

AMBIGUIDADE NOMINAL E INTEGRAÇÃO DE DADOS CIENTÍFICOS

uma análise computacional em larga escala

NOMINAL AMBIGUITY AND THE INTEGRATION OF SCIENTIFIC DATA

a large-scale computational analysis

Patricia Mascarenhas Dias

Universidade do Estado de Minas Gerais, UEMG

E-mail: patriciamdias@gmail.com

Thiago Magela Rodrigues Dias

Centro Federal de Educação Tecnológica de Minas Gerais, CEFET-MG

E-mail: thiagomagela@cefetmg.br

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

Washington Luis R. de Carvalho Segundo

Instituto Brasileiro de Informação em Ciência e Tecnologia, IBICT

E-mail: washingtonsegundo@ibict.br

RESUMO

A identificação inequívoca de autores constitui um elemento central para a confiabilidade das análises bibliométricas e cientométricas, especialmente em contextos de integração de dados em larga escala. Este estudo investiga o fenômeno da homonímia na Plataforma Lattes, analisando seus impactos na atribuição de autoria, na construção de redes de colaboração e na interoperabilidade entre bases científicas. A metodologia baseia-se na implementação de um pipeline computacional capaz de processar currículos acadêmicos, extrair campos nominais, normalizar diferentes formas de representação e gerar bases analíticas em múltiplas dimensões, incluindo nome completo, nome de citação e assinaturas abreviadas. Os resultados evidenciam elevada concentração de homonímia em combinações nominais frequentes, bem como aumento significativo da ambiguidade em representações abreviadas. Adicionalmente, observa-se que a ausência de identificadores persistentes, como ORCID, intensifica os riscos de associação indevida entre registros. O estudo contribui para o avanço das estratégias de desambiguação autoral e destaca a necessidade de integração entre métodos computacionais e identificadores persistentes para garantir a qualidade dos dados científicos.

Palavras-chave: Homonímia, Desambiguação autoral, Identificadores persistentes, Bibliometria, Redes de colaboração.

ABSTRACT

The unambiguous identification of authors is central to the reliability of bibliometric and scientometric analyses, particularly in contexts of large-scale data integration. This study investigates the phenomenon of homonymy on the Lattes Platform, analyzing its impacts on authorship attribution, the construction of collaboration networks, and interoperability between scientific databases. The methodology is based on the implementation of a computational pipeline capable of processing academic curricula vitae, extracting name fields, normalizing different forms of representation, and generating analytical datasets across multiple dimensions, including full names, citation names, and abbreviated signatures. The results reveal a high concentration of homonymy in common name combinations, as well as a significant increase in ambiguity in abbreviated representations. In addition, the absence of persistent identifiers, such as ORCID, is observed to intensify the risks of incorrect association between records. The study contributes to the advancement of author disambiguation strategies and highlights the need to integrate computational methods and persistent identifiers to ensure the quality of scientific data.

Keywords: Homonymy, Author disambiguation, Persistent identifiers, Bibliometrics, Collaboration networks.

REFERÊNCIAS

- DIAS, T. M. R. 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.
- LANE, J. Let's make science metrics more scientific. **Nature**, v. 464, n. 7288, p. 488-489, 2010. DOI: <https://doi.org/10.1038/464488a>.