



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”, “reproduzível” 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 órgãos 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.