CAPÍTULO 4

EXPLORANDO A CIÊNCIA ABERTA: DESAFIOS E PERSPECTIVAS DOS COLETORES, AGREGADORES E COLECIONADORES

DENISE FUKUMI TSUNODA

ALEX SEBASTIÃO CONSTÂNCIO

4.1 INTRODUÇÃO

Na era digital, a Ciência Aberta (*Open Science*) emergiu como um pilar para impulsionar o avanço do conhecimento científico e fomentar a inovação em diversas áreas do saber. A rápida evolução das tecnologias da informação e comunicação (TICs) transformou a maneira como os dados são coletados, analisados, compartilhados e utilizados, promovendo um ambiente propício para a disseminação de práticas colaborativas e transparentes na pesquisa científica.

A Ciência Aberta, muitas vezes definida como o movimento para tornar os resultados da pesquisa científica acessíveis a todos, desde acadêmicos até o público em geral, tem sido promovida pelo reconhecimento de que o conhecimento científico é um bem público que deve ser compartilhado e utilizado para o benefício da sociedade como um todo. O portal da Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp, 2024) define Ciência Aberta como sendo:

> [...] o conjunto de políticas e ações de disseminação do conhecimento, em geral por meios digitais, para que todos os resultados de uma pesquisa sejam acessíveis a todos, passíveis de reutilização e de reprodução.

E ainda complementa que tais resultados incluem "publicações, dados, metodologias e processos computacionais usados no desenvolvimento da pesquisa" (Fapesp, 2024).

A Ciência Aberta inclui o acesso aberto a publicações científicas, dados de pesquisa e infraestruturas, bem como a participação ativa de todos os interessados, incluindo cidadãos não cientistas. A crescente disponibilidade de dados abertos, juntamente com o acesso facilitado (inclusive on-line) a ferramentas de análise de dados e visualização, tem permitido que cientistas de todo o mundo colaborem em projetos de pes-

quisa interdisciplinares antes inviabilizados, de forma a acelerar o ritmo das descobertas científicas e aumentar a qualidade e a confiabilidade das metodologias adotadas e resultados alcançados.

Ainda, a Ciência Aberta tem se apresentado como um catalisador para a inovação, estimulando e oportunizando o desenvolvimento de novas tecnologias, produtos e serviços que podem ter um impacto significativo na economia e na sociedade. Ao compartilhar dados e conhecimentos científicos, a comunidade científica pode colaborar com setores industriais, governamentais e não governamentais para enfrentar desafios complexos, a exemplo de segurança e saúde pública, segurança alimentar, desigualdade social e econômica, sustentabilidade ambiental e segurança cibernética e privacidade de dados. Outrossim, a Ciência Aberta fomenta os esforços, em parceria, da academia com os setores industriais, governamentais e não governamentais.

Portanto, com a era digital, a Ciência Aberta foi transformada de uma mera abordagem metodológica para uma filosofia que fomenta a transparência, a colaboração e a democratização do conhecimento científico.

Naturalmente, as infraestruturas e ferramentas científicas abertas cooperam com a Ciência Aberta para tornar a ciência mais acessível, transparente e colaborativa. A Organização das Nações Unidas para a Educação, a Ciência e a Cultura (Unesco, 2011) destaca a importância da Ciência Aberta para enfrentar desafios ambientais, sociais e econômicos complexos, favorecendo o bem-estar humano, a sustentabilidade ambiental e o desenvolvimento social e econômico sustentável.

Figura 4.1 - Ciência Aberta Unesco

Fonte: Unesco, 2021.

No Brasil, por exemplo, o Portal Brasileiro de Publicações Científicas em Acesso Aberto (Oasisbr)[1] iniciativa do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), reúne aproximadamente 4 milhões de documentos científicos disponíveis para pesquisa, sendo um pouco mais de 1,7 milhões de artigos, 2 milhões de teses e dissertações, quase 6 mil conjuntos de dados de pesquisa e 63 mil livros e capítulos

[1] OASISBR. Disponível em: https://oasisbr.ibict.br/vufind/. Acesso em: 28 maio 2024.

de livros, demonstrando o compromisso do país com a disseminação do conhecimento científico (Ibict, 2024).

A Unesco (2011) recomenda que os Estados-membros adotem medidas para promover a Ciência Aberta, incluindo infraestruturas e serviços científicos abertos que sejam orientados para as necessidades dos cientistas e outros públicos, desenvolvendo funcionalidades adaptadas às suas práticas e apresentando interfaces de fácil uso.

As infraestruturas para a promoção da Ciência Aberta são essenciais para apoiar a pesquisa e a colaboração científica em um ambiente aberto e acessível. Aqui estão alguns exemplos:

- a) repositórios de dados abertos: plataformas que armazenam dados de pesquisa para que possam ser acessados, reutilizados e compartilhados por qualquer pessoa. Exemplos incluem o Figshare[2], uma plataforma web para gerenciamento e disseminação de dados de pesquisa acadêmica, aceitando diversos tipos de arquivos com visualização no navegador, originalmente concebida para pesquisadores, expandiu para atender uma demanda mais ampla, permitindo compartilhamento, citação e descoberta de resultados de pesquisa. A plataforma garante preservação de dados a longo prazo e oferece soluções também para editoras e instituições, seguindo padrões do setor; o Zenodo[3], um repositório digital multidisciplinar de acesso aberto, iniciativa voltada à Ciência Aberta (Open Science) e Dados Abertos (Open Data), desenvolvido pelo consórcio europeu OpenAIRE e CERN – European Organization for Nuclear Research, que permite que investigadores, projetos e instituições que não disponham de um repositório institucional

[2] FIGSHARE. Disponível em: https://figshare.com/. Acesso em: 28 maio 2024.
[3] ZENODO. Disponível em: https://zenodo.org/. Acesso em: 28 maio 2024.

ou temático adequado possam partilhar e disseminar seus resultados científicos, independente da área de conhecimento; e o Dríade, ou ainda *Dryad*[4] um repositório disciplinar internacional com foco em dados científicos e clínicos, que prioriza as áreas de Ciência e Tecnologia, Ciências Ambientais e Ecologia e Biologia Evolucionária, entre outros;

- b) plataformas de publicações de acesso aberto: além do já citado Oasisbr, o *Directory of Open Access Journals*[5] (DOAJ) é um diretório on-line que indexa e oferece acesso a revistas científicas de acesso aberto revisadas por pares e visa aumentar a visibilidade e a facilidade de uso dessas revistas, garantindo a qualidade e a transparência das publicações, além de indexar artigos gratuitos para ler, baixar e distribuir, promovendo o acesso aberto ao conhecimento científico; o SciELO[6] (*Scientific Electronic Library Online*) que mantém uma coleção de revistas científicas de acesso aberto da América Latina, Caribe, Espanha e Portugal; e arXiv[7] um repositório de preprints de acesso aberto mantido pela Universidade Cornell, focado principalmente em física, matemática, ciência da computação e disciplinas afins, que permite que os pesquisadores submetam seus trabalhos (acessíveis gratuitamente) antes da revisão por pares, entre outros benefícios;
- c) ferramentas de gestão de pesquisa: a exemplo do *Open Science Framework*[8] (OSF), que ajuda pesquisadores a planejar, executar e compartilhar seus trabalhos de forma aberta; e do *ScienceOpen*[9], uma plataforma de descoberta

[4] DRYAD. Disponível em: https://datadryad.org/stash. Acesso em: 28 maio 2024.
[5] DOAJ. Disponível em: https://doaj.org/. Acesso em: 28 maio 2024.
[6] SciELO. Disponível em: https://scielo.org/pt/. Acesso em: 28 maio 2024.
[7] arXiv. Disponível em: https://arxiv.org/. Acesso em: 28 maio 2024.
[8] OSF. Disponível em: https://osf.io/. Acesso em: 28 maio 2024.
[9] SCIENCEOPEN. Disponível em: https://www.scienceopen.com/. Acesso em: 28

e comunicação de pesquisa que também oferece serviços de gestão de dados e funcionalidades que facilitam a publicação e compartilhamento de dados de pesquisa, artigos e revisões por pares, além de fornecer métricas de impacto e visibilidade;

- d) bibliotecas digitais: a exemplo do *Project Gutemberg*[10], um projeto voluntário que oferece acesso gratuito a uma vasta coleção de livros eletrônicos (mais de 70 mil) em domínio público, principalmente clássicos da literatura mundial, disponíveis em vários formatos de leitura; da Europeana[11], que oferece acesso a milhões de itens digitalizados de bibliotecas, arquivos e museus de toda a Europa; e da Biblioteca Digital Mundial[12] (*World Digital Library*), uma iniciativa da Biblioteca do Congresso dos EUA e da Unesco, que oferece acesso gratuito a uma vasta coleção de documentos culturais de todo o mundo, incluindo manuscritos, mapas, livros, partituras, gravações e fotografias em vários idiomas, promovendo a compreensão internacional e o intercâmbio cultural, entre outras iniciativas;
- e) iniciativas de ciência cidadã: plataformas que envolvem o público geral na coleta de dados e no processo de pesquisa, como o *Zooniverse*[13], que hospeda diversos projetos de ciência cidadã em áreas como astronomia, biologia, clima, humanidades e outras, permitindo que cidadãos participem de projetos de pesquisa ajudando a classificar dados,

maio 2024.

[10] *Project Gutenberg*. Disponível em: https://www.gutenberg.org/. Acesso em: 28 maio 2024.

[11] EUROPEANA. Disponível em: https://www.europeana.eu/pt. Acesso em: 28 maio 2024.

[12] *World Digital Library*. Disponível em: https://www.loc.gov/collections/world-digital-library/about-this-collection/. Acesso em: 28 maio 2024.

[13] ZOONIVERSE. Disponível em: https://www.zooniverse.org/. Acesso em: 28 maio 2024.

identificar padrões e transcrever documentos históricos; o *iNaturalist*[14], uma plataforma de biodiversidade que permite aos cidadãos registrar observações de plantas e animais, além de colaborar na identificação das espécies, com o objetivo de criar um banco de dados global de biodiversidade que pode ser usado para pesquisa científica e conservação; e o *Open Street Map*[15], um projeto colaborativo para criar um mapa livre e editável do mundo, onde qualquer pessoa pode contribuir com dados geográficos, sendo utilizado por pesquisadores, ONGs e desenvolvedores em todo o mundo para projetos de mapeamento e análise espacial, dentre outras finalidades.

As mencionadas infraestruturas são projetadas para estimular a transparência, reprodutibilidade e eficiência na pesquisa científica, minimizar as desigualdades de acesso ao desenvolvimento científico e democratização do conhecimento.

As infraestruturas dedicadas ao fomento da Ciência Aberta são suportadas por alguns componentes específicos conhecidos como coletores, agregadores e colecionadores, que realizam tarefas bem definidas, contribuindo para a execução de etapas para a disseminação eficiente e sustentável de informações acadêmicas.

Para democratizar o acesso ao conhecimento científico, a coleta, a agregação e o armazenamento de dados desempenham papéis fundamentais para garantir os principais elementos da Ciência Aberta: transparência, acessibilidade e disponibilidade de forma gratuita, conforme ilustrado na Figura 4.2.

[14] *iNaturalist*. Disponível em: https://www.inaturalist.org/. Acesso em: 28 maio 2024.
[15] *OpenStreetMap*. Disponível em: https://www.openstreetmap.org. Acesso em: 28 maio 2024.

Figura 4.2 - Leis fundamentais da Ciência Aberta

Fonte: German National Library of Science and Technology (2018).

O presente capítulo explora as ferramentas e métodos utilizados para reunir, organizar e disponibilizar dados de pesquisa de maneira eficiente e acessível.

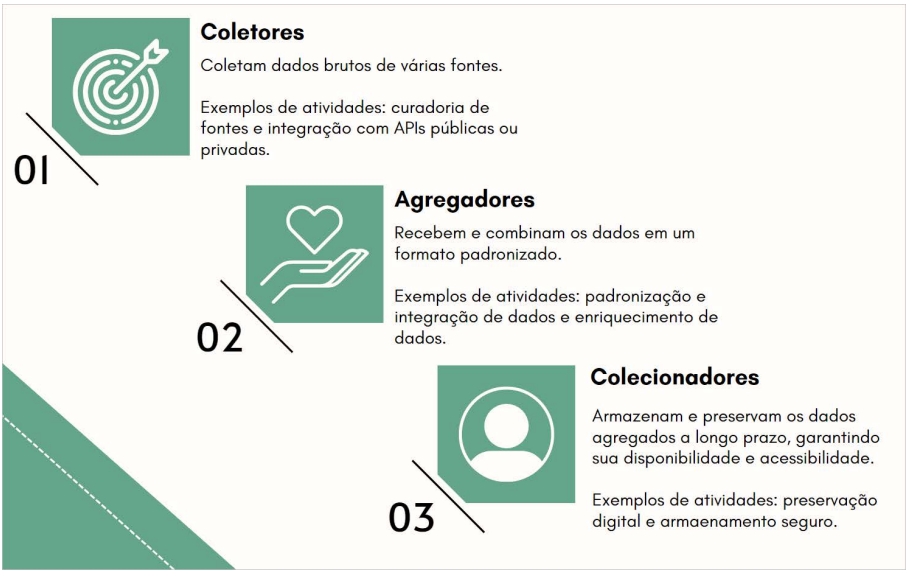
4.2 COLETORES, AGREGADORES E COLECIONADORES

Os coletores, agregadores e colecionadores de dados são componentes automatizados, essenciais do ecossistema de promoção da Ciência Aberta, uma vez que viabilizam a coleta de dados científicos de múltiplas fontes, permitindo sua integração e armazenamento para fácil acesso e reutilização.

A Figura 4.3 representa um fluxo proposto pelos autores, que estabelece a relação entre coletores, agregadores e colecionadores. O fluxo,

na verdade, define um ciclo, uma vez que é repetido de forma ininterrupta ao longo da existência de uma determinada coleção. Esse ciclo demonstra como os dados são processados e gerenciados desde sua coleta inicial até sua preservação e uso futuro, destacando a importância de cada fase para a integridade e valor dos dados ao longo do tempo, bem como algumas atividades inerentes a cada etapa.

Figura 4.3 - Representação gráfica do fluxo de coletores, agregadores e colecionadores



Fonte: Elaborado pelos autores (2024).

Na sequência, cada um desses elementos será apresentado, com destaque para suas funcionalidades e contribuições para a Ciência Aberta.

4.2.1 COLETORES

Os coletores ampliam o acesso aos dados científicos, permitindo que pesquisadores e interessados tenham acesso a dados e informações de diferentes fontes e domínios científicos. Além disso, eles facilitam a descoberta de dados, tornando-os disponíveis para análise, reutilização

e compartilhamento, promovendo, dessa forma, a transparência e a colaboração na pesquisa científica.

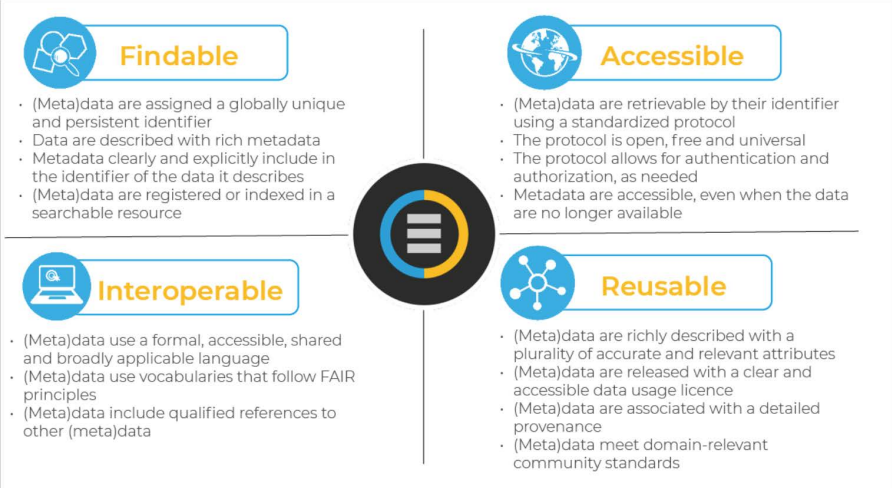
Assim, os coletores podem ser entendidos como sistemas ou ferramentas projetadas para coletar dados em fontes diversas, incluindo bancos de dados públicos, repositórios de pesquisa, redes sociais, sites da *web* e dispositivos sensores, normalmente assegurando precisão e atualização.

Esses agentes podem usar técnicas como *web scraping*, APIs (*Application Programming Interfaces*), mineração de dados e *crowdsourcing* para extrair informações relevantes de diferentes fontes e consolidá-las em um único local.

O processo de coleta requer a capacidade de avaliar o relacionamento semântico entre os diversos pacotes de dados e informações reunidos. Enquanto o ser humano goza naturalmente da capacidade de classificar diferentes documentos ou informações, as máquinas dispõem da capacidade de executar processos repetitivos em alta velocidade. Dado o volume atual de produção científica, o uso de máquinas para a tarefa de coleta é praticamente obrigatório.

Visando estabelecer um conjunto de critérios para uniformizar o processo de coleta, foram adotados os princípios FAIR: *Findable, Accessible, Interoperable and Reusable* (Wilkinson, 2016). Esses princípios, que operam como diretrizes para a elaboração de um coletor, estão apresentados na Figura 4.4 e detalhados na sequência.

Figura 4.4 - Princípios *Findable, Accessible, Interoperable, and Reusable* (FAIR)



Fonte: *Cambridge Crystallographic Data Centre*[16] (CCDC) (2024).

Os princípios FAIR foram desenvolvidos para melhorar a gestão e o compartilhamento de dados científicos. Em resumo, "*Findable*" (F) requer identificadores únicos e metadados claros para facilitar a localização dos dados, "Acc*essible*" (A) garante que os dados estejam disponíveis de forma aberta e acessível on-line, "*Interoperable*" (I) demanda o uso de formatos e padrões comuns para facilitar a integração de dados e "*Reusable*" (R) incentiva a preparação dos dados para permitir sua reutilização por diferentes usuários. Esses princípios, formalizados em 2016 após um workshop na Holanda, têm como objetivo a promoção de uma ciência mais aberta, colaborativa e eficiente, facilitando a replicabilidade dos resultados e promovendo avanços mais robustos no conhecimento científico global.

Desta forma, os requisitos definidos são (Wilkinson, 2016):

a) para ser Localizável (*Findable*):

[16] *Cambridge Crystallographic Data Centre. Fair data principles.* Disponível em: www.ccdc.cam.ac.uk/solutions/about-the-csd. Acesso em: 25 jun. 2024.

F1 - Aos (meta)dados são atribuídos um identificador persistente e global;

F2 - Os dados são descritos com metadados ricos (definidos por R1 abaixo);

F3 - Os metadados incluem clara e explicitamente o identificador dos dados que descrevem;

F4 - Os (meta)dados estão registados ou indexados em um recurso pesquisável;

b) para ser Acessível (*Accessible*):

A1 - Os (meta)dados são recuperáveis pelo seu identificador utilizando um protocolo de comunicações normalizado;

A1.1 - o protocolo é aberto, livre e universalmente implementável;

A1.2 - o protocolo permite um procedimento de autenticação e autorização, quando necessário;

A2 - os metadados são acessíveis, mesmo quando os dados já não estão disponíveis;

c) para ser Interoperável (*Interoperable*):

I1 - Os (meta)dados utilizam uma linguagem formal, acessível, partilhada e amplamente aplicável para a representação do conhecimento;

I2 - Os (meta)dados utilizam vocabulários que seguem os princípios FAIR;

I3 - Os (meta)dados incluem referências qualificadas a outros (meta)dados;

d) para serem Reutilizáveis (*Reusable*):

R1 - Os meta(dados) são ricamente descritos com uma pluralidade de atributos exatos e relevantes;

R1.1 - os (meta)dados são divulgados com uma licença de utilização de dados clara e acessível;

R1.2 - os (meta)dados estão associados a uma proveniência pormenorizada;

R1.3 - os (meta)dados cumprem as normas comunitárias relevantes para o domínio.

A título de exemplo, o DataONE[17] (*Data Observation Network for Earth*) é uma infraestrutura distribuída que permite o acesso a uma rede de repositórios de dados científicos, promovendo a preservação, acesso, uso e reúso de dados ambientais e ecossistêmicos. Ele implementa um ciclo e vida completo de dados, e a coleta é uma das etapas desse ciclo de vida (Michener, 2012). No Brasil, o GO FAIR Brasil atua em todos os domínios do conhecimento e é sediado no Ibict. Mais detalhes podem ser encontrados na obra "Princípios FAIR aplicados à gestão de dados de pesquisa" (Sales et *al.*, 2021).

4.2.2 AGREGADORES

Os agregadores são sistemas ou plataformas que reúnem e organizam dados coletados de várias fontes em um formato padronizado e interoperável. Eles realizam tarefas como curadoria, normalização e enriquecimento dos dados, tornando-os mais acessíveis e utilizáveis. Os agregadores frequentemente aplicam metadados comuns e padrões de formatação para facilitar a integração e o compartilhamento de dados entre diferentes sistemas e usuários. A relevância dos agregadores reside na sua capacidade de agregar e organizar dados dispersos, tornando mais fácil para os pesquisadores localizar, acessar e combinar informações

[17] *DATAONE*. Disponível em: https://www.dataone.org/. Acesso em: 25 jun. 2024.

de diferentes fontes. Com a definição de padronização, os agregadores viabilizam a interoperabilidade e a reutilização de dados, facilitando a colaboração e a análise de dados em larga escala.

Em 2023, foi publicado o artigo "*CORE: A Global Aggregation Service for Open Access*" (Knoth et *al.*, 2023) que detalha o *COnnecting REpositories* (CORE), um "serviço acadêmico amplamente utilizado", fornecendo acesso à maior coleção de publicações de pesquisa de acesso aberto do mundo, adquirida de uma rede global de repositórios e periódicos. O serviço foi criado, inicialmente, com a finalidade de permitir a mineração de textos e dados da literatura científica e apoiar a descoberta científica. Atualmente é utilizado em propósitos diversos no ensino superior, setor privado, organizações sem fins lucrativos e pelo público em geral.

Segundo os autores Knoth et *al.* (2023), os serviços possibilitam inovações como a detecção de plágio em organizações terceirizadas e têm sido fundamental no movimento global pelo acesso aberto, facilitando o acesso livre ao conhecimento científico. O artigo descreve ainda o crescimento do conjunto de dados do CORE, os desafios na coleta de artigos de milhares de provedores de dados e as soluções desenvolvidas. Também discute os serviços e ferramentas criados a partir dos dados agregados e examina vários casos de uso do CORE.

Os autores ainda destacam que diversas instituições renomadas utilizam o CORE, incluindo a Universidade de *Cambridge* e o arXiv.org. Esses usos abrangem desde a recomendação de artigos relevantes até a verificação de plágio. Por exemplo, a colaboração com o Turnitin, líder global em *software* de detecção de plágio, usa o CORE *FastSync* para ampliar significativamente sua base de dados de conteúdo. Além disso, o CORE *Recommender*, ativo em mais de 70 repositórios, melhora a acessibilidade dos resultados de pesquisa, sugerindo artigos similares e promovendo a ampla disseminação de trabalhos científicos. Esses esforços resultam na ampliação da visibilidade e alcance do conteúdo científico global.

Comparando o CORE com outros Agregadores de Acesso Aberto (*OpenAIRE*, *BASE*, *Paperity*, *SHARE*), o trabalho de Knoth *et al.* (2023) aponta que o OpenAIRE armazena textos completos, mas não disponibiliza para download e oferece menos registros de metadados e links de OA em comparação com o CORE; o BASE possui maior quantidade de registros de metadados (acima de 300 milhões), mas sem textos completos hospedados; e o *Paperity* e o SHARE apresentam semelhanças em termos de ausência de hospedagem de textos completos e fornecimento limitado de metadados e links.

O CERN[18] apoia o desenvolvimento e a manutenção do INSPIRE, um agregador que facilita a descoberta e o acesso de artigos acadêmicos, pré-publicações, atas de conferências e outras publicações científicas no campo da física de alta energia.

4.2.3 COLECIONADORES

Embora "coleção" denote um conjunto específico e organizado de itens relacionados por um tema ou propósito, e "repositório" se refira a um sistema de armazenamento e gerenciamento de uma ampla gama de documentos e dados digitais, neste capítulo ambos serão tratados como tendo significados semelhantes, dado que um repositório pode abrigar várias coleções. Assim, os colecionadores são sistemas ou repositórios que armazenam e mantêm coleções de dados científicos ao longo do tempo e podem incluir repositórios institucionais, bancos de dados disciplinares, arquivos de dados de pesquisa e outras plataformas de longo prazo.

Os colecionadores são responsáveis por preservar a integridade e a acessibilidade dos dados, garantindo que permaneçam disponíveis e utilizáveis a longo prazo. A relevância dos colecionadores reside na sua

[18] CERN *Open Science*. Disponível em: https://openscience.cern/infrastructure. Acesso em: 24 jun. 2024.

capacidade de preservar e compartilhar a longo prazo dados científicos, garantindo que possam ser acessados e utilizados por pesquisadores atuais e futuros. Eles promovem a transparência e a confiabilidade da pesquisa, ao mesmo tempo em que protegem os dados contra perda, corrupção ou obsolescência.

Diversos exemplos de colecionadores, ou *data repositories*, que são utilizados para armazenar e compartilhar dados científicos internacionais poderiam ser citados, a exemplo de:

- GenBank[19]: banco de dados de sequências de DNA e RNA mantido pelo *National Center for Biotechnology Information* (NCBI), parte dos Institutos Nacionais de Saúde dos Estados Unidos (NIH);
- *PubMed Central* (PMC)[20]: repositório gratuito de artigos de revistas científicas na área biomédica e de ciências da vida, mantido pelo NCBI;
- *Dryad*[21]: repositório digital que hospeda dados de pesquisa em ciências da vida, especialmente dados associados a publicações científicas;
- *Figshare*[22]: plataforma onde pesquisadores podem fazer upload e compartilhar seus dados de pesquisa em diferentes formatos, como figuras, conjuntos de dados e vídeos;
- *Zenodo*[23]: repositório de dados de pesquisa gratuito e aberto, parte do projeto *OpenAIRE*, que permite o depósito de dados de qualquer disciplina acadêmica;

[19] *GenBank*. Disponível em: https://www.ncbi.nlm.nih.gov/genbank. Acesso em: 29 jan. 2025.
[20] *PubMed Central*. Disponível em: https://pmc.ncbi.nlm.nih.gov/. Acesso em: 29 jan. 2025.
[21] *Dryad*. Disponível em: https://datadryad.org/stash. Acesso em: 29 jan. 2025.
[22] *Figshare*. Disponível em: https://figshare.com/. Acesso em: 29 jan. 2025.
[23] *Zenodo*. Disponível em: https://zenodo.org/. Acesso em: 29 jan. 2025.

- *ArXiv*[24]: servidor de pré-impressão onde pesquisadores podem fazer upload de artigos científicos nas áreas de física, matemática, ciência da computação e outras disciplinas.

No Brasil, dentre os diversos, podem ser citados:

- BDTD (Biblioteca Digital de Teses e Dissertações)[25]: Mantida pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), é uma base de dados que reúne teses e dissertações defendidas em instituições de ensino superior brasileiras;
- Scielo (*Scientific Electronic Library Online*)[26]: Uma biblioteca eletrônica que abrange uma coleção selecionada de periódicos científicos brasileiros e de outros países da América Latina;
- Lattes[27]: Embora não seja estritamente um repositório de dados, a Plataforma Lattes, mantida pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), é um sistema utilizado por pesquisadores para registrar e disponibilizar informações sobre suas atividades acadêmicas e produção científica;
- Brasiliana Digital[28]: Um repositório digital que preserva e disponibiliza documentos históricos, literários e científicos sobre o Brasil, promovido pela Fundação Biblioteca Nacional.
- Rede Cariniana[29]: Um sistema de bibliotecas digitais de acesso aberto que oferece acesso a coleções digitais de instituições de pesquisa e universidades brasileiras.

[24] ArXiv. Disponível em: https://arxiv.org/. Acesso em: 25 jan. 2025.
[25] BDTD. Disponível em: https://bdtd.ibict.br. Acesso em: 29 jan. 2025.
[26] SciELO: Disponível em: https://scielo.org/. Acesso em: 25 jan. 2025.
[27] Lattes. Disponível em: https://lattes.cnpq.br/. Acesso em: 25 jan. 2025.
[28] Brasiliana Digital. Disponível em: https://bndigital.bn.gov.br/. Acesso em: 25 jan. 2025.
[29] Brasiliana Digital. Disponível em: https://bndigital.bn.gov.br/. Acesso em: 25 jan. 2025.

Ainda no Brasil, o Ibict desempenha um papel central na promoção da informação científica e tecnológica do país. Além de manter a já citada BDTD, o Ibict também está envolvido em outros projetos e iniciativas que apoiam a gestão e o acesso à informação científica no país. Alguns desses projetos incluem:

- Portal Brasileiro de Publicações Científicas (periodicos.capes.gov.br)[30]: portal que agrega periódicos científicos brasileiros e estrangeiros de acesso aberto, proporcionando visibilidade à produção científica nacional;
- Diretório de Repositórios Brasileiros[31]: diretório que cataloga e promove a visibilidade de repositórios digitais brasileiros que armazenam teses, dissertações, artigos científicos e outros tipos de documentos;
- Biblioteca Digital Brasileira de Computação[32]: repositório digital que abriga a produção científica na área de computação no Brasil.

O Ibict atua ainda como um facilitador da integração e do acesso à informação científica e tecnológica no Brasil, promovendo a colaboração entre instituições e apoiando a infraestrutura de informação necessária para a pesquisa acadêmica e científica.

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, garantindo, assim, que continuem disponíveis muito além do tempo de vida de um

[30] Portal Brasileiro de Publicações Científicas. Disponível em: https://www.periodicos.capes.gov.br/. Acesso em: 25 jan. 2025.

[31] Diretório de Repositórios Brasileiros. Disponível em: https://www.gov.br/ibict/pt-br/assuntos/informacao-cientifica/repositorios-digitais/repositorios-brasileiros-1. Acesso em: 25 jan. 2025.

[32] Biblioteca Digital Brasileira de Computação. Disponível em: http://www.lbd.dcc.ufmg.br/bdbcomp/. Acesso em: 25 jan. 2025.

projeto. O re3data[33] - *Community Driven Open Reference for Research Data Repositories* (COREF) é um catálogo mundial de repositórios institucionais confiáveis com o objetivo de auxiliar 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 (Medeiros, 2021).

Finalmente, os coletores, agregadores e colecionadores constituem uma infraestrutura para a disseminação e compartilhamento de dados científicos em plataformas abertas. Essas funções trabalham em conjunto para promover a transparência, colaboração e reutilização de dados, impulsionando o avanço do conhecimento científico e da inovação, por isso o objetivo deste capítulo é o de fornecer uma análise detalhada do papel dessas funções na Ciência Aberta e explorar sua implementação e impacto.

O Quadro 4.1 apresenta um resumo comparativo dos coletores, agregadores e colecionadores.

Quadro 4.1 - Resumo comparativo

Aspecto	Coletores	Agregadores	Colecionadores
Responsabilidades	Coletar dados brutos de várias fontes; Assegurar a qualidade e integridade dos dados; Monitorar e atualizar dados regularmente.	Indexar e organizar dados provenientes de diferentes coletores; Facilitar a busca e recuperação eficiente de dados; Garantir a interconectividade entre diferentes fontes de dados.	Armazenar e preservar dados científicos; Oferecer acesso e ferramentas para reutilização de dados; Manter a sustentabilidade e a acessibilidade a longo prazo.

[33] re3data. Disponível em: https://www.re3data.org/. Acesso em: 25 jun. 2024.

Aspecto	Coletores	Agregadores	Colecionadores
Funções	Captura de dados em tempo real ou em intervalos regulares. Implementação de protocolos de coleta de dados. Validação e verificação dos dados coletados.	Centralização de dados de múltiplas fontes. Normalização e padronização de dados para facilitar o acesso. Disponibilização de interfaces de busca e consulta.	Fornecimento de um repositório seguro e confiável para dados científicos. Facilitação do compartilhamento de dados entre pesquisadores. Implementação de metadados para descrever e catalogar dados.
Expectativas	Precisão e confiabilidade na coleta de dados. Conformidade com normas e padrões de coleta de dados. Atualizações regulares para garantir dados atuais e relevantes.	Alta capacidade de integração e interoperabilidade entre diferentes sistemas de dados. Eficiência na busca e recuperação de dados. Disponibilidade de dados de alta qualidade e bem-organizados.	Acesso fácil e permanente aos dados. Suporte a formatos e tipos de dados variados. Garantia de que os dados estejam disponíveis para reutilização e replicação de estudos científicos.
Exemplo	Sensores ambientais que monitoram a qualidade do ar. Dispositivos médicos que coletam dados de saúde. Aplicativos móveis que coletam dados de usuários.	Google Scholar que indexa artigos científicos. PubMed que agrega literatura biomédica. Europe PMC que centraliza artigos de ciências da vida.	Zenodo que armazena dados de pesquisa. Figshare que facilita o compartilhamento de dados. Kaggle que oferece datasets para análise de dados.

Fonte: Elaborado pelos autores (2024).

Assim, os coletores, agregadores e colecionadores suportam a organização e preservação dos dados científicos e as tecnologias disruptivas, a exemplo da inteligência artificial (IA), representando um avanço significativo para a Ciência Aberta, conforme apresentado e discutido na próxima seção.

4.3 TECNOLOGIAS COM INTELIGÊNCIA ARTIFICIAL

Os dados abertos e a inteligência artificial (IA) têm o potencial de apoio e aprimoramento mútuos. Se, por um lado, os dados abertos podem ser utilizados como insumo para o treinamento dos sistemas de IA, por outro lado, a IA pode adicionar valor aos dados abertos por meio da extração automatizada de relações sinérgicas entre diferentes documentos. No contexto específico dos coletores, agregadores e colecionadores, os modelos de IA oferecem recursos particularmente benéficos.

Os coletores têm o propósito primário de extrair dados brutos de fontes variadas, mas esses dados precisam apresentar conteúdo e fonte confiáveis. Os modelos de Redes Neurais Artificiais (RNA) são particularmente potentes em identificar relações complexas entre os dados e, por este motivo, realizar a classificação automática. Na função específica de coleta, essa classificação poderia ser utilizada para enquadrar determinada fonte em categorias de interesse, como “confiável”, “não-confiável” ou “potencialmente confiável” etc.

Os agregadores constituem a segunda etapa do processo, com a função de organizar os conteúdos em categorias, como assuntos, temas, áreas ou qualquer outra de interesse. Novamente a capacidade dos modelos de IA de produzir modelos preditivos tem o poder de categorizar documentos e relacionar semanticamente diferentes estudos, reunindo aqueles que tratam de temáticas similares.

Agregadores baseados em IA surgem como um recurso incomparável para organizar e classificar automaticamente vastos volumes de conteúdo científico, cooperando para a recuperação acurada e rápida de pesquisas de interesse.

Um exemplo prático é o CORE (Knoth, 2023), agregador baseado em Mineração de Dados e Mineração de Texto, que concentra milhares de provedores de dados para oferecer coleções de conteúdo extraídas diretamente de documentos em PDF, compondo um único e vastíssimo conjunto de dados.

Finalmente, os colecionadores têm o propósito de preservar o conteúdo científico e oferecer meios de extração facilitada. Modelos de IA podem ser utilizados para digitalizar automaticamente artigos, modelos, projetos e tabelas de documentos físicos para uma forma digital processável, tornando antigos manuscritos imunes à ação do tempo e indexáveis para consulta e recuperação.

Ao mesmo tempo, revoluções científicas podem invalidar antigos conceitos que merecem ser mantidos por efeitos históricos, mas que devem ter sua aplicabilidade restrita. Novamente, a capacidade de relação semântica que os modelos de IA atuais oferecem podem auxiliar na identificação de concepções conflitantes e na elaboração de linhas de tempo para a análise evolucionária de temas científicos.

A *Foster Open Science*[34], organização europeia para a disseminação da Ciência Aberta, oferece uma interessante taxonomia de tarefas[35] promovidas por tecnologias de IA que podem potencializar coletores, agregadores e colecionadores. A taxonomia é ativa e opera como um seletor de recursos a respeito de cada uma das diversas tarefas (por exemplo, "Busca semântica"), representando fonte de informação especializada em cada tema.

O advento dos modelos de linguagem generativos modificou profundamente a expectativa e a percepção do que já é possível se fazer

[34] FOSTER. Disponível em: https://www.fosteropenscience.eu/. Acesso em: 25 jun. 2024.

[35] FOSTER. *Open Science*. Disponível em: https://www.fosteropenscience.eu/foster-taxonomy/open-science-tools. Acesso em: 25 jun. 2024.

com o apoio da Inteligência Artificial, incluindo o que esperar para os próximos anos.

Em geral, a exposição dos sistemas de IA a um volume maior e a uma variedade maior de dados de treinamento e teste aumenta o poder preditivo dos modelos gerados. Os dados abertos podem ser fontes de grandes quantidades de informações diversas para esses sistemas, por exemplo, dados de segurança pública, saúde, registros climáticos, entre outros.

Os grandes modelos de linguagem (LLM, *Large Language Models*) provaram sua capacidade de recuperar conhecimento valioso a partir de enormes volumes de texto, mas os modelos mais recentes já apresentam capacidade multimodal, permitindo a extração de informação vasta presente em imagens e tabelas (Digital Science, 2023).

Texto, imagem e dados tabulares constituem um volume quase total de toda a informação científica publicada. Considerando que esse volume pode estar acessível para modelos multimodais, pode-se imaginar muito mais do que motores de busca semânticos, partindo-se para o próximo passo que é a ciência apoiada por Inteligência Artificial.

Isso significa dizer que mais do que meras ferramentas para localizar mais acuradamente conteúdos, a próxima geração será a de agentes inteligentes para a ciência. Essa abordagem completamente nova tem o potencial de economizar semanas de estudo e triagem de material, fazendo as pesquisas avançarem a um ritmo jamais imaginado.

Um exemplo de motor de busca semântico baseado em Inteligência Artificial e que atua como um copiloto científico é o *SciSpace*[36]. O *SciSpace* opera com um modelo conversacional baseado em perguntas. Por exemplo, o pesquisador coloca sua questão de pergunta e um conjunto de

[36] SCISPACE. Typeset. Disponível em: https://typeset.io/. Acesso em: 25 jun. 2024.

artigos com alta similaridade semântica com o assunto são selecionados. Modelo de operação similar é oferecido pelo *Elicit*[37].

Adicionalmente, o *SciSpace* também oferece o recurso de responder perguntas a respeito de um artigo específico, recurso já oferecido pelo *ChatGPT*[38] da *OpenAI*, que permite extrair um resumo, metodologia, resultados e dados específicos (por exemplo, o valor de *p-value* identificado em estudos estatísticos) diretamente do texto.

Em um momento mais avançado de uma pesquisa, os modelos atuais já operam como assistentes de análise de dados, escrevendo código em linguagens de programação como *Python*, que convertem esforços de horas em segundos. Esse código pode ser especificado para consumir os dados de uma pesquisa em andamento, possibilitando o acompanhamento dos dados e das conclusões em um ritmo aceleradíssimo para os padrões pré-modelos de linguagem.

Em todos esses cenários, a disponibilidade de dados e relatórios de pesquisa abertos representa o insumo primordial desse tipo de ferramenta. Primeiramente, porque os modelos precisam ser treinados. Nesse sentido, a disponibilidade de dados (textuais e não-textuais) sempre figurou como um dos principais desafios enfrentados pelos pesquisadores na área de Aprendizado de Máquina (subárea da Inteligência Artificial), situação que se tornou ainda mais crítica com o advento do aprendizado profundo, que requer volumes impressionantes de dados para produzir seus resultados.

No momento em que a Ciência Aberta promove a disponibilidade do conteúdo científico, tem-se uma valiosa matéria-prima para o treinamento dos novos agentes. Uma vez treinados, esses agentes passam a operar retroalimentando o sistema, consumindo materiais científicos para a descoberta de novos conhecimentos. Isso estabelece uma outra relação

[37] ELICIT. Disponível em: https://elicit.com/. Acesso em: 25 jun. 2024.
[38] OpenAI. ChatGPT. Disponível em: https://chatgpt.com/. Acesso em: 25 jun. 2024.

com a Inteligência Artificial, na qual ela mesma começa a impulsionar a própria ciência.

Assim, é possível observar o efeito sinérgico da Ciência Aberta, que não apenas promove a elaboração de modelos, mas também contribui para a produção contínua de mais conhecimento científico.

4.4 CONSIDERAÇÕES FINAIS

Na era digital, a Ciência Aberta tem se mostrado fundamental para o avanço do conhecimento científico e a promoção da inovação em diversas áreas. A rápida evolução das tecnologias da informação e comunicação transformou a maneira como os dados são coletados, analisados, compartilhados e utilizados, promovendo um ambiente propício para práticas colaborativas e transparentes na pesquisa científica. Ao tornar os resultados da pesquisa acessíveis a todos, a Ciência Aberta reconhece o conhecimento científico como um bem público, essencial para o benefício da sociedade. Essa abordagem inclui acesso aberto a publicações científicas, dados de pesquisa e a participação ativa de todos os interessados, possibilitando colaborações globais e interdisciplinares.

Além de acelerar as descobertas científicas, a Ciência Aberta estimula a inovação ao facilitar o desenvolvimento de novas tecnologias, produtos e serviços, e promove parcerias entre academia, indústria e governos para enfrentar desafios complexos. Assim, a Ciência Aberta transcende uma simples metodologia para se tornar uma filosofia que valoriza a transparência, a colaboração e a democratização do conhecimento científico.

O CORE é um exemplo de ferramenta que tem sido amplamente utilizada para o avanço do acesso aberto, facilitando a descoberta e uso de conhecimento científico, uma vez que as suas ferramentas e serviços

apoiam diversas aplicações inovadoras, como detecção de plágio e recomendação de documentos.

Os coletores, agregadores e colecionadores são essenciais para a preservação de vastos volumes de dados científicos, garantindo sua acessibilidade a longo prazo para pesquisadores, comunidades acadêmicas e demais interessados. Enquanto essas plataformas facilitam o acesso a dados e informações valiosos, a convergência com a inteligência artificial abre novos horizontes para a Ciência Aberta. A integração de algoritmos avançados de IA permite análises mais sofisticadas e automatizadas desses dados, revelando padrões complexos e insights anteriormente inacessíveis. Essa parceria entre tecnologia de dados e IA não apenas potencializa a descoberta científica, mas também fortalece os princípios de transparência, colaboração e reprodutibilidade que fundamentam a Ciência Aberta, promovendo um ambiente de pesquisa mais dinâmico e inclusivo.

REFERÊNCIAS

DIGITAL SCIENCE. The state of open data 2023. [*S. l.*]: **Digital Science,** 2023. DOI: https://doi.org/10.6084/m9.figshare.24428194.v1. Disponível em https://digitalscience.figshare.com/articles/report/The_State_of_Open_Data_2023/24428194?file=43138708. Acesso em: 24 jun. 2024.

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO. **Open Science @ FAPESP.** São Paulo: FAPESP, 2024. Disponível em: https://www.fapesp.br/openscience/. Acesso em: 20 fev. 2024.

GERMAN NATIONAL LIBRARY OF SCIENCE AND TECHNOLOGY. **The Open Science training handbook.** [*S. l.*]: Foster, 2018. Disponível em: https://www.fosteropenscience.eu/content/open-science-training-handbook. Acesso em: 24 jun. 2024.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. **Oasisbr:** portal brasileiro de publicações e dados científicos em Acesso Aberto. Brasília, DF: Ibict, 2024. Disponível em: https://oasisbr.ibict.br/vufind/. Acesso em: 23 maio 2024.

KNOTH, Petr; HERRMANNOVA, Drahomira; CANCELLIERI, Matteo, ANASTASIOU, Lucas; PONTIKA, Nancy; Pearce, Samuel; GYAWALI, Bikash; PRIDE, David. A Global aggregation service for Open Access papers. **Sci Data,** [*S. l.*], v. 10, n. 366, p. 1-19, 2023. https://doi.org/10.1038/s41597-023-02208-w. Disponível em: https://www.nature.com/articles/s41597-023-02208-w#citeas. Acesso em: 19 set. 2024.

MEDEIROS, Claudio Bauzer. Ciência Aberta – colaboração sem barreiras para o avanço do conhecimento. **Revista da Sociedade Brasileira da Computação,** n. 46. 2021. DOI: https://doi.org/10.5753/compbr.2021.46.4411. Disponível em: https://journals-sol.sbc.org.br/index.php/comp-br/article/view/4411f. Acesso em: 24 jun. 2024.

MICHENER, William K.; JONES, Matthew B. Ecoinformatics: supporting ecology as a data-intensive science. **Trends in ecology & evolution,** [*S.*

l.], v. 27, n. 2, p. 85-93, 2012. DOI: 10.1016/j.tree.2011.11.016. Disponível em: https://pubmed.ncbi.nlm.nih.gov/22240191/. Acesso em: 19 set. 2024.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS PARA A EDUCAÇÃO, A CIÊNCIA E A CULTURA. **Recomendação da UNESCO sobre Ciência Aberta.** Paris: Unesco, 2021. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000215520. Acesso em: 17 set. 2024.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS PARA A EDUCAÇÃO, A CIÊNCIA E A CULTURA. **UNESCO 2011.** Paris: Unesco, 2011. DOI: https://doi.org/10.54677/XFFX3334. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000379949_por. Acesso em: 12 maio 2024.

SALES, Luana Farias; VEIGA, Viviane Santos de Oliveira; HENNING, Patrícia; SAYÃO, Luís Fernando. **Princípios FAIR aplicados à gestão de dados de pesquisa.** Rio de Janeiro: IBICT, 2021. 292p. Disponível em: https://ridi.ibict.br/bitstream/123456789/1182/2/IBICT_Principios%20FAIR%20aplicados%20a%20gest%c3%a3o%20de%20dados%20de%20pesquisa_2021.pdf. Acesso em: 24 jun. 2024.

WILKINSON, Mark D. et *al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data,** [*S. l.*], v. 3, n. 1, p. 1-9, 2016. DOI: https://doi.org/10.1038/sdata.2016.18. Disponível em: https://www.nature.com/articles/sdata201618#citeas. Acesso em: 19 set. 2024.

COMO CITAR ESTE CAPÍTULO:

TSUNODA, Denise Fukumi; CONSTÂNCIO, Alex Sebastião. Explorando a Ciência Aberta: desafios e perspectivas dos coletores, agregadores e colecionadores. *In*: DRUCKER, Debora Pignatari; CIUFFO, Leandro; SAYÃO, Luis Fernando; SHINTAKU, Milton; VIDOTTI, Silvana Aparecida Borsetti Gregorio (org.) **Infraestruturas de suporte à Ciência Aberta.** Brasília, DF: Editora Ibict, 2025. p. 94-122. DOI: 10.22477/9786589167754.cap4.