10° EBBC
Encontro Brasileiro de Bibliometria e Cientometria

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

A DISTRIBUIÇÃO DE NOMES DE PESQUISADORES NA PLATAFORMA LATTES: frequência e unicidade em larga escala

Olívia Rico Teodoro

https://orcid.org/0009-0001-3439-4461
olivia.rico@aluno.ufabc.edu.br
Universidade Federal do ABC (UFABC)

Jesús P. Mena-Chalco

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

Resumo:
A identificação de pesquisadores em sistemas de informação depende, entre outros fatores, da forma como os nomes são registrados, estruturados e utilizados em processos de busca, vinculação e desambiguação autoral. Este trabalho analisa a distribuição de frequência dos nomes de pesquisadores cadastrados na Plataforma Lattes, considerando um corpus de 9.549.171 CVs. Foram examinadas três representações nominais: nome completo, primeiro e último nome. Os resultados indicam que a maioria dos nomes completos aparece uma única vez na base, revelando elevado grau de unicidade dessa representação. Em contraste, primeiros e últimos nomes apresentam maior concentração em poucos valores frequentes, o que reduz sua capacidade discriminatória quando analisados isoladamente. A distribuição observada é marcada por assimetria e presença de cauda longa, especialmente no caso dos nomes completos, fornecendo evidências empíricas para estudos sobre identificação nominal e desambiguação de pesquisadores em bases científicas de larga escala.

Palavras-Chave: Nomes. Frequência. Plataforma Lattes.

EIXO TEMÁTICO:
Mapas e Redes na Ciência

MODALIDADE:
Pecha Kucha

DOI: 10.22477/x.ebbc.1201

COMO CITAR ESTE ARTIGO:

TEODORO, Olívia Rico; MENA-CHALCO, Jesús P. A distribuição de nomes de pesquisadores na Plataforma Lattes: frequência e unicidade em larga escala. *In*: EBBC - Encontro Brasileiro de Bibliometria e Cientometria. 10. **Anais [...]**. Curitiba, 2026. DOI: 10.22477/x.ebbc.1201. Disponível em: https://doi.org/10.22477/x.ebbc.1201.

1 INTRODUÇÃO

Os nomes próprios constituem elementos centrais de identificação em sistemas de informação científica. Em bases bibliográficas, currículos acadêmicos, repositórios institucionais e plataformas de avaliação, o nome do pesquisador é frequentemente utilizado em processos de busca, vinculação de registros e reconhecimento de trajetórias acadêmicas. No contexto da ciência brasileira, o estudo sistemático dos nomes registrados em bases acadêmicas ainda é incipiente. A onomástica, área dedicada ao estudo dos nomes próprios, oferece um arcabouço conceitual relevante para compreender padrões de identificação nominal em grandes bases de dados. Entretanto, em sistemas científicos de larga escala, a análise dos nomes também se relaciona diretamente a problemas práticos de identificação autoral, homonímia, variação nominal e desambiguação de pesquisadores.

A identificação baseada em nomes apresenta desafios relevantes, pois primeiros nomes e últimos nomes podem ser altamente frequentes, enquanto combinações completas tendem a apresentar maior diversidade. Além disso, variações ortográficas, abreviações, partículas nominais, nomes compostos e inconsistências de registro podem afetar a vinculação correta de autores e a integração de informações provenientes de diferentes fontes. Dessa forma, a análise da frequência, da concentração e da unicidade dos nomes fornece evidências úteis para compreender os limites e as potencialidades do uso de atributos nominais em sistemas de informação científica.

A base utilizada neste estudo é a Plataforma Lattes que constitui um repositório abrangente da trajetória acadêmica de pesquisadores, estudantes e profissionais envolvidos em atividades científicas e tecnológicas, sendo amplamente utilizada em estudos cientométricos (Lane, 2010). O objetivo central deste trabalho é anali-

sar a distribuição de frequência dos nomes registrados nessa base, considerando três representações nominais: nome completo, primeiro nome e último nome. Busca-se, assim, identificar padrões de unicidade e concentração nominal que possam contribuir para estudos sobre identificação de pesquisadores, desambiguação autoral e desenvolvimento de métodos mais robustos de integração de registros em bases científicas de larga escala.

2 BASE DE DADOS E PROCEDIMENTOS METODOLÓGICOS

A base utilizada neste estudo foi construída a partir da coleta de Cvs da Plataforma Lattes realizada em 17 de fevereiro de 2026, totalizando 9.549.171 CVs disponíveis publicamente. Os registros foram processados localmente para extração do campo Nome, correspondente ao nome completo informado pelo próprio usuário no currículo. Assim, a unidade de análise deste estudo corresponde à string nominal registrada no currículo, e não a uma validação independente da identidade civil ou autoral do pesquisador.

Antes da análise, os nomes foram submetidos a um processo de normalização textual que incluiu conversão para caixa alta, remoção de acentos e eliminação de espaços redundantes. Esse procedimento teve como objetivo reduzir variações ortográficas superficiais e garantir maior consistência nas contagens de frequência. Não foram removidas partículas nominais, como “de”, “da”, “do”, “das” e “dos”, pois elas compõem a forma textual registrada pelo usuário e sua exclusão poderia alterar a estrutura do nome completo. A partir do nome completo foram derivadas duas representações adicionais: o primeiro nome, definido como o primeiro elemento textual da string normalizada, e o último nome, definido como o último elemento textual.

A distribuição de frequência dos nomes foi analisada por meio de estatísticas descritivas, incluindo número de valores distintos, frequência máxima, proporção de ocorrências únicas e proporção acumulada dos nomes mais frequentes. Esse tipo de análise é amplamente utilizado em estudos de distribuição de fenômenos sociais, frequentemente caracterizados por forte assimetria e caudas longas (Lotka, 1926; Newman, 2005). O estudo não incluiu validação manual dos nomes nem tratamento individual de inconsistências nominais, como abreviações, pseudônimos, erros de digitação ou alterações de nome ao longo da trajetória acadêmica. Portanto, os resultados devem ser interpretados como uma caracterização da distribuição das strings nominais registradas na Plataforma Lattes.

3 RESULTADOS

A análise dos nomes completos revela elevado grau de diversidade nominal. Entre os 9.549.171 registros analisados, foram identificados 8.748.066 **nomes completos** distintos. Desses, 8.395.909 aparecem uma única vez, correspondendo a 95,97% dos nomes distintos e a 87,92% de todos os registros da base. Esses valores indicam que, quando considerado em sua forma completa, o nome apresenta baixa repetição exata na Plataforma Lattes. A frequência máxima observada foi de 23.553 ocorrências e correspondeu à string técnica “CV”, associada a registros nos quais o nome não pôde ser recuperado adequadamente durante o processo de extração. Por não representar um nome próprio de pesquisador, esse valor foi identificado como artefato de processamento e interpretado separadamente. Excluindo esse caso específico, observa-se que a concentração entre os nomes completos mais frequentes é muito baixa: os 10 nomes completos mais frequentes representam apenas 0,27% da base, enquanto os 100 mais frequentes representam 0,35%.

A análise dos **primeiros nomes** mostra um padrão distinto. Foram identificados 283.669 primeiros nomes, dos quais 185.797 aparecem uma única vez, representando 65,5% dos primeiros nomes distintos. Contudo, esses primeiros nomes únicos correspondem a apenas 1,95% dos registros totais, indicando que a maior parte dos pesquisadores possui primeiros nomes compartilhados por muitos registros. O primeiro nome mais frequente na base é Maria, com 278.868 ocorrências. Os 10 primeiros nomes mais frequentes representam 12,3% da base, enquanto os 100 mais frequentes representam 44,6% dos registros. Esses resultados indicam uma concentração expressiva em poucos primeiros nomes.

Os **últimos nomes** apresentam concentração ainda maior. Foram identificados 231.406 últimos nomes distintos, dos quais 96.817 aparecem uma única vez. Entretanto, esses últimos nomes únicos representam apenas 1,01% dos registros totais. O último nome mais frequente na base é Silva, com 797.969 ocorrências. Os 10 últimos nomes mais frequentes representam 27,34% da base, enquanto os 100 mais frequentes representam 60,74%. Em conjunto, os resultados mostram uma distribuição fortemente assimétrica: primeiros e últimos nomes concentram grande parcela dos registros em poucos valores, enquanto os nomes completos apresentam maior dispersão e proporção elevada de ocorrências únicas. Esse padrão indica que a combinação completa dos elementos nominais possui maior capacidade discriminatória do que seus componentes analisados isoladamente.

4 CONCLUSÃO

Os resultados evidenciam padrões distintos na estrutura nominal da base analisada. Enquanto primeiros e últimos nomes apresentam forte concentração em poucos valores frequentes, o nome completo apresenta elevado grau de unicidade. Esse padrão indica que, embora componentes nominais isolados sejam pouco

discriminatórios em uma base de grande escala, sua combinação tende a reduzir substancialmente a repetição exata entre registros.

A distribuição observada caracteriza-se por forte assimetria e presença de cauda longa, padrão frequentemente encontrado em fenômenos sociais e informacionais (Newman, 2005). No contexto da identificação científica, esse resultado é relevante porque mostra que diferentes representações nominais possuem capacidades discriminatórias distintas. Nomes completos apresentam menor concentração relativa, enquanto primeiros e últimos nomes, analisados isoladamente, concentram parcela expressiva dos registros em poucos valores.

Este estudo contribui para caracterizar empiricamente a distribuição das strings nominais registradas na Plataforma Lattes e oferece subsídios para pesquisas sobre identificação de pesquisadores em bases científicas de larga escala. Os resultados são particularmente úteis para estudos de desambiguação de autores, integração de registros bibliográficos e desenvolvimento de métodos de identificação em sistemas de informação científica (Milojević, 2013). Como limitação, destaca-se que não foi realizada validação manual dos nomes nem análise individual de inconsistências nominais, de modo que os resultados devem ser interpretados como uma descrição da estrutura de frequência dos nomes registrados, e não como uma validação da identidade autoral dos indivíduos.

REFERÊNCIAS

LANE, J. Let's make science metrics more scientific. **Nature**, v. 464, n. 7288, p. 488-489, 2010. ISSN 0028-0836. Disponível em: https://doi.org/10.1038/464488a. Acesso em: 28 mar. 2026.

LOTKA, A. J. The frequency distribution of scientific productivity. **Journal of the Washington Academy of Sciences**, v. 16, n. 12, p. 317-323, 1926. ISSN 0043-0439.

MILOJEVIĆ, S. Accuracy of simple, initials-based methods for author name disambiguation. **Journal of Informetrics**, v. 7, n. 4, p. 767-773, 2013. ISSN 1751-1577. Disponível em: https://doi.org/10.1016/j.joi.2013.06.006. Acesso em: 28 mar. 2026.

NEWMAN, M. E. J. Power laws, Pareto distributions and Zipf's law. **Contemporary Physics**, v. 46, n. 5, p. 323-351, 2005. ISSN 0010-7514. Disponível em: https://doi.org/10.1080/00107510500052444. Acesso em: 28 mar. 2026.