10° EBBC
Encontro Brasileiro de Bibliometria e Cientometria

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

HOMONÍMIA E IDENTIFICADORES PERSISTENTES: implicações para redes de colaboração e integração de bases

Thiago Magela Rodrigues Dias

https://orcid.org/0000-0001-5057-9936
thiagomagela@cefetmg.br
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)

Patrícia Mascarenhas

https://orcid.org/0000-0002-8448-6874
patriciamdias@gmail.com
Universidade do Estado de Minas Gerais (UEMG)

Guilherme Mascarenhas Dias

https://orcid.org/0009-0001-7760-4692
Guilhermemdias2501@gmail.com
Instituto Federal de Minas Gerais (IFMG)

Resumo:
Este artigo analisa a presença de homônimos na Plataforma Lattes a partir de um pipeline computacional aplicado a arquivos curriculares, com foco em nomes completos e nomes de citação. O estudo parte do pressuposto de que a homonímia compromete a atribuição correta de autoria e a construção de redes de colaboração. Metodologicamente, o pipeline realiza leitura dos currículos, extração de identificadores e campos nominais, normalização textual, geração de bases analíticas por nome, mensuração da presença de ORCID e DOI. Como resultado, o desenho analítico permite quantificar a ambiguidade nominal em diferentes níveis, comparar a vulnerabilidade relativa entre nome completo e nome de citação.

Palavras-Chave: Homonímia. Desambiguação autoral. Identificadores persistentes.

EIXO TEMÁTICO:
Bases e Fontes de Dados

MODALIDADE:
Resumo expandido

DOI: 10.22477/x.ebbc.1192

COMO CITAR ESTE ARTIGO:

DIAS, Thiago Magela Rodrigues; MASCARENHAS, Patrícia; DIAS, Guilherme Mascarenhas. Homonímia e identificadores persistentes: implicações para redes de colaboração e integração de bases. *In*: EBBC - Encontro Brasileiro de Bibliometria e Cientometria. 10. **Anais [...]**. Curitiba, 2026. DOI: 10.22477/x.ebbc.1192. Disponível em: https://doi.org/10.22477/x.ebbc.1192.

1 INTRODUÇÃO

A identificação inequívoca de autores é uma condição central para a qualidade das análises bibliométricas, cientométricas e de redes de colaboração. Em bases curriculares e bibliográficas extensas, a simples coincidência nominal entre indivíduos distintos produz um problema metodológico recorrente: a homonímia. Esse fenômeno compromete tanto a atribuição correta de autoria quanto a integração automática entre diferentes fontes de informação, sobretudo quando os registros não incorporam identificadores persistentes de pessoa. Em consequência, operações aparentemente triviais, como consolidar a produção científica de um pesquisador ou reconstruir suas relações de coautoria, tornam-se sujeitas a erros de associação, fragmentação de identidade e fusão indevida de perfis.

No contexto brasileiro, a Plataforma Lattes ocupa posição estratégica como infraestrutura nacional de informação curricular e científica (Lane, 2010). Sua relevância decorre não apenas da amplitude de cobertura de pesquisadores, docentes, discentes e profissionais da ciência, mas também de sua utilização recorrente em estudos sobre produção acadêmica, avaliação de pesquisa, mobilidade, formação e colaboração científica. Essa centralidade, contudo, amplia a necessidade de compreender as limitações inerentes ao uso dos campos nominais presentes nos currículos, especialmente quando o objetivo é integrar os dados do Lattes com repositórios, bases bibliográficas, índices de citações e sistemas de informação científica.

O problema se torna ainda mais agudo porque a identidade autoral não se expressa de maneira única. Um mesmo indivíduo pode aparecer pelo nome completo, por formas abreviadas, por nomes para citação bibliográfica e por assinaturas parciais utilizadas em publicações.

Essa instabilidade amplia a ambiguidade e dificulta procedimentos automáticos de deduplicação e desambiguação. Ao mesmo tempo, identificadores persistentes como o ORCID, e identificadores de objetos como o DOI, oferecem mecanismos capazes de reduzir incertezas, favorecer interoperabilidade e aumentar a confiabilidade das integrações entre bases.

Esse problema assume relevância ainda maior quando se consideram sistemas internacionais de avaliação científica e rankings acadêmicos, como QS World University Rankings e Times Higher Education (THE) World University Rankings, que utilizam dados bibliométricos e métricas de produção científica como base para a classificação de instituições. Nesses contextos, erros na identificação autoral decorrentes de homonímia podem gerar distorções na atribuição de publicações, impactando indicadores institucionais, redes de colaboração e medidas de visibilidade científica. Dessa forma, a ambiguidade nominal não afeta apenas análises acadêmicas, mas pode influenciar diretamente processos de avaliação, financiamento e posicionamento estratégico de instituições de ciência, tecnologia e inovação.

Diante desse cenário, este artigo tem como objetivo analisar, em larga escala, a ocorrência de homonímia na Plataforma Lattes, comparando diferentes formas de representação nominal (nome completo, nome de citação e assinaturas abreviadas) e discutindo o papel dos identificadores persistentes na mitigação da ambiguidade autoral.

2 METODOLOGIA

Este estudo adota abordagem quantitativa, exploratória e descritiva, baseada na análise automatizada de arquivos curriculares da Plataforma Lattes (Figura 1).

O universo do estudo corresponde ao conjunto de currículos da Plataforma Lattes processados pelo pipeline, considerando arquivos curriculares disponíveis em formato XML. A coleta de dados foi realizada de forma massiva, abrangendo aproximadamente 9.5 milhões de currículos coletados em 12/2025. No que se refere ao recorte temporal, a análise considera o estado dos currículos no momento da coleta.

Figura 1 – *Pipeline* de tratamento e análise dos dados.

Fonte. Os autores (2026).

A primeira etapa do pipeline consiste, portanto, na leitura dos arquivos e na recuperação dos documentos internos com parser desenvolvido (Dias, 2016). Em seguida, realiza-se a extração dos principais campos analíticos: identificador do currículo, nome completo, nome(s) de citação bibliográfica, presença de ORCID e elementos de produção bibliográfica.

A etapa seguinte envolve a normalização textual dos campos nominais. Essa normalização é necessária para reduzir ruídos derivados de capitalização, acentuação, espaçamento irregular, formas redundantes de pontuação e pequenas variações

gráficas. Para os nomes completos, aplica-se padronização de caixa, remoção de acentos e compressão de espaços. Para os nomes de citação, além dessas rotinas, são harmonizados os espaços ao redor de vírgulas.

Uma vez normalizados, os dados são organizados em duas bases principais: uma base analítica por currículo e uma base expandida por nome de citação. A primeira preserva uma linha por currículo e reúne variáveis de identidade, produção e adoção de identificadores persistentes. A segunda desdobra os nomes de citação em múltiplas linhas, permitindo observar a ambiguidade no nível das assinaturas bibliográficas. Paralelamente, o *pipeline* gera uma assinatura simplificada derivada do nome completo normalizado, útil para simular cenários em que bases externas registram autores por formas abreviadas.

Com essas bases, procede-se às agregações analíticas em três dimensões: nome completo, nome de citação e assinatura. Para cada forma normalizada, são calculados o número total de ocorrências, o número de currículos distintos, a quantidade de ORCIDs distintos, o total de ocorrências com e sem ORCID, o número de variantes brutas associadas e marcadores de homonímia.

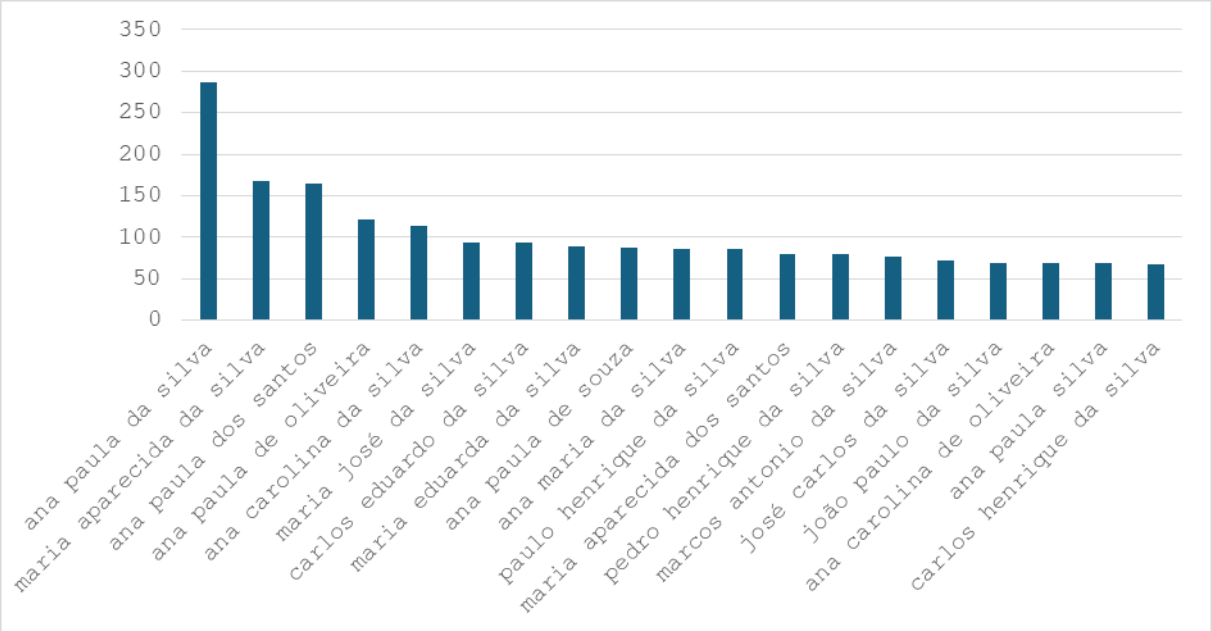
Além disso, o *pipeline* realiza uma análise de variantes ortográficas e formais, mapeando quantas formas brutas convergem para a mesma forma normalizada. Essa etapa permite demonstrar que a ambiguidade não decorre apenas da coincidência entre pessoas distintas, mas também da instabilidade das formas de nome utilizadas nos registros curriculares e bibliográficos. Em seguida, os dados de produção científica são resumidos por currículo, incluindo total de itens de produção, total de artigos publicados, total de artigos com DOI e proporção de artigos identificados por DOI. Essa dimensão é importante porque desloca a discussão da identidade do pesquisador para a integrabilidade de sua produção.

Por fim, o *pipeline* implementa uma classificação heurística de risco de integração automática. Cada currículo é enquadrado em classes de risco baixo, moderado, alto ou extremo, considerando a combinação entre repetição de nome completo, repetição de nome de citação, repetição de assinatura, volume de variantes e ausência de ORCID. O resultado é um conjunto de dados analíticos, preparados para a construção de tabelas comparativas, curvas de distribuição, gráficos de concentração, rankings de homonímia e análises relacionando ambiguidade nominal, identificadores persistentes e produção científica.

3 RESULTADOS

A análise das formas completas de nome revela forte recorrência de determinadas combinações nominais no conjunto analisado. A Figura 1 apresenta os nomes completos com maior frequência entre os currículos processados pelo *pipeline*. Observa-se que determinadas combinações nominais, como variantes associadas ao sobrenome Silva, concentram um número expressivo de ocorrências.

Figura 1 – Frequência de nomes completos no conjunto analisado.



Fonte. Os autores (2026).

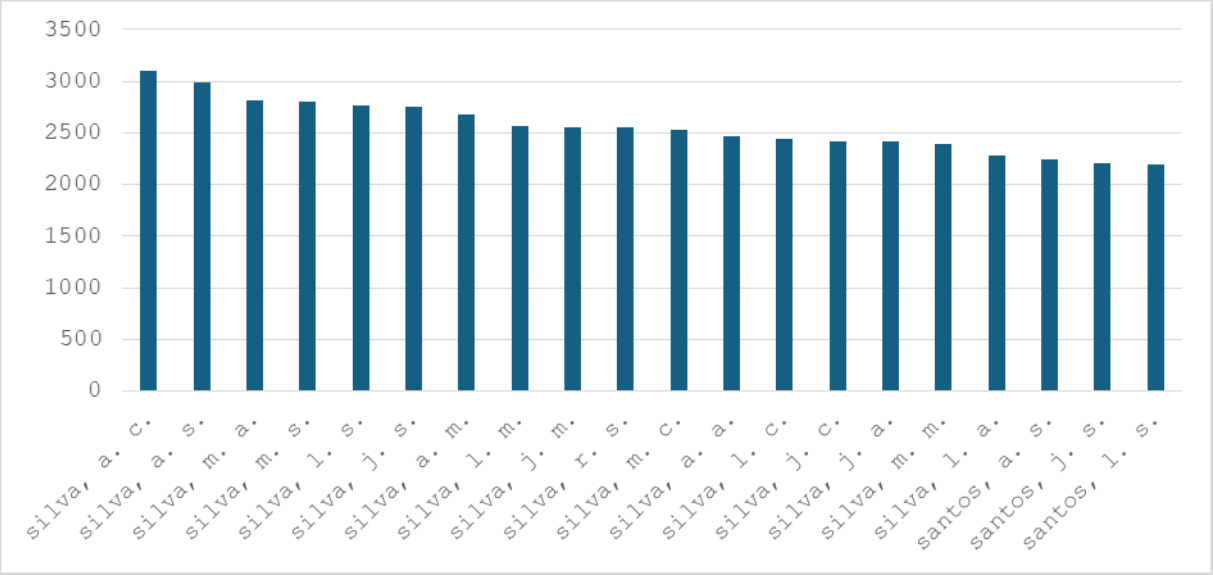
Esse padrão evidencia um fenômeno clássico em bases de dados autorais: a concentração da homonímia em sobrenomes altamente frequentes. Em contextos de identificação exclusivamente textual, tais ocorrências aumentam o risco de associação indevida entre registros pertencentes a indivíduos distintos.

A presença reiterada de combinações como “Ana Paula de Oliveira”, “Maria Aparecida da Silva” ou “Carlos Eduardo da Silva” ilustra como a combinação entre prenomes comuns e sobrenomes de alta frequência populacional cria zonas de elevada ambiguidade nominal.

A análise das assinaturas abreviadas evidencia um padrão ainda mais concentrado do que aquele observado para os nomes completos. A Figura 2 apresenta a distribuição de frequência das assinaturas mais recorrentes, derivadas do sobrenome e das iniciais do nome completo.

Observa-se que formas como “Silva A. S.”, “Silva A. M. G.” ou “Silva J. S.” apresentam frequências extremamente elevadas. Esse resultado é consistente com a literatura sobre identificação autoral em bases bibliográficas, na qual assinaturas abreviadas tendem a gerar níveis de ambiguidade significativamente maiores do que os nomes completos.

Figura 2 – Distribuição dos nomes de citações informados.



Fonte. Os autores (2026).

Essa característica decorre da compressão informacional produzida pela abreviação, que reduz múltiplas identidades distintas a uma mesma forma textual simplificada. Em ambientes de indexação científica, onde a autoria frequentemente aparece apenas como sobrenome e iniciais, essa redução aumenta substancialmente a probabilidade de colisões nominais.

De maneira geral, os resultados indicam que a homonímia constitui um fenômeno estrutural na base curricular analisada, manifestando-se de forma particularmente intensa em sobrenomes altamente frequentes. A análise comparativa entre nomes completos, nomes de citação e assinaturas abreviadas demonstra que a ambiguidade nominal tende a aumentar à medida que a representação textual da identidade autoral é comprimida.

4 CONSIDERAÇÕES

Os resultados permitem sustentar a hipótese de que a homonímia constitui uma limitação estrutural para o uso de dados curriculares da Plataforma Lattes em

análises automatizadas de produção científica, colaboração e integração entre fontes. O problema não se restringe aos nomes completos, mas tende a se intensificar quando se observam os nomes de citação bibliográfica e as assinaturas abreviadas, isto é, precisamente as formas mais comuns em sistemas bibliográficos e repositórios científicos.

Esses efeitos tornam-se particularmente relevantes quando considerados no contexto de sistemas internacionais de avaliação científica, nos quais métricas bibliométricas são utilizadas para compor rankings institucionais e indicadores de desempenho. A presença de erros na identificação autoral pode resultar na atribuição indevida ou na omissão de produções científicas, afetando diretamente a visibilidade e o posicionamento de instituições em rankings globais, além de impactar análises comparativas entre países, áreas do conhecimento e organizações de CT&I.

O estudo também reforça o papel central dos identificadores persistentes. A presença de ORCID reduz a incerteza associada à identificação da pessoa. Contudo, a mera existência desses identificadores em parte dos registros não elimina o problema, pois sua distribuição tende a ser desigual. Assim, a ausência de ORCID em contextos de repetição nominal deve ser entendida como fator de risco para integração indevida, particularmente em procedimentos automáticos que cruzam currículos, publicações e grafos de coautoria.

REFERÊNCIAS

DIAS, Thiago Magela Rodrigues. **Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes**. 2016. Doutorado em Modelagem Matemática e Computacional-Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016. Disponível em: https://sig.cefetmg.br/sigaa/verArquivo?idArquivo=2033874&key=d8d1d2008e1ebe20f0f136527af3a222. Acesso em: 10 mar. 2026.

LANE, Julia. Let's make science metrics more scientific. **Nature**, v. 464, n. 7288, p. 488-489, 2010. DOI: https://doi.org/10.1038/464488a. Acesso em: 20 jan. 2026.