24º ENANCIB
Encontro Nacional de Pesquisa em Ciência da Informação
Perspectivas Contemporâneas na Ciência da Informação
• Vitória - ES • Ancib • PPGCI/UFES

**XXIV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – XXIV ENANCIB**

**ISSN 2177-3688**

**GT 8 – Informação e Tecnologia**

# INTEGRAÇÃO ENTRE A PLATAFORMA LATTES E O OPENALEX:
MODELO DE ENRIQUECIMENTO DE DADOS PARA O BRCRIS

# *INTEGRATION BETWEEN LATTES PLATFORM AND OPENALEX:*
*DATA ENRICHMENT MODEL FOR BRCRIS*

**Fabio Lorensi do Canto** – Universidade Federal de Santa Catarina (UFSC)
**Thiago Magela Rodrigues Dias** – Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
**Washington Luís Ribeiro de Carvalho Segundo** – Instituto Brasileiro de informação em Ciência e Tecnologia (IBICT)
**Raulivan Rodrigo da Silva** – Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
**Marcel Garcia de Souza** – Instituto Brasileiro de informação em Ciência e Tecnologia (IBICT)

**Modalidade: Trabalho Completo**

**Resumo:** Descreve um modelo de integração de dados entre a Plataforma Lattes e o OpenAlex por meio do cruzamento de identificadores persistentes de autores e de publicações. Foram listados mais de 1 milhão e 166 mil registros DOI e 250 mil registros ORCID de todos currículos da plataforma Lattes. Foi utilizada uma infraestrutura computacional em nuvem na Amazon Web Services para processamento de mais de 250 milhões de registros de publicações extraídos do OpenAlex. As listas de DOI e ORCID do Lattes foram cruzadas com os dados das publicações do OpenAlex, identificando-se 98,8% dos registros DOI e 52,2% de registros ORCID. O conjunto final resultado dos dois cruzamentos foi deduplicado e carregado na plataforma BrCris, enriquecendo e diminuindo inconsistências. Constatou-se que o modelo de integração utilizado baseado em identificadores persistentes possibilita o enriquecimento de conjuntos de dados a serem disponibilizados em plataformas abertas, especialmente sistemas CRIS nacionais.

**Palavras-chave:** bases de dados científicas; interoperabilidade de dados; tratamento de dados; sistemas CRIS.

**Abstract:** It describes a data integration model between the Lattes Platform and OpenAlex by cross-referencing persistent identifiers of authors and publications. More than 1,166,000 DOI and 250,000 ORCID records from all Lattes platform curricula were listed. A cloud computing infrastructure on Amazon Web Services was used to process more than 250 million publication records extracted from OpenAlex. The lists of DOIs and ORCIDs from Lattes were cross-referenced with the publication data from OpenAlex, identifying 98.8% of the DOI and 52.2% of the ORCID records. The final dataset

resulting from the two cross-references was deduplicated and loaded into the BrCris platform, enriching and reducing inconsistencies in data related to the Brazilian scientific ecosystem. It was found that the integration model used, based on persistent identifiers, enables the enrichment of datasets to be made available on open platforms, especially national CRIS systems.

**Keywords:** Scientific databases; Data interoperability; Data preparation; CRIS systems;

## 1 INTRODUÇÃO

Documentos técnicos e científicos são indexados em bases de dados e repositórios, garantindo a preservação, a visibilidade e a recuperação da informação. Outra função das fontes especializadas é a possibilidade de geração de indicadores de ciência e tecnologia (Waltman, 2016), representando o padrão e o desempenho de tipologias documentais, de autoria e vinculação institucional, de escopo temático e geográfico, impacto e citação, inovação, visualização e acesso, entre outras variáveis relativas à produção científica (Kong *et al*., 2019).

Em nível internacional, o CrossRef, OpenCitations, OpenAIRE Research Graph, o Latindex e o DOAJ são alguns exemplos de tradicionais fontes abertas de informação em CT&I. Recentemente, o OpenAlex vem despontando como relevante fonte aberta de dados sobre a produção científica mundial, sobretudo devido a sua ampla cobertura, ao nível de abertura e de tratamento e padronização de dados (Priem *et al*., 2022).

No Brasil, se destacam as plataformas Lattes, os dados abertos da Capes, o Portal Brasileiro de Publicações e Dados Científicos em Acesso Aberto (Oasisbr) (Gibbon *et al*., 2023) e a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) e, mais recentemente, o BRCris (Vidal *et al*., 2023).

Integrar e analisar conjuntos de dados destas e de outras fontes é um desafio que exige um grande poder computacional (Puuska *et al*., 2020). A capacidade de armazenamento necessária ultrapassa a que um sistema gerenciador de banco de dados tradicional consegue suportar, exigindo soluções avançadas, tais como a infraestrutura de um lago de dados (*data lake*) (Giebler *et al*., 2019) com processamento em nuvem, aliado ao uso de técnicas de análise de dados e de identificadores persistentes, com processos de deduplicação e desambiguação de registros (Dappert *et al*., 2017).

Nesse contexto, surge a seguinte pergunta de pesquisa: como a integração de dados entre a Plataforma Lattes e o OpenAlex pode melhorar a consistência e a confiabilidade dos registros de publicações indexadas na plataforma BRCris?

Visando a responder essa questão, este trabalho tem como objetivo apresentar e avaliar um modelo de integração de dados entre a Plataforma Lattes e o OpenAlex, utilizando identificadores persistentes de publicações (DOI) e de autores (ORCID) para o cruzamento de registros de publicações entre as duas fontes. Grandes conjuntos de dados relativos à produção científica nacional são extraídos das duas fontes. O armazenamento e processamento é realizado em um lago de dados científicos com infraestrutura computacional em nuvem na Amazon Web Services. O conjunto final é resultante do modelo é analisado, visando a melhorar a consistência e a confiabilidade dos registros incorporados no projeto BRCris.

## 2 REVISÃO TEÓRICA

### 2.1 Sistemas CRIS

Sistemas CRIS (*Current Research Information Systems*) são plataformas que agregam, organizam e disponibilizam dados relacionados à pesquisa em CT&I, usualmente no contexto de ecossistemas nacionais ou regionais. Essas plataformas abrangem diversas entidades, desde pesquisadores e publicações até informações sobre projetos e financiamentos (Van Leeuwen *et al*., 2019).

Sistemas CRIS são relevantes para a gestão de dados acadêmicos e científicos. Permitem a coleta, organização e análise de informações relacionadas a pesquisas, facilitando o acesso e a disseminação de conhecimento. Por meio de um CRIS, instituições acadêmicas podem acompanhar o progresso de projetos, avaliar a produtividade de pesquisadores e garantir a transparência nos processos de financiamento e publicação. Isso contribui para a integridade científica e a otimização dos recursos disponíveis (Svertsen, 2019).

Os sistemas CRIS desempenham um papel de destaque na promoção da colaboração entre pesquisadores e instituições. Suportam ainda a criação de métricas e indicadores de avaliação e de desempenho, fundamentais para a tomada de decisões estratégicas e políticas de pesquisa. Assim, os sistemas CRIS não apenas melhoram a eficiência da gestão das

atividades científicas, mas também potencializam o impacto, a visibilidade e a distribuição de recursos para pesquisa (Puuska *et al*., 2020).

Diversos países vêm implementando sistemas CRIS e promovendo a integração em plataformas regionais. Na Europa, destacam-se países como Alemanha, Reino Unido, França, Espanha, Holanda, Suécia e Itália. Na América Latina, a plataforma La Referencia integra doze plataformas regionais, incluindo o CRIS brasileiro, BrCris.

## 2.2 BRCris

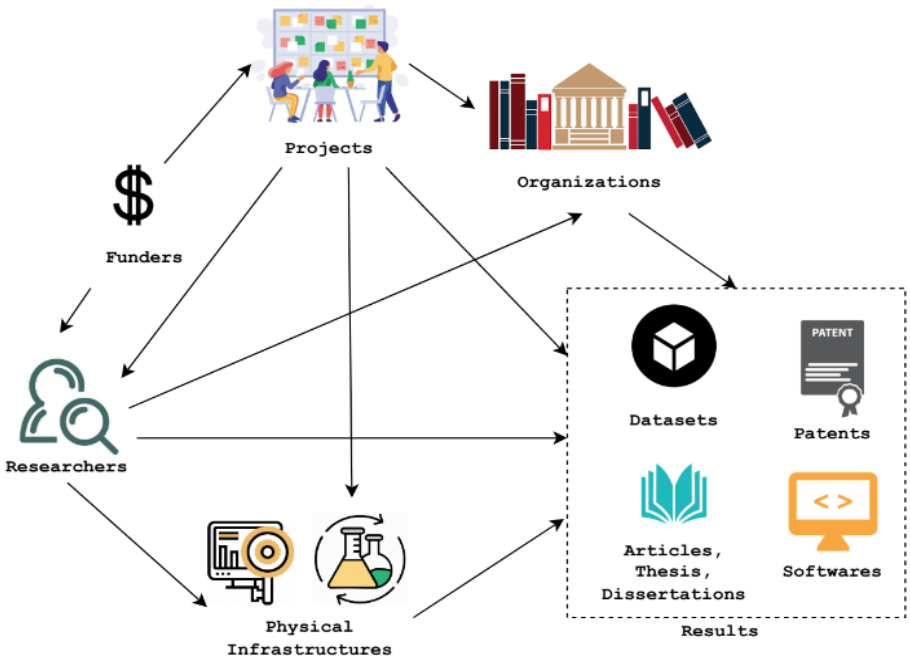
O Ecossistema de Informação da Pesquisa Científica Brasileira (BrCris), desenvolvido e lançado em 2023 pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), é uma plataforma computacional de integração e prospeção de dados científicos que visa estabelecer um modelo organizacional único de informação científica de todo o ecossistema de pesquisa nacional (Vidal *et al*., 2023).

O BrCris integra dados e informações de diferentes fontes, proporcionando análises amplas e consolidadas da produção científica nacional. Baseado em um modelo relacional, as principais entidades abrangidas são pesquisadores, instituições, publicações (teses, dissertações, artigos), patentes e programas de pós-graduação (Vidal *et al*., 2023).

A integração de dados provenientes de diversas fontes é um dos aspectos fundamentais do BrCris. Ao coletar e consolidar dados de fontes como a Plataforma Lattes, Diretório de Instituições, Diretório de Grupos de Pesquisa, entre outros, a plataforma cria um repositório amplo e atualizado, refletindo com significativa precisão o ecossistema de produção científica nacional. Uma abordagem integrada como a proposta facilita o acesso a informações mais abrangentes e confiáveis, contribuindo para uma compreensão mais completa do ecossistema de pesquisa no Brasil.

Além de simplificar a coleta e integração de dados, a Plataforma BrCris oferece ferramentas de visualização, indicadores, relatórios e outros recursos. Através de visualizações de dados intuitivas e painéis de indicadores personalizáveis, os pesquisadores podem explorar tendências, identificar colaborações e avaliar o impacto da ciência brasileira de maneira mais precisa e abrangente. A Plataforma BrCris emerge, portanto, como uma ferramenta importante para os estudos métricos no contexto da produção científica nacional e regional (Dias *et al*., 2021).

**Figura 1 –** Entidades e relacionamentos do BrCris




Fonte: Vidal *et al*. (2023).

O BrCris é alimentado por diversas fontes, incluindo repositórios, bibliotecas digitais de teses e dissertações, revistas eletrônicas de acesso aberto, bem como livros e capítulos de livros disponíveis no OasisBr (Gibbon *et al*., 2023). Visa, dessa forma, estabelecer um modelo unificado de organização da informação científica, abrangendo todo o ecossistema de pesquisa no Brasil (Figura 1).

Uma das principais fontes de dados do BrCris é a Plataforma Lattes do CNPq, um amplo sistema que reúne informações acadêmicas e profissionais de pesquisadores, cientistas e acadêmicos vinculados a instituições do Brasil (Dias *et al*., 2021).

## 2.3 Plataforma Lattes

A Plataforma Lattes é uma base de dados com mais de 8.6 milhões de currículos. O pesquisador declara a sua formação, produção acadêmica, participação em congressos e projetos, premiações acadêmicas, entre outras atividades de natureza acadêmica e científica. É a principal plataforma de currículo para pesquisadores brasileiros, utilizada como fonte oficial em processos seletivos, concursos para docentes, editais de concessão de bolsas e de

financiamentos de pesquisa, entre outros procedimentos de avaliação acadêmica e científica no Brasil (Dias *et al*., 2023).

Agências governamentais vêm se empenhando na criação de serviços de interoperabilidade entre o ORCID, a Plataforma Lattes, repositórios científicos de acesso aberto e plataformas de financiamento de pesquisas.

Além disso, o Lattes é uma fonte de dados para pesquisas quantitativas sobre a produção científica, artística e tecnológica brasileira. A sua maior abrangência representa uma vantagem em comparação ao uso das bases de dados tradicionalmente utilizadas em estudos bibliométricos e cientométricos, tais como a Web of Science, a Scopus e a Dimensions.

Por outro lado, o uso de dados da Plataforma Lattes em estudos em larga escala apresenta diversos desafios, tais como a ausência de padronização de determinados campos, erros de digitação em campos textuais (título, resumo e palavras-chave), o atraso na inserção de dados de novas publicações por parte dos pesquisadores, entre outros.

Diante disso, o uso dos dados do Lattes para fins de pesquisas por ser maximizado a partir do cruzamento de dados com outras fontes, resultando em conjuntos com maior nível de tratamento de confiabilidade.

## 3 PROCEDIMENTOS METODOLÓGICOS

Foi realizada a carga do conjunto de dados de trabalhos (*works*) do OpenAlex em uma infraestrutura em nuvem da Amazon Web Services (AWS), com gerenciamento por meio de uma arquitetura *serverless* e x86_64 e processamento de dados em *cluster* Spark com a ferramenta EMR.

Devido ao tamanho do conjunto, aproximadamente 3TB, e da complexidade do modelo de dados, com dezenas de campos de metadados, foi utilizado um *script* Python para selecionar os campos desejados e ignorar os desnecessários, especialmente campos de estrutura complexa, como por exemplo o '*abstract_inverted_index*' e o '*n-grams*'. Essa técnica resultou em um conjunto menor e de mais simples processamento, contendo apenas campos fundamentais para esta e as pesquisas decorrentes do mesmo conjunto.

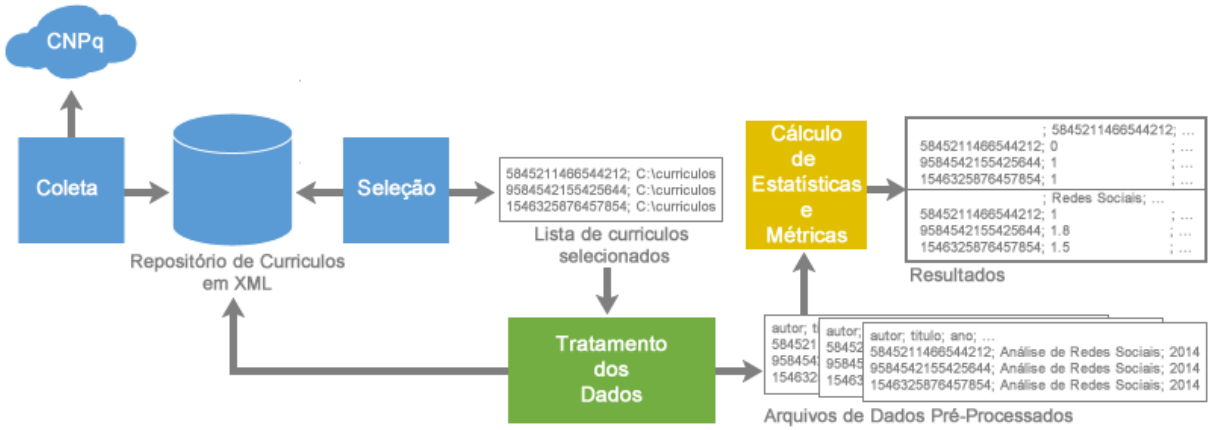
Os dados processados foram armazenados em um lago de dados científicos, no formato *json (JavaScript Object Notation)* nativo adotado pelo OpenAlex (Carvalho Segundo *et al*., 2023).

Simultaneamente, utilizou-se o *framework* LattesDataXplorer (Dias, 2016) para extrair dados de currículos da Plataforma Lattes. Este framework compreende um conjunto de técnicas e métodos para coletar, selecionar, processar e analisar dados presentes em currículos armazenados na plataforma. A Figura 2 apresenta uma visão geral do LattesDataXplorer.

O LattesDataXplorer é responsável por englobar todo o conjunto de técnicas e métodos para a coleta, tratamento e análise dos dados curriculares cadastrados na Plataforma Lattes. Ele é composto por componentes responsáveis por todo o processo de coleta e tratamento dos dados.

O módulo de coleta é executado para extrair os currículos registrados na Plataforma Lattes. Nesta fase, uma solicitação é feita diretamente à plataforma, permitindo a extração e o armazenamento local dos currículos em formato XML. Com os currículos baixados armazenados localmente, é possível manipular os dados de forma flexível e explorar todo o potencial oferecido por esses currículos.

**Figura 2 –** Arquitetura geral do LattesDataXplorer




Fonte: Extraído de Dias (2016).

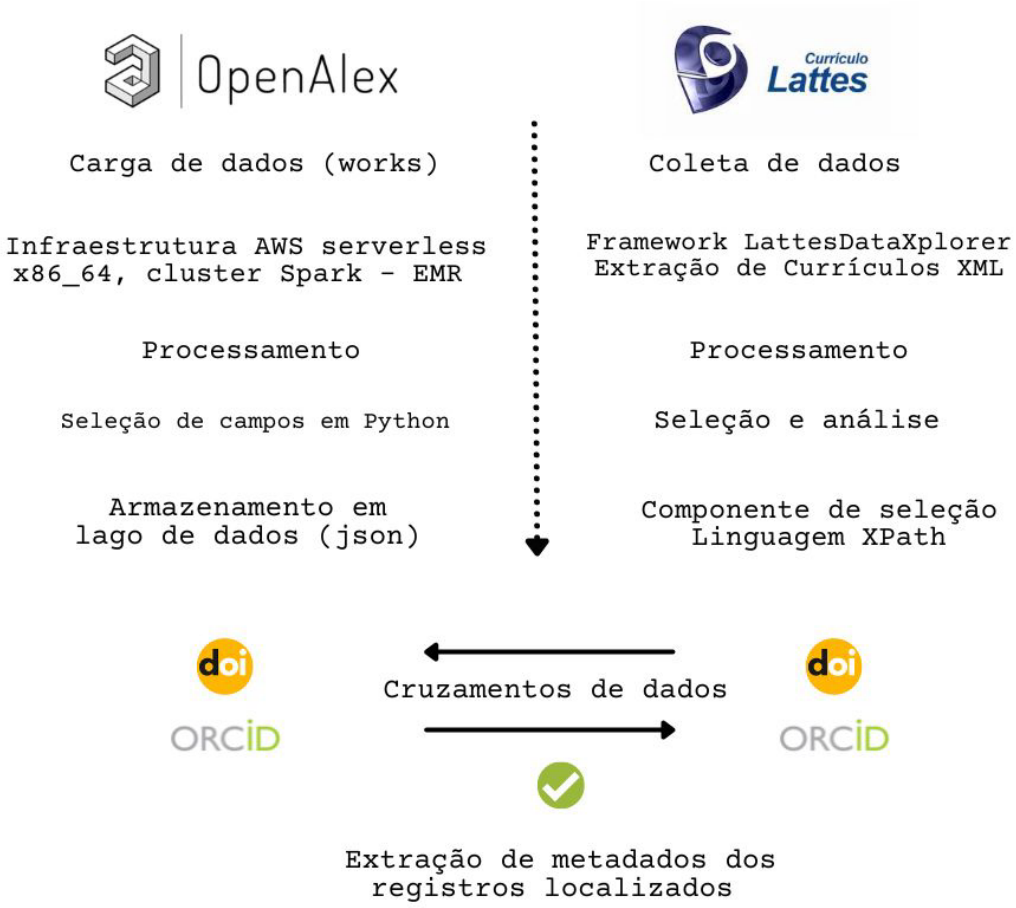
Para analisar grupos específicos de perfis, como professores de programas de pós-graduação ou de uma determinada instituição, utiliza-se o componente de Seleção para formar subgrupos baseados nas informações presentes nos registros. A linguagem de consulta XPath (*XML Path Language*) é empregada para realizar a tarefa de seleção, permitindo a construção de expressões que processam documentos XML de forma semelhante ao uso de

expressões regulares. Assim, é possível agrupar um conjunto de perfis com base em parâmetros predeterminados.

Do conjunto de currículos foram extraídos 2.614.670 registros DOI de publicações cadastradas em todos currículos. Esses registros foram padronizados no mesmo formato utilizado pelo OpenAlex, que adota o URL completo. Os DOIs padronizados foram então deduplicados, resultando em uma lista de 1.166.231 registros únicos.

Além disso, todos os registros ORCID dos indivíduos com currículos Lattes foram extraídos, obtendo-se uma lista de 254.805 registros distintos. As duas listas, de DOIs e de ORCID, foram então cruzadas com o conjunto de mais de 250 milhões de trabalhos (*works*) do OpenAlex na infraestrutura em nuvem, extraindo-se os metadados dos registros localizados (Figura 3).

**Figura 3 –** Fluxo de integração dos conjuntos de dados analisados




Fonte: Elaborado pelos autores (2024).

Os dados com identificadores persistentes, como as publicações com DOI e os indivíduos que registraram seus ORCIDs, são coletados dos currículos da Plataforma Lattes. Após a preparação dos dados da OpenAlex, esses identificadores são utilizados para consultas no conjunto de dados da OpenAlex, viabilizando a recuperação e a integração dos dois conjuntos de dados, que então são deduplicados gerando um conjunto final de dados.

## 4 RESULTADOS

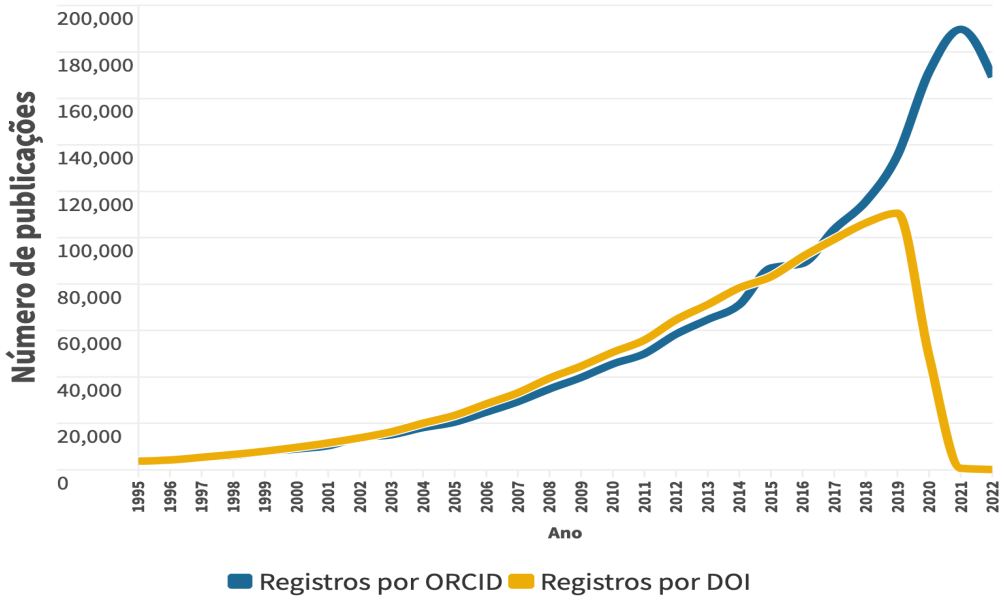
Primeiramente foram localizados 1.824.912 registros de trabalhos no OpenAlex a partir da lista de ORCIDs de pesquisadores com currículo Lattes. Além disso, foram localizados 1.152.582 registros de trabalhos a partir da lista de 1.166.231 DOIs deduplicados cadastrados no Lattes, o que corresponde a um solapamento de 98,8% de registros DOI entre as duas fontes. Esse resultado indica uma excelente cobertura do OpenAlex em relação as publicações com DOI brasileiras, provavelmente decorrente do uso do CrossRef como fonte de metadados de publicações com DOI pelo OpenAlex.

Ambos os conjuntos de dados foram extraídos em formato JSON, nativo do OpenAlex, e deduplicados por meio identificador de trabalhos do OpenAlex (*word_id*). O conjunto final restou composto por 2.350.894 registros distintos de publicações. Esse número corresponde a aproximadamente 75% de todos os registros de trabalhos do OpenAlex que tem pelo menos um autor vinculado a uma instituição brasileira (*country_code=br*).

### 4.1 Publicações por ano

Da análise dos anos do conjunto final de publicações, foi possível perceber que a curva de aumento de publicações identificadas por ano é similar nos dois conjuntos até 2020 (Figura 4).

**Figura 4 -** Registros de publicações por ano identificadas a partir dos dois identificadores



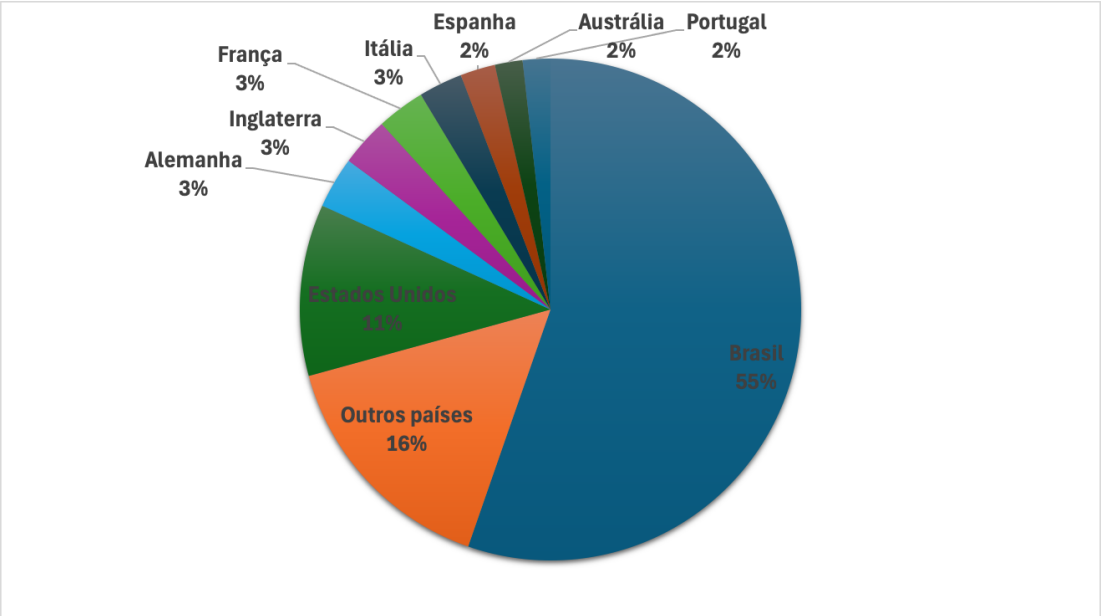

Fonte: Elaborado pelos autores (2024).

Após esse período, contudo, o conjunto extraído por meio dos registros ORCID segue em crescimento significativo até 2023, enquanto que os trabalhos recuperados por meio de DOI tem queda acentuada. Isso pode ser resultado da demora dos pesquisadores em atualizar seus currículos com os dados dos trabalhos recentemente publicados.

### 4.2 Países de origem de autores

Da análise dos países de origem dos autores das publicações do conjunto final identificou-se mais da metade de autoria brasileira, seguido de autores dos Estados Unidos, Inglaterra, França, Itália, Espanha, Austrália e Portugal. Outros países são a origem de 16% dos autores das publicações (Figura 5).

**Figura 5 -** Países de origem dos autores das publicações identificadas




**Fonte:** Elaborado pelos autores (2024).

A presença de registros de pesquisadores estrangeiros no conjunto de dados decorre, provavelmente, de dois fatores. O primeiro é a presença de pesquisadores estrangeiros com vínculos com instituições do Brasil como titulares de currículos Lattes. O segundo fator referente às publicações de pesquisadores brasileiros em coautoria com autores estrangeiros, que não possuem currículo Lattes, mas foram recuperados no OpenAlex.

De toda forma, considera-se esses dados de autoria relevantes no contexto do BrCris para fins de análise de redes de coautoria, colaboração internacional e enriquecimento de registros ORCID não obtidos no Lattes ou em outras fontes.

### 4.3 Registros ORCID identificados

Resultado importante foi obtido a partir da análise dos registros ORCID identificados no conjunto final. Foram identificados 133.070 autores com ORCID a partir da lista de 254.805 ORCIDs extraídos do Lattes, o que representa 52,2% de solapamento de ORCID entre as duas fontes.

Por outro lado, foram identificados mais 634.332 registros ORCID no conjunto final, vinculados aos coautores dos titulares de currículos Lattes. Esse conjunto ainda não foi devidamente analisado, mas as hipóteses é que se tratam de registros de coautores sem

currículo Lattes, sobretudo estrangeiros, e coautores com currículo Lattes, mas que não informaram o ORCID nesta plataforma.

### 4.4 Revistas identificadas

Uma contribuição significativa para enriquecer os dados das publicações nos currículos Lattes refere-se às revistas científicas nas quais esses artigos foram publicados. Como determinados campos relacionados às revistas científicas são de preenchimento livre na Plataforma, aumenta a possibilidade de inconsistências nos dados extraídos. Portanto, ao considerar conjuntos de dados consistentes e validados como os da OpenAlex, é possível enriquecer o conjunto de registros da Plataforma Lattes de forma eficaz.

Nesse sentido, foram identificados 46.085 registros de ISSNs distintos vinculados às publicações identificadas no conjunto final. Os dados dessas ISSNs foram tratados para fins de enriquecimento da entidade 'Revistas' no BrCris, que apresenta uma quantidade significativa de inconsistências nos dados oriundos do Lattes em razão de erros de preenchimento, múltiplos ISSNs para uma mesma revista, entre outras incongruências.

A Tabela 1 apresenta as revistas com a maior quantidade de artigos vinculados, independentemente de sua data de publicação.

**Tabela 1 -** Revistas com maior quantidade de publicações identificadas

| ISSN | Título | Publicações |
|---|---|---|
| 1932-6203 | PloS one | 8.700 |
| 0103-8478 | Ciência rural | 5.654 |
| 0102-311X | Cadernos de Saúde Pública | 4.613 |
| 1413-8123 | Ciência & Saúde Coletiva | 4.212 |
| 0100-4042 | Química Nova | 3.964 |
| 1516-3598 | Revista Brasileira de Zootecnia | 3.731 |
| 0103-5053 | Journal of the Brazilian Chemical Society | 3.579 |
| 0102-0935 | Arquivo Bras. de Medicina Veterinária e Zootecnia | 3.479 |
| 1679-0359 | Semina. Ciências Agrárias | 3.229 |
| 0100-204X | Pesquisa Agropecuária Brasileira | 3.216 |
| 1678-4227 | Arquivos de Neuro-Psiquiatria | 3.039 |
| 0031-9007 | Physical review letters | 2.970 |
| 1175-5326 | Zootaxa | 2.935 |
| 1678-4170 | Arquivos Brasileiros de Cardiologia | 2.914 |
| 2045-2322 | Scientific reports | 2.809 |
| 1984-0446 | Revista Brasileira de Enfermagem | 2.744 |

| ISSN | Título | Publicações |
|---|---|---|
| 0034-8910 | Revista de Saúde Pública | 2.686 |
| 0250-9776 | Diffusion and Defect Monograph Series | 2.661 |
| 0255-5476 | Materials Science Forum | 2.661 |
| 1422-6375 | Journal of Metastable and Nanocrystalline Materials | 2.661 |
| 1550-2368 | Physical Review D | 2.635 |
| 0074-0276 | Memórias do Instituto Oswaldo Cruz | 2.539 |
| 0100-879X | Brazilian Journal of Medical and Biological Research | 2.531 |
| 1098-0121 | Physical Review B | 2.450 |
| 0100-0683 | Revista Brasileira de Ciência do Solo | 2.417 |
| 0037-8682 | Revista da Sociedade Brasileira de Medicina Tropical | 2.397 |

**Fonte:** Elaborado pelos autores (2024).

O periódico com a maior quantidade de artigos vinculados é o "PloS one" (ISSN 1932-6203), com 8.700 publicações. Em seguida, destacam-se "Ciência Rural" (ISSN 0103-8478) com 5.654 publicações, "Cadernos de Saúde Pública" (ISSN 0102-311X) com 4.613 publicações, e "Ciência & Saúde Coletiva" (ISSN 1413-8123) com 4.212 publicações. Outros periódicos de destaque incluem "Química Nova" (ISSN 0100-4042) com 3.964 publicações e a "Revista Brasileira de Zootecnia" (ISSN 1516-3598) com 3.731 publicações.

No total, a tabela lista 25 revistas, abrangendo diversas áreas do conhecimento, como saúde, química, agricultura e ciência dos materiais, que apresentam um panorama das principais fontes de divulgação da produção nacional.

### 4.5 Registros de instituições identificados

Da análise do conjunto final foi possível também enriquecer os dados relativos à vinculação institucional dos autores. Isso porque na plataforma Lattes é informado apenas o nome da instituição ao qual o titular do currículo é vinculado, sem campo específico para o identificador persistente de organizações ROR (*Research Organization Registry*).

Foi possível identificar os registros ROR de 30.690 instituições indicadas pelos autores como vínculo institucional no momento de publicação (Tabela 2). Entre as instituições com maior quantidade de publicações vinculadas estão a Universidade de São Paulo (USP), a Universidade Estadual Paulista (Unesp), a Universidade Estadual de Campinas (UNICAMP) e a Universidade Federal do Rio Grande do Sul (UFRGS).

**Tabela 2 -** ROR das instituições com maior quantidade de registros

| ROR | Nome da instituição | nº de publicações |
|---|---|---|
| 036rp1748 | Universidade de São Paulo | 410.537 |
| 00987cb86 | Universidade Estadual Paulista | 171.157 |
| 04wffgt70 | Universidade Estadual de Campinas | 147.485 |
| 041yk2d64 | Universidade Federal do Rio Grande do Sul | 134.922 |
| 03490as77 | Universidade Federal do Rio de Janeiro | 121.720 |
| 0176yjw32 | Universidade Federal de Minas Gerais | 116.454 |
| 02feahw73 | Centre National de la Recherche Scientifique | 98.954 |
| 02k5swt12 | Universidade Federal de São Paulo | 94.218 |
| 01b78mz79 | Universidade Federal de Santa Maria | 69.259 |
| 041akq887 | Universidade Federal de Santa Catarina | 68.125 |
| 00892tw58 | University of Adelaide | 66.021 |
| 04jhswv08 | Fundação Oswaldo Cruz | 65.405 |
| 05syd6y78 | Universidade Federal do Paraná | 60.774 |
| 00qdc6m37 | Universidade Federal de São Carlos | 56.996 |
| 03srtnf24 | Universidade Federal do Ceará | 54.028 |
| 0409dgb37 | Universidade Federal de Viçosa | 53.056 |
| 047908t24 | Universidade Federal de Pernambuco | 50.326 |
| 02xfp8v59 | Universidade de Brasília | 47.387 |
| 0482b5b22 | Brazilian Agricultural Research Corporation | 45.757 |
| 0198v2949 | Universidade do Estado do Rio de Janeiro | 41.689 |
| 04bqqa360 | Universidade Estadual de Maringá | 41.459 |
| 02rjhbb08 | Universidade Federal Fluminense | 40.889 |
| 01ggx4157 | European Organization for Nuclear Research | 38.463 |
| 039shy520 | Institute of High Energy Physics | 36.875 |
| 04wn09761 | Universidade Federal do Rio Grande do Norte | 36.870 |
| 05msy9z54 | Universidade Federal de Pelotas | 35.879 |
| 03k3p7647 | Universidade Federal da Bahia | 35.296 |
| 00p9vpz11 | Universidade Federal da Paraíba | 33.730 |
| 0039d5757 | Universidade Federal de Goiás | 33.043 |
| 01585b035 | Universidade Estadual de Londrina | 32.611 |

**Fonte:** Elaborado pelos autores (2024).

Esses registros são importantes pois irão viabilizar futuros processos de desambiguação e de deduplicação de nomes de instituições no BrCris.

## 5 CONSIDERAÇÕES FINAIS

Modelos de integração com grandes quantidades de dados de diferentes fontes precisam lidar com desafios técnicos e computacionais, que incluem a capacidade das

infraestruturas, a compatibilidade entre diferentes padrões e a inconsistência de dados. Esses modelos, contudo, são indispensaveis para a gestão de amplas infraestruturas de dados científicos, tais como os sistemas CRIS.

O modelo de integração das Plataformas Lattes e OpenAlex por meio dos identificadores persistentes de publicações (DOI) e de autores (ORCID) resultou em uma significativa correspondência entre registros das duas fontes. Isso sugere que o modelo utilizado foi adequado, pois permitiu o enriquecimento do conjunto final de dados por meio de inferências com baixa margem de erro.

Os resultados demonstram que a abordagem proposta, de integração entre as duas fontes por meio dos dois identificadores persistentes foi exitosa, pois manteve o vínculo inequívoco com a plataforma Lattes (fonte principal de identificação da entidade pessoa no BRCris), agregando os registros com uma camada enriquecida e outros relacionamentos pré-definidos no modelo do OpenAlex (coautoria, fontes das publicações, vinculação institucional).

A utilização de uma infraestrutura em nuvem facilitou o processo computacionalmente mais caro do modelo, de carga e processamento de todo conjuntos de trabalhos do OpenAlex, com mais de 250M de registros e dezenas de campos de metadados.

Ao utilizar avançadas de computação em nuvem e técnicas de análise de dados, foi possível superar os desafios de armazenamento e processamento de grandes quantidades de dados de diferentes fontes e em diferentes padrões, resultando em um conjunto final de dados enriquecido e consistente.

Como atividade futura, pretende-se analisar o conjunto de dados a partir dos ORCIDs, identificando os pesquisadores vinculados, visando o enriquecimento da entidade 'Pessoa' no BrCris, considerando o baixo percentual de pesquisadores do Lattes que possuem ou informam seus registros de ORCID na plataforma.

## REFERÊNCIAS

CARVALHO SEGUNDO, W. L. R. de; PINTO, A. L.; CANTO, F. L. do; NEUBERT, P. Projeto Laguna: infraestrutura de um lago de dados científicos em acesso aberto. **BiblioCanto**, Natal, RN, v. 9, n. 2, p. 133–138, 2023. DOI: 10.21680/2447-7842.2023v9n2ID33825. Disponível em: https://periodicos.ufrn.br/bibliocanto/article/view/33825. Acesso em: 10 jul. 2024.

GIBBON, C. A.; MIRANDA, A. C. D.; CARVALHO SEGUNDO, W. L. R; SILVA, L. S; MORAES, M. H. Ciência Aberta brasileira e o portal Oasisbr: análise dos descritores da produção científica. **BiblioCanto**, Natal, RN, v. 9, n. 2, 2023. Disponível em: https://periodicos.ufrn.br/bibliocanto/article/view/34331. Acesso em: 14 jul. 2024.

GIEBLER *et al*. Leveraging the Data Lake: Current State and Challenges. *In*: BIG DATA ANALYTICS AND KNOWLEDGE DISCOVERY**:** DaWaK 2019. **Lecture Notes in Computer Science**. [*S.l.*]: Springer, Cham, 2019. Disponível em: https://doi.org/10.1007/978-3-030-27520-4_13. Acesso em: 10 jul. 2024.

DAPPERT, A., FARQUHAR A., KOTARSKI, R.; HEWLLET, K. Connecting the persistent identifier ecosystem: building the technical and human infrastructure for open research, **Data science journal,** n. 16, v. 28, 2017. Disponível em: https://doi.org/10.5334/dsj-2017-028. Acesso em: 10 jul. 2024.

DIAS, T. M. R. **Um estudo sobre a produção científica brasileira a partir de dados da Plataforma Lattes**. 2016. 181 p. (Tese de Doutorado) - Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016.

DIAS, T. M. R; MENA-CHALCO, J. P.; CARVALHO SEGUNDO, W. L. R. de; PINTO, A. L.; MOREIRA, T. H. J. BRCRIS: plataforma para integração, análises e visualização de dados técnicos-científicos. **Informação & Informação**, Londrina, v. 27, n. 3, p. 622–638, 2023. DOI: 10.5433/1981-8920.2022v27n3p622. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/47215. Acesso em: 10 jul. 2024.

HENNING, P. C. *et al*. Go fair e os princípios fair: o que representam para a expansão dos dados de pesquisa no âmbito da ciência aberta. **Em Questão**, Porto Alegre, v. 25, n. 2, 2019. doi: http://dx.doi.org/10.19132/1808-5245252.389-412.

KONG, X.; SHI, Y.; YU, S.; LIU, J.; XIA, F. Academic social networks: modeling, analysis, mining and applications. **Journal of network and computer applications**, [S.l.], v. 132, p. 86-103, 2019. Disponível em: https://dl.acm.org/doi/abs/10.1016/j.jnca.2019.01.029. Acesso em: 27 mar. 2023.

PRIEM, J., PIWOWAR, H., & ORR, R. OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. **ArXiv**, 2022. Disponível em: https://doi.org/10.48550/arXiv.2205.01833. Acesso em: 10 mar. 2024.

PUUSKA, Hanna-Mari; NIKKANEN, Joonas; ENGELS, Tim; GUNS, Raf; IVANOVIĆ, Dragan; PÖLÖNEN, Janne. Integration of national publication databases – towards a high-quality and comprehensive information base on scholarly publications in Europe. **ITM Web of Conferences**, v. 33, p. 02001, 2020. Disponível em: https://doi.org/10.1051/itmconf/20203302001. Acesso em: 7 jul. 2024.

SIVERTSEN, G. Developing Current Research Information Systems (CRIS) as data sources for studies of research. *In*: GLÄNZEL, W.; MOED, H. F.; SCHMOCH, U.; THELWALL, M. (ed.). **Springer handbook of science and technology indicators**. [*S.l.*]: Springer Handbooks,

Springer, Cham, 2019. p. 667-683. Disponível em: https://doi.org/10.1007/978-3-030-02511-3_25. Acesso em: 27 jun. 2024.

VAN LEEUWEN, T. N.; VAN WIJK, E.; WOUTERS, P. F. Bibliometric analysis of output and impact based on CRIS data: a case study on the registered output of a Dutch university. **Scientometrics**, v. 106, n. 1, p. 1–16, 2016. Disponível em: https://scholarlypublications. universiteitleiden.nl/handle/1887/78132. Acesso em: 7 jul. 2024.

VIDAL, L. H. C.; CARVALHO SEGUNDO, W. L. R.; MENA-CHALCO, J. P.; GABRIEL JUNIOR, R. F. BrCris: Visualização de Indicadores Científicos. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 23., 2023, Aracaju. **Anais [...].** Aracaju: ANCIB, 2023. Disponível em: https://www.ancib.org.br/enancib/index.php/enancib/xxxiiienancib/ paper/viewFile/1893/1398. Acesso em: 10 jul. 2024.

WALTMAN, L. A review of the literature on citation impact indicators. **Journal of informetrics**, v. 10, n. 2, p. 365–391, maio 2016. Disponível em: http://dx.doi.org/10.1016/j. joi.2016.02.007. Acesso em: 10 jul. 2024.