https://doi.org/10.51408/issi2025_057

A Research Entities Disambiguation Methodology Tested on Brazilian Researchers Database

Alysson Fernandes Mazoni[1], Estevão Fernandes Macedo[2], Luís Fabiano Farias Borges[3], Esteban Fernandez Tuesta[4]

[1]*afmazoni@unicamp.br*
University of Campinas, Department of Science and Technology Policy, Institute of Geosciences,
R. Carlos Gomes, 250, Cidade Universitária, Campinas, São Paulo (Brazil)

[2]*estevao.macedo@usp.br*
University of São Paulo, Interdisciplinary Group of Modeling Complex Systems, School of Arts,
Sciences and Humanities, Rua Arlindo Béttio, 1000, Ermelino Matarazzo, São Paulo,
São Paulo (Brazil)

[3]*luis.borges@capes.gov.br*
University of Campinas, Department of Science and Technology Policy, Institute of Geosciences,
R. Carlos Gomes, 250, Cidade Universitária, Campinas, São Paulo (Brazil)
Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Setor Bancário Norte,
Quadra 2, Bloco L, Lote 6, Brasília (Brazil)

[4]*tuesta@usp.br*
University of São Paulo, Interdisciplinary Group of Modeling Complex Systems, School of Arts,
Sciences and Humanities, Rua Arlindo Béttio, 1000, Ermelino Matarazzo, São Paulo,
São Paulo (Brazil)

Abstract

This paper proposes a methodology for disambiguating research entities in databases, with a focus on matching authors, institutions, and publications across various systems. The study examines the OpenAlex and Lattes databases (Brazil's national researcher registry), aiming to enhance the quality and coverage of both databases. Persistent identifiers, such as DOIs and ORCIDs, are utilized to link entities, while co-authorship and affiliation data assist in the matching process. The Levenshtein distance metric is employed to compare names and titles for accuracy. The proposed method is straightforward to implement in tabular databases, making it an effective solution for research information systems. By improving the linkage of authors and publications, this methodology enhances bibliometric research and data curation on platforms like Lattes and OpenAlex. The results illustrate the potential of integrating local and comprehensive databases to address issues of ambiguous names and incomplete metadata.

Introduction

There are several research entities disambiguation systems and algorithms (Ferreira, Gonçalves, & Laender, 2012; Levin, Krawczyk, Bethard, & Jurafsky, 2012; Sanyal, Bhowmick, & Das, 2021; Xu, Shen, Li, & Fu, 2018). They are used usually inside research information systems in order to allow bibliographic studies. In commercial databases such as Scopus (Boyle & Sherman, 2006) authors are required to fill their data and a persistent identifier is created and maintained that way. That should account for authors and institutions disambiguation. However, the references cited are not always part of the input and tend to not have their metadata correctly disambiguated. Curatorship of the data is instrumental to their use and much of this work is hidden as a commercial product inside vendors' platforms (Mongeon & Paul-Hus, 2016).

The recent attempts at comprehensive research information databases such as DataCite, OpenAlex, OpenAIRE try to overcome this using several advanced matching algorithms, many of them based on machine learning (Kim & Kim, 2018; Qian, Hu, Cui, Zheng, & Nie, 2011; Rehs, 2021). There are weak spots in these approaches mainly because of coverage that produces incorrect identification (Rehs, 2021).

It is highly likely, as shown in this work, that a projection of the content of these databases on local scientific databases can overcome these problems in regards to ambiguous names, lack of persistent identifiers, among others. In this use case, we used the Lattes database (Mena-Chalco & Junior, 2009) (the Brazilian registry of researchers) as a base to cross their production to OpenAlex. A nearly full matching of researchers, institutions and publications would allow bibliometric research across the Lattes database, that is not fully linked currently and would also correct and improve on the OpenAlex database for it would increase coverage and metadata quality.

In this work, a methodology to match research entities is proposed to disambiguate them inside research information databases. The methodology uses persistent identifiers as clues and their entities links, such as co-authorships and affiliations, to propagate these clues. The clues are then combined using distance metrics for names and titles. The methodology is tested matching research entities from the Brazilian registry of researchers (Curriculum Lattes) against similar entities in the OpenAlex database. The results indicate high precision and ease of implementation and use for tabular databases, which is a common ground for research information systems.

Theoretical Background

Research information systems

A research information system is a sort of database that manages information about scientific activities and production. Institutional databases that keep information on their research such as thesis, monographs, books and articles are examples of such systems (de Castro & Puuska, 2023). Granting agencies usually maintain registries of their supported projects and sometimes try to maintain registries for their research products (Alshamaila et al., 2024). Namely, such systems allow for bibliometric and scientometric research on the cosmos they cover.

There are also databases that purport to be comprehensive about science, in the sense that they aim to cover all knowledge production indexed according to some criteria, such as Scopus, Web of Science, OpenAlex, Dimensions (Turgel & Chernova, 2024).

Such systems are important for the maintaining institutions to be aware of their own function and priorities. In that direction, also government agencies that evaluate scientific research are always in great need of such information, in many cases, keeping such databases as public policy: such is the case of the Lattes platform

(Digiampietri et al., 2012) and other national efforts, CVUY (Simón, Fontáns, & Aguirre-Ligüera, 2013).

Persistent identifiers

The information about scientific activity is spread around several entities, such as authors, journals, institutions, publications, among others. Such entities are usually mapped to data entities in databases in order to create a data model that would allow translating questions about scientific activity into queries or filterings on such databases. Some data models about research are very complex and mature, aiming at traceability of research entities such as the OpenAIRE graph (Vichos et al., 2022) linked to the OpenAIRE database (Manghi, Manola, Horstmann, & Peters, 2010).

The linking that allows the creation of such models demands the identification of research entities, ideally using unambiguous identifiers. The most egregious of them:

- DOI, ARK - for publications (Freire, Manguinhas, Isaac, & Charles, 2023)
- ROR - for institutions (Welke & Krause, 2024)
- ORCID - for researchers (Schnieders et al., 2022)
- ISSN - for journals (Bequet, 2022)

And others that could be linked, such as patent identifiers, companies registries in the national offices, etc.

Usually databases of research information systems collect their data from forms manually filled by researchers and staff or by collecting other online databases. The diversity of sources and the intrinsic variable nature of human filled information creates challenges for matching entities across databases and inside the same databases. The variations in titles, in the writing of names given its multiple components, abbreviations and translated names for research institutions, all that adds to the importance of persistent identifiers.

Distance metrics to identify names

When matching of titles is needed, it is common to use metrics that compare pieces of text as strings (Slavin, Andreeva, & Putincev, 2022). These metrics use character variation as bases, the most common of them being the Levenshtein distance (Öztürk, Ertürk, Casale-Brunet, Ribeca, & Mattavelli, 2024; Sadiah, Iryani, Zuraiyah, Wahyuni, & Zaddana, 2024). Although widely used to compare human names (Kiawkaew, Kaothanthong, & Theeramunkong, 2023), it is not quite appropriate for this application, given that names are usually given in several alternative forms omitting family or given names or replacing them by initials.

An attempt to match names using such a metric would create artificial distances and proximities that make the matching less likely in many cases. To tend to these limitations, this work uses an adaptation of the Levenshtein metric on portions of the name, taking into account only names that appear and initials, also their relative order.

Data model for scientific production

Following the inspiration from OpenAlex and OpenAIRE (Manghi et al., 2010; Vichos et al., 2022) this paper adopts a relational model for the entities aiming at their disambiguation. In this model, an author is linked to its publications by an authorship relation. The authorship contains an affiliation to an institution. Each work is connected to authorships as indicated in Figure 1.

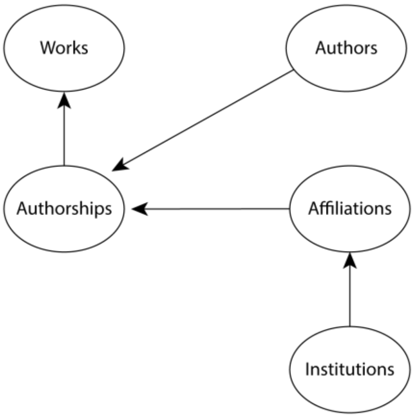

Figure 1. Data model for production.

Methodological Proposal and Rationale

Authors' identification

Given it is an official registry linked to the national identification code, it is safe to assume that the Lattes registry is unique and unambiguous about researchers. Starting from this, the matching proceeds to find them as authors in OpenAlex. The list publications provided for every author inside Lattes is not guaranteed to be complete, however, it is in the best interest of researchers to fill the list completely given that the evaluation from the national agencies are based on the Lattes information.

The publications metadata is manually filled and can contain errors and publications can be multiply registered by its coauthors. The strategy consists of selecting the publications with DOI for every author. Thus, several publications will share authors. That creates a list of possible authors from both sources (Lattes and OpenAlex) for every DOI.

Every list of possible authors is a matching pool for a comparison with a distance metric for names. That way, the number of authors to be compared is limited to a small number of coauthors. It is possible to assume that the most similar name is the correct matching if a certain distance threshold is assumed as the minimum acceptable. The improved metrics for names improves the possibility of finding author names. The threshold can be easily checked ordering the most likely author matching in each case from the worst distance to the best.

The method consists of finding a persistent identifier that collects an entity that must be matched in two databases. Using the smaller subset of possible matches filtered by the identifier, it is possible to evaluate against the distance metric for names.

In the case of the two databases here used, we find publications on the Lattes databases with explicit DOIs. There will be an unambiguous author associated with it. We want to locate this author inside OpenAlex. The found DOIs link to a list of coauthors of each publication. The coauthors are possible candidates of matching for the original author taken from Lattes. This way, the name matching happens on the smaller subset of possible coauthors. The best candidate according to the metric, given a certain threshold, is pointed as the match from Lattes to OpenAlex.

A point that should be observed is that, given a certain number of common DOIs between two candidate matching names, a certain threshold is appropriate to guarantee correct matching. However, with a larger number of DOIs in common, one could use a smaller threshold for the distance metric for names. That is so because more clues about the right candidate allow for a looser criterion for the matching. That way, the best threshold to use is a function of the number of common DOIs.

By checking the Lattes database, we can see that the largest possible number of common DOIs between authors' names is 649. We selected 0,1 distance (10% of length of name different from one to another) as the threshold for the worst possible case (just 1 DOI in common) and 0,4 (40% of difference) as the threshold for the situation with 649 DOIs in common. That leads to an inequality (with the threshold as t and the number of common DOIs as n):

$$t < \frac{64.5 + 0.3n}{648}$$

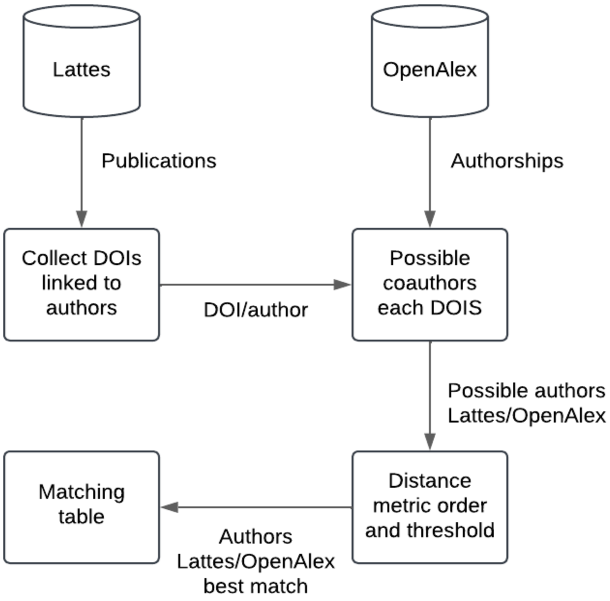
Lattes
OpenAlex
Publications
Authorships
Collect DOIs linked to authors
DOI/author
Possible coauthors each DOIS
Possible authors Lattes/OpenAlex
Matching table
Authors Lattes/OpenAlex best match
Distance metric order and threshold

Figure 2. Data model for production.

The Lattes Database and its Application

In the late 1990s, Brazil’s national funding agency recognized the need for a new approach to evaluating researchers' credentials. To address this, it first established a ‘virtual community’ comprising federal agencies and researchers to design and develop the Lattes infrastructure. This database provides high-quality data on approximately millions of researchers and thousands of institutions registered within it (Lane, 2010). Lane (2010) argues that the Brazilian experience with the Lattes Database (http://lattes.cnpq.br/english) exemplifies best practices in research assessment and creates appropriate incentives for researchers and academic institutions to utilize the database.

Results and Conclusion

We collected the data for all authors in Lattes that are PhDs and have published in the last 5 years. From their declared publications, we found all the ones with DOIs and applied our methodology.

The number of Lattes identifiers (representing researchers) with at least one valid DOI to apply our methodology is 154,474. The method identifies 151,318 authors in the OpenAlex database. Contrary to expected, this is not due to authors not being found, but to multiple people assigned to just one author in the OpenAlex database. The number of matches with an exact correspondence of one to one is 148,431. This amounts to 6,043 people in Lattes being identified as 2,887 authors in OpenAlex. This strange phenomenon is mostly due to its disambiguation algorithm (Barrett, 2023) that compares names and takes fuzzy clues such as collaborations and areas of research in order to identify people. It relies ultimately on ORCID, but such an identifier is absent in many cases.

Figure 3 presents the number of DOIs used in the matching of the authors. Table 1 presents the number of incorrectly assigned to a single author identifier.

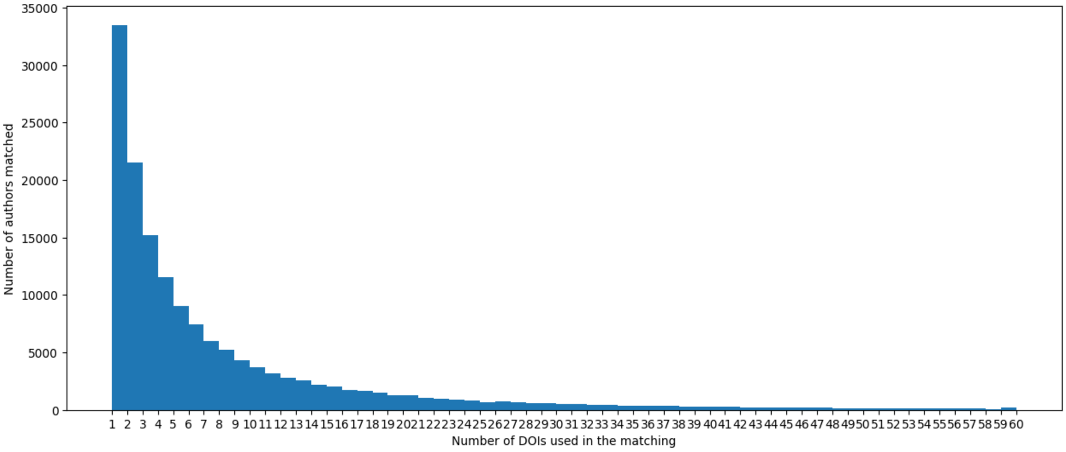

Figure 3. DOIs used in the matching of authors.

Table 1. Total of assigned identifiers to a single author identified after matches our methodology.

Total of identifiers	*Total of assigned*
1	148,431
2	2,645
3	220
4	17
5	5

The results show that the disambiguation algorithm has succeeded in a high percentage of the authors but plays on the risky side of assigning just one author identifier to a few people in some cases instead of a more conservative approach to keep doubtful cases split.

The use of the OpenAlex database however, given its open nature, allows for corrections such as the one here presented. A large deal of information can now be extracted for the authors that are uniquely identified. The smaller percentage of misidentified people can now be studied and the room for improvements based on local databases is paved. The methodology of combining persistent identifiers with an adapted metric in a dynamic threshold can be expanded to improve the database and its applications by adding other local databases, such as institutional data, for example.

References

Alshamaila, Y., Alsawalqah, H., Habib, M., Al-Madi, N., Faris, H., Alshraideh, M., Aljarah, I., et al. (2024). An intelligent rule-oriented framework for extracting key factors for grants scholarships in higher education. *International Journal of Data and Network Science*, *8*(2), 1325–1340.

Barrett, J. (2023). Openalex name disambiguation. Retrieved from https://github.com/ourresearch/openalex-name-disambiguation/tree/main/V3

Bequet, G. (2022). From the Cradle to the Digital Vault: Tracking the Path of E-journals. *Serials Librarian*, *82*(1–4), 199–204.

Boyle, F., & Sherman, D. (2006). Scopus™: The Product and Its Development. *The Serials Librarian*, *49*(3), 147–153.

de Castro, P., & Puuska, H.-M. (2023). Research Information Management Systems: Covering the whole research lifecycle (Vol. 95, pp. 257–265). Presented at the EPiC Series in Computing.

Digiampietri, L. A., Mena-Chalco, J. P., Pérez-Alcázar, J. J., Tuesta, E. F., Delgado, K. V., Mugnaini, R., & Silva, G. S. (2012). Minerando e Caracterizando Dados de Currículos Lattes. Retrieved from https://sol.sbc.org.br/index.php/brasnam/article/view/6868

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. *ACM SIGMOD Record*, *41*(2), 15–26.

Freire, N., Manguinhas, H., Isaac, A., & Charles, V. (2023). Persistent Identifier Usage by Cultural Heritage Institutions: A Study on the Europeana.eu Dataset. In O. Alonso, H. Cousijn, G. Silvello, M. Marrero, C. Teixeira Lopes, & S. Marchesin (Eds.), *Linking Theory and Practice of Digital Libraries* (pp. 341–348). Cham: Springer Nature Switzerland.

Kiawkaew, T.-A., Kaothanthong, N., & Theeramunkong, T. (2023). A Practical Technique for Thai-English Word Mapping Using Phonetic Rules: Person Name Matching Case Study. *2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1–6). Presented at the 2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). Retrieved January 25, 2025, from https://ieeexplore.ieee.org/document/10354663/?arnumber=10354663

Kim, J., & Kim, J. (2018). The impact of imbalanced training data on machine learning for author name disambiguation. *Scientometrics*, *117*(1), 511–526.

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology*, *63*(5), 1030–1047.

Manghi, P., Manola, N., Horstmann, W., & Peters, D. (2010). An Infrastructure for Managing EC Funded Research Output – The OpenAIRE Project□ –.

Mena-Chalco, J. P., & Junior, R. M. C. (2009). scriptLattes: An open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, *15*(4), 31–39. SpringerOpen.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228.

Öztürk, Ü., Ertürk, U. G., Casale-Brunet, S., Ribeca, P., & Mattavelli, M. (2024). Efficient Neural Clustering and Compression of Strings Through Approximate Euclidean Embeddings of the Levenshtein Distance. *2024 Data Compression Conference (DCC)* (pp. 575–575). Presented at the 2024 Data Compression Conference (DCC). Retrieved January 25, 2025, from https://ieeexplore.ieee.org/document/10533750/?arnumber=10533750

Qian, Y., Hu, Y., Cui, J., Zheng, Q., & Nie, Z. (2011). Combining machine learning and human judgment in author disambiguation. *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1241–1246). Presented at the CIKM '11: International Conference on Information and Knowledge Management, Glasgow Scotland, UK: ACM. Retrieved January 20, 2025, from https://dl.acm.org/doi/10.1145/2063576.2063756

Rehs, A. (2021). A supervised machine learning approach to author disambiguation in the Web of Science. *Journal of Informetrics*, *15*(3), 101166.

Sadiah, H. T., Iryani, L. D., Zuraiyah, T. A., Wahyuni, Y., & Zaddana, C. (2024). Implementation of Levenshtein Distance Algorithm for Product Search Query Suggestions on Koro Pedang Edutourism E-Commerce. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *42*(2), 188–196.

Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2021). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*, *47*(2), 227–254.

Schnieders, K., Mierz, S., Boccalini, S., Meyer zu Westerhausen, W., Hauschke, C., Hagemann-Wilholt, S., & Schulze, S. (2022). ORCID coverage in research institutions—Readiness for partially automated research reporting. *Frontiers in Research Metrics and Analytics*, *7*.

Simón, L., Fontáns, E., & Aguirre-Ligüera, N. (2013). El currículum vitae como fuente de datos en los estudios métricos. Retrieved from http://sedici.unlp.edu.ar/handle/10915/38075

Slavin, O., Andreeva, E., & Putincev, D. (2022). Application of modified Levenshtein distance for classification of noisy business document images. *Fourteenth International Conference on Machine Vision (ICMV 2021)* (Vol. 12084, pp. 78–85). Presented at the Fourteenth International Conference on Machine Vision (ICMV 2021), SPIE. Retrieved January 25, 2025, from https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12084/120840B/Application-of-modified-Levenshtein-distance-for-classification-of-noisy-business/10.1117/12.2623437.full

Turgel, I. D., & Chernova, O. A. (2024). Open Science Alternatives to Scopus and the Web of Science: A Case Study in Regional Resilience. *Publications*, *12*(4), 43. Multidisciplinary Digital Publishing Institute.

Vichos, K., De Bonis, M., Kanellos, I., Chatzopoulos, S., Atzori, C., Manola, N., Manghi, P., et al. (2022). A Preliminary Assessment of the Article Deduplication Algorithm Used for the OpenAIRE Research Graph (Vol. 3160). Presented at the CEUR Workshop Proceedings.

Welke, B., & Krause, B. (2024). Automatically generated Research Profiles for Experts, Institutions and Working Groups (Vol. 249, pp. 112–119). Presented at the Procedia Computer Science.

Xu, J., Shen, S., Li, D., & Fu, Y. (2018). A Network-embedding Based Method for Author Disambiguation. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 1735–1738). Presented at the CIKM '18: The 27th ACM International Conference on Information and Knowledge Management, Torino Italy: ACM. Retrieved January 20, 2025, from https://dl.acm.org/doi/10.1145/3269206.3269272