ACADEMIA BRASILEIRA DE CIÊNCIAS

MCMXVI

ABERTURA E GESTÃO DE DADOS:
DESAFIOS PARA A CIÊNCIA BRASILEIRA

AGOSTO • 2020

ACADEMIA BRASILEIRA DE CIÊNCIAS
MCMXVI

ABERTURA E GESTÃO DE DADOS:
DESAFIOS PARA A CIÊNCIA BRASILEIRA

INTEGRANTES:

Alberto Henrique Frade Laender
Membro Titular da ABC - Universidade Federal de Minas Gerais (UFMG)

Claudia Maria Bauzer Medeiros
Membro Titular da ABC - Universidade Estadual de Campinas (Unicamp)

Iscia Lopes-Cendes
Membro Titular da ABC - Universidade Estadual de Campinas (Unicamp)

Mauricio Lima Barreto
Membro Titular da ABC - Fundação Oswaldo Cruz (Fiocruz)

Marie-Anne Van Sluys
Membro Titular da ABC - Universidade de São Paulo (USP)

Ulisses Barres de Almeida
Membro Afiliado da ABC, 2018-2022 - Centro Brasileiro de Pesquisas Físicas (CBPF)

Em abril de 2018, a Academia Brasileira de Ciências (ABC) sediou o workshop Gerenciamento de Dados Científicos na América Latina e Caribe, uma parceria entre a ABC, o Museu do Amanhã e o *World Data System* do Conselho Internacional de Ciência (ISC-WDS, na sigla em inglês). Os acadêmicos presentes foram unânimes em reconhecer a necessidade de que os temas do workshop, e outros relacionados, deveriam ser aprofundados no âmbito da Academia e no contexto do país. Nesta perspectiva é que foi proposta, e acatada pela Diretoria da ABC, a criação do Grupo de Estudos Abertura e Gestão de Dados com o objetivo de discutir e apresentar iniciativas que possam auxiliar o Brasil a vencer os desafios existentes relacionados à produção, abertura e gestão de grandes volumes de dados no exercício dos diversos campos científicos no país. Em meados de 2019, então, o grupo começou a trabalhar neste primeiro texto. Cada seção apresenta considerações sobre um tópico específico levantado pelos membros, que poderá ser expandido em atividades futuras. O objetivo do documento, além de introduzir estes temas, é alimentar o debate e subsidiar posicionamentos da ABC. É importante ressaltar que várias instituições e organizações internacionais estão oficializando suas políticas de dados para os próximos anos, destacando-se a Estratégia Europeia de Dados, publicada pela Comissão Europeia em fevereiro de 2020[1].

[1] European Commission. European Strategy for Data. https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf.

CIÊNCIA ABERTA, DADOS ABERTOS: VISÃO GERAL

Vivemos em um mundo digital, no qual a comunicação e o acesso à informação foram profundamente alterados após a virada do século. A internet, capilarizada pela sociedade, promoveu transformações radicais em todas as áreas. Estas transformações ocorreram em sinergia com inovações disruptivas impulsionadas pelas tecnologias de informação e comunicação. Dentre os fundamentos deste mundo digital, temos o uso intensivo de volumes massivos de dados (também conhecidos como *big data*). Esses são coletados nas mais diversas áreas e, embora nem sempre produzidos diretamente por projetos de pesquisa, podem com frequência ser utilizados para promover o avanço do conhecimento e a melhoria da qualidade de vida. O aumento na produção e uso de dados, dentro e fora do âmbito científico e nas mais diferentes esferas, tem promovido intensos debates sobre sua importância, potencialidades, desafios e impactos nos diversos setores de uma sociedade conectada e global, apoiada no uso de tecnologias digitais cada vez mais potentes e especializadas.

De fato, esta crescente capacidade de processamento e armazenamento de dados, aliada à crescente interconectividade, torna possível que qualquer pessoa no planeta possa ter acesso virtualmente ilimitado a qualquer informação online. Isso abre novas possibilidades e demandas para o compartilhamento de dados e informações em larga escala, antes inexistentes, com evidentes implicações na democratização global do conhecimento, educação e capacitação (*capacity building*) para o século XXI. Todas essas questões ganham maior relevância e tornam-se ainda mais complexas no contexto da pesquisa científica, que também tem gerado grandes volumes de dados, para os quais o mais amplo acesso e integração tornam-se condições fundamentais para a completa exploração do seu potencial científico, ampliando as possiblidades de geração de novos conhecimentos e, ao mesmo tempo, reduzindo o custo específico de produzi-los.

Dados abertos tornaram-se, assim, um instrumento importante para o desenvolvimento científico. A gestão e o compartilhamento desses dados vêm se tornando, cada vez mais, condição basilar para o aumento da cooperação internacional em pesquisa e, portanto, do progresso científico. Ressalte-se que a noção de dado aberto não exige sempre que os dados propriamente ditos sejam incondicionalmente abertos e disponíveis para qualquer pessoa. A abertura de dados nem sempre pode ser total, por questões relacionadas à ética cientifica, à privacidade, ou à propriedade intelectual. No entanto, a descrição dos dados (metadados) é obrigatoriamente aberta – por exemplo, informação sobre suas características, localização e como acessá-los.

Na maioria das disciplinas científicas, a produção de dados de qualidade sempre foi o ativo mais importante e, frequentemente, mais custoso no processo de produção do conhecimento. Nessa perspectiva, nos últimos anos e no contexto da denominada ciência aberta, o acesso, compartilhamento e reuso de dados gerados por qualquer grupo de pesquisa passaram a ser estimulados por agências de financiamento. Mais recentemente, revistas científicas passaram a solicitar a guarda dos resultados e dos dados que subsidiam os achados de cada artigo submetido, com vistas a permitir, dentre outros, a verificação ou a reprodutibilidade dos achados científicos, testes de novas hipóteses, ou meta estudos com integração de dados de diversas fontes. De um lado, a disponibilização dos dados de pesquisa em repositórios abertos oferece a possibilidade de que uma coleta seja efetivamente utilizada por outros, não havendo necessidade de novas coletas. Por outro lado, este processo gera imensos desafios que se estendem desde aqueles relacionados à harmonização de diferentes bases de dados até questões de privacidade, principalmente quando se refere a informações pessoais. Essa harmonização é necessária para tentar contornar, entre outros, a heterogeneidade e a fragmentação dessas diferentes bases. Em seu conjunto, geram desafios éticos e de privacidade que necessitam de regras e infraestruturas especiais para sua proteção, acesso e uso.

Neste contexto, observa-se um grande esforço de convergência internacional acerca das recomendações e políticas relacionadas ao acesso, compartilhamento e reuso de dados para pesquisa. A guarda e o acesso a esses dados ocorre em repositórios acessíveis via *web*, criados especialmente para tal finalidade. Seu conteúdo é disponibilizado para potenciais usuários do mundo inteiro, aumentando a visibilidade da pesquisa associada e possibilitando o surgimento de colaborações multinacionais.

Surgem, assim, novas práticas de produção do conhecimento que utilizam modelagens computacionais e algoritmos sofisticados, resultando em uma ciência intensiva no uso de dados (a chamada *data-intensive science*) e na colaboração por meio do uso de plataformas de pesquisa baseadas em computação de alto desempenho. Isto é refletido no termo *eScience*, um paradigma em que avanços de pesquisa e tecnologia em computação se aliam à

pesquisa em outros domínios do conhecimento, possibilitando descobertas em todos os campos envolvidos. Essa onda tem provocado imenso impacto em todas as áreas do conhecimento, trazendo necessidades de readequações epistemológicas (por exemplo, *data-driven versus hypothesis-driven science*), metodológicas (como questões inferenciais na análise de grandes volumes de dados, por exemplo) e operacionais (novas infraestruturas para armazenamento e processamento de grandes volumes de dados e governança para acesso compartilhado), que afetam o modo de produzir o conhecimento pelas mais diversas áreas da ciência. Neste contexto, o papel e as implicações do uso da inteligência artificial na produção de conhecimento científico tornam-se cada vez mais relevantes. Vale ressaltar que o sucesso das técnicas da inteligência artificial depende fortemente do trabalho conjunto entre especialistas em computação (que dominam tais técnicas) e pesquisadores de outras áreas (que proveem o conhecimento sobre os dados a serem processados).

AVALIAÇÃO DAS TENDÊNCIAS INTERNACIONAIS SOBRE A ABERTURA E A GESTÃO DE DADOS, INCLUINDO OS MODELOS ADOTADOS PELAS INSTITUIÇÕES DE PESQUISA E AGÊNCIAS DE FOMENTO

A entrada no século XXI marca uma série de declarações de países sobre a importância da disponibilização mais ampla de resultados científicos coletados por pesquisadores, oriundos de financiamento com recursos públicos, para o avanço da ciência e do conhecimento e para a descoberta de novas soluções para o seu desenvolvimento econômico e social. Conforme salientado pelo *UK Reseach Council*, "dados de pesquisa financiados com recursos públicos são um bem público, produzidos no interesse público, e devem ser tornados disponíveis de forma aberta, com o menor número de restrições possíveis, de forma responsável e oportuna"[2], ilustrando bem o espírito que vem norteando todas as ações nesta direção.

Na prática, os primeiros países a implementar oficialmente políticas de abertura de dados como parte integrante das boas práticas da pesquisa foram Austrália, Reino Unido e Países Baixos, entre 2003 e 2004. Desde então, as agências de fomento desses países passaram a definir políticas nacionais de gestão de dados e a reservar anualmente uma parcela dos seus recursos financeiros para assegurar a criação e manutenção de centros dedicados à gestão e à curadoria desses dados. Parte desses recursos continua sendo anualmente dedicada ao treinamento de pesquisadores, técnicos e bibliotecários (que se tornaram "bibliotecários de dados") na criação e manutenção de repositórios nacionais ou institucionais de dados de pesquisa. Em paralelo, na mesma década, as principais agências de fomento dos Estados Unidos (EUA) lançaram um conjunto de iniciativas para promover a abertura de dados dos projetos por elas financiados.

A implementação de políticas para dados de pesquisa abertos se propagou aos poucos, sendo hoje considerada prática obrigatória por todas as agências de fomento públicas e privadas da América do Norte, Oceania e Europa Ocidental. Na Ásia, destacam-se ações recentes da China, Japão e Coréia do Sul. Na África, além da África do Sul, vários países já estão iniciando trabalhos nessa direção. Na América do Sul, algumas iniciativas vêm sendo adotadas desde meados da década de 2010, destacando-se, em âmbito nacional, as ações brasileiras e chilenas.

Ao mesmo tempo, começaram a surgir entidades internacionais com a missão de apoiar e estimular a abertura de dados como forma de acelerar o progresso científico, promover a colaboração em pesquisa, facilitar o reuso (e consequente economia de recursos) e garantir a reprodutibilidade. Destacam-se, dentre outros, o *World Data System*[3] (WDS), um órgão interdisciplinar do Conselho Internacional de Ciência (ISC, na sigla em inglês) cujos membros são instituições e sociedades científicas, o CODATA[4] (comitê de dados do ISC), e a *Research Data Alliance*[5] (RDA), que congrega pesquisadores, técnicos e agências de fomento de 137 países com o objetivo de propor e implementar métodos e padrões associados a dados abertos. WDS, CODATA e RDA têm desenvolvido uma série de iniciativas em todo o mundo visando estimular o amplo compartilhamento de dados de pesquisa.

Os modelos e políticas de gestão de dados variam entre países dentro de um espectro que tem dois extremos: centralizado e distribuído. No modelo centralizado, alguns repositórios nacionais concentram os dados de pesquisa e centros nacionais de dados e/ou computação científica fornecem infraestrutura computacional e de apoio a

[2] UK Research and Innovation. Common principles on data policy. https://www.ukri.org/funding/information-for-award-holders/data-policy/common-principles-on-data-policy/.
[3] World Data System. https://www.icsu-wds.org/.
[4] CODATA. https://codata.org.
[5] Research Data Alliance. https://www.rd-alliance.org/.

pesquisadores e instituições. No distribuído, cada instituição acadêmica (ou até mesmo cada laboratório ou departamento) define suas políticas e repositórios, adotando, via de regra, padrões para o armazenamento e a troca de dados, em geral propostos pela ISC (por meio de seus órgãos já mencionados) ou pela RDA. Embora não exista um modelo totalmente centralizado, países como a África do Sul e os Países Baixos têm características mais próximas deste padrão. Na Austrália, uma única instituição (*Australian Research Data Commons*-ARDC[6]) serve como ponte de comunicação de dados entre as centenas de instituições acadêmicas, além de prover serviços de dados e promover eventos e treinamento de pessoal para todo o país, em todos os níveis. Já os EUA são o principal exemplo de um modelo totalmente descentralizado.

O conceito de dados de pesquisa abertos (*open research data*) é frequentemente considerado como sinônimo de ciência aberta, que, em setembro de 2017, foi declarada como uma das prioridades dos países do G7[7]. Ainda, segundo a definição de um estudo de julho de 2018 das Academias de Ciências, Engenharia e Medicina dos EUA, a ciência aberta é formada pelo tripé "acesso aberto, dados abertos, processos abertos"[8].

No mesmo período, a Comissão Europeia lançou as bases para o plano de financiamento à pesquisa para o período 2021-2027 (chamado *Horizon Europe*), que será apoiado em 3 pilares, um dos quais a ciência aberta[9]. No contexto europeu, ciência aberta combina acesso aberto, dados abertos e colaboração internacional em pesquisa. Reforça-se, em todos esses casos, a indissociabilidade de ciência em si e dos dados processados e produzidos pela ciência. Devido ao planejamento da *Horizon Europe*, os países-membro da União Europeia estão lançando programas para fomentar o acesso aberto (artigos) e criar redes de repositórios de dados abertos. Um exemplo é a iniciativa francesa de ciência aberta, lançada em julho de 2018[10].

Dentro do *Horizon Europe*, destaca-se a iniciativa multinacional *European Open Science Cloud*[11] (EOSC). Iniciada em 2017 para fornecer infraestrutura de pesquisa aos países da União Europeia, será utilizada como a base computacional das iniciativas de ciência aberta do *Horizon Europe*. Neste sentido, abrangerá compartilhamento de dados e de ferramentas de software para todos os membros da União.

Todas as iniciativas de dados abertos devem ser consideradas sob dois ângulos: as práticas adotadas pelos pesquisadores e as políticas preconizadas ou incentivadas por agências de fomento e instituições acadêmicas públicas e privadas. Geralmente, os cientistas seguem as práticas ou recomendações de suas respectivas áreas de pesquisa, como, por exemplo, na escolha de repositórios, de padrões de armazenamento ou de metadados. Já as instituições criam a infraestrutura computacional e de pessoal para apoiar seus pesquisadores e servem de ponto de contato para as agências financiadoras. Estas, por sua vez, definem diretrizes associadas aos dados resultantes das pesquisas por elas financiadas. O primeiro passo nesta direção é a obrigatoriedade da criação, por parte dos projetos, dos chamados Planos de Gestão de Dados, que descrevem os dados associados às pesquisas e como os pesquisadores pretendem disponibilizá-los e preservá-los para usos futuros. Vale destacar que, a partir de janeiro de 2019, quase todas as agências de fomento europeias passaram a seguir as recomendações de um formato único para esses planos, que são obrigatórios para qualquer projeto que venha a ser financiado.

Internacionalmente, há um consenso com relação à necessidade de que os dados sigam recomendações FAIR (acrônimo que significa *Findable*, *Accessible*, *Interoperable* e *Reusable*). Em outras palavras, não basta que os dados sejam abertos, eles precisam ser encontráveis e acessíveis para reuso. Originária dos Países Baixos, a iniciativa FAIR vem permeando todas as ações mundiais de dados científicos abertos, sendo hoje disseminada por uma coalisão de instituições acadêmicas, agências de fomento e governos federais de dezenas de países.

[6] Australian Research Data Commons. https://ardc.edu.au/.
[7] G7 Ministerial Meeting on Science Communiqué (2017). http://www.g7italy.it/sites/default/files/documents/G7%20Science%20Communiqu%c3%a9/index.pdf.
[8] National Academies of Sciences, Engineering and Medicine. Open Science by Design - Realizing a Vision for 21st Century Research. http://nap.edu/25116.
[9] European Commission. Research and Innovation Policy - Open Science. https://ec.europa.eu/research/openscience/index.cfm.
[10] National Plan for Open Science. https://cache.media.enseignementsup-recherche.gouv.fr/file/Recherche/50/1/SO_A4_2018_EN_01_leger_982501.pdf.
[11] European Commission. European Open Science Cloud. https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.

REPOSITÓRIOS DE DADOS TEMÁTICOS ABERTOS

As iniciativas relatadas na seção anterior refletem políticas multinacionais ou de Estado associadas à ciência como um todo. Nessa mesma direção, e muito anteriores a tais iniciativas, ao final dos anos 80, surgiram esforços internacionais para criação de bases de dados de referência para domínios específicos do conhecimento – por exemplo, em física, astronomia, biodiversidade ou genômica. Nesta última, destaca-se o *International Nucleotide Sequence Database Collaboration*[12] (INSDC), em que um acordo firmado entre órgãos internacionais gestores de bancos de dados de sequências (nos EUA e na Europa, e posteriormente no Japão) determina que qualquer sequência depositada em um dos bancos componentes seja imediatamente compartilhada com os demais, de modo que o mundo rapidamente tenha acesso às informações do genoma humano, vírus, bactérias patogênicas e genoma de várias espécies de interesse agronômico ou médico.

O repositório europeu desta colaboração (o *European Nucleotide Archive*-ENA) é alimentado por laboratórios de pesquisa de 25 países e está contido na iniciativa Elixir[13]. Complementam o INSDC, o *National Center for Biotechnology Information*[14] (NCBI), nos EUA, e o *DNA Data Bank of Japan*[15] (DDBJ), no Japão.

O *Global Biodiversity Information Facility*[16] (GBIF) é outro exemplo de um esforço multinacional de mais de duas décadas para a criação e manutenção de um repositório temático de dados abertos para pesquisas em biodiversidade. Ainda na área de ciências da vida, o Global Genome Biodiversity Network (GGBN)[17] congrega informações de coleções biológicas diversas. Dados de todos os países do mundo são fornecidos por pesquisadores individuais e instituições de pesquisa, validados centralmente e, a seguir, publicados online.

A PANDEMIA DE COVID-19 E O MOVIMENTO MUNDIAL PELO COMPARTILHAMENTO DE DADOS ABERTOS

A pandemia de COVID-19 mostrou muito rapidamente a cientistas de todas as áreas a premência do compartilhamento de dados, principalmente clínicos, epidemiológicos, na área de "omics" e em ciências sociais. Se o compartilhamento nas três primeiras áreas está sendo motivado por problemas diretamente associados à saúde, existe a necessidade de conectá-los com dados socioeconômicos, geográficos e educacionais para que se possa entender problemas regionais e planejar o atendimento adequado à população e, principalmente, o futuro pós-pandemia.

Muitas das iniciativas de compartilhamento partem de grupos de pesquisadores ou instituições de pesquisa. No entanto, a comunidade acadêmica mundial tem se unido para estabelecer recomendações para que se priorize o compartilhamento desses dados, resguardando-se aspectos éticos e de privacidade[18]. Neste sentido, a disponibilização de dados de pesquisa de qualidade está sendo caracterizada como essencial para a produção de conhecimento – e muitas vezes priorizada pelos cientistas, que estão disponibilizando os dados antes mesmo de produzir publicações associadas. De certa forma, esta necessidade vem acelerando o reconhecimento, no mundo acadêmico, da importância de dados abertos. A pandemia e essas iniciativas contribuem também para que a sociedade dê reconhecimento ao trabalho naturalmente colaborativo em ciência.

Dentre os muitos projetos internacionais que se debruçam sobre compartilhamento de dados da COVID-19 podem ser citados o documento da Academia Africana de Ciências (AAS, na sigla em inglês), que dedica uma seção inteira ao

[12] International Nucleotide Sequence Database Collaboration. http://www.insdc.org.
[13] ELIXIR Core Data Resources. https://elixir-europe.org/platforms/data/core-data-resources.
[14] National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/.
[15] DNA Data Bank of Japan. https://www.ddbj.nig.ac.jp.
[16] Global Biodiversity Information Facility. https://www.gbif.org/.
[17] Global Genome Biodiversity Network. http://www.ggbn.org/ggbn_portal/.
[18] Almeida, B.A, Doneda, D, Ichihara, M.Y, Netto, M.B, Matta, G.C, Rabello, E.T, Gouveia, F.C, Barreto, M. Preservação da privacidade no enfrentamento da COVID-19: Dados pessoais e a pandemia global. Cien Saude Colet [periódico na internet] (2020/Abr). Está disponível em: http://www.cienciaesaudecoletiva.com.br/artigos/preservacao-da-privacidade-no-enfrentamento-da-covid19-dados-pessoais-e-a-pandemia-global/17570?id=17570.

compartilhamento ético de dados da COVID-19[19], e as recomendações sobre compartilhamento desses dados elaboradas pela *Research Data Alliance*[20] por encomenda da Comissão Europeia.

POTENCIAIS APLICAÇÕES DE DADOS ABERTOS EM ALGUMAS OUTRAS ÁREAS CIENTÍFICAS

- Medicina e Saúde Pública

O compartilhamento de dados na medicina tem sido muito discutido e promovido, principalmente quando se considera a incorporação da genômica na prática clínica e a informatização dos resultados de exames diversos. Para que dados genômicos sejam adequadamente aplicados às práticas médicas correntes, é essencial que sejam disponibilizados para a comunidade médica segundo padrões definidos. No entanto, para isso, é fundamental que os preceitos éticos intrinsicamente relacionados à disponibilização e ao compartilhamento de dados de grupos de indivíduos (voluntários sadios e pacientes) sejam respeitados. Projetos internacionais têm sido estabelecidos com o objetivo de propor parâmetros para o compartilhamento ético, responsável e eficiente de dados genômicos e outros relacionados à saúde humana. Podem ser citados como exemplo o *Human Variome Project*[21] e a *Global Alliance for Genomics and Health*[22], entre outros.

Estas iniciativas têm produzido documentação que serve como referência para projetos nacionais e locais e, entre tais documentos, é importante citar o *framework* proposto[23] para compartilhamento responsável de dados genômicos e relacionados à saúde, que serve como referência para as discussões éticas sobre o assunto. Com a disseminação cada vez maior da medicina genômica e de sua crescente aplicação na prática médica, a criação de repositórios públicos deve crescer mundialmente e iniciativas nacionais serão mais frequentes, como a Iniciativa Brasileira de Medicina de Precisão[24] (BIPMed, na sigla em inglês), que centraliza e disponibiliza os dados médicos e genômicos da população brasileira.

Sistemas de saúde geram quantidades cada vez maiores de dados sobre os indivíduos que utilizam os seus serviços. Em países que têm estes serviços estruturados em sistemas nacionais de saúde, os dados correspondentes vêm sendo organizados e mantidos em sistemas integrados, demonstrando de forma crescente a sua utilidade para a produção de conhecimento. Algumas nações, como o Reino Unido, têm incluído a utilização destas bases de dados em suas estratégias nacionais de pesquisa. Isso tem acontecido através do financiamento de grandes programas para a criação de infraestruturas e para o desenvolvimento de métodos que permitam a utilização deste conjunto complexo e gigantesco de dados, com vistas a transformá-los em conhecimento. No Brasil, algumas iniciativas específicas vêm sendo desenvolvidas, destacando-se o Centro de Integração de Dados e Conhecimentos para Saúde[25] (Cidacs) da Fundação Oswaldo Cruz (Fiocruz), em Salvador, Bahia, que dispõe de infraestrutura e governança para vincular e processar dados pessoais identificados provenientes de grandes bases de dados nacionais, como Cadastro Único, mortalidade, nascimentos, entre outros[26].

- Ciências Agrárias

Ciências agrárias e cadeias produtivas de produção de alimentos vêm, igualmente, lucrando muito com a abertura de dados, como com a permissão de análise das bases disponibilizadas pela Organização das Nações Unidas para Alimentação e Agricultura[27] (FAO, na sigla em inglês). Acessíveis por todos (órgãos governamentais, pesquisadores,

[19] African Academy of Sciences. On COVID-19: Ethics, Governance and Community engagement in times of crises. https://www.aasciences.africa/sites/default/files/2020-04/COVID-19%20Ethics%2C%20Governance%20and%20Community%20engagement%20in%20times%20of%20crises%2020April2020.pdf.
[20] Research Data Alliance. RDA COVID-19 Guidelines and Recommendations. https://www.rd-alliance.org/group/rda-covid19-rda-covid19-omics-rda-covid19-epidemiology-rda-covid19-clinical-rda-covid19-0.
[21] Human Variome Project. https://www.humanvariomeproject.org/.
[22] Global Alliance for Genomics and Health. https://www.ga4gh.org/.
[23] Global Alliance for Genomics and Health. Framework para Compartilhamento Responsável de Dados Genômicos e Relacionados à Saúde. https://www.ga4gh.org/wp-content/uploads/Framework-Portuguese-translation.pdf.
[24] Brazilian Initiative on Precision Medicine. https://bipmed.org/.
[25] Centro de Integração de Dados e Conhecimentos para Saúde. https://cidacs.bahia.fiocruz.br/.
[26] Barreto, ML, Ichihara, MY, Almeida, BA, Barreto, ME, Cabral, L, Fiaccone, RL, Carreiro, RP1, Teles, CAS1, Pitta, R, Penna, GO, Barral-Netto, M, Ali, MS, Barbosa, G1, Denaxas, S9, Rodrigues LC, and Smeeth L. The Center for Data and Knowledge Integration for Health (CIDACS): Linking Health and Social Data in Brazil. International Journal of Population Data Science (2019) 4:2:04 Doi 10.23889/ijpds.v4i2.1140.
[27] Food and Agriculture Organization of the United Nations. http://www.fao.org/home/en/.

empresas e cidadãos), permitem tomadas de decisão em vários níveis, desde questões de política econômica até o planejamento de qual produto agrícola cultivar em determinada época ou região. Combinadas a dados genômicos abertos, tais bases permitem inovação e maior eficiência na produção e comercialização de bens agrícolas. Bases públicas de dados de imagens de satélite (por exemplo, disponibilizadas pelo Instituto Nacional de Pesquisas Espaciais-INPE e pela Empresa Brasileira de Pesquisa Agropecuária-Embrapa) ou de dados agrometeorológicos (como do Instituto Nacional de Meteorologia-INMET) contribuem para auxiliar o trabalho de pequenos agricultores, inclusive pedidos de financiamento para o plantio. Existem, também, centros que preservam coleções biológicas de interesse agronômico, como o CENARGEN[28], o CIMMYT[29] e o IRRI[30], todos com acesso aberto a suas coleções de sementes e variedades agrícolas ancestrais e modernas.

- Física e Astronomia

A física e a astronomia também são campos da ciência fortemente dirigidos pelos dados, o que é particularmente sentido na física de altas energias, um campo em que, talvez, este problema tenha historicamente se colocado primeiro. Não por acaso, a internet teve sua origem no contexto das soluções de dados desenvolvidas para a área[31]. Sem as soluções de computação e tecnologia da informação, desenvolvidas por instituições como o CERN[32] e abertas para toda a comunidade científica (outro bom exemplo é a plataforma de dados ROOT,[33] desenvolvida em 1994 como ferramenta pioneira para ciências de dados), o progresso hoje atingido neste campo de pesquisa seria impensável. O CERN mantém, também, a principal plataforma para abertura de dados na área, o portal CERN OpenData[34].

No caso específico da astronomia, as observações do céu são conduzidas em todas as bandas do espectro eletromagnético, das ondas de rádio aos raios-gama, e, mais recentemente, no domínio dos chamados multi-mensageiros, que incluem os raios-cósmicos, os neutrinos e as ondas gravitacionais. O volume de dados típicos das varreduras do céu realizadas há uma década é equivalente a apenas uma noite de observação dos principais instrumentos disponíveis atualmente. E, em menos de dez anos, a produção de dados por grandes instrumentos astronômicos crescerá de duas a três ordens de magnitude. Pela natureza transiente de muitas das fontes astrofísicas, os levantamentos astronômicos devem ainda se preocupar com a dimensão temporal das observações, que implicam monitoramentos periódicos do céu, dando-nos uma visão multidimensional dos catálogos astronômicos modernos. Tudo isso implica numa gestão dos dados que depende fortemente da abertura, integração e interoperabilidade dos mesmos. Grande parte dos esforços da área nesta direção é conduzida pela International Virtual Observatory Alliance (IVOA)[35], uma federação global de instituições e pesquisadores responsáveis pela definição dos protocolos de dados e serviços de dados em astronomia que são adotados mundialmente.

- Ciências Humanas e Sociais

Também as ciências humanas e sociais trazem vários exemplos interessantes e pioneiros de dados abertos. Um dos primeiros e principais sistemas abertos de gestão de dados científicos no mundo, o *Dataverse*[36], criado na década de 90, teve sua origem no projeto *Virtual Data Center* (VDC) da Universidade Harvard, motivado pelas necessidades de compartilhamento de dados em ciências sociais quantitativas. Até hoje, este software é mantido e distribuído pelo *Institute for Quantitative Social Science* (IQSS), também de Harvard. Adotado por centenas de instituições de pesquisa e centros de dados para ciência aberta, este software é usado atualmente para armazenamento e gerenciamento de dados em todos os domínios do conhecimento. A própria Universidade Harvard hospeda uma das maiores coleções de dados de pesquisa em ciências sociais do mundo, muitos dos quais abertos e acessíveis através desse software. Exemplos importantes de grandes coleções na área incluem o repositório holandês de dados arqueológicos (com fotos, mapas, vídeos e milhares de relatórios de escavações, que podem ser acessados por meio do centro de dados

[28] Embrapa Recursos Genéticos e Biotecnologia. https://www.embrapa.br/recursos-geneticos-e-biotecnologia.
[29] International Maize and Wheat Improvement Center. https://www.cimmyt.org/.
[30] International Rice Research Institute. https://www.irri.org/.
[31] European Organization for Nuclear Research (CERN). The birth of the Web. https://home.cern/science/computing/birth-web.
[32] European Organization for Nuclear Research (CERN). Open source for open science. https://home.cern/science/computing/open-source-open-science.
[33] ROOT Data Analysis Framework. https://root.cern/.
[34] European Organization for Nuclear Research (CERN). OpenData Portal. http://opendata.cern.ch/.
[35] International Virtual Observatory Alliance. http://ivoa.net/.
[36] Harvard Dataverse. https://dataverse.harvard.edu/.

de pesquisa daquele país[37]) e um diretório de mais de 200 coleções alemãs de dados em artes e ciências humanas e sociais disponibilizado pelo projeto europeu *Digital Research Infrastructure for the Arts and Humanities*[38] (DARIAH).

IMPACTOS SOCIAIS E APLICAÇÕES DE DADOS ABERTOS PARA EDUCAÇÃO E CAPACITAÇÃO (CAPACITY BUILDING), COMO RESPOSTA AOS OBJETIVOS DE DESENVOLVIMENTO SUSTENTÁVEL DA ONU

Os benefícios da abertura de dados não são apenas de natureza estritamente científica. A difusão de dados científicos em larga escala pode também potencializar a capacidade da ciência como um instrumento para o desenvolvimento sustentável, habilitando a distribuição democrática dos seus avanços técnicos, culturais e sociais por todo o planeta e em múltiplos âmbitos.

A ciência assume papel como força cultural e torna evidente seu potencial e responsabilidade para com os desafios da sociedade do século XXI. Tais desafios foram recentemente formulados no grupo de 17 Objetivos de Desenvolvimento Sustentável[39] (ODS), com os quais a comunidade internacional, incluindo o Brasil, se comprometeu em 2015, quando da adoção, pela Organização das Nações Unidas (ONU), da Agenda 2030 para o Desenvolvimento Sustentável[40].

A abertura e a disseminação de dados científicos em larga escala guardam em si potencial relevante para que se alcancem os ODS. Isso se faz notar à medida que se reconhece que os dados (e a informação deles extraída) formam o pilar sobre o qual novos conhecimentos e soluções podem ser construídos para os desafios da sociedade. Por outro lado, a abertura e disseminação são a condição *sine qua non* para que os benefícios advindos de suas aplicações sejam democraticamente distribuídos entre os povos e alcancem os diferentes interesses e setores da sociedade, permitindo um desenvolvimento equânime e equilibrado entre as nações. Pode-se dizer, de fato, que os dados resultantes da aplicação do conhecimento humano e da atividade científica constituem verdadeiro patrimônio cultural de toda a humanidade e, como tal, devem ser colocados à disposição e ao serviço de todos. Ressalte-se igualmente que uma parcela importante dos grandes volumes de dados é produzida por meio de financiamento público, fato que naturalmente demanda sua ampla acessibilidade.

São muitos os modos por meio dos quais dados abertos podem contribuir aos objetivos que constituem a Agenda 2030 da ONU. Duas modalidades, porém, são fundamento comum para muitos destes objetivos e se encontram no potencial educativo e de capacitação (*capacity building*) que advêm da abertura e amplo acesso aos dados. É claro o quanto a educação científica pode se beneficiar do uso de dados abertos. A capacitação (*capacity building*), por sua vez, é um mecanismo cada vez mais relevante que pode permitir a regiões menos desenvolvidas do globo usufruir e se beneficiar do progresso, dando-lhes a autonomia necessária para que se tornem protagonistas e usuários ativos, e não apenas beneficiários passivos, das novas tecnologias e conhecimentos.

IMPACTO DA UTILIZAÇÃO MASSIVA DE DADOS NA FORMA DE PRODUZIR CONHECIMENTOS E NOS MÉTODOS DE DIFERENTES CAMPOS CIENTÍFICOS

A utilização massiva de dados cria possibilidades de mudanças na forma de produção de conhecimento, dentro e fora da academia, ao possibilitar novas e mais eficientes maneiras de se planejar, conduzir, institucionalizar, disseminar e avaliar a pesquisa. Espera-se que a capacidade de vinculação e interseção de conjuntos de dados provenientes de fontes diferentes aumente a precisão, o poder preditivo e a generabilidade das descobertas científicas e ajude os pesquisadores a identificar futuras direções de investigação. A disponibilidade de grandes volumes de dados fornece um incentivo à busca por novos procedimentos e ferramentas computacionais para obtenção, armazenamento, organização e análise desses dados, podendo trazer aperfeiçoamentos na forma de produzir e de dar maior transparência a todo o processo de criação do conhecimento científico.

[37] Data Archiving and Networked Services. https://easy.dans.knaw.nl/ui/home.
[38] Digital Research Infrastructure for the Arts and Humanities. Collection Registry. https://www.dariah.eu/tools-services/tools-and-services/tools/collection-registry/.
[39] Organização das Nações Unidas. 17 objetivos para transformar o nosso mundo. https://nacoesunidas.org/pos2015/.
[40] Organização das Nações Unidas. Agenda 2030. https://nacoesunidas.org/pos2015/agenda2030/.

Em cada campo científico existe o grande desafio de se transformar quantidades cada vez maiores de dados em conhecimento. Isto implica na adequação de métodos científicos tradicionais em métodos alternativos que sejam capazes de analisar tal volume. Bons exemplos são a genética e a genômica. A genética é uma ciência bem estabelecida que estuda a transmissão das características hereditárias de genes específicos. Também é um dos alicerces para a genômica, que resulta da capacidade de sequenciamento da molécula de DNA e do desenvolvimento da biologia computacional. A genômica irá permitir estudar não mais genes isolados, como na genética, mas toda a estrutura genética de um ser vivo, de uma população ou de todas as espécies – como na inciativa internacional *Earth Biogenome Project*[41], lançada em 2018.

Porém, além das questões metodológicas e operacionais nas diversas áreas da ciência, devemos também destacar os debates de ordem epistemológica que a emergência do *big data* tem suscitado. Desde pelo menos o século XIX, a ciência se desenvolve utilizando uma estratégia de produzir e testar hipóteses. Entretanto, o advento dos grandes volumes de dados, e algumas metodologias a eles associadas, alavancou a denominada ciência *data-driven*, uma nova forma de indutivismo, em lugar da abordagem dedutiva ou *hypothesis-driven*[42]. Estes debates estão se desdobrando em cada área científica e certamente vêm tendo impactos na forma com que cada uma destas áreas aborda os seus problemas.

CONSIDERAÇÕES SOBRE OS BENEFÍCIOS E POSSÍVEIS RISCOS ADVINDOS DA ABERTURA DE DADOS E DA IMPLANTAÇÃO DE POLÍTICAS PÚBLICAS DE GESTÃO DE DADOS, INCLUINDO QUESTÕES LEGAIS E ÉTICAS RELATIVAS À SEGURANÇA E À PRIVACIDADE DOS DADOS PESSOAIS

Uma parte importante dos dados existentes são relacionados a pessoas e incluem: registros diversos que cada indivíduo faz durante a vida (dados administrativos), exames de saúde, em mídias sociais, em diferentes recursos da *web*, etc. São, em geral, gerados com objetivos diversos, porém, como muitos são guardados em meio digital, tornam-se importantes fontes de informação sobre as pessoas e podem ter usos diferentes dos objetivos originais de sua coleta, sejam estes lícitos ou não. Consideremos, por exemplo, os escândalos recentes envolvendo o uso de dados de mídias sociais para direcionamento de posições políticas em diversas situações e países.

Na pesquisa, utilizam-se dados pessoais nas mais diferentes áreas do conhecimento, como a medicina, a saúde pública e a epidemiologia, as ciências sociais, a economia, a história, dentre outras. Neste sentido, uma das principais questões é: como proteger essa imensa massa de dados existentes em meio digital, preservando os direitos dos indivíduos? Em especial, a privacidade tem sido o desafio para muitas sociedades. Em 2017, um grande avanço foi dado pela União Europeia, cujo Parlamento, após muitos anos de discussão, aprovou a Regulamentação Geral de Proteção de Dados (GDPR, na sigla em inglês), que passou a vigorar em 2018. Seguindo estes passos, o Congresso Nacional aprovou, também em 2018, a chamada Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709/18), que dispõe sobre a proteção de dados pessoais no Brasil e que está prevista para entrar em vigor em agosto de 2020.

Por fim, a segurança e proteção de dados pessoais representa grande esforço de pesquisa em diversos campos científicos, seja criando algoritmos criptográficos mais robustos, seja implementando maneiras de processamento de dados pessoais sem riscos para a identidade dos envolvidos. Neste contexto, tem-se o tema da anonimização, ou seja, quando dados pessoais sofrem processo de retirada de todos os possíveis identificadores. O problema é a possibilidade de reidentificação, tema que tem gerado crescente interesse, visto que métodos computacionais comumente utilizados, como o aprendizado de máquina, aumentam a probabilidade de reidentificação de indivíduos com base em variáveis não identificadoras mantidas nas bases anonimizadas[43].

[41] Earth Biogenome Project. https://www.earthbiogenome.org/.
[42] Kell DB(1), Oliver SG. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. Bioessays. 2004 Jan;26(1):99-105.
[43] Doneda D, Almeida BA, Barreto ML. Uso e proteção de dados pessoais na pesquisa científica. Revista Direito Público 2019; 16(90):179-194.

DIAGNÓSTICO DA SITUAÇÃO DE ABERTURA E GESTÃO DE DADOS NO BRASIL ATÉ O PRESENTE MOMENTO

Já existem várias iniciativas brasileiras para criação de repositórios de dados de pesquisa voltados a dados abertos. Cada um deles segue padrões diferentes de divulgação dos dados e usa sistemas diferentes de armazenamento e gerenciamento desses dados. A maioria está sendo criada por iniciativas individuais, em que grupos de pesquisa publicam, por meio de um site de projeto, dados ou metadados da pesquisa realizada pelo grupo. Desta forma, a preocupação com a interoperabilidade, ou uso de padrões mundiais de metadados, é crucial. Os parágrafos a seguir ilustram algumas das iniciativas institucionais no Brasil.

A maior parte desses projetos está ligada à área ciências da vida e da terra, como o Cidacs-Fiocruz[44] ou a *Brazilian Initiative for Precision Medicine* (BIPMed), sediada na Unicamp. Enquanto o Cidacs centraliza várias plataformas nacionais de dados sociais e de saúde, a BIPMed é o nó latinoamericano de uma rede mundial de dados genômicos. Já o Sistema de Informação sobre a Biodiversidade Brasileira (SIBBr), sediado no LNCC, é uma plataforma online que disponibiliza acervos de dados de biodiversidade, fornecidos por várias instituições de pesquisa e universidades. Destaca-se, também, o INPE na área de monitoramento da composição da superfície do território nacional, seu uso e as informações climáticas dependentes de componentes oceanográficos e atmosféricos.

Iniciativas institucionais para criação de redes de dados abertos são recentes, por exemplo, a coordenada pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) em parceria com a Rede Nacional de Ensino e Pesquisa (RNP), que visa criar uma rede brasileira de dados de pesquisa. A Embrapa também já deu início à sua rede de repositórios, que deverá possibilitar a criação e conexão de repositórios em todas as suas unidades de pesquisa.

Destaca-se, igualmente, a rede de repositórios de dados de pesquisa abertos do consórcio criado, em 2017, pelas seis universidades públicas do Estado de São Paulo (Universidade de São Paulo-USP, Universidade Estadual Paulista-Unesp, Universidade Estadual de Campinas-Unicamp, Universidade Federal de São Paulo-Unifesp, Universidade Federal do ABC-UFABC e Universidade Federal de São Carlos-UFSCar), pelo Instituto Tecnológico de Aeronáutica (ITA) e pelo CNPTIA-Embrapa. Cada participante desta rede paulista gerencia seus repositórios de forma independente, mas todos são acessíveis por meio de um portal de metadados único, que segue padrões propostos pela RDA. Esta rede foi criada para atender às diretivas da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) para Ciência Aberta. Enquanto a rede da Embrapa visa disponibilizar os dados associados à pesquisa da empresa, o objetivo das demais é tornar visíveis quaisquer dados de pesquisa produzidos pelas instituições participantes.

SUGESTÕES PARA A IMPLANTAÇÃO DE UM MODELO DE ABERTURA DE DADOS E POLÍTICAS DE GESTÃO DE DADOS PARA O BRASIL

Toda política de gestão de dados deve considerar, no mínimo, três aspectos: a governança, a infraestrutura computacional (software e hardware) necessária para o armazenamento e disponibilização dos dados, e o treinamento e formação de pessoal que garantam a execução da política. O fator treinamento e formação é crucial para o sucesso da política e requer investimentos continuados. Pessoal, neste contexto, abrange tanto os pesquisadores, quanto os profissionais que garantirão o funcionamento estável e duradouro dos repositórios. Já existe um farto material de treinamento disponível na *web*, destacando-se o produzido pelo Reino Unido, Austrália e Países Baixos. A RDA vem disponibilizando material adicional de recomendações e padrões que estão sendo adotados em todo o mundo. Uma tendência mundial importante é a de possibilitar a citação dos dados (independentemente de estarem ou não ligados a artigos científicos). Para isso, existe hoje a possibilidade de ser criado um identificador único para cada arquivo, reconhecido mundialmente, à semelhança do DOI para artigos científicos. As recomendações FAIR, mencionadas anteriormente, exigem que dados tenham tal identificador. As duas principais organizações mundiais de geração de DOI para dados são o *DataCite* e o *Handler Registry* (parte do Handler.org). Em ambos os casos, cada identificador gerado tem um custo, assim como os DOI de artigos em periódicos.

Desta forma, é importante ressaltar que a abertura de dados envolve custos inerentes, como para geração e manutenção de identificadores, aquisição e manutenção de hardware e software, manutenção e atualização de

[44] Barreto, ML, Ichihara, MY, Almeida, BA, Barreto, ME, Cabral, L, Fiaccone, RL, Carreiro, RP1, Teles, CAS1, Pitta, R, Penna, GO, Barral-Netto, M, Ali, MS, Barbosa, G1, Denaxas, S9, Rodrigues LC, and Smeeth L. The Center for Data and Knowledge Integration for Health (CIDACS): Linking Health and Social Data in Brazil. International Journal of Population Data Science (2019) 4:2:04 Doi 10.23889/ijpds.v4i2.1140.

infraestrutura física, e treinamento e formação de pessoal. A escolha do modelo e das políticas a serem adotadas deve levar estes fatores em consideração.

Há, ainda, inúmeros desafios computacionais intrínsecos ao cenário mundial de pesquisa: multinacional, multi-institucional, multidisciplinar, dirigida por dados. Este cenário implica em múltiplos desafios, tais como: novos protocolos de coordenação, seja para a coleta de dados (por exemplo, no caso da astronomia), seja para o seu processamento; uma grande capacidade de processamento (por vezes em tempo real), frequentemente aliada à transmissão de volumes massivos de dados para centros espalhados por todo o mundo; e a criação e gestão de grandes centros para armazenamento e acesso aos dados, sobre os quais pesam fortes demandas de acessibilidade e interoperabilidade. Soma-se a estes a necessidade crescente do uso de tecnologias e métodos baseados em inteligência artificial para análise dos dados, os quais, cada vez mais, se tornam por demais volumosos para que sejam analisados, inspecionados diretamente ou supervisionados por pessoas.

Como última recomendação, e à semelhança de muitas universidades importantes em todo o mundo, deveríamos exigir, como parte da formação dos nossos doutorandos, treinamento básico em princípios de abertura de dados. Um ponto elementar para dar início a este treinamento seria, por exemplo, ensiná-los a preparar planos de gestão de dados, como parte integrante de um projeto de pesquisa. Para isto, várias universidades brasileiras, como USP, UFABC, Unesp e Unicamp, já criaram modelos básicos, em português, com um guia de perguntas e respostas que auxilia pesquisadores a pensar a gestão de seus dados de pesquisa. Tais modelos estão disponíveis no site DMPTool[45], que contribui para que cientistas criem planos de gerenciamento de dados.

CONCLUSÕES

A ciência é uma das atividades humanas capazes de unir os povos em busca de objetivos comuns, cruzando barreiras de naturezas diversas para integrar culturas e inspirar a inovação e o desenvolvimento. Hoje, mais do que nunca, estas características globais da atividade científica se confundem com sua própria dinâmica e *modus operandi*. A revolução imposta pelas tecnologias de informação e comunicação vem promovendo mudanças profundas e rápidas na dinâmica e no processo científico, renovando antigos e criando novos desafios, alguns dos quais cobrarão respostas imediatas. O papel e o impacto que a ciência terá na sociedade de amanhã dependerá das decisões tomadas no presente quanto à gestão e ao armazenamento das informações coletadas.

Ao buscar orientar este rápido progresso técnico para o desenvolvimento humano, há que se atentar especialmente para os riscos e desequilíbrios que podem advir de mudanças rápidas sem reflexão. É evidente, por exemplo, que o progresso, se não orientado por meio da cooperação internacional e da busca de objetivos comuns, pode se tornar responsável por um crescimento da desigualdade entre as nações[46], como de fato se observou em muitas ocasiões, inclusive durante o século XX. Do ponto de vista cultural, a inundação do cenário científico, e da sociedade em geral, por grandes volumes de dados e informação – de forma rápida e, por vezes, descontrolada – pede reflexões e soluções que permitam salvaguardar uma epistemologia adequada do conhecimento na era da informação. Paradoxalmente, o excesso de informação, se mal gerido, pode se opor ao avanço do conhecimento, sendo a razão crítica, a capacidade de síntese e o contexto à base do conhecimento substituídos pela fatualidade crua e superficial dos dados[47]. Neste sentido, a ciência pode dar uma contribuição cultural relevante à sociedade da era digital, por meio de sua própria orientação positiva, servindo de modelo cultural a outros setores.

EQUIPE TÉCNICA DA ABC:

Marcos Cortesão Barnsley Scheuenstuhl
Secretário Executivo de Relações Internacionais

Vitor Vieira de Oliveira Souza
Assessor de Projetos

[45] DMPTool. https://dmptool.org/.
[46] Cf. Papa Paulo VI. Discurso de Abertura do Encontro Unispace da ONU. 1968.
[47] Cf. Henry Kissinger. Ordem Mundial. 2013.