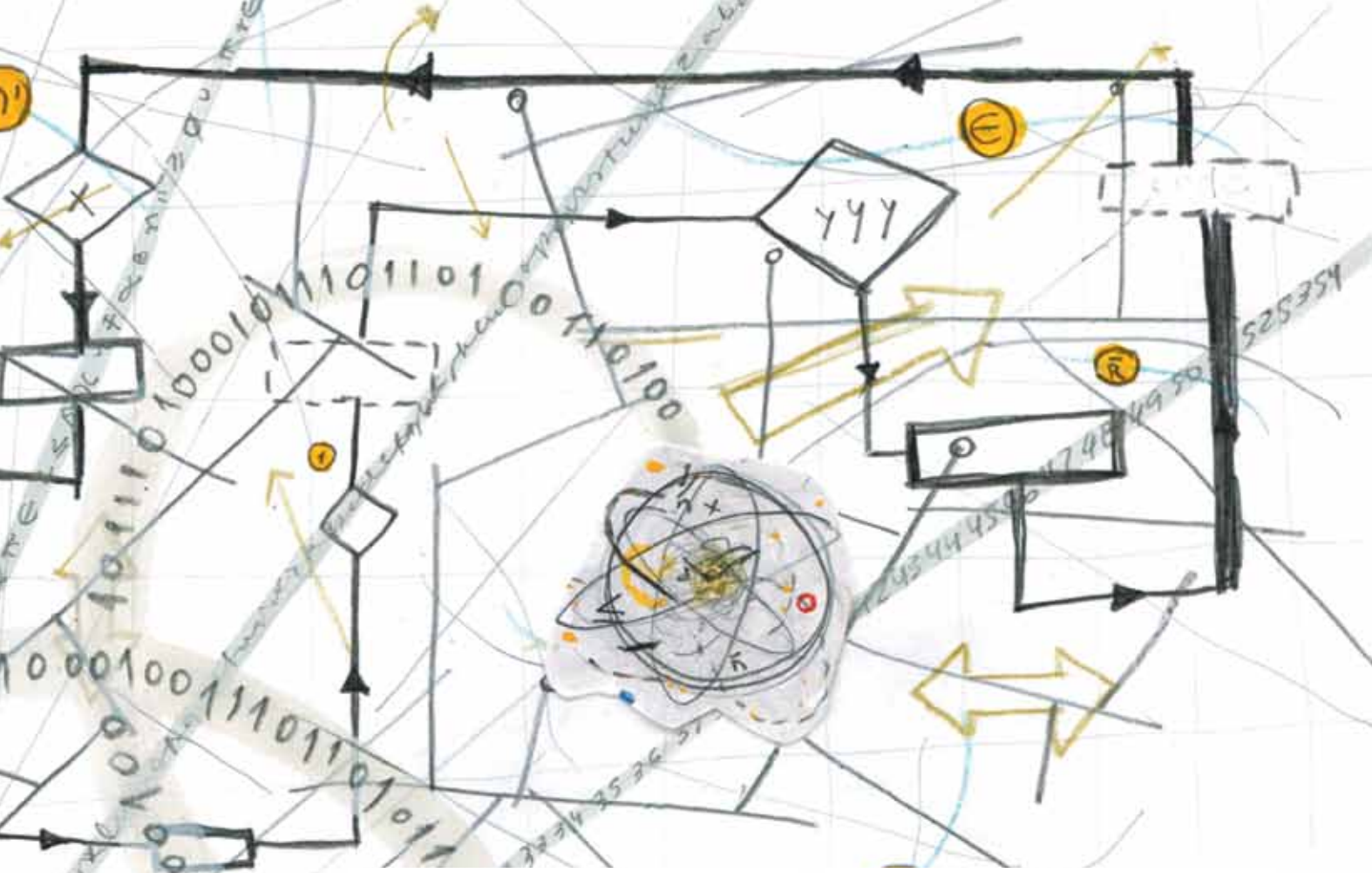
Ciência transparente

Cada vez mais os pesquisadores são solicitados a armazenar os dados primários de seus estudos em repositórios públicos

Fabrício Marques

Certas transformações ocorrem de forma tão gradativa que só é possível perceber o seu alcance em um momento adiantado do processo. Um desses momentos que parecem cristalizar mudanças aconteceu em março, com a decisão das sete revistas científicas *PLoS* (sigla para Public Library of Science) de condicionar a aceitação de novos artigos à divulgação por seus autores, em repositórios públicos, dos chamados dados de pesquisa, aquela massa de informações primárias que, uma vez analisada e interpretada, serve de base para as conclusões do *paper*. A nova regra da *PLoS*, que se enquadra em uma ampla mobilização de agências de fomento, cientistas e universidades para dar mais transparência à publicação de resultados de pesquisa, não chega a ser propriamente uma novidade. A maioria dos periódicos já recomenda aos autores que disponibilizem os dados e esta recomendação virou exigência há tempos em revistas de genética e de bioinformática, cujos estudos geram gigantescos volumes de informação sobre sequências de DNA e proteínas. Em 2013, o

ILUSTRAÇÕES BEL FALLEIROS



Escritório de Política Científica e Tecnológica do governo norte-americano enviou um memorando às principais agências de fomento estabelecendo o acesso aberto a resultados de pesquisa financiada com dinheiro público, incluindo a oferta dos dados primários em repositórios, salvo restrições de confidencialidade e privacidade pessoal – mas faltou estipular prazos para a ideia sair do papel.

A decisão da *PLoS* parece criar um ponto de inflexão nessa tendência. "Nosso ponto de vista é simples. Garantir o acesso aos dados subjacentes deve ser parte intrínseca do processo de publicação científica", justificou Theodora Bloom, diretora editorial da *PLoS Biology*, da *PLoS Computational Biology* e da *PLoS Genetics*. Com mais de 30 mil artigos publicados no ano passado, as revistas *PLoS* foram criadas ao longo da década de 2000 por uma instituição sem fins lucrativos seguindo um modelo inovador. Publicam artigos apenas *on-line* e em acesso aberto – ou seja, podem ser consultadas por qualquer pessoa, pela internet, sem cobrar por isso –, mas, graças a um corpo de revisores de primeira linha, alcançaram um fator de impacto comparável aos de publicações tradicionais. A *PLoS Medicine*, por exemplo, teve um fator de impacto de 15,2 em 2012 – significa dizer que, em média, cada um de seus artigos publicados entre 2010 e 2011 teve 15,2 citações em periódicos indexados em 2012. A concorrente *Nature Medicine*, do grupo Nature, teve no mesmo período fator de impacto de 24,3. "Como a *PLoS* é uma referência internacional, a sua decisão vai contribuir para disseminar a ideia do depósito dos dados de pesquisa e criar uma demanda adicional para repositórios e também modelos que financiem essa demanda", diz Abel Packer, coordenador da biblioteca SciELO Brasil, um programa especial da FAPESP criado em 1998 que reúne quase 300 publicações científicas do Brasil de acesso aberto.

As novas regras da *PLoS* geraram dúvidas e algum alvoroço. Dez dias após o início da implementação, seus editores pediram desculpas por pontos ambíguos e esclareceram que nada mudou em relação à natureza dos dados que precisam estar descritos nos artigos – a única preocupação nova é apontar em que banco ou repositório podem ser encontrados dados primários (os arquivos do próprio pesquisador não são uma opção), caso os revisores do artigo ou outros pesquisadores interessados no assunto precisem avaliá-los. Dados primários são entendidos pela *PLoS* como aqueles que abastecem tabelas e análises estatísticas publicadas no artigo e são indispensáveis para que outros pesquisadores consigam reproduzir os mesmos achados. Dados protegidos por razões de segurança ou de privacidade não estão incluídos na exigência.

As mudanças despertaram reações de quem enxerga nas regras um novo ônus para os pes-

É preciso conhecer a origem do dado e saber como ele foi produzido para reutilizá-lo com propriedade

quisadores. O geneticista David Crotty, editor do programa de publicação de revistas científicas da Oxford University Press, escreveu em seu *blog*, no portal The Scholarly Kitchen, que a mudança poderá reduzir o número de artigos submetidos às revistas *PLoS*. "Se a publicação em uma revista *PLoS* exige que você faça semanas de trabalho adicional para organizar seus dados em uma forma reutilizável ou pelo menos reconhecível, sem falar no custo de hospedar os dados e no esforço para encontrar um repositório adequado, então por que não publicar o artigo em um jornal diferente e eliminar os custos e os gastos de tempo?", indagou Crotty. Não se trata, observa Abel Packer, de fazer trabalho adicional, pois a mudança de paradigma é bem mais profunda. "Estamos falando de novas práticas, nas quais os dados já são organizados durante a realização da pesquisa, de modo que possam ser disponibilizados nos repositórios e sejam inteligíveis e reutilizáveis para outros usuários", afirma.

O armazenamento de dados científicos em repositórios e sua reutilização é uma das preocupações do recém-lançado Programa FAPESP de Pesquisa em eScience, expressão que resume o desafio de pesquisa para organizar, classificar e garantir acesso ao gigantesco volume de dados gerados continuamente em todos os campos de pesquisa, a fim de extrair novos conhecimentos e fazer análises abrangentes e originais. "Não se deve imaginar que basta o pesquisador fazer o *download* de um dado contido num repositório para utilizá-lo num novo estudo", diz Claudia Bauzer Medeiros, professora do Instituto de Computação da Universidade Estadual de Campinas (Unicamp) e coordenadora-adjunta de programas especiais eScience da FAPESP. "O compartilhamento de dados para reutilização ou reprodução de experimentos exige conhecer sua origem e entender como foram produzidos, associando à informação métodos, algoritmos ou técnicas adotados, e ainda ter acesso ao *software* necessário para processá-los, o que torna o processo bastante complexo. Sem isso, pode não ser possível reproduzir o experimento original ou reutilizar o dado em uma outra pesquisa", afirma a professora, que lembra que a primeira chamada de propostas do programa eScience está aberta até 28 de abril. Um dos alvos do programa é a pesquisa relacionada a repositórios de dados. "Esperamos que os projetos apresentados, que deverão envolver pesquisa conjunta em computação e em outras áreas do conhecimento, contribuam na criação de metodologias e modelos de dados para criar repositórios, e em formas mais eficientes de descrever o conteúdo e estruturá-lo, para poder recuperá-lo", explica. "Não é suficiente, por exemplo, descrever dados por palavra-chave. Se um pesquisador quiser reutilizar aquele dado para um propósito diferente dificilmente o encontrará por palavra-chave", afirma.

Este tipo de esforço de pesquisa inspirou o Nature Publishing Group, que edita a revista *Nature*, a lançar uma nova revista a partir de maio. Trata-se da *Scientific Data*, publicação *on-line* em acesso aberto voltada para descrever não novos achados científicos, mas sim os *datasets* (conjuntos de dados) de pesquisas consideradas valiosas. O objetivo é promover a documentação, intercâmbio e reutilização dos dados que sustentam as pesquisas, em modo aberto, para acelerar o passo das descobertas científicas. Para atingir este objetivo, os editores da revista introduziram um metadado (dados a respeito de

1 Banco de dados da Organização Europeia para Pesquisa Nuclear (Cern), em Genebra

2 Sala de controle de satélite meteorológico operado pela Agência Espacial Europeia e a Organização Europeia para Exploração de Satélites Meteorológicos, em Darmstadt, Alemanha

3 Arquivo de imagens de sensoriamento remoto do Serviço Geológico dos Estados Unidos

FOTOS **1** CERN **2** YSANGKOK / WIKIMEDIA **3** USGS

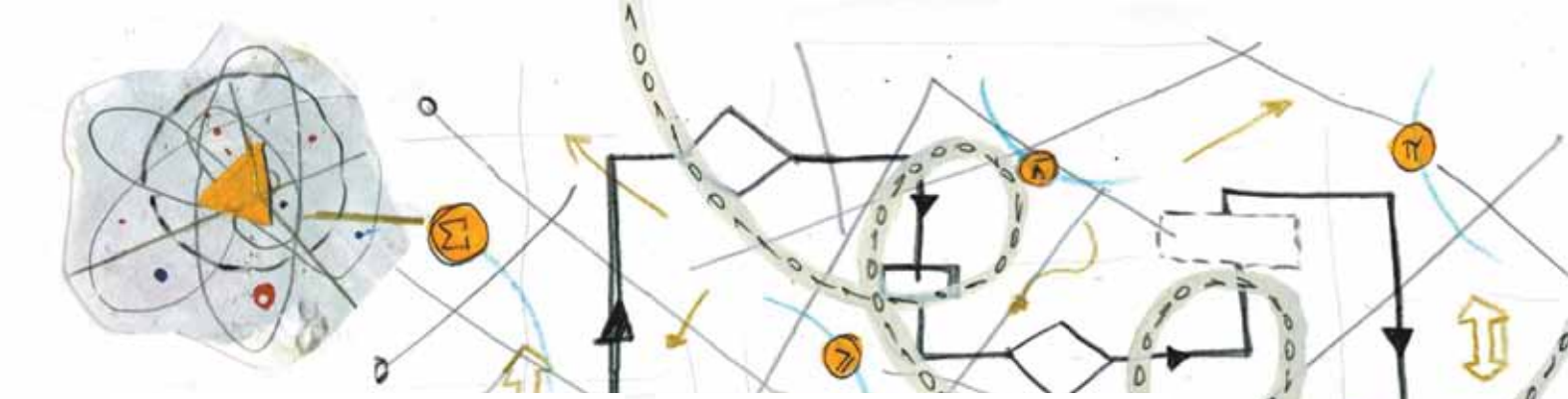
outros dados) chamado *data descriptor*. Na revista, estes metadados devem fornecer descrições detalhadas de conjuntos de dados em ciências da vida, biomédicas e ambientais, com foco exclusivo em como eles foram produzidos, por quem e como podem ser reutilizados por pesquisadores independentes. "Os metadados dotam os dados científicos com identidade e documentação padronizada e com capacidade de serem acessíveis nas buscas, interoperáveis com diferentes sistemas na *web*, reutilizáveis em outras pesquisas e citáveis", diz Abel Packer.

Descobertas científicas que acabam não sendo validadas assombram cientistas, publicações e empresas

O princípio da reprodutibilidade das pesquisas é o dínamo mais importante para a criação dos repositórios de dados de pesquisa. Uma quantidade não desprezível de descobertas científicas acaba não sendo confirmada após sua publicação, por problemas que incluem erros e fraudes, mas que também se estendem a falsos resultados positivos ou negativos obtidos de boa fé. O problema assombra pesquisadores e revistas científicas, obrigados a cancelar a publicação de trabalhos cujos resultados soavam promissores, e tornou-se um pesadelo para empresas farmacêuticas e de biotecnologia. Segundo reportagem recente da revista *The Economist*, pesquisadores de uma empresa de biotecnologia, a Amgen, constataram que era possível reproduzir apenas seis entre 53 estudos considerados "marcos" na pesquisa do câncer.

"Além de verificar a validade dos resultados, o acesso e reutilização de dados também permitem fazer novas pesquisas e estudos comparativos combinando dados de origens diferentes", diz Abel Packer. "Para as agências de fomento, trata-se de um avanço importante, pois permite gerar mais conhecimento a partir de um mesmo investimento." A experiência mostra que os pesquisadores têm dificuldades em manter os dados primários disponíveis ao longo do tempo. Um artigo publicado em dezembro na revista *Current Biology* mostrou que as informações que servem de base a artigos científicos vão se perdendo ao longo do tempo. Os autores esquadrinharam 516 artigos da área de ecologia publicados entre 1991 e 2011 – e foram ver o que aconteceu com os dados primários. Constataram que os artigos publicados nos dois anos anteriores estavam acessíveis, mas as chances de isso acontecer com os publicados anteriormente caíam a uma taxa de 17% ao ano. "Cedo ou tarde, o *software* que permite acessar um arquivo ou banco de dados vai ficar obsoleto. Há uma área da pesquisa em computação, chamada curadoria, voltada para preservar os aparatos computacionais e garantir não apenas a qualidade, mas também a preservação de dados para seu uso futuro, ou pelo menos os considerados mais valiosos", diz Claudia Bauzer Medeiros.

Um desafio ainda em aberto é desenvolver um modelo para financiar os serviços ligados a essa nova etapa. "As taxas eventualmente cobradas pelos repositórios não são altas, mas alguém tem que financiar. Atualmente, instituições e programas de pesquisa, por exemplo na área de genética e de proteínas, criaram repositórios desse tipo e financiam o armazenamento e a disponibilização dos dados", diz Abel Packer, referindo-se a casos como o do GenBank, banco de dados de sequências de DNA e de aminoácidos do Centro Nacional de Informação Biotecnológica dos EUA. Num esforço para organizar mais de 600 repositórios e desenvolver metodologias, foram criados dois catálogos que trabalham de forma cooperativa. Um deles é o Biosharing.org, sediado na Universidade de Oxford, que reúne a lista de repositórios de dados de ciências biológicas, como DNA e proteínas. O segundo é o Registry of Research Data Repositories, financiado pela Fundação de Pesquisa da Alemanha, que compila os repositórios das demais ciências, inclusive as sociais.



A biblioteca SciELO Brasil deverá ter até o final do ano uma política definida para o arquivamento em repositórios dos dados das pesquisas publicadas em suas revistas seguindo padrões internacionais. "Estamos estudando se é mais recomendável criarmos um repositório próprio da biblioteca SciELO além das alianças com repositórios já existentes", afirma Abel Packer. No Brasil a criação de repositórios de dados científicos ainda é incipiente. Um exemplo pioneiro de banco de dados advindo de projetos científicos é o criado para o Sistema de Informação Ambiental (SinBiota), que reúne e integra as informações produzidas em projetos vinculados ao Programa Biota-FAPESP. O SinBiota permite analisar a distribuição das espécies catalogadas no território paulista sobre uma base cartográfica digital. "Quem está organizando um movimento de dados abertos é o Ministério do Planejamento, mas a preocupação é com dados públicos governamentais, não dados de pesquisa", observa Hélio Kuramoto, tecnologista sênior do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) e um estudioso do movimento de acesso aberto a pesquisas científicas. Várias universidades brasileiras, inclusive as três estaduais de São Paulo, criaram repositórios para reunir sua produção científica – um grande avanço, mas que ainda não considera o armazenamento dos dados que amparam tais pesquisas.

Entre as revistas científicas brasileiras, um exemplo raro com política de publicação semelhante à da *PLoS* é o da *Brazilian Political Science Review* (*BPSR*), vinculada à Associação Brasileira de Ciência Política. A *BPSR* é uma revista de acesso aberto, publicada exclusivamente em inglês em formato eletrônico. Desde o ano passado, os autores de artigos cujo conteúdo se baseia em métodos quantitativos são solicitados a disponibilizar, no próprio *site* da revista, os bancos de dados que embasaram o *paper* e também os chamados *codebooks*, dicionários que permitem a identificação das variáveis empregadas nos bancos de dados. A medida elevou os custos de manutenção da revista, que necessita de ajuda de um profissional para manter o repositório. "O princípio que orientou a adoção desta iniciativa é um princípio básico da ciência, que consiste em viabilizar a replicação dos procedimentos que levaram às conclusões obtidas em um trabalho de investigação. Caso o leitor queira refazer os cálculos de modo a saber se as conclusões estão corretas, é preciso que os dados que lhe serviram de base estejam disponíveis publicamente, isto é, sem que o leitor precise contar com a boa vontade dos autores da pesquisa para fornecê-los", explica Marta Arretche, professora da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo (USP) e coeditora da revista, juntamente com Janina Onuki, professora do Instituto de Relações Internacionais da USP. Outra motivação é a possibilidade de ampliar a repercussão dos artigos publicados na revista. Marta Arretche cita estudo feito sobre o *Journal of Peace Research*, também da área de ciência política e relações internacionais. O estudo concluiu que artigos do periódico que disponibilizam dados primários têm duas vezes mais citações do que os demais.

"Uma terceira motivação está relacionada ao custo de produzir bancos de dados, que é muito alto. Os repositórios permitem coletivizar esses custos e aumentar as oportunidades de entrada a um tema de pesquisa", diz a professora, que é coordenadora do Centro de Estudos da Metrópole (CEM), um dos 17 Centros de Pesquisa, Inovação e Difusão (Cepid) financiados pela FAPESP. O CEM, a partir dos anos 2000, tornou-se conhecido por produzir e disseminar dados georreferenciados sobre ass principais metrópoles brasileiras, disponibilizando em seu *site*, gratuitamente, diversas bases de dados. Segundo a professora, a maior parte dos autores da revista lida bem com a exigência da oferta dos dados. "Eles têm alguns receios legítimos, como a possibilidade de que alguém use os dados sem dar o devido crédito, embora a revista deixe claro que é necessário citar a fonte. Já pensamos em exigir que os usuários se identifiquem como requisito para ter acesso aos dados, mas isso feriria o espírito do acesso aberto à publicação científica. Outros autores gostariam de utilizar intensamente as informações antes de disponibilizá-las. Há, de fato, uma tensão entre o princípio da replicação e o princípio da autoria, mas o da replicação tem prevalecido", afirma. ■

"Há uma tensão entre o princípio da replicação de dados e o princípio da autoria", diz Marta Arretche