



10° EBBC

Encontro Brasileiro de
Bibliometria e Cientometria



ALÉM DO ORCID AUTODECLARADO NO LATTES: inferência de identificadores via DOI-OpenAlex e seus limites

Wellington Barbosa Rodrigues

<https://orcid.org/0000-0002-6465-6012>
wellington.rodrigues@usp.br
Universidade de São Paulo (USP)

Rogério Mugnaini

<https://orcid.org/0000-0001-9334-3448>
mugnaini@usp.br
Universidade de São Paulo (USP)

Cecilia Tomassini

<https://orcid.org/0000-0002-7989-841X>
ctomassini@csic.edu.uy
Universidad de la República (UDEAR)

Jesús Pascual Mena-Chalco

<https://orcid.org/0000-0001-7509-5532>
jesus.mena@ufabc.edu.br
Universidade Federal do ABC (UFABC)

Resumo:

O estudo analisa a integração entre a Plataforma Lattes e a OpenAlex para identificação de pesquisadores brasileiros, avaliando a cobertura do ORCID autodeclarado e a inferência por meio da combinação DOI-OpenAlex com heurística nominal. A partir de 307 bolsistas PQ-SR, observa-se baixa cobertura de ORCID no Lattes (45,9%), mas elevada recuperação via heurística (94,5% com ao menos um ORCID). Embora a estratégia amplie a identificação, surgem ambiguidades, evidenciando limites metodológicos e a necessidade de cautela em análises cientométricas integradas.

Palavras-Chave: Identificadores persistentes; ORCID; Plataforma Lattes; OpenAlex; Heurística nominal.

EIXO TEMÁTICO:

Bases e Fontes de Dados

MODALIDADE:

Resumo expandido

DOI: 10.22477/x.ebbc.1017

COMO CITAR ESTE ARTIGO:

RODRIGUES, Wellington Barbosa; TOMASSINI, Cecilia; MUGNAINI, Rogério; MENA-CHALCO, Jesús Pascual. Além do ORCID autodeclarado no Lattes: inferência de identificadores via DOI-OpenAlex e seus limites. In: EBBC - Encontro Brasileiro de Bibliometria e Cientometria. 10. Anais [...]. Curitiba, 2026. DOI: 10.22477/x.ebbc.1017. Disponível em: <https://doi.org/10.22477/x.ebbc.1017>.



1 INTRODUÇÃO

A identificação de pesquisadores em múltiplas bases científicas constitui um dos desafios metodológicos centrais da bibliometria contemporânea, especialmente quando se busca integrar diferentes infraestruturas de informação para além dos indicadores tradicionais de produtividade e citações (Britto; Robledo; Zuluaga, 2026). Cada base apresenta características próprias quanto à cobertura, aos critérios de indexação e às formas de identificação das produções científicas (Nowakowska, 2025). Nesse contexto, a reconciliação de identidades de autores entre sistemas distintos permanece um problema técnico e epistemológico relevante, uma vez que a qualidade das análises bibliométricas depende diretamente da precisão na identificação dos pesquisadores.

No contexto brasileiro, a Plataforma Lattes constitui a principal infraestrutura curricular da comunidade científica, amplamente utilizada em processos de avaliação, concessão de bolsas e formulação de políticas públicas (Mena-Chalco et al., 2025). Entretanto, sua estrutura foi concebida principalmente para o registro curricular individual, não tendo sido originalmente projetada para mineração massiva de dados ou interoperabilidade automatizada com sistemas internacionais. Em contraste, a OpenAlex apresenta-se como uma infraestrutura aberta e estruturada orientada à organização de informações sobre trabalhos científicos, integrando dados provenientes de múltiplas fontes e operando com identificadores persistentes que permitem consultas programáticas e análises em larga escala (Priem; Piwowar; Orr, 2022). Assim, enquanto o Lattes é centrado no currículo do pesquisador, a OpenAlex organiza seus registros a partir dos trabalhos científicos, evidenciando diferenças estruturais entre as duas bases.

Um dos principais mecanismos de interoperabilidade entre sistemas científicos é

o uso de identificadores persistentes, como o ORCID para pesquisadores e o DOI para documentos (Juty et al., 2020). Esses identificadores reduzem ambiguidades nominais e favorecem o cruzamento automatizado de dados. Contudo, sua adoção ainda não é universal, o que limita a cobertura das análises e dificulta a integração entre sistemas que não disponibilizam mecanismos de busca baseados nesses identificadores (Santos; Maimone, 2023). Na ausência de ORCID, tornam-se necessárias estratégias alternativas de reconciliação, como heurísticas de similaridade nominal, que ampliam a cobertura de identificação, embora introduzam riscos de ambiguidade e potenciais falsos positivos.

Diante desse cenário, este estudo investiga os limites e potencialidades da identificação de pesquisadores brasileiros na integração entre a Plataforma Lattes e a OpenAlex. Nesse sentido, busca-se responder ao seguinte problema de pesquisa: em que medida a integração entre essas bases, por meio do uso de DOI e heurísticas de pareamento nominal, é capaz de ampliar a cobertura e a confiabilidade na identificação de pesquisadores brasileiros? Para isso, analisa-se um conjunto de currículos de bolsistas de Produtividade em Pesquisa do CNPq, nível Sênior (PQ-SR), examinando a cobertura do ORCID autodeclarado no Lattes e a possibilidade de inferência de identificadores por meio da integração DOI-OpenAlex associada a uma heurística de pareamento nominal. Ao discutir os ganhos de cobertura, os níveis de precisão e as ambiguidades decorrentes dessa estratégia, o trabalho contribui para o debate sobre interoperabilidade, qualidade de dados e avaliação responsável da ciência.

2 PROCEDIMENTOS METODOLÓGICOS

Um total de 309 PQ-SR foram identificados a partir do Mapa de Fomento em Ciência, Tecnologia e Inovação¹. Optou-se por esse grupo por se tratar de pesqui-

¹ <http://www.bi.cnpq.br/painel/mapa-fomento-cti/>, acessado em: 21/02/2026.

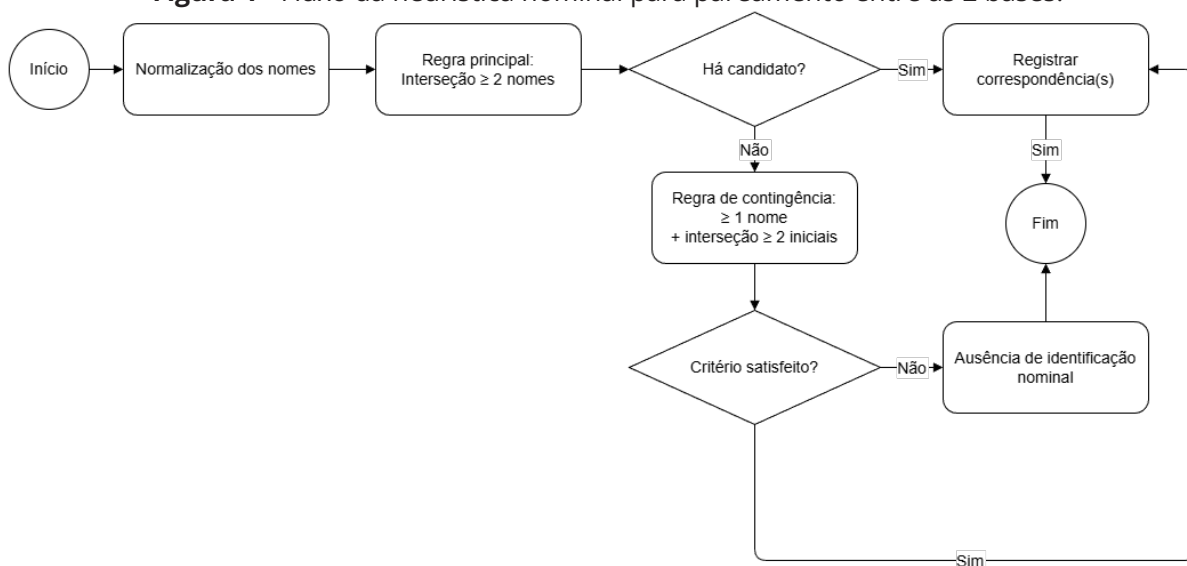
sadores considerados referência na ciência nacional, com produção científica consolidada e volume de publicações que favorece sua identificação em bases internacionais, além de abranger distintas áreas do conhecimento.

Os currículos Lattes foram baixados no mês de fevereiro de 2026 utilizando o software scriptLattes (Mena-Chalco; César Júnior, 2009). A partir desses, rotinas automatizadas em Python foram utilizadas para identificar a presença de ORCID em cada currículo e extrair os identificadores DOI associados às produções registradas.

Os DOIs identificados foram consultados individualmente por meio da API da OpenAlex para recuperação dos metadados dos trabalhos, incluindo listas de autores. Com base nessas informações, aplicou-se uma heurística de pareamento nominal para identificar, entre os autores registrados na OpenAlex, aquele correspondente ao pesquisador titular do currículo Lattes.

A heurística, representada na Figura 1, baseia-se na comparação textual entre o nome completo extraído do currículo Lattes e os nomes dos autores retornados pela OpenAlex. Inicialmente, os nomes são normalizados (remoção de acentos, conversão para minúsculas, tratamento de hífen e padronização de espaços) e decompostos em tokens alfabéticos, desconsiderando termos de ligação frequentes. A correspondência principal considera como candidatos os autores que compartilham ao menos dois tokens com o nome do Lattes, sendo adicionalmente aplicada uma restrição segundo a qual todos os tokens do nome na OpenAlex devem estar contidos no conjunto de tokens do nome no Lattes.

Figura 1 - Fluxo da heurística nominal para pareamento entre as 2 bases.



Fonte: Autores (2026)

Quando nenhum candidato é identificado por essa regra, aplica-se um procedimento de contingência baseado na presença de ao menos um token em comum e na interseção mínima de duas iniciais nominais, contemplando casos de abreviação de nomes. A heurística pode resultar em um ou mais candidatos para um mesmo trabalho, sendo todos mantidos para análise.

Os procedimentos descritos permitiram organizar o corpus segundo a presença ou ausência de ORCID nos currículos e aplicar a heurística nominal a todos os casos, possibilitando comparar os resultados obtidos por correspondência via ORCID e por pareamento nominal. Na seção seguinte apresentam-se os resultados quantitativos dessa identificação.

3 RESULTADOS

Dos 309 PQ-SR, 2 não apresentam ORCID registrado nem ao menos um trabalho com DOI informado no currículo Lattes, sendo, portanto, desconsiderados das análises subsequentes. Assim, a base efetiva de análise é constituída por 307 pesquisadores com produção identificável por DOI.

Entre os 307 bolsistas analisados, 141(45,9%) possuem ORCID explicitamente informado no currículo Lattes, enquanto 166 (54,1%) não apresentam esse identificador. Esses resultados indicam que, ao depender do autopreenchimento, a cobertura de ORCID no Lattes para esse grupo permanece inferior à metade dos pesquisadores, o que reforça limitações já apontadas na literatura quanto à adoção não universal de identificadores persistentes (Juty et al., 2020).

No subconjunto de pesquisadores com ORCID informado no Lattes, observa-se elevada concordância com os identificadores recuperados pela heurística na OpenAlex. Dos 141 casos, 138 (97,9%) apresentam correspondência, sem divergências, e apenas 3 não têm ORCID recuperado. Esses resultados indicam precisão do procedimento e ausência de falsos positivos no conjunto validado, uma vez que o pareamento nominal ocorre exclusivamente entre titular do currículo e autores associados ao trabalho na OpenAlex.

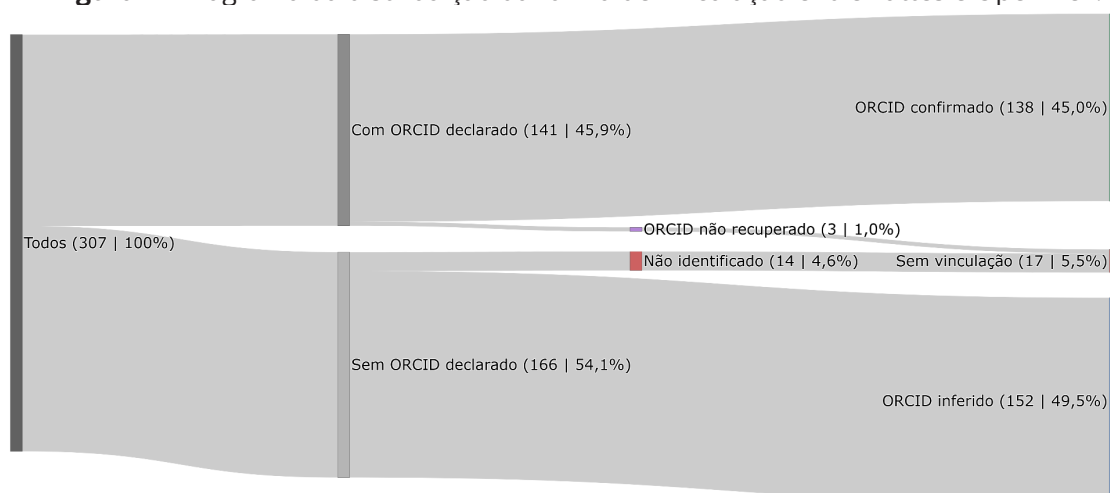
Ao considerar o conjunto de pesquisadores que não possuem ORCID declarado no currículo Lattes, totalizando 166 casos, observa-se que a heurística permite recuperar ao menos um ORCID para 152 deles, o que corresponde a 91,6% desse subconjunto, enquanto 14 pesquisadores (8,5%) permanecem sem qualquer ORCID inferido.

Esse resultado evidencia elevado potencial de recuperação de ORCID por meio do uso de DOI, mesmo na ausência de declaração explícita no Lattes, sugerindo que a integração com dados da OpenAlex, aliada a um procedimento de desambiguação nominal aplicado em nível de registro, é capaz de reduzir de forma substantiva a lacuna de identificação observada na base original, que apresenta limitações já documentadas na literatura quanto à qualidade e atualização dos currículos (Mena-Chalco *et al.*, 2025). Cabe destacar, entretanto, que essa elevada taxa de re-

cuperação pode estar associada às características do recorte analisado, composto por pesquisadores com trajetória consolidada e maior presença de produções com DOI, o que pode favorecer sua identificação em bases como a OpenAlex.

A Figura 2 apresenta o fluxo de identificação e inferência de ORCID para os 307 pesquisadores analisados. Do total, 141 (45,9%) possuem ORCID declarado no currículo Lattes, dos quais 138 (45,0% do total) foram confirmados pela heurística na OpenAlex, enquanto 3 (1,0%) não foram recuperados. Entre os 166 (54,1%) pesquisadores sem ORCID declarado, a heurística permitiu inferir ao menos um ORCID para 152 (49,5% do total), enquanto 14 (4,6%) permaneceram não identificados. Considerando ambos os grupos, há ainda 17 casos (5,5%) sem qualquer tipo de vinculação, evidenciando tanto o ganho de cobertura proporcionado pela estratégia quanto a persistência de lacunas na identificação.

Figura 2 - Diagrama da distribuição da forma de vinculação entre Lattes e OpenAlex.



Fonte: Autores (2026)

A análise em nível de pesquisador permite qualificar esses resultados em termos de desempenho do procedimento. Considerando o subconjunto de pesquisadores com ORCID declarado no Lattes ($n = 141$), a heurística recuperou o identificador esperado em 138 casos (97,9%), o que pode ser interpretado como a revocação em relação à referência. Dentre esses, 124 (87,9%) correspondem a situações de

concordância estrita, nas quais o único ORCID inferido coincide com o informado no currículo, enquanto 14 (10,1%) apresentam múltiplos candidatos, ainda que incluindo o identificador esperado.

No grupo sem ORCID declarado ($n = 166$), a heurística recuperou ao menos um identificador em 152 casos (91,6%), evidenciando seu potencial de ampliação de cobertura. Considerando a base total ($n = 307$), a taxa de ambiguidade — definida pela presença de mais de um ORCID associado a um mesmo pesquisador — atinge 28 casos (9,1%), indicando que, embora o método apresente elevada capacidade de recuperação, parte das associações requer cautela quanto à sua desambiguação. Em trabalhos futuros, serão exploradas estratégias de desambiguação desses ORCIDs, como a atribuição do identificador mais frequente entre os candidatos associados a cada pesquisador.

No que se refere aos identificadores da OpenAlex, 303 pesquisadores (98,7%) apresentam ao menos um ID inferido, e 152 (49,5%) possuem mais de um ID associado. Diferentemente do ORCID, a presença de múltiplos IDs na OpenAlex não implica necessariamente erro metodológico, uma vez que a base apresenta casos de fragmentação autoral. Assim, a elevada proporção de múltiplos IDs na OpenAlex pode refletir tanto ambiguidade nominal quanto características estruturais da própria base, devendo ser explorada em maior profundidade em investigações futuras².

4 CONCLUSÕES

Os resultados evidenciam que a dependência exclusiva do ORCID autodeclarado na Plataforma Lattes implica limitação significativa de cobertura, o que pode com-

² Para testar a robustez do procedimento, pretende-se aplicar os mesmos métodos, em etapas futuras, a outra base de currículos semelhante à Lattes: o CVUy do Uruguai, gerido pela Agência Nacional de Pesquisa e Inovação (ANII), que reúne mais de 16 mil currículos de acesso público.

prometer a identificação de pesquisadores em análises bibliométricas. A estratégia de inferência baseada na integração DOI–OpenAlex associada à heurística nominal demonstrou elevado potencial de recuperação, permitindo identificar ao menos um ORCID para a maioria dos pesquisadores sem identificador declarado. No subconjunto validado, observou-se elevada taxa de concordância entre o ORCID informado no Lattes e aquele recuperado pela heurística, sugerindo bom desempenho do procedimento no subconjunto analisado do procedimento quando aplicado em nível de registro bibliográfico.

Entretanto, a ocorrência de múltiplos ORCIDs associados a um mesmo pesquisador evidencia a principal limitação da abordagem heurística, indicando que o aumento de cobertura pode vir acompanhado de ambiguidades. No caso dos identificadores da OpenAlex, a presença de múltiplos IDs reflete tanto limitações do pareamento nominal quanto características estruturais da própria base – fragmentação autoral.

Por fim, destaca-se que o recorte amostral adotado limita a generalização dos resultados, introduzindo também um possível viés de seleção, uma vez que os bolsistas PQ-SR constituem um grupo de pesquisadores com trajetória consolidada, maior volume de produção e maior presença de trabalhos com DOI, e que a integração entre Lattes e OpenAlex foi realizada exclusivamente por meio do DOI, restringindo o universo de análise às produções que possuem esse identificador. Estudos futuros podem explorar estratégias baseadas em comparação de títulos ou combinações híbridas de metadados. Assim, a inferência por DOI não substitui o uso de identificadores persistentes, mas atua como estratégia complementar para ampliar a cobertura, desde que acompanhada de validação e controle de ambiguidade.

REFERÊNCIAS

BRITTO, Angelo; ROBLED, Sebastian; ZULUAGA, Martha. BibFusion: a Python package to integrate, deduplicate, and harmonize exported bibliographic records from Scopus and Web of Science for bibliometric analysis. **Iberoamerican Journal of Science Measurement and Communication**, v. 6, n. 1, p. 1–21, 2026. DOI: 10.47909/ijsmc.342. Disponível em: <https://ijsmc.pro-metrics.org/index.php/i/article/view/342>. Acesso em: 25 jun. 2026.

JUTY, Nick; WIMALARATNE, Sarala; SOILAND-REYES, Stian; KUNZE, John; GOBLE, Carole; CLARK, Tim. Unique, persistent, resolvable: Identifiers as the foundation of FAIR. *Data Intelligence*, v. 2, n. 1-2, p. 30-39, 2020. DOI: 10.1162/dint_a_00025. Disponível em: <https://direct.mit.edu/dint/article/2/1-2/30/9992/Unique-Persistent-Resolvable-Identifiers-as-the>. Acesso em: 25 jun. 2026.

MENA-CHALCO, Jesús Pascual; RODRIGUES, Wellington Barbosa; WENDT, Lucas George; GOUVEIA, Fabio Castro; DIGIAMPIETRI, Luciano. Sobre os currículos Lattes desatualizados: caracterização e implicações. *In: XXV ENANCIB: Rio de Janeiro, Brasil, 2025, Rio de Janeiro. Anais [...]*. Rio de Janeiro: ENANCIB, 2025. Disponível em: <https://enancib.ancib.org/enancib/article/download/199/199>. Acesso em: 25 jun. 2026.

MENA-CHALCO, Jesús Pascual; CESAR JUNIOR, Roberto Marcondes. scriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31–39, dez. 2009. DOI: 10.1007/BF03194511. Disponível em: <https://link.springer.com/article/10.1007/BF03194511>. Acesso em: 25 jun. 2026.

NOWAKOWSKA, Marzena. A comprehensive approach to preprocessing data for bibliometric analysis. **Scientometrics**, v. 130, p. 5191–5225, 2025. DOI: 10.1007/s11192-025-05415-x. Disponível em: <https://link.springer.com/article/10.1007/s11192-025-05415-x>. Acesso em: 25 jun. 2026.

PRIEM, Jason; PIWOWAR, Heather; ORR, Richard. OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. **arXiv**, 2022. DOI: 10.48550/arXiv.2205.01833. Disponível em: <https://arxiv.org/abs/2205.01833>. Acesso em: 25 jun. 2026.

SANTOS, Thamyres Vieira dos; MAIMONE, Giovana Deliberali. Recuperação da informação e visibilidade científica sob a perspectiva de identificadores persistentes. **Revista Ibero-Americana de Ciência da Informação**, [S. l.], v. 16, n. 3, p. 518–542, 2023. DOI: 10.26512/rici.v16.n3.2023.44053. Disponível em: <https://periodicos.unb.br/index.php/RICI/article/view/44053>. Acesso em: 25 jun. 2026.