



10° EBBC

Encontro Brasileiro de
Bibliometria e Cientometria



CONFIABILIDADE DE DADOS CIENTÍFICOS EM INFRAESTRUTURAS DE PESQUISA: estratégias de qualificação no ecossistema BrCris

Thiago Magela Rodrigues Dias

<https://orcid.org/0000-0001-5057-9936>

thiagomagela@cefetmg.br

Centro Federal de Educação Tecnológica
de Minas Gerais (CEFET-MG)

Fábio Lorensi do Canto

<https://orcid.org/0000-0002-8338-1931>

fabio.lc@ufsc.br

Universidade Federal de Santa Catarina
(UFSC)

Washington L. R. Carvalho Segundo

<https://orcid.org/0000-0003-3635-9384>

washingtonsegundo@ibict.br

Instituto Brasileiro de Informação em
Ciência e Tecnologia (Ibict)

Resumo:

A crescente disponibilidade de grandes volumes de dados científicos tem impulsionado o desenvolvimento de infraestruturas digitais voltadas à integração, organização e análise da produção científica. No entanto, a heterogeneidade das fontes de informação e as inconsistências nos metadados representam desafios relevantes para a confiabilidade das análises bibliométricas e cientométricas. Neste contexto, o Ecossistema de Informação da Pesquisa Científica Brasileira (BrCris) surge como uma infraestrutura nacional destinada à integração de dados provenientes de diferentes sistemas científicos. Este trabalho analisa os principais desafios relacionados à qualidade dos dados integrados ao BrCris e apresenta estratégias metodológicas voltadas à deduplicação, desambiguação e validação de registros científicos.

Palavras-Chave: BrCris. Qualidade dos dados. Deduplicação. Desambiguação.

EIXO TEMÁTICO:

Técnicas, Ferramentas e
Infraestruturas

MODALIDADE:

Resumo expandido

DOI: 10.22477/x.ebbc.1096

COMO CITAR ESTE ARTIGO:

DIAS, Thiago Magela Rodrigues; CANTO, Fabio Lorensi do; SEGUNDO, Washington L. R. Carvalho. Confiabilidade de dados científicos em infraestruturas de pesquisa: estratégias de qualificação no ecossistema BrCris. In: EBBC - Encontro Brasileiro de Bibliometria e Cientometria. 10. **Anais [...]**. Curitiba, 2026. DOI: 10.22477/x.ebbc.1096. Disponível em: <https://doi.org/10.22477/x.ebbc.1096>.



1 INTRODUÇÃO

A produção científica atualmente é fortemente mediada por infraestruturas digitais responsáveis pela coleta, organização e disseminação de dados de pesquisa. Nesse contexto, a disponibilidade de grandes volumes de dados científicos possibilita o desenvolvimento de análises bibliométricas e cientométricas cada vez mais sofisticadas, contribuindo para a compreensão das dinâmicas de produção, circulação e impacto do conhecimento científico.

Entretanto, a crescente integração de múltiplas fontes de informação científica também tem ampliado desafios relacionados à qualidade, consistência e confiabilidade dos dados utilizados nesses processos analíticos. Problemas como duplicidade de registros, inconsistências nos metadados, ausência de identificadores persistentes e heterogeneidade na representação de autores e instituições podem comprometer significativamente a precisão das análises e gerar distorções nos indicadores científicos.

No contexto latino-americano, o Brasil tem uma parcela relevante na produção científica internacional, especialmente em nichos estratégicos associados à sua economia, biodiversidade e saúde pública. O país é reconhecido como líder na produção de conhecimento na América Latina (Collazo-Reyes, 2014), além de atuar como atrator de talentos e centro regional de pesquisa (Saravia e Miranda, 2004). Tal protagonismo foi acompanhado por iniciativas estruturantes que buscam aprimorar a governança da informação científica.

Diante desse cenário, o desenvolvimento do Ecossistema de Informação da Pesquisa Científica Brasileira (BrCris¹) representa um esforço estratégico voltado à integração e organização de dados científicos provenientes de diferentes infra-

¹ <https://brcris.ibict.br>

estruturas informacionais, tais como a Plataforma Lattes, o OpenAlex e o Oasisbr (Souza *et al.*, 2025; Pinto *et al.*, 2021). Ao consolidar informações sobre pesquisadores, instituições, publicações e grupos de pesquisa, o BrCris possibilita a construção de um ambiente integrado para análise e visualização da ciência brasileira.

Contudo, a integração de dados provenientes de sistemas heterogêneos implica desafios relevantes relacionados à governança da informação científica. No âmbito do BrCris, os principais problemas de qualidade dos dados podem ser agrupados em categorias recorrentes como Duplicidade de Registros, Inconsistência de Identificadores Persistentes, Erros de Preenchimento e Padronização e Desatualização e Incompletude de Registros.

Diante desse cenário, emerge o seguinte problema de pesquisa: como garantir a confiabilidade de dados científicos integrados a partir de múltiplas fontes heterogêneas, considerando a presença de inconsistências estruturais, duplicidades e ambiguidades nos metadados?

A relevância deste estudo justifica-se pela crescente dependência de dados científicos integrados para a realização de análises bibliométricas e cientométricas, bem como para a formulação de políticas públicas baseadas em evidências. A presença de inconsistências nos dados pode comprometer a validade dessas análises, tornando fundamental o desenvolvimento de estratégias capazes de melhorar a qualidade e a confiabilidade das informações científicas em infraestruturas nacionais.

Nesse contexto, o objetivo deste estudo é analisar de que maneira estratégias de deduplicação, desambiguação e validação de metadados podem contribuir para aumentar a confiabilidade dos dados científicos integrados ao ecossistema BrCris.

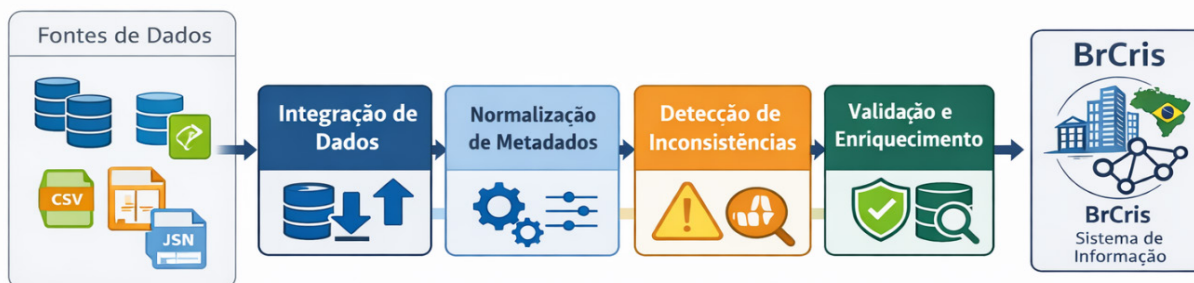
2 METODOLOGIA

Do ponto de vista científico, este estudo caracteriza-se como uma pesquisa de natureza aplicada, com abordagem qualitativa e caráter exploratório-descritivo. A investigação adota como estratégia metodológica o estudo de caso do ecossistema BrCris, com foco na análise dos processos de qualificação de dados científicos implementados na plataforma.

A unidade de análise do estudo corresponde ao pipeline de tratamento e qualificação de dados científicos do BrCris, considerando suas etapas de ingestão, normalização, detecção de inconsistências, deduplicação, desambiguação e validação de metadados. (Figura 1).

A análise foi conduzida a partir da observação sistemática dessas etapas, buscando identificar de que forma os procedimentos adotados contribuem para a melhoria da consistência, integridade e interoperabilidade dos dados científicos. Para isso, foram considerados os efeitos das estratégias implementadas sobre a redução de redundâncias, a identificação de entidades e a padronização dos metadados.

Figura 1 – Fluxo de dados do BrCris.



Inicialmente, os dados provenientes das diferentes bases são coletados por meio de rotinas automatizadas de extração e importação. Nessa etapa, os registros científicos são convertidos para um modelo comum de representação de meta-

dados, permitindo a consolidação e integração das informações em um ambiente unificado de análise.

Na etapa seguinte, realiza-se a normalização estrutural dos dados, na qual são aplicados procedimentos de padronização de campos textuais, harmonização de identificadores e consolidação de entidades informacionais. Esse processo envolve a aplicação de dicionários controlados para nomes de instituições, programas de pós-graduação e áreas do conhecimento, reduzindo variações ortográficas e inconsistências na representação das entidades.

Após a normalização inicial, são executados mecanismos de detecção de inconsistências nos registros científicos. Nessa fase são empregados algoritmos de comparação textual e técnicas de similaridade entre registros, capazes de identificar possíveis duplicidades e ambiguidades em nomes de autores, instituições e publicações. Esses algoritmos analisam atributos como títulos de trabalhos, identificadores persistentes, afiliações institucionais, dentre outros atributos, permitindo a identificação de registros potencialmente redundantes ou inconsistentes.

Uma vez identificadas as inconsistências, são aplicados processos de deduplicação e desambiguação de entidades informacionais. A deduplicação busca consolidar registros que representam uma mesma produção científica ou entidade institucional, enquanto a desambiguação tem como objetivo resolver ambiguidades associadas à identificação de autores, produções e instituições. Para isso, são utilizadas estratégias baseadas na combinação de regras heurísticas, análise de similaridade textual e verificação de identificadores persistentes como DOI, ORCID e ISSN.

Posteriormente, os registros passam por uma etapa de validação e enriquecimento dos metadados. Nesse momento, as informações são comparadas com fon-

tes externas de referência e infra estruturas internacionais de dados científicos, permitindo verificar a consistência dos registros e complementar metadados ausentes. Esse processo contribui para aumentar a interoperabilidade do sistema e aprimorar a rastreabilidade das informações científicas.

Por fim, os dados qualificados são disponibilizados para análise por meio de ferramentas, possibilitando a geração de indicadores, visualizações e redes de colaboração científica. A integração com softwares especializados de análise de redes e visualização de dados permite não apenas explorar a produção científica brasileira sob diferentes perspectivas analíticas, mas também identificar novas inconsistências que podem ser retroalimentadas ao processo de curadoria e melhoria contínua dos dados.

3 RESULTADOS

A implementação das estratégias de qualificação de dados no ecossistema BrCris permitiu avanços significativos na consistência, integridade e interoperabilidade das informações científicas integradas ao sistema. Os resultados observados evidenciam que a adoção de processos estruturados de deduplicação, desambiguação, padronização e validação de metadados contribui diretamente para a melhoria da confiabilidade das análises bibliométricas e cientométricas derivadas da plataforma.

Os resultados podem ser analisados a partir de quatro dimensões principais: (i) redução de duplicidades de registros, (ii) melhoria na identificação de pesquisadores, publicações e instituições, (iii) aumento da consistência e padronização dos metadados e (iv) ampliação da capacidade analítica por meio da integração com ferramentas bibliométricas.

i) A aplicação de algoritmos de deduplicação permitiu identificar registros redundantes provenientes da integração de diferentes bases de dados científicas. Essas duplicidades frequentemente surgem devido a variações na grafia de títulos, autores ou instituições, bem como pela ausência de identificadores persistentes. Os procedimentos de deduplicação aplicados no BrCris analisam múltiplos atributos dos registros científicos, incluindo títulos de publicações, listas de autores, identificadores persistentes e informações de afiliação institucional. A combinação de regras heurísticas e métodos automatizados de similaridade textual possibilita identificar registros potencialmente duplicados e consolidá-los em representações únicas.

Como resultado, observou-se uma redução significativa da redundância informacional na base integrada, contribuindo para a melhoria da precisão das métricas de produção científica e evitando a superestimação de indicadores de produtividade acadêmica.

ii) Outro resultado relevante refere-se ao aprimoramento da identificação de pesquisadores, publicações e instituições científicas. A integração de identificadores persistentes, como ORCID para autores, DOI para publicações e ROR para instituições, permitiu reduzir ambiguidades na representação das entidades informacionais.

A combinação dessas identificações com processos automatizados de análise de similaridade entre nomes e afiliações institucionais possibilitou identificar variantes de nomes que anteriormente eram tratadas como entidades distintas. Esse processo é particularmente relevante no contexto brasileiro, onde registros provenientes de diferentes bases de dados frequentemente apresentam variações na grafia de nomes institucionais e de autores.

A integração entre os dados da Plataforma Lattes e da base OpenAlex demonstrou resultados expressivos na melhoria da identificação de pesquisadores, permitindo ampliar a precisão na vinculação entre autores, publicações e instituições. Como consequência, as redes de colaboração científica geradas a partir do BrCris tornaram-se mais representativas das relações reais entre os atores do sistema científico.

iii) Os resultados também evidenciam melhorias significativas na padronização dos metadados integrados ao sistema. A adoção de dicionários controlados para instituições, programas de pós-graduação e áreas do conhecimento permitiu reduzir inconsistências textuais e variações semânticas nos registros científicos.

Antes da inserção definitiva no banco de dados do BrCris, os registros passam por processos de validação que comparam as informações com esses vocabulários controlados. Esse mecanismo permite identificar e corrigir automaticamente variações ortográficas, abreviações não padronizadas e inconsistências estruturais nos metadados.

iv) A integração do BrCris com ferramentas especializadas de análise bibliométrica e cientométrica representa outro resultado relevante das estratégias implementadas. A possibilidade de exportação de conjuntos de dados estruturados para ferramentas como VOSviewer, Gephi e Visão permite explorar a produção científica brasileira sob diferentes perspectivas analíticas.

Essas ferramentas possibilitam a geração de redes de colaboração científica, mapas de coocorrência de termos, análises de impacto e visualizações interativas da estrutura da ciência brasileira.

4 CONCLUSÕES

A análise desenvolvida neste estudo evidencia que a qualidade dos dados científicos constitui um elemento central para a confiabilidade das análises bibliométricas e cientométricas em infraestruturas de informação científica. Ao examinar o pipeline de qualificação de dados implementado no ecossistema BrCris, o trabalho contribui para a literatura ao explicitar um conjunto estruturado de estratégias voltadas à deduplicação, desambiguação e validação de metadados em ambientes caracterizados pela integração de múltiplas fontes heterogêneas.

Como principal avanço científico, destaca-se a sistematização dessas estratégias no contexto de uma infraestrutura nacional de informação científica, evidenciando como a combinação de técnicas automatizadas de processamento de dados com mecanismos de padronização e validação pode contribuir para o fortalecimento da consistência, integridade e interoperabilidade dos registros científicos. Ao analisar essas estratégias sob a perspectiva de suas implicações para análises bibliométricas e cientométricas, o estudo amplia a compreensão sobre os desafios e soluções associados à governança de dados científicos em larga escala.

No entanto, este trabalho apresenta algumas limitações. A análise foi conduzida a partir de uma abordagem qualitativa e exploratória, sem a incorporação de métricas quantitativas que permitam mensurar de forma precisa o impacto das estratégias adotadas sobre a qualidade dos dados.

REFERÊNCIAS

COLLAZO-REYES, Francisco. Growth of the number of indexed journals of Latin America and the Caribbean: the effect on the impact of each country. **Scientometrics**, v. 98, n. 1, p. 197-209, 2014. DOI: <https://doi.org/10.1007/s11192-013-1036-2>.



PINTO, Adilson Luiz; DIAS, Thiago Magela Rodrigues; SILVA, Vivian Santos; GOMES, Josir Cardoso; SEGUNDO, Washington Luis R. de Carvalho. BrCris como um sistema de recomendação científico-tecnológica. In: Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, 21., 2021. **Anais [...]** Rio de Janeiro: PPGCI; IBICT; UFRJ, 2021. Disponível em: <https://enancib.ancib.org/index.php/enancib/xxienancib/paper/view/335>. Acesso em: 25/02/2026.

SARAVIA, Nancy Gore; MIRANDA, Juan Francisco. Plumbing the brain drain. **Bulletin of the World Health Organization**, v. 82, p. 608-615, 2004. Disponível em: <https://www.scielo.org/article/bwho/2004.v82n8/608-615/>. Acesso em: 07/03/2026.

SEGUNDO, Washington Luis R. de Carvalho; DIAS, Thiago Magela Rodrigues; SOUZA, Marcel Garcia de. BrCris: facilitando o caminho para a Ciência Aberta através da exportação de conjuntos de dados. **Ciência da Informação**, 2025. DOI: <https://doi.org/10.18225/ci.inf.v54i2.7205>.