoperáveis e reusáveis (do inglês FAIR, que é a ênfase do artigo de Campos, Borges e Moreira).

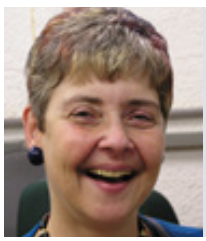
Ciência Aberta também inclui a colaboração com não cientistas, como descrito no artigo de Dallaqua, Fazenda e Faria, e que é exemplificada por um estudo de caso em desflorestamento e Ciência Cidadã. Um fator importante é a privacidade dos dados pessoais – e os desafios descritos no artigo de Machado. Fechando os textos sobre as bases do movimento, o artigo de Carro discute aspectos de hardware aberto. Finalmente, o artigo de Nunes e Ferreira apresenta uma grande rede de repositórios de dados abertos, pioneira na América Latina, que utiliza vários dos conceitos descritos nos artigos anteriores. Encerrando o número, o texto de Packer e Viterbo apresenta um pilar importante para o tema, que são as publicações abertas.


Referências

1. Unesco. Intergovernmental Meeting of Experts (Category II) Related to a Draft UNESCO Recommendation on Open Science. Document SC-PCB-SPP/2021/OS-IGM/WD3, May 2021 https://unesdoc.unesco.org/ark:/48223/pf0000378381.locale=en



Agradeço aos professores Alirio Sá e Raimundo Macêdo da UFBA a oportunidade de organizar este número do CB, e do muito que aprendi lendo os artigos convidados.


**CLAUDIA BAUZER MEDEIROS** é Professora Titular do Instituto de Computação da Unicamp. Desenvolve pesquisa em gestão de grandes volumes de dados distribuídos e heterogêneos, para aplicações do mundo real, em várias áreas, especialmente agricultura, saúde e mudanças climáticas. Ex-presidente da SBC (2003-2006), coordena o grupo de Ciência Aberta da Academia Brasileira de Ciências e co-coordena as iniciativas em Ciência Aberta da Fapesp.

ARTIGO

# CIÊNCIA ABERTA – COLABORAÇÃO SEM BARREIRAS PARA O AVANÇO DO CONHECIMENTO

**POR**
Claudia Bauzer Medeiros
cmbm@ic.unicamp.br

## Ciência Aberta – introdução e implementação computacional

O que é Ciência Aberta? Vamos separar a "especificação" da "implementação". Embora não haja uma definição fixa para o termo "Ciência Aberta", ele costuma ser usado para denotar o conjunto de políticas, iniciativas e ações para disseminar e compartilhar conhecimento, geralmente por meio digital, para que todos os resultados associados à pesquisa científica se tornem acessíveis a todos. O conceito principal desta especificação é *colaboração sem barreiras* – geográficas, temporais, culturais, sócio-econômicas ou políticas.

A implementação exige pesquisa e desenvolvimento em Computação.

A colaboração é facilitada criando repositórios institucionais públicos que disponibilizam abertamente os resultados e processos de uma pesquisa, tornando-a "acessível", "reprodutível" e "reutilizável" – três conceitos importantes e que exigem muita pesquisa em Computação.

Que resultados são esses? Incluem publicações, dados, algoritmos, processos computacionais, software, especificações de design de hardware e metodologias usadas para conduzir um determinado projeto. Além das recomendações da Unesco[6], dois documentos são importantes para entender Ciência Aberta [4,5], que precisa ser praticada desde o início do planejamento de um projeto, ao longo dele e mesmo após terminado [5] em um ecossistema complexo [4] que detalha como ela funciona em um ciclo virtuoso contínuo de colaboração, no qual pesquisadores do mundo inteiro compartilham a sua pesquisa (os resultados, as atividades e a documentação associada). Os pilares computacionais do movimento, também chamados de E-infrastructure, compreendem Hardware, Software, Redes de Computadores e Repositórios, e exigem engajamento e treinamento de cientistas da computação e de todas as outras áreas do conhecimento.

## Dados abertos – uma barreira e uma oportunidade

O compartilhamento de dados - e dados abertos em particular – apresenta inúmeros desafios de implementação, como por exemplo descrito em [3]. Por incrível que pareça, embora todos usemos e produzamos dados continuamente, existem compreensões muito diferentes do que sejam dados e do que significa "disponibilização para reuso". Como consequência, o compartilhamento de dados tornou-se uma enorme barreira digital no ecossistema da Ciência Aberta. Vamos ignorar a definição do que pode ser um dado – a qual, aliás, varia bastante dependendo do domínio científico – e nos concentrar na sua disponibilização em repositórios – como, por exemplo, os da rede de repositórios de dados de pesquisa do estado de SP [7].

Há três conceitos fundamentais associados a dados abertos – Planos de Gestão de Dados, Metadados e Repositórios, além de todos os aspectos de qualidade e curadoria, e dúvidas levantadas pela nova Lei Geral de Proteção de Dados, como descrito neste número do CB.

**Plano de Gestão de Dados (PGD)** – trata-se de um texto que descreve os dados que serão usados e produzidos por um projeto, e como estes últimos serão gerenciados – como, quando, onde e por quanto tempo. Um PGD tornou-se parte obrigatória de projetos de pesquisa em um grande número de países; no Brasil, a Fapesp instituiu sua obrigatoriedade em 2017. O papel de um PGD em um projeto é triplo. Em primeiro lugar, ajuda a planejar como devemos gerenciar e armazenar os dados ao longo de todo seu ciclo de vida. Em segundo lugar, auxiliam sua documentação e evoluem com a execução de um projeto (sendo inclusive chamados de "documentos vivos"). Finalmente, garantem que os dados irão perdurar e poderão ser utilizados por um longo tempo e não serão perdidos. Outra grande vantagem é considerar explicitamente, neste

planejamento, aspectos éticos, legais e de propriedade intelectual dos dados. A principal ferramenta usada no mundo inteiro para fazer um PGD é a dmptool [2] disponibilizada livremente online nos EUA e também na Europa, em versões e variações para instituições e países. Esta ferramenta é recurso excelente para que todos aprendam os inúmeros aspectos importantes relacionados à gestão de dados, pelas várias opções de ajuda que ela disponibiliza.

Todos sabemos o que são **metadados** (registros que descrevem o conteúdo de arquivos). A questão é qual padrão usar para documentar um determinado arquivo ou conjuntos de arquivos. Há uma enorme quantidade de padrões de metadados para domínios diferentes do conhecimento. Seu objetivo é facilitar a indexação e documentação dos dados nos repositórios. A base de todos os padrões é o Dublin Core [1], criado para documentar a catalogação de documentos da Biblioteca do Congresso Americano. Há padrões específicos para, por exemplo, catalogação de amostras físicas, objetos astronômicos, mudanças climáticas, ciências sociais e, inclusive, mais de mil padrões de metadados para pesquisas em Humanidades. Se tais padrões são criados para garantir interoperabilidade, a indexação de dados abertos cria o problema de interoperabilidade entre metadados.

**Repositórios** são locais que armazenam e disponibilizam, usando metodologias específicas, conjuntos de arquivos, dispondo de um catálogo que facilita sua indexação e acesso (o catálogo dos metadados dos arquivos). Essa disponibilização precisa seguir uma série de protocolos de comunicação para garantir acesso de pessoas e máquinas. Na Ciência Aberta, repositórios precisam ser gerenciados de forma institucional, para que os arquivos neles armazenados sejam catalogados e preservados de forma correta, segundo padrões mundiais, para garantir que continuem disponíveis muito além do tempo de vida de um projeto. O re3data [8] é um catálogo mundial de repositórios institucionais confiáveis, elencando em outubro de 2021 uma lista de 2900 repositórios reconhecidos mundialmente. O objetivo do re3data é ajudar pesquisadores a identificar locais onde possam depositar seus dados de pesquisa. Para ser incluído no re3data, um repositório precisa obedecer a várias regras, como curadoria e preservação institucionais, e normas claras sobre acesso e responsabilidades.

## Dados abertos no Brasil – o que a SBC pode fazer?

O Brasil tem um longo histórico de dados abertos para pesquisa – por exemplo, no IBGE ou no INPE. Este último é pioneiro mundial na disponibilização de dados abertos para pesquisas em mudanças climáticas, com 1.3 PTbytes de imagens de satélite publicamente acessíveis em seus repositórios. O quarto plano nacional do governo aberto brasileiro, iniciado em 2018, previa ações para dados abertos, que foram conduzidas (e ainda o são) por orgãos como EMBRAPA, FIOCRUZ, IBICT e RNP. Além das iniciativas federais, universidades (como a UFC ou a UFG) já estão criando seus repositórios de dados. A Fapesp promoveu a criação da

rede pública de repositórios de dados de pesquisa, envolvendo as universidades públicas do estado de SP e o ITA, inaugurada em 2019 [7].

Em 2022, estamos no 5o plano nacional do governo aberto, ainda em fase preliminar, cujas ações precisam de participação dos pesquisadores em Computação e da SBC. É preciso envolvimento de cientistas da computação e também profissionais e pesquisadores da Ciência da Informação (cuja formação envolve boas práticas de catalogação, padronização e arquivamento). Estes dois grupos precisam colaborar, conjuntamente com pesquisadores dos vários domínios, para que, juntos, possamos ter sucesso na implantação da Ciência Aberta no Brasil.


Referências

1. DCMI – Dublin Core Metadata Initiative. https://dublincore.org
2. DMPTOOL – Data Management Plan Tool. https://dmptool.org
3. A. LAENDER, C. B. MEDEIROS, I LOPES-CENDES, M. L, BARRETO, M-A. VAN SLUYS, U. B. de ALMEIDA. Abertura e Gestão de Dados: Desafios para a Ciência Brasileira. Academia Brasileira de Ciências, 2019. - http://www.abc.org.br/wp-content/uploads/2020/09/ABC-Abertura-e-Gest%C3%A3o-de-Dados-desafios-para-a-ci%C3%AAncia-brasileira.pdf
4. C. B. MEDEIROS, B.R. DARBOUX, J.A. SANCHEZ, M.L. MENEGUETTI, J.K. SHINWARI, J.C. MONTOYA, I. SMITH, A. McCRAY and K. VERMEIR. IAP Input into the Unesco Open Science Recommendation. Interacademies Report, June 2020 https://www.interacademies.org/sites/default/files/2020-07/Open_Science_0.pdf
5. NATIONAL ACADEMIES OF SCIENCES, ENGINEERING AND MEDICINE. Open Science by Design: Realizing a Vision 21st Century Research. The National Academies Press, 018. https://doi.org/10.17226/25116 .
6. Unesco. Intergovernmental Meeting of Experts (Category II) Related to a Draft UNESCO Recommendation on Open Science. Document SC-PCB-SPP/2021/OS-IGM/WD3, May 2021 https://unesdoc.unesco.org/ark:/48223/pf0000378381.locale=en
7. https://www.metabuscador.uspdigital.usp.br
8. https://www.re3data.org

.



**CLAUDIA BAUZER MEDEIROS** é Professora Titular do Instituto de Computação da Unicamp. Desenvolve pesquisa em gestão de grandes volumes de dados distribuídos e heterogêneos, para aplicações do mundo real, em várias áreas, especialmente agricultura, saúde e mudanças climáticas. Ex-presidente da SBC (2003-2006), coordena o grupo de Ciência Aberta da Academia Brasileira de Ciências e co-coordena as iniciativas em Ciência Aberta da Fapesp.

ARTIGO

# SOFTWARE LIVRE: PRÉ-REQUISITO PARA A CIÊNCIA ABERTA

**POR**

Christina von Flach e Fabio Kon

flach@ufba.br, kon@ime.usp.br

A ciência sempre se pautou pela disseminação do conhecimento, com destaque para a reprodutibilidade de experimentos. É claro que há limites práticos – e mesmo éticos – para o que pode ser compartilhado e reproduzido, mas expandir esses limites é um objetivo de interesse para cientistas de todas as áreas. Esse é o foco do movimento pela ciência aberta, que tem tido bastante repercussão nos últimos anos.

A Ciência Aberta requer que as ferramentas e os instrumentos necessários para a prática científica estejam disponíveis para todos, de modo que os experimentos possam ser reproduzidos e os resultados verificados por terceiros. No Século 21, o software consolidou-se como onipresente no conjunto de ferramentas usado por pesquisadores das Ciências Exatas e Biomédicas e tem uso crescente nas Ciências Sociais e Humanas. Software é usado para limpeza, processamento e visualização de dados, bem como para criar modelos e realizar previsões. Algoritmos especializados são codificados na forma de bibliotecas, scripts e metadados. Sem compartilhar todos esses artefatos, é muito difícil e custoso reproduzir a pesquisa científica e validar sua correção.

Assim, para alcançar a abertura plena na Ciência, é essencial que bibliotecas de software, scripts, pacotes e aplicativos usados e desenvolvidos por cientistas estejam prontamente disponíveis para qualquer pesquisador interessado em reproduzir ou estender uma determinada parte da pesquisa. Em particular, quando se trata de financiamento público, é altamente desejável que o software produzido no contexto da pesquisa (**research software**) seja **Software Livre** para garantir que seus benefícios sejam disponibilizados para uma vasta comunidade e de forma abrangente.

O **Movimento do Software Livre** teve início em 1983 quando o projeto GNU foi fundado como um meio de desenvolver um corpo robusto de software, incluindo um sistema operacional, ferramentas e aplicativos associados que estariam completamente disponíveis para todos os interessados. Em 1998, o termo **Open Source Software** passou a ser utilizado como uma estratégia para difundir, para empresas, essa nova forma colaborativa de desenvolver e compartilhar software. Atualmente, o software livre é reconhecido e adotado pelas principais empresas de TI do mundo, que usam e produzem software sob licenças de software livre/aberto. Milhares de componentes de software de alta qualidade estão disponíveis como software livre, desde drivers e sistemas operacionais até bibliotecas, arcabouços e aplicativos.

A **Free Software Definition** (https://www.gnu.org/philosophy/free-sw.html) estabelece o que é Software Livre em termos de quatro liberdades dadas aos seus usuários: (1) para executar o software para qualquer propósito, (2) para estudar e modificar o software (o acesso ao código fonte é uma pré-condição para isso), (3) para redistribuir cópias do software e (4) para distribuir suas versões modificadas do software. Essas quatro liberdades alavancam os princípios da **Ciência Aberta.** Por exemplo, a permissão para executar o software para qualquer finalidade permite que pesquisadores utilizem o software existente sem ter que comprá-lo ou construí-lo do zero para realizar seus próprios estudos. A permissão para estudar e modificar o software na forma de código-fonte oferece suporte à reprodutibilidade e replicabilidade, divulgando software de pesquisa, artefatos relacionados e o conhecimento embutido neles. Também aumenta a transparência **(fluxos de trabalho visíveis)**, auditabilidade e confiabilidade (os resultados podem ser verificados por terceiros e qualquer pessoa pode detectar e corrigir um erro ou um recurso malicioso). A permissão para redistribuir cópias possibilita o compartilhamento de **pacotes de replicação** que, além dos dados brutos, fornecem o código necessário para seus experimentos de análise e interpretação em diferentes ambientes. Finalmente, a permissão para distribuir versões modificadas do software permite que os pesquisadores desenvolvam seu próprio trabalho, reutilizando e expandindo o fluxo de trabalho, a base de código ou a ferramenta de alguém, de modo a compartilhar o novo conhecimento para o benefício de outros.

As **licenças de software livre** correspondem à definição mencionada anteriormente e, como tal, alavancam a Ciência Aberta. Para que seja

compartilhado como Software Livre (*Free and Open Source Software*), um trecho de código deve idealmente ser distribuído sob uma licença formalmente aprovada pela **Open Source Initiative** (https://opensource.org/licenses), de preferência uma que seja rotulada como "popular e amplamente utilizada ou com comunidades fortes", tais como a *GNU General Public License* (GPL) ou a *Mozilla Public License 2.0.*

Projetos de código aberto bem organizados adotam um modelo de desenvolvimento que resulta em código de alta qualidade que se adapta rapidamente a diferentes situações. Várias práticas estabelecidas são usadas e reconhecidas como contribuições importantes para a manutenção de software de alta qualidade. Estas incluem o uso de repositórios públicos, sistemas de controle de versão, colaboração por pares, revisão de código, testes automatizados, formatos e interfaces padrão e documentação relevante, e podem ser consideradas boas práticas para a Ciência Aberta.

Os sistemas de controle de versão permitem que várias versões do mesmo software sejam mantidas e, possivelmente, referenciadas por experimentos de pesquisa e artigos científicos. Repositórios abertos podem ser usados para compartilhar vários tipos de ativos de pesquisa, como algoritmos, dados, código, relatórios e fluxos de trabalho, apoiando a reprodutibilidade, reduzindo a redundância e promovendo a colaboração científica aberta. A colaboração entre pares é frequente e habilitada por meio de acesso compartilhado ao código-fonte e vários canais de comunicação. A revisão de código promove a melhoria da qualidade do software por meio de compartilhamento, colaboração e revisão por pares e pode ser aplicada a outros ativos de pesquisa. Os testes automatizados aumentam a confiabilidade e a facilidade de manutenção, além de promover agilidade no desenvolvimento de novos recursos. O uso de formatos de dados e de interfaces padronizados facilita a integração com outros sistemas e desencoraja a dependência de fornecedores específicos. A documentação frequente e contínua é uma prática recomendada para manter os guias do usuário, os manuais e os outros documentos relevantes atualizados em relação à versão mais recente do software. Essas, e outras práticas usadas pelas comunidades de software livre, podem ser consideradas boas práticas para a Ciência Aberta, uma vez que oferecem suporte contínuo à disponibilidade e reúso de ativos, colaboração entre pares, transparência do fluxo de trabalho e confiabilidade.

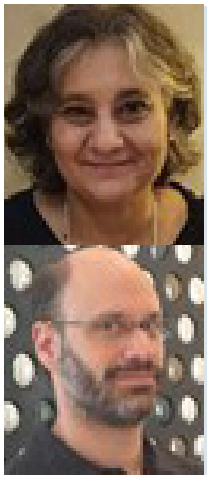
O software desenvolvido ou utilizado no contexto de uma pesquisa científica deve ser facilmente localizável, acessível, interoperável e reutilizável (https://bit.ly/3mGZass). Nesse contexto, os ecossistemas de software livre também promovem princípios da Ciência Aberta. Software livre pode ser localizado em repositórios com base em identificadores e descritores, utilizando diversos critérios como palavras-chave, linguagem de programação, versão do software, entre outros. A acessibilidade é encorajada em software livre disponível em repositórios abertos, com licenças de compartilhamento explícitas e bem definidas e documentação associada. A definição de interfaces de programação e formatos de entrada/saída e o uso de padrões promovem a interoperabilidade

e o reúso por vários grupos de pesquisa em todo o mundo. Além disso, no ecossistema da Ciência Aberta, o software deve ser sustentável, citável e reconhecido como uma valiosa contribuição da pesquisa científica, tão importante quanto os artigos, dados e metadados. A adoção do modelo de software livre promove a sustentabilidade do software no longo prazo. No entanto, esforços e incentivos são necessários para que os créditos para o software se tornem mais específicos e rastreáveis na ciência.

O Software Livre tem sido o principal componente da pesquisa em Ciência da Computação nas últimas décadas. Há centenas de exemplos de sistemas, de bibliotecas e de ferramentas que têm promovido o rápido desenvolvimento da pesquisa em Computação. Exemplos conhecidos incluem o sistema operacional Gnu/Linux, a ferramenta de composição tipográfica LaTeX e bibliotecas e linguagens para computação estatística, como R. Por fim, atualmente, há uma grande coleção de bibliotecas livres como Scikit-learn, TensorFlow e PyTorch que contribuem para o rápido desenvolvimento dos campos de Inteligência Artificial e Aprendizagem de Máquina.

Outras Ciências também se beneficiam com o software livre. A genômica, por exemplo, desenvolveu o software de código aberto EMBOSS, um pacote de análise para atender as necessidades de biólogos moleculares. Já ROOT é uma infraestrutura de suporte à análise de dados usada em pesquisas de Física Experimental de Altas Energias. É um software livre que nasceu no CERN e tem recebido contribuições de desenvolvedores de todo o mundo. Atualmente, mais de 1 Exabyte de dados científicos são armazenados em arquivos ROOT. O bóson de Higgs foi encontrado com o suporte do software livre ROOT.

O Software Livre é fundamental para a Ciência Aberta. Esta mensagem deve ser transmitida a cientistas, agências de fomento à pesquisa, organizações científicas e governamentais.

**CHRISTINA VON FLACH** é Professora Associada do Instituto de Computação da Universidade Federal da Bahia. Atua na área de Engenharia de Software tendo publicado mais de 100 artigos científicos em congressos e periódicos. Recentemente tem trabalhado com temas relacionados à Educação em Engenharia de Software, Sustentabilidade e Aspectos Sócio-técnicos de Ecossistemas de Software, e Ciência Aberta.

**FABIO KON** é Professor Titular de Ciência da Computação na Universidade de São Paulo. Atua na área de Sistemas Distribuídos, Engenharia de Software e Ciência de Dados tendo publicado cerca de 200 artigos científicos em congressos e periódicos. Recentemente tem trabalhado na área de Cidades Inteligentes e gostaria de ter mais tempo para tocar seu vibrafone.

ARTIGO

# PRINCÍPIOS FAIR: GESTÃO DE DADOS PARA HUMANOS E MÁQUINAS

**POR**

Maria Luiza M. Campos, Vânia Borges, João Luiz R. Moreira

mluiza@ppgi.ufrj.br, vjborges30@ufrj.br e j.luizrebelomoreira@utwente.nl

Os princípios FAIR [6], publicados em 2016, constituem um conjunto de 13 boas práticas que visam orientar a publicação de dados, em particular de pesquisas científicas, de modo a serem localizáveis(em inglês, *Findable*), acessíveis (em inglês, *Accessible*), interoperáveis (em inglês, *Interoperable*) e reusáveis (em inglês, Reusable). São 4 princípios para o F, 3 para o A, 3 para o I, e 3 para o R. Iniciativas recentes consideram vital garantir que os dados sejam FAIR, no sentido original do acrônimo, e, também, no sentido de "***F****ederated,* ***AI***- ***R****eady*", ou seja, "Dados Federados e Prontos para Inteligência Artificial (IA)", portanto legíveis e acionáveis por máquinas [1]. Dados legíveis significam dados em forma digital que possam ser lidos tanto por humanos quanto por máquinas, ou seja dados formatados com padrões de interoperabilidade sintática como, por exemplo, CSV, XML, e JSON. Dados acionáveis por máquina significam dados em forma digital no qual aplicações de IA possam 'raciocinar', a exemplo de dados na forma de grafos de conhecimento.

A iniciativa GO FAIR[2] vem complementar outras iniciativas já em andamento no Brasil, voltadas para a abertura de dados de pesquisa, a exemplo da participação brasileira na *Research Data*

1 https://www.go-fair.org/wp-content/uploads/2020/03/VODAN-IN-Manifesto.pdf
2 https://www.go-fair.org/

*Alliance* (RDA)[3] , da rede de repositórios da FAPESP (apelidado de 'metabuscador'[4]), do LattesData[5], e do suporte da Rede Nacional de Pesquisa (RNP) à implantação piloto de repositórios em universidades e centros de pesquisa[6]. É importante ressaltar que implementações associadas aos princípios FAIR foram impulsionadas em diversas iniciativas globais a partir de 2016, após a publicação original na revista Nature [6]. Os princípios FAIR são endossados pelo G7, pelo G20, pelo Fórum Econômico Mundial, e pela Comissão Europeia (entre outros), sendo a base da nuvem europeia de ciência aberta (EOSC) [7].

O Escritório de Apoio de Coordenação GO FAIR Brasil[8], liderado pelo IBICT com o suporte da UFRJ, FIOCRUZ, RNP e EMBRAPA, tem a responsabilidade de fomentar a adoção dos princípios FAIR no país, disseminando, apoiando e coordenando as atividades para sua implementação [4]. Além disso, provê suporte às necessidades das redes de implementação (RI) GO FAIR ativas, alinhado com as recomendações do RDA Brasil. Até o momento, 6 redes vêm sendo estruturadas, cada uma liderada por uma instituição chave em seu domínio: Saúde (FIOCRUZ), Agro (EMBRAPA), Biodiversidade (Jardim Botânico do Rio de Janeiro), Humanidades (IBICT), Ciências Nucleares (CNEN) e Ensino, Ciência, Tecnologia e Inovação (UNESP).

3 https://www.rd-alliance.org/groups/rda-brazil
4 https://metabuscador.uspdigital.usp.br
5 https://www.gov.br/cgu/pt-br/governo-aberto/noticias/2020/2/cnpq-e-ibict-lancam-lattes-data
6 https://www.rnp.br/noticias/rnp-implementara-federacao-de-repositorios-para-dados-de-pesquisa-ate-2020
7 https://eosc-portal.eu/
8 https://www.go-fair-brasil.org/

Embora os princípios FAIR sejam boas práticas de gestão de dados empregadas na Computação, sua grande diferença para outras iniciativas é o 'empacotamento' (de forma complementar) dessas boas práticas, e o direcionamento tanto para o uso humano quanto para o uso por máquinas. Os princípios FAIR refletem a convergência de uma série de linhas de pesquisa da Computação e da Metaciência, assim como de Organização do Conhecimento e da Ciência da Informação. Desse modo, observa-se que princípios relacionados à gestão de metadados são diretrizes típicas da Administração de Dados, enquanto princípios sobre protocolos de comunicação padronizados para acesso e reuso de dados são práticas comuns da Engenharia de Software, e os princípios sobre representação de conhecimento com dados distribuídos são apoiados por técnicas em dados interligados e IA.

Para a implementação dos princípios FAIR, desde 2020 a comunidade GO FAIR passou a trabalhar em um *framework* baseado nas experiências obtidas por grupos de trabalhos das diferentes redes de implementação. Conhecido como *framework* de *FAIRificação*, visa contribuir com os desafios identificados em processos de geração de dados FAIR, sendo constituído por três elementos-chave [5]. O primeiro elemento de análise são os metadados, estruturados de acordo com o domínio específico da pesquisa, legíveis e acionáveis por máquina, sendo por isso denominados de Metadados para Máquinas (em inglês, *Metadata for Machine* - M4M). O segundo elemento refere-se

à seleção da estratégia de implementação FAIR a ser adotada na pesquisa. Para auxiliar essa escolha, tem-se como referência o Perfil de Implementação FAIR (em inglês, *FAIR Implementation Profile* – FIP) que relaciona as estratégias, padrões e ferramentas adotadas por diferentes comunidades. O terceiro elemento do *framework* corresponde à tecnologia para publicar os metadados e serviços de acesso aos dados. O FAIR *Data Point* (FAIR DP) é um repositório de metadados que promove a visibilidade dos dados de pesquisa através de uma hierarquia de descritores padrão, possibilitando acesso automatizado, de acordo com o esquema de autorização definido (aberto ou tão restrito quanto necessário, empregando algum mecanismo de autenticação [2]).

Complementar a esses elementos do *framework* (M4M, FIP e FAIR DP), outros elementos se mostram essenciais quando se trata de enfrentar os desafios de apoiar a interoperabilidade semântica de dados, intra e inter domínios. Conforme pontuado por Guizzardi [1], os princípios FAIR, associados à interoperabilidade, necessitam não apenas de vocabulários, mas sim de ontologias de domínio bem fundamentadas. Essas ontologias devem obedecer práticas de engenharia de ontologias e modelagem conceitual baseadas em categorias básicas consistentes definidas por ontologias de fundamentação, de forma a permitir a correta associação entre elementos de conceitualizações de diferentes sistemas. A pesquisa brasileira nessa área é, hoje, referência mundial, com um histórico de longa data em eventos e grupos de pesquisa que são reconhecidos internacionalmente, como o Ontobras[9] e o Núcleo de Estudos em Modelagem Conceitual e Ontologias (NEMO) da UFES[10].

Como uma evolução dos princípios inicialmente voltados para recursos de dados, a comunidade FAIR passou a discutir e considerar um escopo mais amplo para o "R" em FAIR, Assim, além do reuso, reprodutibilidade requer que as operações sobre os dados representadas como fluxos de atividades nos assim chamados *workflows* científicos, sejam também gerenciados e aderentes aos princípios FAIR.

Muitas das soluções tecnológicas para apoiar os princípios FAIR vêm sendo experimentadas no contexto da *RI Virus Outbreak Data Network* (VODAN), criada no início da pandemia da COVID-19 e concebida para desenvolver uma infraestrutura distribuída de apoio à interoperabilidade de dados sobre surtos virais correntes e futuros [3]. No Brasil, o VODAN BR faz parte da rede GO FAIR Brasil Saúde, com colaboração multi-institucional com a UFRJ, a UNIRIO, e hospitais parceiros.

Ao considerar estratégias para a gestão de dados de pesquisa no Brasil, é importante adotar uma abordagem evolutiva, na qual *FAIRness* ainda representa um caminho a ser trilhado, combinando mudança de cultura, de políticas e de serviços de apoio ao pesquisador e, não menos importante, de suporte tecnológico. A iniciativa GO FAIR se estrutura neste sentido, através de seus 3 pilares, GO CHANGE, GO TRAIN e GO BUILD, orga-

9 https://www.inf.ufrgs.br/ontobras/en/14th-seminar-on-ontology-research-in-brazil/
10 https://nemo.inf.ufes.br/

nização também adotada pelo Escritório GO FAIR Brasil. Como primeiro passo, a cultura de implantação de repositórios vem se disseminando, com ampla participação de bibliotecários e de pesquisadores da Ciência da Informação. Ao mesmo tempo, pesquisadores brasileiros de áreas diversas participam de grupos de trabalho e de interesse associados às RI FAIR, assim como à iniciativa da RDA, transformando as propostas ali concebidas em políticas institucionais e treinamentos para seus pares. De forma complementar, mas já baseado em resultados de pesquisas anteriores no Brasil nas áreas de gestão de metadados, engenharia de ontologias, tecnologias e padrões da web semântica e dados interligados, segue-se investindo na construção de infraestruturas de apoio aos dados FAIR. Em especial, face aos desafios para se chegar ao amplo reuso e interoperabilidade, há que se desenvolver mecanismos de apoio baseados em IA, particularmente em técnicas de Aprendizado de Máquina e Processamento de Linguagem Natural.


Referências

1.GUIZZARDI, G. Ontology, ontologies and the "I" of FAIR. Data Intelligence 2(2020), 181–191.

2.LANDI A et al. The "A" of FAIR – As Open as Possible, as Closed as Necessary. Data Intelligence, 2020, 2(1–2) 47–55.

3.MONS, B. The VODAN IN: support of a FAIR-based infrastructure for COVID-19. European Journal of Human Genetics. 2020; 28. 1-4.

4.SALES, L. et al. GO FAIR Brazil: a challenge for Brazilian data science. Data Intelligence,2019,1(1) 238-245.

5.SCHULTES, E. et al. Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. ER 2020 Workshops, Austria, Proceedings. Springer Science and Business Media Deutschland GmbH. 2020. p. 138-147.

6.WILKINSON, M. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016; 3:160018.



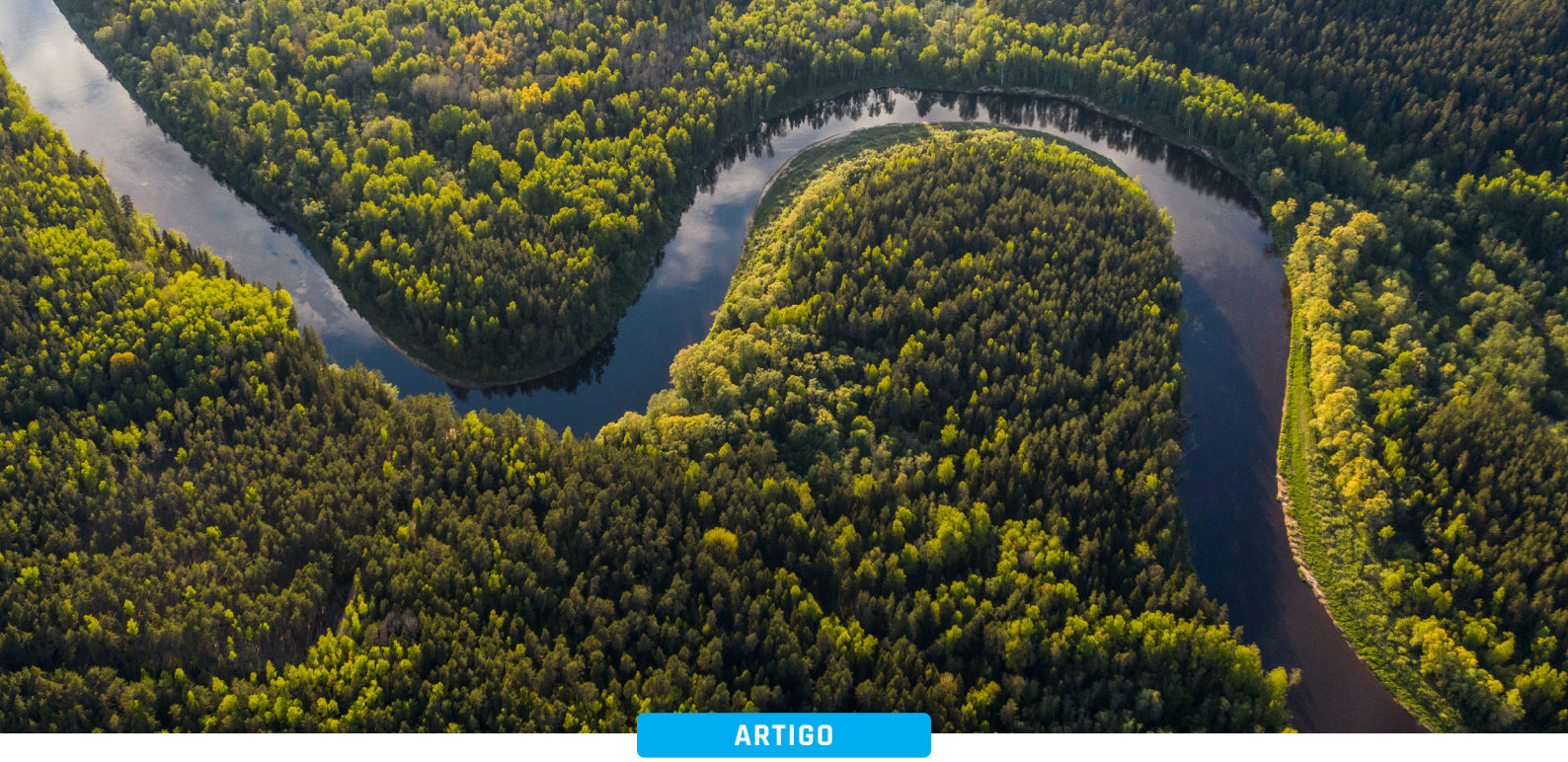
**MARIA LUIZA MACHADO CAMPOS** é Professora Associada no Instituto de Computação da Universidade Federal do Rio de Janeiro e atua na coordenação executiva do GO FAIR Brasil como representante internacional e na coordenação tecnológica do Projeto VODAN BR Seus principais temas de pesquisa estão associados à integração de informações heterogêneas, abordando principalmente gestão de metadados, engenharia de ontologias, modelagem conceitual, dados abertos e web semântica.



**JOÃO LUIZ RABELO MOREIRA** é Professor Assistente na Universidade de Twente e atua na coordenação executiva do GO FAIR Brasil, como representante na Europa. Sua pesquisa de doutorado seguiu a linha de interoperabilidade semântica, com ênfase na integração de dados de serviços inteligentes de emergência suportados por tecnologias de IoT no contexto de cidades inteligentes. Obteve seu pós-doutorado pela Universidade VU Amsterdam com pesquisa em engenharia de ontologias para workflows científicos com princípios FAIR.



**VÂNIA BORGES** é Militar da Reserva da Marinha, onde atuou por 9 anos na Diretoria de Saúde da Marinha no acompanhamento e controle das atividades para o desenvolvimento do Sistema Informatizado de Gestão de Saúde. Doutoranda pelo PPGI-UFRJ, atua como coordenadora de desenvolvimento no Projeto VODAN BR. Seus temas de interesse estão associados à interoperabilidade entre bases heterogêneas, gestão de metadados, ontologias, modelagem conceitual e web semântica.



ARTIGO

# CIDADÃOS COMUNS AJUDANDO NO COMBATE AO DESMATAMENTO DAS FLORESTAS TROPICAIS

POR

*Fernanda B. J. R. Dallaqua, Álvaro L. Fazenda, Fabio A. Faria*
*fernanda.dallaqua@unifesp.br, alvaro.fazenda@unifesp.br, e ffaria@unifesp.br,*

A criação de plataformas de Ciência Cidadã está incluída nas práticas de Ciência Aberta [1,2]. Esta área (do inglês, Citizen Science) utiliza a participação de voluntários não-especializados/comuns em diferentes tarefas de pesquisas tais como a coleta, a análise e a classificação de dados para resolução de problemas técnicos e científicos.

Ciência Cidadã é uma área que tem atraído bastante atenção de pesquisadores na academia devido a grande quantidade de dados gerados, que tendem a apresentar boa qualidade aliada ao baixo custo para sua obtenção. Acredita-se que a Ciência Cidadã é benéfica tanto para a comunidade científica quanto para os próprios voluntários envolvidos nos projetos e à sociedade como um todo [3].

Para os cientistas, a ajuda recebida em suas tarefas (coleta, análise e classificação) resulta em rápida obtenção de grandes quantidades de dados que têm valores inestimáveis para o avanço de suas pesquisas e, consequentemente, para a ciência [3].

Já os voluntários, além de adquirirem experiência no processo científico, são reconhecidos por suas contribuições e se sentem satisfeitos em integrarem um projeto com relevância científica e social [3].

Finalmente, o benefício para a sociedade vem da criação de uma estreita conexão entre a academia e o público, tornando-o mais consciente sobre complexos problemas existentes na sociedade que antes não eram tão acessíveis e, por sua vez, acabam muitas vezes resultando no engajamento dos voluntários na busca por soluções para esses desafios enfrentados (desmatamento de florestas e poluição do meio ambiente) [3].

Infelizmente, a cada ano são perdidos milhões de hectares de florestas tropicais através de desmatamento e degradação do território nacional. De acordo com um dos mais conhecidos e bem sucedidos programas de monitoramento, o PRODES (Programa de Monitoramento da Floresta Amazônica Brasileira por Satélite), o desmatamento na Amazônia Legal Brasileira (ALB) foi de 10.851 km² no período de agosto/2019 a julho/2020, correspondendo a um aumento de 7,13% quando comparado ao período anterior [4].

O desmatamento é causado por diversos fatores econômicos, como agropecuária, garimpo e extração ilegal de madeira. Tais atividades podem trazer consequências irreversíveis e catastróficas como perda da biodiversidade, aumento da emissão dos gases do efeito estufa, mudanças climáticas, desertificação, escassez de água potável, aumento de doenças e até surgimento de pandemias [5].

Como a conservação das florestas tropicais é urgente e extremamente necessária, programas de monitoramento e fiscalização foram criados por agências governamentais e instituições sem fins lucrativos. Esses programas utilizam, na maioria das vezes, imagens de sensoriamento remoto (imagens sobre a superfície terrestre obtidas à distância, muitas vezes a partir de satélites) e técnicas de processamento de imagens, inteligência artificial e fotointerpretação de especialistas para analisar, identificar e quantificar mudanças na cobertura florestal [6].

Em abril/2019 foi lançado o projeto *ForestEyes* [1], exemplificado na Figura 1, com o objetivo de gerar dados complementares, auxiliando os especialistas dos programas de monitoramento e órgãos de fiscalização. É hospedado na plataforma *Zooniverse* e tem como objetivo aliar Ciência Cidadã e Inteligência Artificial para ajudar no monitoramento do desmatamento de florestas

1 Site do projeto ForestEyes https://fafaria.wixsite.com/fabiofaria/amazon-deforestation

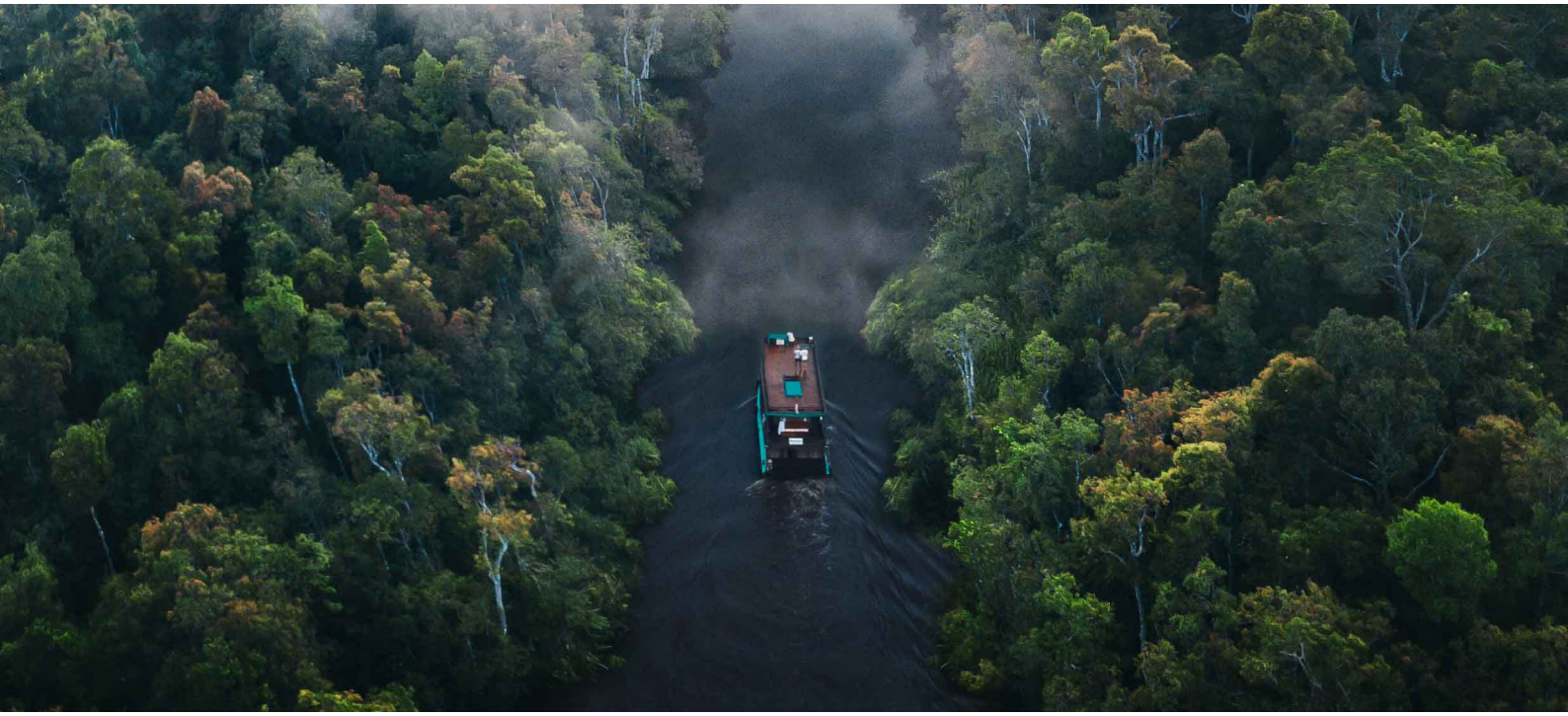

FIG. 01 | **UM DIAGRAMA SIMPLIFICADO DO PROJETO FORESTEYES.**

tropicais. Os voluntários/cidadãos comuns classificam partes delimitadas de imagens obtidas por satélite, chamadas de segmentos, em floresta ou não-floresta. Essas contribuições são analisadas e utilizadas como dados de entrada para um sistema inteligente, baseado em técnicas de aprendizagem de máquina, que irá por fim, reconhecer automaticamente a existência de locais de desmatamento em novas imagens de uma abrangente região florestal [7]. No futuro, o sistema poderá gerar sinais de alerta para as autoridades competentes ou gerar dados auxiliares para os programas oficiais de monitoramento.

O projeto *ForestEyes* [2] contou, até o momento, com a participação de 644 voluntários espalhados pelo mundo, os quais realizaram mais de 86.000 contribuições ou classificações sobre 5.408 segmentos de imagens em 6 diferentes campanhas que compreenderam uma pequena área do estado de Rondônia/Brasil para os anos base de 2013, 2016 e 2017. Com essas contribuições, o sistema inteligente atingiu acurácias médias de 80% em regiões de áreas com desmatamento consolidado ou recente, quando comparado com dados oficiais do PRODES, programa criado e mantido pelo Instituto Nacional de Pesquisas Espaciais (INPE).

O projeto mostrou que os voluntários são capazes de classificar segmentos de imagens da região da Floresta Amazônica com precisão, incluindo a identificação de tarefas consideradas difusas, ruidosas ou difíceis, resultando na criação de conjuntos de treinamento mais robustos para alimentar as técnicas de Inteligência Artificial baseadas em aprendizagem de máquina. Portanto, os cientistas cidadãos podem ajudar de forma efe-

2. Agradecimentos - INCT (a Internet do Futuro para Cidades Inteligentes, CNPq 465446/2014-0), CAPES, CNPq (408919/2016-7) e FAPESP (14/50937-1, 15/24485-9 e 18/23908-1).



tiva o combate ao desmatamento, tornando-se mais conscientes do grande desafio enfrentado pela sociedade e melhorando os processos de sistemas de monitoramento das florestas tropicais.


Referências:

1.Towards a UNESCO Recommendation on Open Science: Building a Global Consensus on Open Science. UNESCO. Disponível em: <https://en.unesco.org/>. Acesso: 10/10/2021.

2.Open Science @FAPESP. FAPESP. Disponível em: <https://www.fapesp.br/openscience/>. Acesso: 10/10/2021.

3.Grey, F. (2009). Viewpoint: The age of citizen cyberscience. Cern Courier, 29.

4.INPE (2021). A taxa consolidada de desmatamento por corte raso para os nove estados da Amazônia Legal em 2020 foi de 10.851km2. Disponível em: <http://www.inpe.br/noticias/noticia.php?Cod_Noticia=5811/>. Acesso: 09/10/2021.

5.Martin, C. (2015). On the Edge: The State and Fate of the World's Tropical Rainforests.Greystone Books Ltd.

6.Luz, E. F.; et al. The ForestWatchers: A Citizen Cyberscience Project for Deforestation Monitoring in the Tropics.Human Computation, v.1, p.137–145,2014.

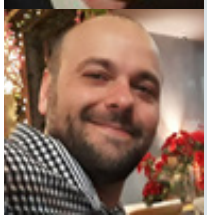
7.Dallaqua, F. B. J. R. (2020). Projeto ForestEyes - Ciência Cidadã e Aprendizado de Máquina na Detecção de Áreas Desmatadas em Florestas Tropicais. PhD thesis, Universidade Federal de São Paulo. Instituto de Ciência e Tecnologia.



**FERNANDA B. J. R. DALLAQUA** é graduada em Ciência e Tecnologia (2014), em Ciência da Computação (2015) e doutora em Ciência da Computação (2020) pelo Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo (ICT/UNIFESP). Atua na área de Aprendizado de Máquina e Sensoriamento Remoto.



**ÁLVARO L. FAZENDA** é Professor Associado do Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo (ICT/UNIFESP), com mestrado (1997) e doutorado em Computação Aplicada pelo Instituto Nacional de Pesquisas Espaciais - INPE (2002). Sua pesquisa tem foco em Programação Paralela e de Alto Desempenho, Sistemas Distribuídos e Ciência Cidadã.


**FABIO A. FARIA** é Professor adjunto do Instituto de Ciência e Tecnologia na Universidade Federal de São Paulo (ICT/UNIFESP), com mestrado e doutorado em Ciência da Computação pelo IC-UNICAMP e estágio de pós-doutoramento no Australian Institute for Machine Learning (AIML) da The University of Adelaide. Atua nas áreas de Inteligência Artificial, Processamento de Imagens e Visão Computacional para aplicações eScience.

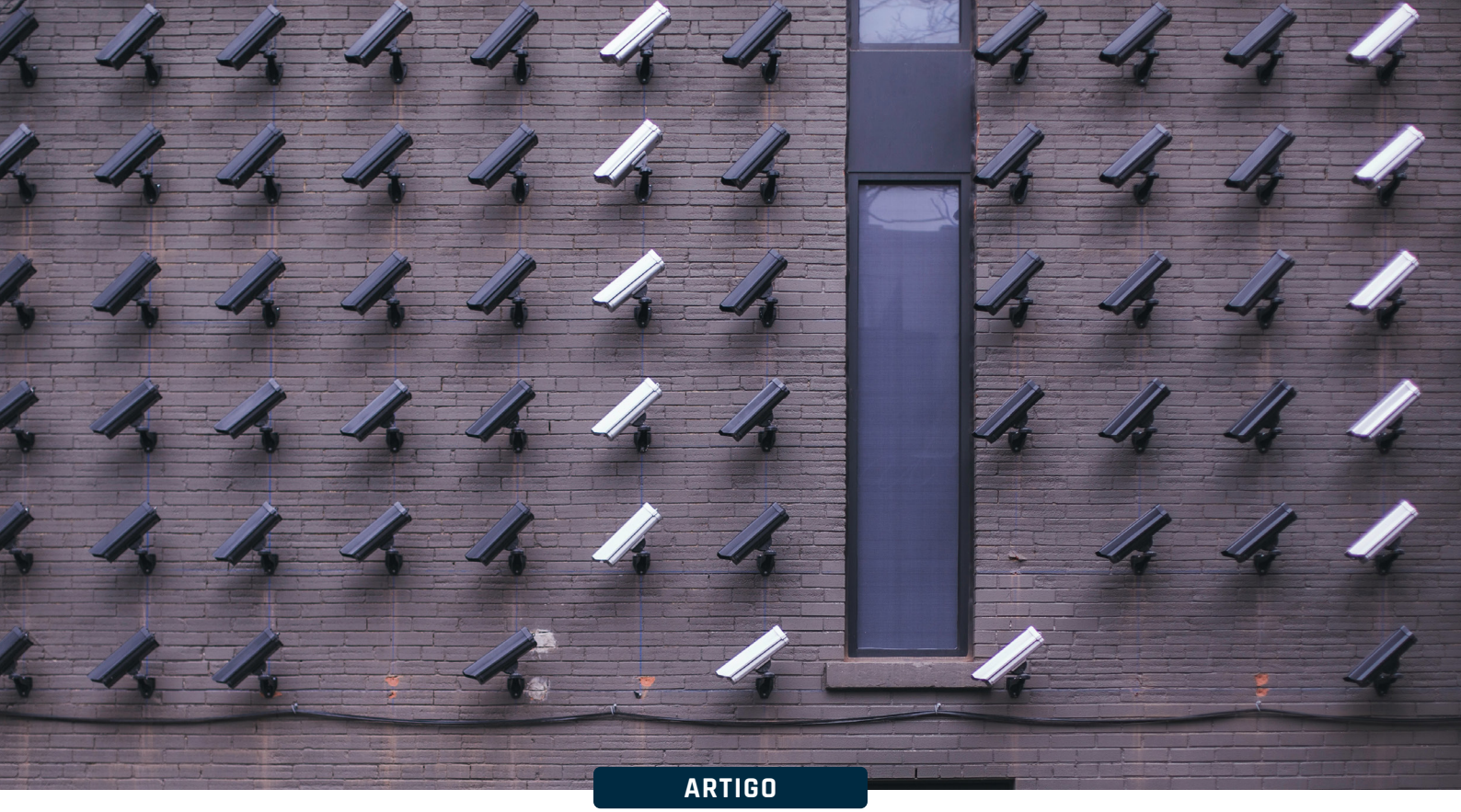
# 42°CSBC

CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO

## EM 2022 O CSBC VOLTA A SER PRESENCIAL.

**De 31 de julho a 05 de agosto**
**a gente se encontra em Niterói, RJ.**

csbc.sbc.org.br/2022 @congresso.csbc congresso.csbc



ARTIGO

# É POSSÍVEL APRENDER SOBRE AS PESSOAS SEM LHES ESCANCARAR A PRIVACIDADE?

**POR**

*Javam Machado*

*javam.machado@dc.ufc.br*

Organizações modernas - públicas ou privadas - têm continuamente coletado dados pessoais. Quando fazemos buscas na internet, nossos interesses são capturados pelo provedor do serviço de busca. Ao fornecer frequentemente o CPF para acumular pontos em um programa de relacionamento de supermercado, da farmácia de preferência ou do posto de gasolina mais próximo, o indivíduo permite associar padrões de consumo à sua pessoa, e assim facilita a construção de seu perfil histórico de consumo. O hábito crescente de fazer compras em grandes varejistas de vendas online por aplicativos ou sítios web possibilita o mapeamento de interesses individuais e o acúmulo de considerável quantidade de informação sobre as pessoas. Da mesma forma, quando uma pessoa usa um aplicativo de serviço de localização, *streaming* de áudio e de vídeo e redes sociais em seu *smartphone*, ela fornece importante volume de informação pessoal para os provedores desses serviços, seja diretamente por meio das suas escolhas ou informações que ela deliberadamente publica, seja pela coleta de dados que esses aplicativos fazem quando se faz uso de outros aplicativos no mesmo *smartphone*. Os deslocamentos do cidadão nas grandes cidades por meio dos serviços de compartilhamento de bicicle-

tas e de carros elétricos é igualmente uma relevante fonte de informação sobre os indivíduos que fazem uso desse tipo de serviço [6].

A despeito de eventuais abusos na coleta de informação sobre os indivíduos, há o consenso de que a aprendizagem sobre o consumo de bens e serviços, sobre o deslocamento nas cidades, buscas na internet e recomendação de acomodações, áudio e vídeo, todas essas coisas podem servir para melhorar sobremaneira a prestação de serviço, otimizar a venda e a distribuição de bens, definir políticas públicas, enfim facilitar a vida moderna das pessoas. Todavia a aprendizagem, a classificação e a identificação de padrões e de tendências podem ser todas realizadas quando se tem acesso a grandes conjuntos de dados, sem, contudo, representar ou re-identificar o indivíduo que contribui com os seus dados pessoais. Sem que seja possível associar o indivíduo a um item do conjunto de dados, é viável oferecer algum grau de privacidade, assim as pessoas estariam mais seguras para contribuir com o coletivo e não ser importunadas por campanhas de marketing direto, por controle de organizações públicas ou privadas, ou por informações tendenciosas que levam a escolhas induzidas. Desta forma, o que é de consenso pode ser atendido enquanto a privacidade das pessoas é minimamente respeitada.

A Lei Geral de Proteção de Dados (LGPD) discrimina dados pessoais e disciplina o tratamento desses dados pelas organizações a fim de dar garantias de privacidade aos indivíduos [2]. Logo no seu Art 2o inciso I, a LGPD afirma que a disciplina da proteção de dados pessoais tem como fundamento o respeito à privacidade. No seu Art. 5o, inciso I, a mesma norma também define dado pessoal como a informação relacionada à pessoa natural identificada ou identificável. Ainda no Art 5o, agora nos incisos III e XI respectivamente, a LGPD define dado anonimizado como sendo o dado relativo ao titular que não possa ser identificado; e anonimização como a utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo. A lei, portanto, não só associa dados pessoais ao conceito de privacidade dos indivíduos, como também estabelece o uso de técnicas de tratamento de dados que impossibilitem a associação dos dados aos seus respectivos titulares, ressalvado o uso para objetivos de segurança, combate à criminalidade e jornalismo, dentre outros assemelhados. Mesmo para uso acadêmico, a LGPD determina em seu Art 7o, inciso IV, que esforços sejam feitos para anonimizar os dados pessoais.

A prática comum de anonimização corrente nas organizações que coletam dados pessoais, definidas na LGPD por controladores, procura substituir os identificadores dos indivíduos por valores artificiais, normalmente chamados de pseudônimos. Assim, ao liberar dados para o público ou compartilhar com parceiros, o controlador muitas vezes substitui CPF e nome por valores resultantes de algoritmos que mapeiam os dados

originais para dados fictícios gerados por funções do tipo hash. Entretanto estudos mostram que esse processo de anonimização é insuficiente para impossibilitar a re-identificação do titular do dado ou seja é incapaz de assegurar que o dado público perde a possibilidade de associação, direta ou indireta, a um indivíduo, como preconiza o Art 5º da LGPD [2].

As estratégias mais promissoras para assegurar a perda de associação do titular ao dado publicado buscam associar técnicas de modificação dos dados originais para gerar um conjunto de dados de publicação que mantém características semelhantes ao original, todavia os dados reais dos titulares não estão completamente representados. As principais técnicas de anonimização são a supressão, a generalização e a perturbação [1].

A técnica de supressão de dados remove valores ou substitui um ou mais valores de um conjunto de dados por algum valor especial, impossibilitando a descoberta dos valores originais por um eventual adversário. A generalização aumenta a incerteza de um adversário ao tentar associar um indivíduo a seus dados. Nessa técnica, os valores dos atributos são automaticamente substituídos por valores semanticamente similares, porém menos específicos. A técnica de perturbação substitui os valores dos atributos originais por valores fictícios, mas semelhantes, de modo que informações estatísticas calculadas a partir dos dados originais não se diferenciam significativamente de informações estatísticas calculadas sobre os dados perturbados. Ao contrário das técnicas de generalização e de supressão, que preservam a veracidade dos dados, a perturbação resulta em um conjunto de dados com valores muitas vezes sintéticos [3].

As técnicas de anonimização descritas até aqui são chamadas sintáticas em oposição às técnicas probabilísticas, dentre elas a mais promitente é a privacidade diferencial [4], que goza de fortes garantias de privacidade para os titulares ancoradas em seu arcabouço de definição formal. A privacidade diferencial pode ser vista como um *middleware* entre um estudioso dos dados e um banco de dados, que responde, de maneira privada, a consultas executadas no banco de dados. O *middleware* que implementa a privacidade diferencial é comumente chamado de mecanismo.

Na privacidade diferencial, o mecanismo é dito aleatório porque a probabilidade de qualquer saída desse mecanismo não varia muito com a presença ou ausência de qualquer titular na base de dados [5]. Essa propriedade é garantida pela adição de ruído aleatório controlado à saída da consulta. Normalmente o controle do ruído segue uma distribuição de probabilidade que se assemelha à distribuição de probabilidade da resposta da consulta ao dado original. Assim, apesar de retornar respostas provavelmente fictícias, essa técnica assegura alta dificuldade de re-identificação do titular, enquanto fornece capacidade de análise e estudo do conjunto de dados consultado.

Neste ponto reunimos elementos para responder afirmativamente à pergunta no início deste texto. Entretanto, temos visto o avanço crescente do processo de coleta de dados, construção de perfis

individuais e aprendizagem sobre as pessoas, sem a contrapartida necessária para lhes assegurar a privacidade, a individualidade, o direito à autonomia de escolhas. Precisamos de ferramentas de anonimização formalmente sólidas e disponíveis para integrar esse processo, fortalecendo a aprendizagem e protegendo os indivíduos. A LGPD vai nesse sentido, mas é apenas o passo inicial.


Referências:

1.BRITO, F.; MACHADO, J. Preservação de privacidade de dados: Fundamentos, técnicas e aplicações. 36a JAI - Jornada de Atualização em Informática, Cap 3, pag 1–40. SBC, Porto Alegre, 2017.

2.Congresso Nacional. Lei Geral de Proteção de Dados - Lei No 13.709 de 14 de Agosto de 2018.

3.DOMINGO-FERRER, J.; SANCHEZ, D.; SORIA-COMAS, J. Database anonymization: Privacy, utility, and microaggregation-based inter-model connections. Synthesis Lectures on Information Security, Privacy, and Trust. Morgan & Claypool, 2016

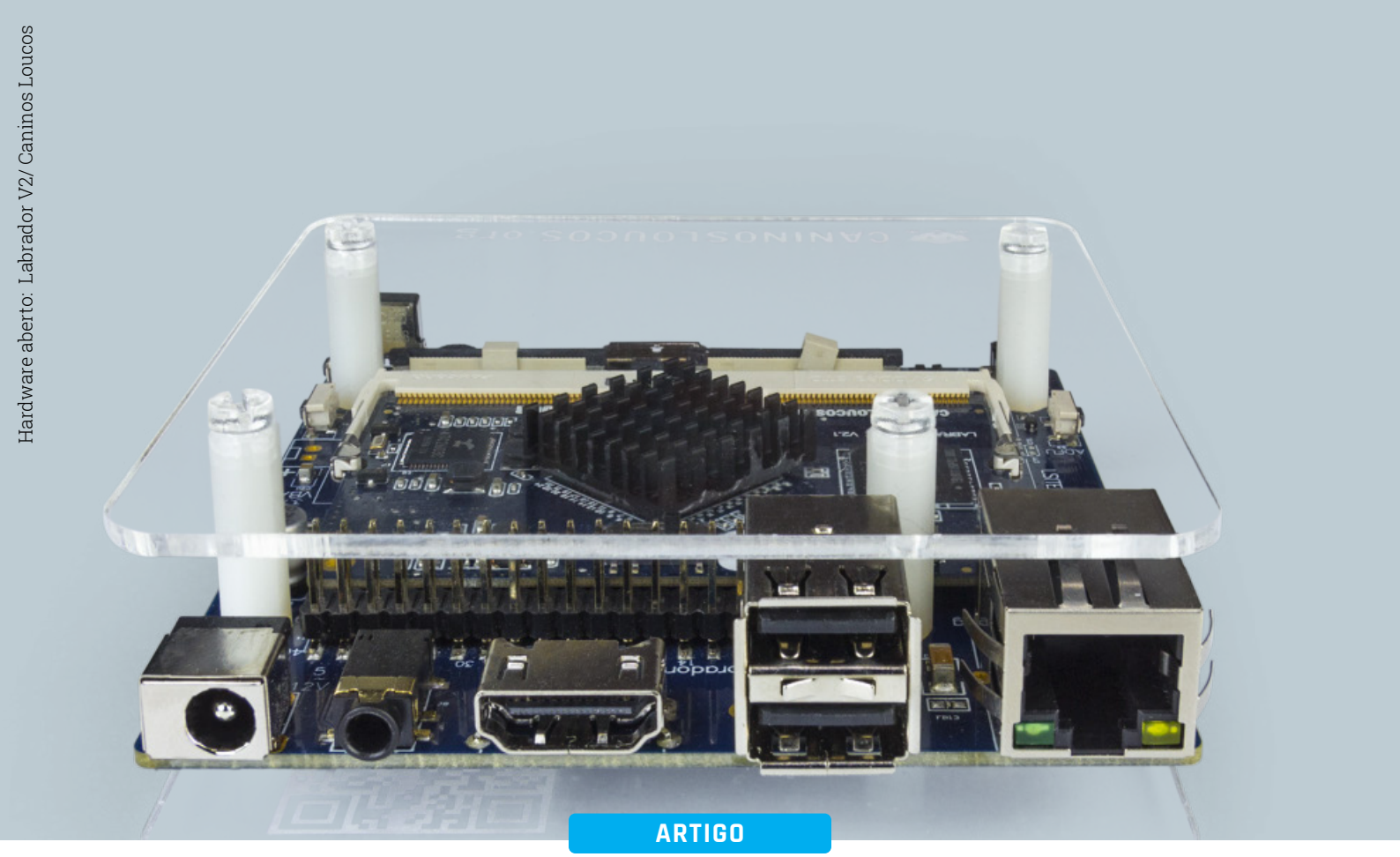
4.DWORK, C. Differential privacy. 33rd International Colloquium on Automata, Languages and Programming, pages 1–12, 2006

5.DWORK, C.; ROTH, A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4):211–407, 2014.

6.MACHADO, J; DUARTE NETO, E. Privacidade de dados de localização: Modelos, técnicas e mecanismos. 40a JAI - Jornada de Atualização em Informática, Cap 3, pag 105–148. SBC, 2021.



**JAVAM MACHADO** é professor titular do Departamento de Computação da UFC, onde fundou e coordena o Laboratório de Sistemas e Bancos de Dados (LSBD). Javam foi coordenador da Comissão Especial de Bancos de Dados da SBC (2017) e pesquisador visitante na Telecom SudParis – FR (2001) e no AT&T Labs-Research – USA (2018 e 2020). No momento, o professor Javam se interessa cientificamente pelas áreas de privacidade de dados e de não-discriminação em técnicas de aprendizagem automática.



Hardware aberto: Labrador V2/ Caninos Loucos

ARTIGO

# HARDWARE ABERTO, UMA ANÁLISE DE POSSIBILIDADES


**POR**

*Luigi Carro*
*carro@inf.ufrgs.br*


Os principais exemplos de *hardware* aberto são processadores, clássicos como CPUs RISC, superescalares, VLIW e mesmo máquinas vetoriais, FPGAs (*Field Programmable Gate Arrays*), placas de divulgação (por exemplo, Arduino) e o próprio CAD que suporta a realização de projetos sobre estas plataformas abertas, que vai desde compiladores até programas de síntese de alto nível, mapeamento, *placement* e *routing*.

## Instâncias de hardware aberto

O exemplo mais completo de processador aberto é o BOOM [1], *Berkeley Out-of-Order Machine*. Este é um processador baseado na arquitetura RISC-V, e é possível modificá-lo incluindo ou retirando instruções, mas também mudando sua organização, de pipeline simples para

fora de ordem de 2, 4 ou mais *issues*. Todo o ambiente de suporte para a síntese e a simulação da CPU é fornecido e por sua vez pode ser modificado, e o processador obtido é extremamente competitivo em relação às CPUs comerciais com as mesmas organizações. Muito importante, o BOOM pode rodar Linux, tem suporte à hierarquia de memórias e à memória virtual, além de compiladores e otimizadores.

Embora o BOOM permita enorme latitude na exploração do espaço de projeto, ele em si não é uma máquina vetorial, extremamente necessária nos problemas baseados em grande volume de dados do mundo atual. Existem também máquinas vetoriais abertas, como o VESPA, um soft-processor vetorial para FPGAs, totalmente aberto e suportado pela universidade de Toronto [2]. Embora um projetista possa projetar num FPGA qualquer tipo de processador, ao se dispor de uma máquina vetorial como soft-processor (que pode ser facilmente mapeada para um FPGA), o foco do projeto pode ser então o desenvolvimento do software que irá trabalhar com a enorme massa de dados. O objetivo do VESPA é a simplificação do projeto de *hardware* de uma máquina massivamente paralela, e todo o suporte de *software* para síntese e simulação pode ser encontrado.

Processadores abertos para sistemas embarcados já foram objetos de livros, como por exemplo o VEX [3], no qual uma máquina VLIW (*very long instruction word*) foi apresentada com arquitetura e organização abertas, bem como o compilador e ferramentas de suporte. Este processador poderia explorar melhor o paralelismo, em nível de instruções, disponível em aplicações embarcadas, com custo energético muito menor que um superescalar. Extensões reconfiguráveis deste processador podem ser vistas em [4], o que garantia sua adaptação a diferentes cenários com máxima eficiência, e com toda a cadeia de ferramentas de síntese, compilação e simulação.

Muitos *soft-processors* foram feitos, na sua maioria, para serem implementados em FPGA primordialmente. Embora existam diferentes arquiteturas comerciais fechadas, existe também o projeto de um FPGA totalmente aberto, seja o hardware seja o software, o *Open Source FPGA Foundation*[5]. O objetivo da OSFPGA é acelerar a adoção de FPGAs como um componente importante do processamento de informações, para execução acelerada de *software* crítico, sem as barreiras de entrada de um determinado fabricante. O grande problema é que, como o FPGA é base para diversos outros projetos, o grau de otimização que esta plataforma tem de fornecer é extremamente elevado, e, portanto, dificilmente alcançável apenas de maneira aberta.

Finalmente, tem-se o Arduino, que é uma plataforma baseada em microcontroladores e todo o suporte de hardware e software para que funcionem quase como *plug&play* [6]. Placas Arduino podem ler entradas digitais e analógicas, como botões, sensores ou mesmo mensagens complexas, e tomar uma ação a partir disto, ativando leds, um motor, ou respondendo à mensagem e gerando novas conexões. Toda a pilha de *software* suporta a programação de diferentes dispositivos. As CPUs envolvidas podem ser microcon-

troladores de 8-bits ou mesmo CPUs mais complexas, mas o grande valor agregado está na interface de entrada e saída com o mundo real (conversores AD/DA embutidos, portas de I/O e rede, etc.) e a enorme facilidade de programação.

**Pontos positivos**

É evidente que os custos não recorrentes de desenvolvimento de *hardware* podem ser largamente minimizados pelo uso de plataformas abertas, como as citadas acima, e outras que não foram listadas. O maior impacto das versões abertas de diferentes arquiteturas de *hardware* se dá na educação. Com a disponibilidade de hardware complexo, os cursos podem oferecer aulas com mais profundidade, e também podem explorar diferentes possibilidades de implementação de um mesmo conceito, como por exemplo, investigando diferentes versões de uma CPU para um projeto visando a baixa energia.

Em termos de pesquisa, as possibilidades são ainda mais interessantes, já que as plataformas abertas permitem fáceis modificações e adaptações a múltiplos cenários de uso. A flexibilidade do *hardware* aberto permite aos alunos e pesquisadores uma flexibilização enorme para experimentação de novos conceitos de maneira muito rápida. Por exemplo, uma nova arquitetura para acelerar o software de um nicho de mercado, com versões com e sem pipeline ou superescalaridade é facilmente obtida pelo conjunto de ferramentas do RISC-V. Como o *hardware* é muito parecido com os efetivamente usados no mercado, o resultado dos estudos e pesquisa está muito mais próximo da realidade do mercado.

**Pontos que dificultam a adoção**

No âmbito empresarial, a adoção de *hardware* aberto é mais complexa. A primeira questão importante é que os custos de NRE (*non recurring engineering)* são mitigados, mas não vão a zero. Enquanto um *software* de arquitetura aberta pode ser rapidamente adaptado e transformado com apenas o custo da mão-de-obra, o *hardware* tem, além dos custos de transformação, os custos de fabricação.

Outro aspecto importante é a diferença de desempenho. Mesmo ao se prototipar uma CPU aberta em um FPGA comercial, a diferença de desempenho é tão grande que talvez as modificações que agregam valor não possam se manifestar em comparação com o HW fabricado em tecnologia comercial mais velha. Por exemplo, pode-se paralelizar um algoritmo para aumentar sua velocidade de execução, mas quando mapeado para um FPGA, este último exige um hardware maior e mais caro que uma CPU rápida não aberta, e, portanto, dificilmente o FPGA será utilizado.

Todo o *hardware* atual, mesmo o aberto, depende de uma enorme cadeia de *software* para seu correto funcionamento, desde os compiladores até os montadores e os ligadores. Uma nova arquitetura de CPU pode ser facilmente feita com o RISC-V, mas o compilador e outras ferramentas não são facilmente modificáveis (embora também abertos). Isto vai contra um tempo rápido de projeto, pois modificar um compilador pode aumentar em muito o tempo de projeto, sem falar na necessidade de uma equipe multidisciplinar para

atuação tanto no *hardware* quanto no *software*. E o projeto deixa de chegar ao mercado de maneira rápida se for necessário programar em *assembler*.

Por fim, é preciso considerar que os produtos de *hardware* não comoditizados, com alto valor agregado, são aqueles em nichos de grande demanda e procura, como por exemplo *network processors* para 5G, processadores vetoriais da Nvidia para processamento gráfico e de realidade aumentada. Para estes nichos não existe uma versão *open source*. Embora existam processadores vetoriais abertos, como citado em [2], o coração das máquinas da Nvidia é o *scheduler* e o suporte à CUDA, que facilitam enormemente a vida do programador. Neste nicho, nem o *hardware* nem o *software* são abertos.

## Conclusão

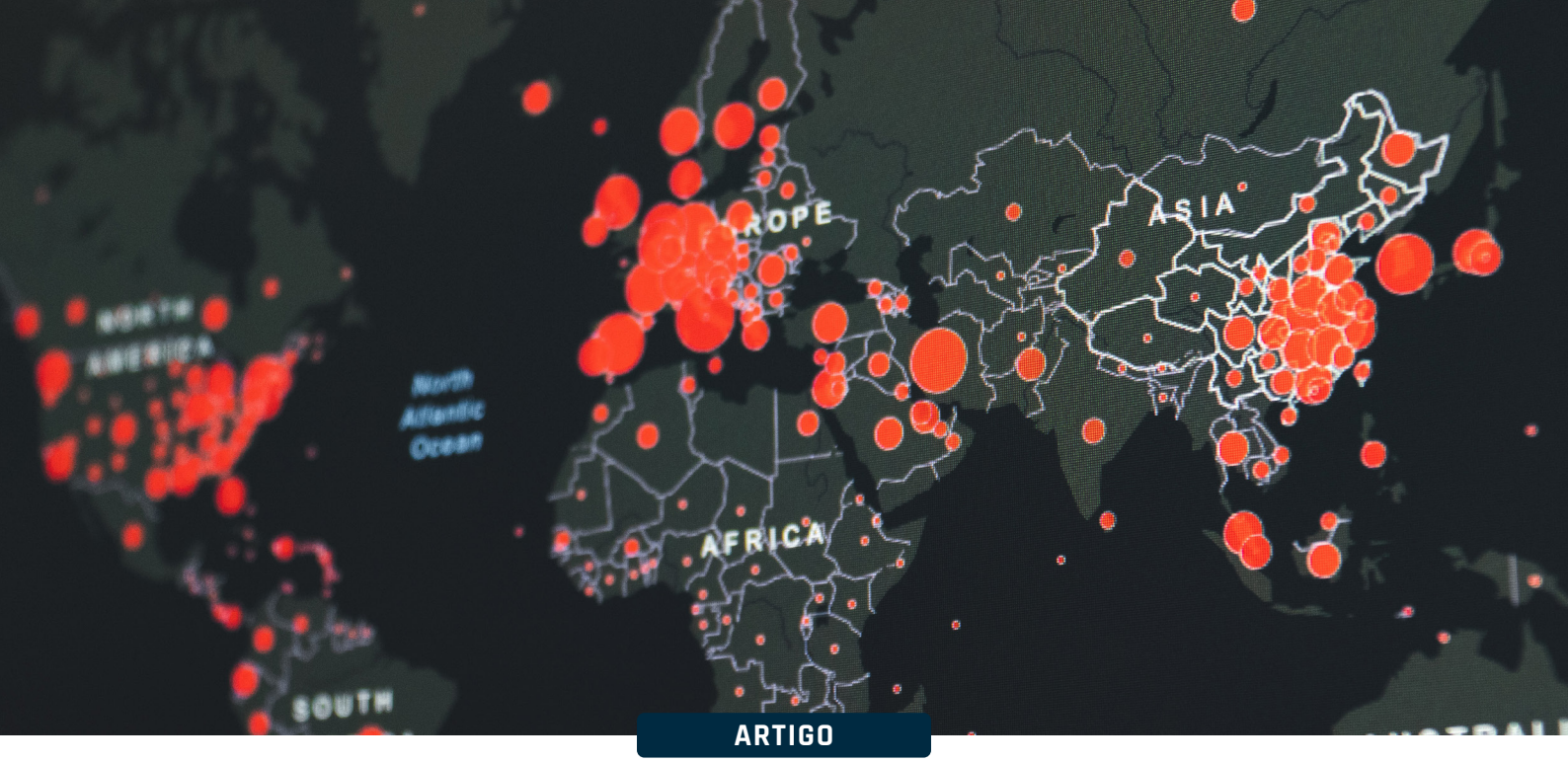
O *hardware* aberto definitivamente tem seu espaço, sobretudo no meio acadêmico. Comercialmente, os desafios ainda são enormes, sobretudo pelo atraso de bases de *hardware* aberto, e pelo custo de fabricação que elas impõem. Os melhores exemplos de open *hardware* têm origem em ambientes universitários ou em empresas menores, e tendem a estar um ou mais passos atrás em relação ao produto comercializado.


Referências

1. https://boom-core.org/
2.https://www.eecg.utoronto.ca/VESPA/
3. Embedded Computing: A VLIW Approach to Architecture, Compilers and Tools
Joseph A. Fisher, Paolo Faraboschi, Cliff Young. ISBN-13: 978-1558607668
Elsevier 2005
4. The r-vex processor: http://rvex.ewi.tudelft.nl/
5. https://osfpga.org/
6. Arduino.cc



**LUIGI CARRO** é professor titular do Instituto de Informática da UFRGS, e pesquisador CNPq-1A. Com formação em Engenharia Elétrica (UFRGS-1985) e Doutorado em Computação (UFRGS-1996), já trabalhou na ST-Microelectronics (1989-90) e foi visiting professor em Montpellier, San Diego, Torino e Delft. Publicou mais de 150 artigos, tendo orientado 25 teses de doutorado e vários mestrados junto ao PPGC-UFRGS. Sua pesquisa atual abrange software e hardware de baixa energia




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

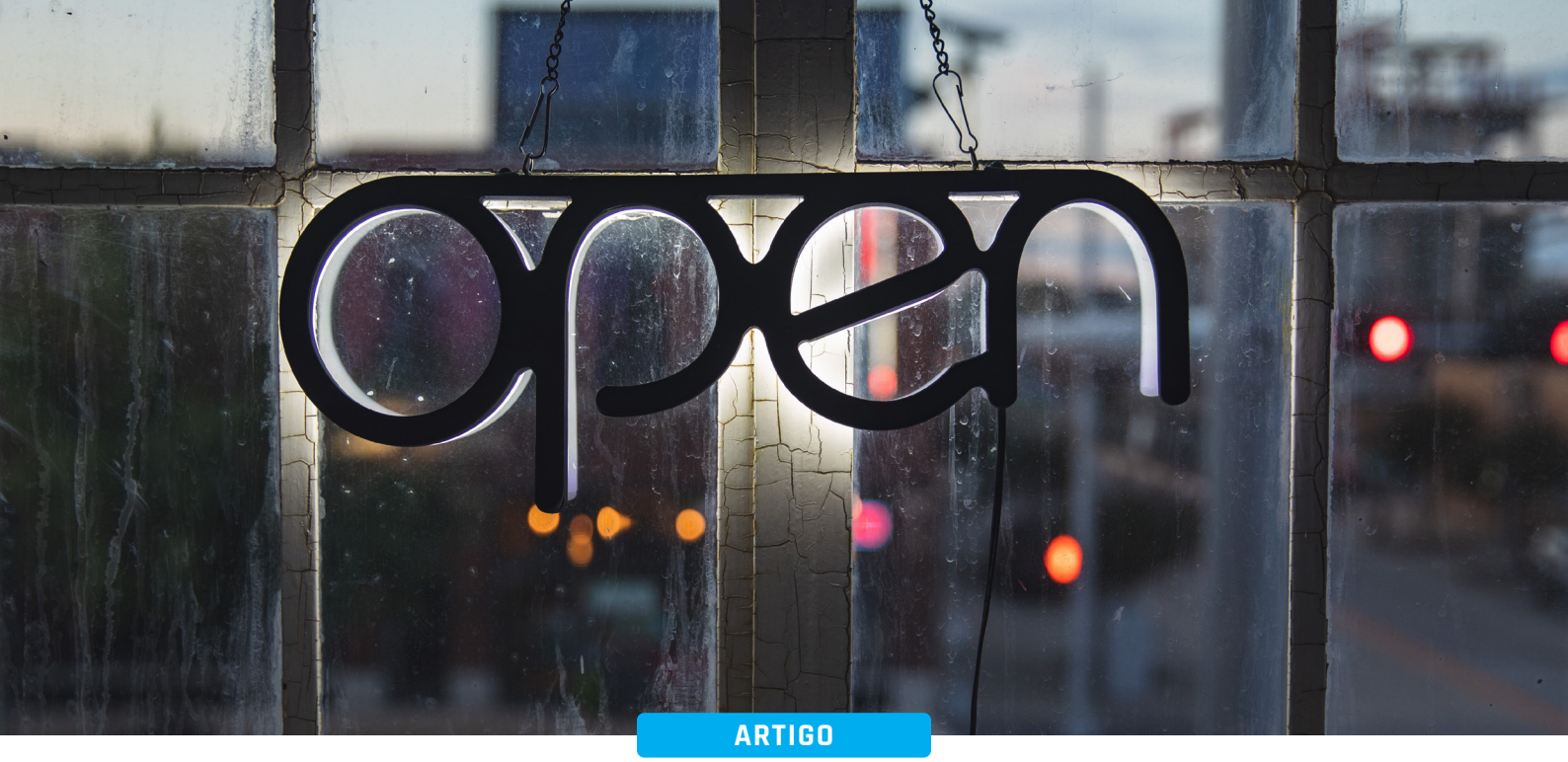
Os autores agradecem à Fapesp e aos integrantes das diversas equipes que viabilizaram o projeto COVID-19 DataSharing/BR.

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.




ARTIGO

# O PIONEIRISMO BRASILEIRO NO ACESSO ABERTO

**POR**

*Abel L. Packer e José Viterbo*

*abel.packer@scielo.org, viterbo@ic.uff.br*

Acesso Aberto é um modelo de publicação na Web de textos e de outros conteúdos científicos livre de custos para leitura e download com diferentes opções de reuso especificadas por licenças *Creative Commons*. Assim, um texto pode ter licença CC-BY, que permite amplo reuso inclusive comercial, desde que os autores e a fonte na qual foi publicado sejam citadas. Já uma licença CC-BY-NC-ND impede o uso comercial do texto e não permite derivativos.

A fundamentação do acesso aberto advém da constatação do conhecimento científico como uma dimensão essencial do desenvolvimento e da solução de grandes problemas da humanidade como as iniquidades, as emergência de pandemias como a COVID-19, a mudança climática e, de forma ampla, os Objetivos de Desenvolvimento Sustentável das Nações Unidas. Daí a concepção do conhecimento científico como um bem público global, e como tal deveria estar disponível para todos. De fato, a promoção do acesso aberto é identificada, por um lado, como um movimento com várias lideranças institucionais do meio acadêmico, como as bibliotecas, e de fundações em prol do desenvolvimento, e, por outro lado, como política pública, com dois destaques. O primeiro é o Programa da FAPESP *Scientific Electronic Library Online* (SciELO) [4] apoiado pela CAPES e CNPq, pioneiro na adoção do acesso aberto

(ver detalhes mais abaixo). O segundo é o chamado Plano S, liderado por um consórcio de fundações de apoio à pesquisa da Europa e algumas fundações dos Estados Unidos, que exige que os artigos das pesquisas que financiam sejam disponibilizados imediatamente em acesso aberto.

Existem duas modalidades principais de acesso aberto identificadas como rotas. Na dourada, (*Golden route*) os artigos são publicados em acesso aberto diretamente pelos periódicos, como é o caso dos periódicos SciELO. O segundo é a verde em que os autores depositam em repositórios institucionais uma versão do manuscrito, em geral já aprovada, mas não editada pelo periódico e, muitas vezes, após um embargo de 6 a 12 meses. Uma terceira rota são os periódicos híbridos, que publicam artigos em acesso aberto condicionado ao pagamento de uma taxa conhecida como *Article Processing Charge* (APC), com valor médio de 2 mil a 3 mil dólares, o que inviabiliza muitos autores de publicarem nelas.

O acesso aberto enfrenta enorme resistência de parte das grandes editoras comerciais que publicam a maioria dos periódicos científicos de qualidade e que desfrutam tradicionalmente de enorme lucro com a venda de assinaturas de acesso. Elas vêm adotando progressivamente o acesso aberto desde que não afete o retorno financeiro que tinham com o acesso por assinaturas. Uma estratégia de avanço do acesso aberto, que vem sendo utilizada em muitos países e instituições dos países desenvolvidos, é o uso dos Acordos Transformativos que permite aos pesquisadores de um assinante de periódicos, como seria os brasileiros no caso da CAPES, fazer uso do valor pago pelas assinaturas como APC.  O Brasil, na esteira do Programa SciELO, é um dos países que relativamente mais publica em acesso aberto. O estabelecimento de um acordo transformativo pela CAPES consolidaria esta condição.

**SciELO**

Lançada em 1998, a coleção SciELO Brasil é resultado do programa especial da FAPESP de apoio à infraestrutura de comunicação de pesquisas. Pioneiro na adoção do acesso aberto, o SciELO é hoje um dos mais importantes programas de comunicação científica em acesso aberto do mundo. O Modelo SciELO de Publicação é multilíngue, estrutura os textos em XML com elementos de dados seguindo padrões, e é adotado por outros 16 países formando a Rede SciELO de coleções de periódicos editados nacionalmente. Em outubro de 2021 a coleção SciELO Brasil indexou  308 periódicos que costumeiramente publicam 22500 artigos de pesquisa por ano. O repositório do SciELO Brasil acumula 450 mil documentos que em 2020 serviram uma média de um milhão de downloads e acessos por mês filtrando robôs e contagem única de acesso por sessão. A Rede SciELO, com 1250 periódicos, publica anualmente 51 mil novos artigos de pesquisa.

.

## A Biblioteca Digital de Acesso Aberto da SBC

A biblioteca digital de acesso aberto para a publicação de artigos selecionados em eventos e periódicos da SBC foi lançada em julho de 2018, inicialmente com a denominação de Portal de Conteúdo da SBC. Visando oferecer uma maior visibilidade internacional e obter-se uma URL mais curta e mnemônica (sol.sbc.org.br), no ano seguinte o repositório passou a ser chamado SBC OpenLib, recebendo o acrônimo de SOL e nova programação visual. Em um primeiro momento, a SOL tinha como principal objetivo disponibilizar em acesso aberto os anais de eventos da SBC que até então vinham sendo publicados na BDBComp [5], mantida pelo Departamento de Computação da UFMG, em bibliotecas digitais de associações científicas internacionais, ou em sites de editoras científicas privadas. Dentre estes, somente a BDBComp oferecia acesso aberto para todos os trabalhos publicados, ao contrário dos outros repositórios, que disponibilizam seus conteúdos somente para associados, assinantes ou compradores, com a possibilidade de os autores optarem pela disponibilização de seus artigos em acesso aberto, mediante pagamento de taxa específica.

A escolha do sistema a ser utilizado na SOL levou em consideração as recomendações de ferramentas fornecidas nas páginas de ajuda do Google Scholar, cuja utilização facilitaria a indexação automática do conteúdo publicado por aquela base. Dentre as indicadas, optou-se pelo Open Journal Systems (OJS), sistema de software livre de instalação fácil, rápida e sem custos, que é mantido e atualizado ativamente pelo *Public Knowledge Project* (PKP), um projeto que recebe recursos de instituições canadenses e americanas e emprega dezenas de desenvolvedores ao redor do mundo [3]. Além dessas características, o OJS é uma ferramenta de simples utilização e bastante popular, sendo utilizada por mais de 3.500 periódicos em 2014, dos quais, mais da metade eram editados e publicados em países em desenvolvimento [2]. Após a instalação de uma instância do OJS exclusivamente para a publicação de anais de eventos, a SOL foi expandida com a instalação de uma instância da ferramenta *Open Monograph Press* (OMP), também um sistema de software livre criado e mantido pelo PKP, desenvolvido para a publicação de livros e monografias em diferentes formatos eletrônicos [1]. Além disso, uma outra instância do OJS foi instalada para receber os periódicos tradicionais da SBC que eram publicados em plataformas externas.

Atualmente, a SOL publica os anais de mais de 110 eventos, entre simpósios, workshops e escolas regionais, o que compreende mais de 11.000 artigos disponibilizados online. A área de livros do repositório já disponibilizou mais de 50 títulos, compreendendo livros em capítulos e relatórios. Além disso, 9 periódicos da SBC, compreendendo mais 850 artigos já publicados, estão disponíveis na respectiva área, entre os quais o *Journal of the Brazilian Computer Society* (JBCS) - que, aliás, durante anos foi publicado na SciELO. Desde julho de 2020, quando a SOL teve 50.000 visua-

lizações de páginas, o número de visualizações seguiu aumentando até atingir 100.000 visualizações em março de 2021, demonstrando o crescente interesse da sociedade em geral pelo conteúdo disponível na plataforma. Como parte das ações em andamento, até o final de 2022 a plataforma será aperfeiçoada com a disponibilização de um mecanismo de busca integrada, que permitirá a consulta a toda a base de artigos, e serão estabelecidos processos para a indexação dos artigos publicados na SOL nas principais bases de referência.


Referências

1. ANDRADE, R. de L. de V.; ARAÚJO, W. J. Aplicação do Open Monograph Press por editoras brasileiras. Anais do XVII Encontro Nacional de Pesquisa em Ciência da Informação (XVII ENANCIB), 2017.

2.MACGREGOR, J.; STRANACK, K.; WILLINSKY, J. The Public Knowledge Project: Open source tools for open access to scholarly communication. In: Opening science. Springer, Cham, 2014. p. 165-175.

3.NDUNGU, M. W. Publishing with Open Journal Systems (OJS): A Librarian's Perspective. Serials Review, v. 46, n. 1, p. 21-25, 2020.

4. PACKER, A.L. The Pasts, Presents, and Futures of SciELO. In: EVE, M. P, GRAY, J. eds. Reassembling Scholarly Communications: Histories, Infrastructures, and Global Politics of Open Access. (Cambridge): The MIT Press, 2020. pp.297-313. https://doi.org/10.7551/mitpress/11885.003.0030

5.SILVA, L. V.; GONÇALVES, M. A.; LAENDER, A. H. F. Evaluating a digital library self-archiving service: The BDBComp user case study. Information Processing & Management, v. 43, n. 4, p. 1103-1120, 2007.




**ABEL L. PACKER** é cofundador do SciELO e Diretor do Programa SciELO / FAPESP e Coordenador de Projetos da FapUNIFESP desde 2010. Foi diretor da BIREME / OPAS /OMS entre 1999 e 2010. É mestre em Biblioteconomia.



**JOSÉ VITERBO** é Professor Associado do Instituto de Computação da Universidade Federal Fluminense, com mestrado em Ciência da Computação pela UFF (2004) e doutorado pela PUC-RIO (2009). Sua pesquisa tem se focado em Inteligência Coletiva e Sistemas Distribuídos com foco em Governo Eletrônico e Transparência Pública. Desde 2013, é Diretor de Publicações da Sociedade Brasileira de Computação (SBC).

# Associe-se ou renove sua associação

A SBC tem atuado fortemente em prol da consolidação e do desenvolvimento da computação no país! A participação das associadas e associados é fundamental para que a SBC continue defendendo os interesses da área!

**Acesse centraldesistemas.sbc.org.br/mom**

## Valores válidos até 31 de dezembro:

- **Efetivo/Fundador:** R$279,90
- **Efetivo/Fundador associado à ACM:** R$ 251,00
- **Professora e Professor da Educação Básica:** R$98,00
- **Estudante Pós-graduação:** R$98,00
- **Estudante Pós-graduação associado à ACM:** R$ 93,00
- **Estudante (graduação, técnico, tecnólogo e ensino básico):** R$24,00
- **Institucional:** 2.683,00*

*Em decorrência da pandemia de COVID-19, o valor da categoria Institucional será, excepcionalmente, de R $2.500,00 (o mesmo praticado em 2020).

Para mais informações, estamos disponíveis no e-mail sbc@sbc.org.br ou no (51) 99252-6018 (também WhatsApp).


conecte-se
com a SBC!


sbc.org.br