GESTÃO DE DADOS

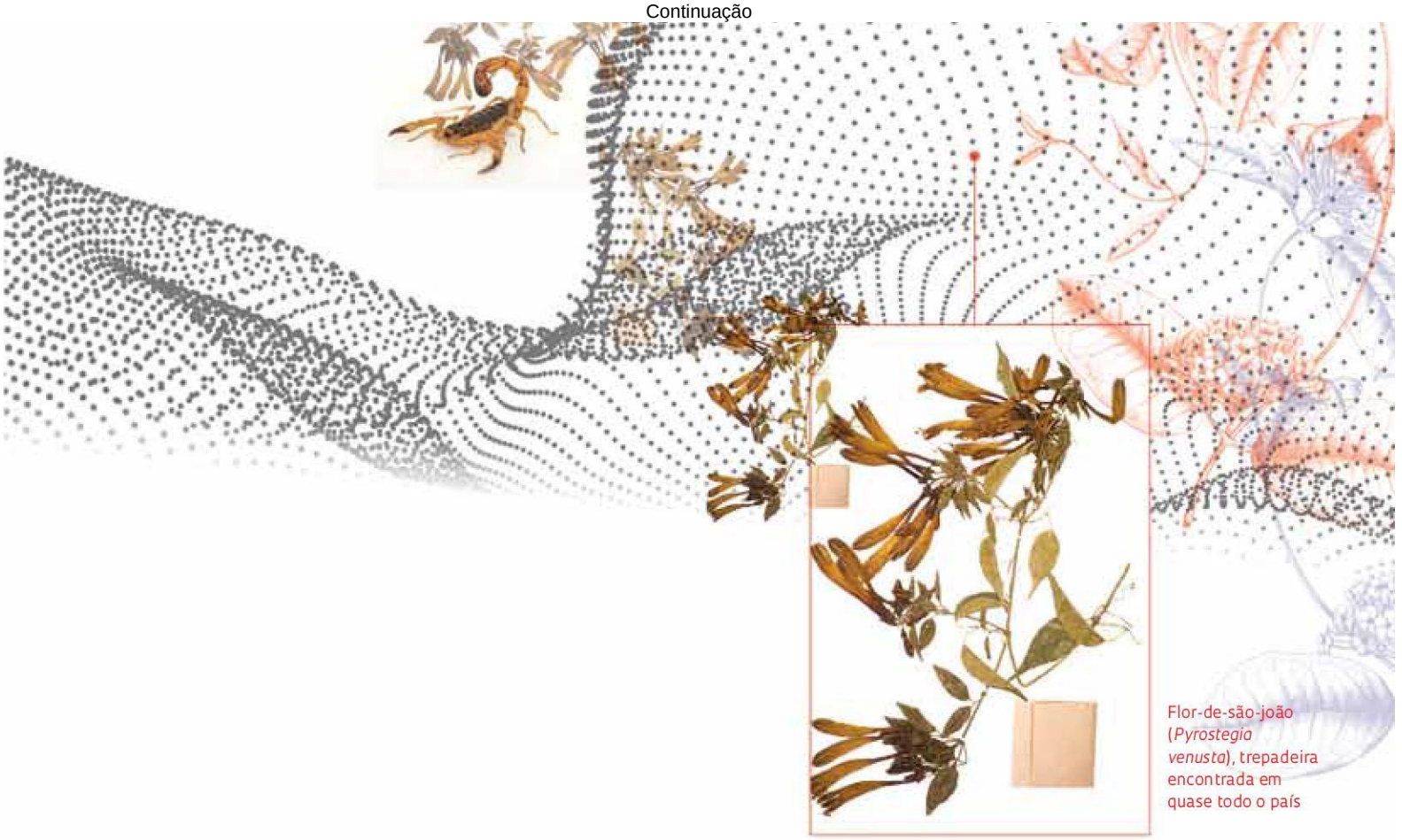
Escorpião-amarelo (*Tityus serrulatus*), das regiões Sudeste, Centro-Oeste e Nordeste

Manaosella cordifolia, árvore da região Norte

Informação com qualidade

Carlos Fioravanti

Borboletas de uma coleção da Universidade Estadual de Campinas



Flor-de-são-joão (*Pyrostegia venusta*), trepadeira encontrada em quase todo o país

Pesquisadores propõem estratégias para tornar mais confiáveis acervos de bases on-line sobre biodiversidade

Quem entra em bases de dados on-line sobre biodiversidade encontra milhões de registros sobre espécies de plantas e animais e as áreas que ocupam ou ocuparam no Brasil e em outros países. Depois da satisfação de encontrar matéria-prima abundante para fundamentar os trabalhos científicos, começam as inquietações: como extrair e filtrar os dados e, principalmente, como saber se são realmente confiáveis? Eventuais erros de nomes de espécies e de localização serão automaticamente indicados e eliminados? Essas questões são importantes porque dados incorretos ou incompletos frequentemente levam a análises inconsistentes.

Pesquisadores da Escola Politécnica da Universidade de São Paulo (Poli-USP) ingressaram no debate sobre o controle de qualidade das informações dos bancos de dados on-line, propondo novas estratégias para resolver problemas observados há uma década. Em 2006, ao ingressar em uma rede de pesquisa em biodiversidade formada por biólogos de 11 países das Américas, o engenheiro elétrico Antônio Mauro Saraiva, professor da Poli-USP, encontrou nomes científicos diferentes para as mesmas espécies, coordenadas geográficas erradas e escassez de detalhes sobre os organismos coletados. Essas informações abastecem bases de dados on-line, a partir das quais se produzem trabalhos científicos sobre distribuição ou abundância de espécies de animais e plantas. "Cinco ou 10 anos depois, os pesquisadores não compreendiam mais os códigos e abreviações que tinham usado nas coletas", observou.

Em 2008, Saraiva começou a discutir com o cientista da computação Allan Koch Veiga como os critérios de organização e qualidade dos bancos de dados poderiam ser aprimorados e padronizados para resultar em informações corretas e completas. Veiga concluiu o doutorado em 2016, sob orientação de Saraiva. Atualmente fazendo um estágio de pós-doutorado na Poli, ele apresentou uma proposta conceitual elaborada pelo grupo coordenado por Saraiva no dia 3 de outubro em Ottawa, no Canadá, para unificar a terminologia e os critérios de avaliação da qualidade das informações dos acervos on-line de animais, microrganismos, plantas e fungos.

FOTOS EDUARDO CESAR E LÉO RAMOS CHAVES **ILUSTRAÇÃO** FREEPIK E FLORABRASILIENSIS

O besouro *Cyrtomon luridus*, parasita de plantas

Paratudo-do-cerrado ou perpétua (*Comphrena macrocephala*), do Cerrado

A mais abrangente entre as cerca de 25 bases mundiais é a Plataforma Global de Informações sobre Biodiversidade (GBIF, www.gbif.org). Criada em 2001, reúne quase 850 milhões de registros de espécies, dos quais 6 milhões provêm do Brasil, um dos cerca de 60 países dessa rede.

"Como não existe uma base conceitual consensual, cada grupo de trabalho nessa área define a qualidade e a avalia de modo diferente, impossibilitando a comparação entre os resultados", comenta Saraiva, que é coordenador do Núcleo de Apoio à Pesquisa em Biodiversidade e Computação da USP (BioComp). O que o grupo da Poli está propondo, em conjunto com especialistas do Canadá, Estados Unidos, Austrália e Dinamarca, é uma linguagem comum que facilite a gestão da qualidade de dados. Segundo Saraiva, um levantamento feito com base na proposta da Poli – realizado por pesquisadores de diversos países de um grupo de trabalho do Biodiversity Information Standards, uma associação científica internacional que desenvolve padrões de qualidade de dados – identificou 100 tipos de testes de verificação de qualidade nas bases de dados. Os testes consistem em programas ou subprogramas e indicam, por exemplo, se as coordenadas geográficas de uma coleta estão corretas. "Mesmo que os programas tenham o mesmo objetivo, não podemos comparar os resultados entre eles porque os critérios que adotam são diferentes", diz ele. "Pretendemos enquadrar todos em um mesmo pano de fundo conceitual, deixando claro os modos de funcionamento de cada um."

Esses conceitos nortearão a plataforma de qualidade de dados que o grupo da Poli deve desenvolver, a partir de 2018, para o Sistema de Informação sobre a Biodiversidade Brasileira (SiBBr, www.sibbr.gov.br). Lançada em 2014, essa base reúne cerca de 10 milhões de registros da ocorrência de 155 mil espécies de animais e plantas do país. "Apesar dos avanços, como a crescente oferta de softwares abertos que permitem a publicação de informações científicas, ainda falta uma política nacional de gerenciamento de dados que estabeleça as ações e os critérios de qualidade", ressalta a bióloga Andrea Nunes, coordenadora-geral de biomas do Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTIC) e diretora nacional do SiBBr.

À medida que avançar, a plataforma de qualidade de dados deverá interagir com as bases que alimentam o SiBBr e estabelecer padrões comuns de funcionamento. "Um critério nem sempre adotado pelas bases de dados é que o ponto de partida da definição de qualidade é o uso que o pesquisador pretende dar à informação", comenta Saraiva, que usa uma analogia para explicar melhor. "Tomates para molho ou para salada podem ter qualidade diferente; para molho os tomates podem ser bem maduros e um pouco amassados, enquanto para saladas precisam ser bem firmes, não muito maduros", compara.

A meta do grupo da Poli é ajudar o pesquisador a definir os critérios de seleção de dados antes de começar uma busca, para não ter de filtrar depois o que interessa em meio a milhares de registros sobre uma espécie ou grupo de espécies, e, além disso, deixar esses critérios expostos, como um guia, para outros usuários. "Se um pesquisador quiser apenas uma lista de espécies de um país, não precisará da coordenada geográfica exata de cada localidade, mas essa informação será indispensável se quiser fazer um estudo sobre a distribuição geográfica de animais ou plantas em uma região", diz Veiga.

VERIFICAÇÃO DE ERROS

A rede *species*Link (www.splink.cria.org.br), uma das bases nacionais de biodiversidade, permite a seleção de informações sobre a ocorrência e a distribuição de espécies de microrganismos, algas, fungos, plantas e animais. Desenvolvida a partir de 2001 com apoio da FAPESP, integrando 12 coleções biológicas do estado de São Paulo, a base se expandiu, principalmente com o Herbário Virtual da Flora e dos Fungos, um dos Institutos Nacionais de Ciência e Tecnologia (CNPq), e hoje reúne registros de 470 coleções do Brasil e de outros países.

Como não existe uma base conceitual consensual, cada grupo define qualidade de modo diferente, diz Saraiva

Essas coleções compartilham cerca de 9 milhões de registros de 125 mil espécies, das quais 2.756 ameaçadas de extinção.

Do total de registros, 68% possuem coordenadas geográficas exatas, no município indicado na coleta, 23% não têm informação sobre a localização da coleta e 8% apresentam dados imprecisos. As coordenadas de 1% do total de registros estão bloqueadas para verificação pelos curadores, os especialistas responsáveis pelos dados de cada coleção. "Se um dado for considerado sensível, como a coordenada geográfica de uma espécie ameaçada de alto valor comercial, a localização ou até o registro completo pode ser bloqueado. Cabe ao curador decidir o que deve ser compartilhado na rede", diz a engenheira de alimentos Dora Canhos, diretora do Centro de Referência em Informação Ambiental (Cria), de Campinas, instituição responsável pelo desenvolvimento e manutenção da rede *species*Link. "Todo erro precisa ser corrigido na origem. Nenhum registro é alterado pelo Cria." Uma vez incorporada à rede, a informação é compartilhada de forma livre e aberta a qualquer interessado.

FOTOS EDUARDO CESAR E LÉO RAMOS CHAVES **ILUSTRAÇÃO** FREEPIK E FLORABRASILIENSIS

MILHÕES DE ESTRELAS

"As equipes que trabalham com os dados dos herbários são insuficientes para limpar os dados, verificar a qualidade e atualizar os nomes científicos", observa o tecnologista Luís Alexandre Estevão da Silva, coordenador do núcleo de computação científica e geoprocessamento do Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Por essa razão, a instituição criou e implantou programas de detecção automática com 81 filtros de verificação de qualidade capazes de relatar, por exemplo, que "a coordenada não coincide com o município informado". "Temos muito a avançar, porque ainda há numerosas duplicatas e diferenças de classificação de plantas nos herbários", diz Silva. Sua equipe desenvolveu e em 2005 implantou o Jabot, um sistema de gerenciamento de coleções científicas de herbários, liberado para outras instituições em 2016 e atualmente adotado por 28 herbários de universidades e centros de pesquisa brasileiros.

"Temos de adotar métodos para analisar a qualidade de dados enquanto são produzidos", afirma a engenheira eletricista Cláudia Bauzer Medeiros, professora do Instituto de Computação da Universidade Estadual de Campinas (Unicamp) e coordenadora do programa eScience da FAPESP. "Ao utilizar dados produzidos por outros, é comum que os pesquisadores não verifiquem a confiabilidade da informação, mesmo sabendo que a validação dos resultados de uma pesquisa depende da qualidade dos dados." Muitas vezes, ela acrescenta, "essa verificação é inviável, por falta de informação sobre a qualidade dos dados".

Ainda que as estratégias de controle de qualidade de dados não estejam integradas e padronizadas, a preocupação com a consistência da matéria-prima da ciência – a informação – é crescente. E não apenas na biologia. O físico colombiano Alberto Molino Benito trabalha há dois anos com sua equipe no Instituto de Astronomia, Geofísica e Ciências Atmosféricas (IAG) da USP no desenvolvimento de programas para extrair informação numérica – de modo automático e com grande precisão – das imagens que começaram a ser captadas pelo telescópio Southern Photometric Local Universe Survey (S-Plus, ou levantamento por fotometria do universo local do hemisfério Sul), construído em Cerro Tololo, no Chile, sob a coordenação do próprio IAG.

"As informações servirão para gerar catálogos de estrelas, galáxias, quasares e asteroides, com suas posições, tamanho, luminosidade, distância da Terra e massa", conta Benito. "Estamos terminando a calibração e validação dos programas para que os pesquisadores não tenham de se preocupar com a qualidade dos dados, quando começar a coleta automática de imagens, no início de 2018." Com um espelho de 80 centímetros de diâmetro, o S-Plus deve concluir a observação do céu do hemisfério Sul em dois anos, reunindo informações sobre a distribuição espacial de milhões de estrelas e galáxias. ■

Artigo científico

VEIGA, A. K *et al.* A conceptual framework for quality assessment and management of biodiversity data. **PLOS ONE**. v. 12 (6), p. e0178731. 2017.