Research Article

Hesham Amin Hamdy El Shamly and Subaveerapandiyan A.*

Author Name Disambiguation in Scholarly Research: A Bibliometric Perspective

https://doi.org/10.1515/opis-2025-0035
Received May 23, 2025; accepted November 16, 2025; published online January 23, 2026

Abstract: The rapid expansion of scholarly publishing has amplified the long-standing challenge of author name ambiguity in academic databases. This issue, manifesting as homonymy and synonymy, undermines the accuracy of bibliometric analyses, author-level metrics, and research evaluation systems. Author Name Disambiguation (AND) has thus emerged as a critical focus area in digital scholarship, with evolving strategies ranging from supervised machine learning and graph-based models to the adoption of persistent digital identifiers like ORCID. Despite notable advancements, significant challenges remain – particularly in linguistically diverse and underrepresented regions – where metadata inconsistencies, transliteration issues, and limited ORCID adoption exacerbate disambiguation errors. This study presents a comprehensive bibliometric analysis of 2,004 publications on AND from 2005 to 2024, sourced from the Scopus database. Using tools such as Biblioshiny and VOSviewer, the analysis identifies publication trends, leading authors and institutions, core sources, co-authorship networks, and thematic evolution in the field. Findings highlight increasing international collaboration, the dominance of computer science-driven methodologies, and the critical role of metadata quality and institutional frameworks. The study concludes with recommendations for inclusive, multilingual, and interoperable disambiguation systems, advocating for cross-disciplinary collaboration to ensure equitable author identification in global scholarly communication.

Keywords: author name disambiguation; bibliometric analysis; digital identifiers (ORCID); name ambiguity in scholarly databases; co-authorship networks; metadata quality in academic publishing

1 Introduction

The exponential growth of scholarly publications in recent decades has intensified challenges in accurately identifying individual authors, particularly within large bibliographic databases. A critical issue is author name ambiguity, which manifests as homonymy – multiple authors sharing identical or similar names – and synonymy – a single author publishing under different name variants (Cappelli et al. 2025; Hussain and Asghar 2017). These ambiguities compromise the accuracy of bibliometric analyses, author-level metrics, and research evaluations (Liu et al. 2014; Tekles and Bornmann 2020).

Author Name Disambiguation (AND) has therefore emerged as a pivotal task to ensure the integrity of digital libraries and scholarly databases (Firdaus et al. 2022; Hussain and Asghar 2017). The task generally involves two sub-problems: Author Assignment (AA), which links new publications to existing author profiles, and Author Grouping (AG), which clusters publications likely authored by the same individual (Cappelli et al. 2025). Although persistent digital identifiers like ORCID have been introduced to resolve these issues, their uneven adoption – especially in under-resourced regions – limits their standalone efficacy (Panigabutra-Roberts 2025; Sebo et al. 2021).

***Corresponding author: Mr. Subaveerapandiyan A.**, Assistant Librarian, Department of Library, Bennett University, Greater Noida, Uttar Pradesh, India, E-mail: subaveerapandiyan@gmail.com
Hesham Amin Hamdy El Shamly, Associate Professor, Department of Surgery, RAK Medical & Health Sciences University, Ras Al Khaimah, UAE, E-mail: hesham@rakmhsu.ac.ae. https://orcid.org/0009-0007-3173-5404

To address these complexities, diverse computational strategies have been developed. Traditional methods leverage supervised and unsupervised machine learning techniques based on metadata features such as co-authorship, affiliation, and publication venues (Kim 2018; Tekles and Bornmann 2020). In recent years, graph-based models and deep learning architectures have shown improved performance by capturing semantic and structural metadata patterns (Firdaus et al. 2022; Hussain and Asghar 2017).

However, the problem becomes more acute in culturally and linguistically diverse contexts. Common surnames, transliteration inconsistencies, and metadata deficiencies – particularly in East and South Asia – further hinder effective disambiguation (Firdaus et al. 2022; Xu and Hu 2024). Additionally, metadata incompleteness, inconsistent identifier integration, and lack of multilingual support in global repositories such as ORCID and Crossref exacerbate the issue (Donathan et al. 2025; Kim 2018).

Given the methodological diversity and the evolving landscape of AND research, a comprehensive bibliometric analysis is essential to assess the field's development, key contributors, and emerging themes. This study addresses this need by conducting a systematic and quantitative review of AND research from 2005 to 2024. By mapping publication trends, leading institutions, collaborative networks, and research hotspots, the study also explores the integration of novel approaches – including blockchain verification, AI-driven disambiguation, and culturally adaptive frameworks (Cappelli et al. 2025; Zhou et al. 2024).

Objectives of the Study:

- To analyze annual publication trends and citation impact in AND research (2005–2024).
- To identify leading journals, conferences, and sources contributing to the domain.
- To evaluate the productivity and influence of key authors and institutions.
- To map co-authorship networks and international collaboration patterns.
- To examine thematic evolution and research hotspots through keyword analysis.

Need for the Study

Despite decades of research, author name ambiguity continues to pose serious challenges for digital libraries, bibliometric tools, and scholarly communication systems. The consequences are significant: inaccurate author profiling, misattributed works, fragmented citation counts, and compromised research evaluation metrics (Cappelli et al. 2025; Firdaus et al. 2022). While advances in supervised learning, graph-based methods, and persistent identifiers like ORCID have yielded progress, these tools remain insufficient, especially in linguistically diverse and underrepresented regions.

First, the rapid expansion of scholarly publications has exacerbated cases of homonymy and synonymy, especially in large-scale databases such as Scopus, DBLP, and PubMed. These result in erroneous clustering or separation of author records, affecting both bibliometric precision and recall (Hussain and Asghar 2017; Tekles and Bornmann 2020). For example, Xu and Hu (2024) demonstrated that Chinese scholars are disproportionately affected due to common names and transliteration practices – a pattern also observed in other Asian regions.

Second, while ORCID has been proposed as a universal identifier, its adoption is inconsistent and metadata quality remains poor. Many ORCID records lack affiliations, publications, or standardized formatting, undermining their effectiveness in author identification (Morgan and Eichenlaub 2018; Panigabutra-Roberts 2025). Moreover, institutional and linguistic barriers hinder broader adoption, particularly in multilingual and low-resource settings (Donathan et al. 2025; Xu and Hu 2024).

Third, current AND techniques often lack global applicability. Cultural variations in name structures, language-specific metadata inconsistencies, and non-uniform naming conventions continue to challenge universal models. This is compounded by incomplete integration across systems like Crossref, ORCID, and national repositories, limiting both discoverability and machine-readability (Donathan et al. 2025; Shi et al. 2025). Although recent technologies – including hybrid attention models and deep graph networks – show promise, their practical effectiveness and interoperability are yet to be fully assessed (Santini et al. 2022; Zhou et al. 2024).

This study is therefore essential for two reasons. First, it systematically reviews the methodological advances in AND and highlights the evolution of research themes, contributors, and collaborative dynamics. Second, it identifies critical gaps and regional disparities, providing insights to guide the development of more inclusive, multilingual, and interoperable AND frameworks. Such evidence is vital for stakeholders – research institutions, data curators, and policy-makers – who seek to improve the accuracy, fairness, and transparency of author attribution across the global scholarly ecosystem.

2 Literature Review

2.1 Understanding the Problem of Author Name Ambiguity

Author name ambiguity remains one of the most pressing challenges in bibliometric and digital library systems. It predominantly arises in two forms: homonymy, where multiple distinct authors share identical or very similar names, and synonymy, where a single author's name appears in multiple variant forms across publications (Cappelli et al. 2025; Manzoor et al. 2022). These issues significantly affect the accuracy of scholarly attribution, citation metrics, and research assessment.

The phenomenon is particularly prevalent in large-scale bibliographic databases, where the same name can correspond to numerous different individuals. For instance, databases like DBLP have been shown to return hundreds of results for common names such as "Ajay Gupta" or abbreviated forms like "A. Gupta," demonstrating the widespread issue of polysemy (Hussain and Asghar 2017). Moreover, inconsistent use of initials, cultural naming conventions, and institutional naming formats contribute to the complexity of the problem (Firdaus et al. 2022).

Studies have shown that ambiguity impacts both precision and recall in literature retrieval systems. For example, synonymy leads to citation splitting, where an author's works are scattered across different profiles, reducing their perceived scholarly output. Conversely, homonymy results in citation merging, where multiple authors' records are mistakenly consolidated, leading to inflated metrics and erroneous attributions (Liu et al. 2014; Sebo et al. 2021).

Digital libraries like PubMed and DBLP have implemented various disambiguation systems to address this issue. However, evaluations indicate that errors persist, particularly in distinguishing authors with common names or across multilingual contexts (Kim 2018; Tekles and Bornmann 2020). For instance, in PubMed, it is estimated that approximately two-thirds of ambiguous author names result from homonymy, especially among authors who share initials and surnames (Manzoor et al. 2022).

Additionally, the user-side experience is compromised by these ambiguities. Log analyses of PubMed queries revealed that over one-third of searches involve author names, and these frequently return incomplete or irrelevant results, forcing users to modify or refine their queries (Islamaj Dogan et al. 2009; Liu et al. 2014).

To better illustrate these challenges, researchers have conceptualized AND into two sub-problems: Author Assignment (AA) – linking new documents to known author profiles, and Author Grouping (AG) – clustering existing documents authored by the same individual (Cappelli et al. 2025). These sub-problems highlight the dual threat posed by both homonymy and synonymy, and the importance of accurate disambiguation for scholarly integrity.

2.2 Challenges in Author Name Disambiguation Across Contexts

Author name disambiguation (AND) faces diverse contextual challenges stemming from linguistic, cultural, and institutional factors that complicate consistent author identification across databases and publication systems. One of the core challenges arises from cultural naming conventions, where surname structures, name orders, and character systems differ significantly. For instance, in East Asian contexts, especially Chinese, names often follow the family-name-first format and may lack Western-style middle names, leading to frequent homonymy and low granularity in author records. Xu and Hu (2024) illustrate this through the "namesake" and "heteronymous" name problems prevalent among Chinese scholars, highlighting the added complexity of transliteration inconsistencies and institutional practices that discourage ORCID adoption, thus impeding retrospective disambiguation efforts.

Multilingualism also introduces challenges in metadata consistency and completeness. Donathan et al. (2025) emphasize that while English monolingual metadata dominate in Crossref records, multilingual and non-English metadata are often incomplete or inconsistently structured, limiting discoverability and accurate author matching. The lack of metadata schema flexibility to accommodate multilingual nuances contributes to metadata degradation across global academic infrastructures.

Furthermore, name changes due to personal life events (e.g., marriage) or career rebranding further compound disambiguation issues, particularly for female scholars or those from regions where name-changing traditions are prevalent (Xu and Hu 2024). The analysis by Mryglod et al (2023) of Ukrainian economic publications reinforces this by showing that poor use of digital identifiers and inconsistent formatting of Cyrillic and Latin names obstruct automatic disambiguation, especially in less-resourced non-Western academic systems.

Metadata inconsistencies, especially in large repositories like ORCID and Crossref, result from diverse submission standards and publisher practices. Delgado et al. (2018) and Donathan et al. (2025) both reveal that metadata errors – ranging from inaccurate author affiliations to duplicated or missing entries – are not just technical oversights but reflective of structural issues in the scholarly publishing ecosystem. Shi et al. (2025) identified 32 types of metadata-related issues linked to cultural and language dimensions, including representation, geography, and naming logic.

To address these problems, scholars propose stronger integration of persistent digital identifiers (e.g., ORCID, ISNI), cross-platform data synchronization, and culturally aware metadata standards (Donathan et al. 2025; Xu and Hu 2024). However, systemic inertia, regional disparities in digital adoption, and gaps in multilingual representation still hinder equitable progress.

2.3 Methodological Approaches to Disambiguation

Author Name Disambiguation (AND) has seen the application of a variety of computational approaches, broadly categorized into supervised, unsupervised, hybrid, and graph-based machine learning methods. Each approach varies in its assumptions, data requirements, and effectiveness across bibliographic contexts.

Supervised techniques rely on labeled datasets where pairs of author name instances are tagged as either belonging to the same author or not. Common algorithms include Random Forest, Logistic Regression, Support Vector Machines, and Naïve Bayes classifiers. Rehs (2021) implemented both Random Forest and Logistic Regression models on over 1.2 million publication pairs and achieved F1 scores of 0.82 and 0.75, respectively. Feature importance analysis played a critical role in enhancing prediction accuracy, using attributes such as co-authorship, affiliation, and publication metadata (Rehs 2021). Similarly, Kim and Kim (2018) found that training models with imbalanced datasets – containing more negative than positive pairs – could still yield robust performance when feature selection was properly optimized (Kim et al. 2018).

Unsupervised methods typically involve clustering techniques that do not require labeled data. These include hierarchical agglomerative clustering, k-means, and density-based methods, often relying on similarity metrics like cosine similarity, Jaccard index, or Levenshtein distance applied to metadata such as author names, co-authors, titles, and affiliations (Protasiewicz and Dadas 2016). Such methods are particularly useful in large-scale scenarios where manual labeling is infeasible, although they may struggle with accuracy in highly ambiguous name sets.

Hybrid frameworks combine rule-based systems with learning algorithms or cluster-based methods. For example, Protasiewicz and Dadas (2016) proposed a two-tier framework comprising rule-based disambiguation followed by hierarchical clustering, enabling incremental author identification in growing datasets. Another hybrid model by Rodrigues and Ralha (2024) integrated BERT for semantic text analysis, Graph Convolutional Networks (GCN) for structural representation, and Graph-Enhanced Hierarchical Agglomerative Clustering (GHAC), achieving a 95 % F1 score and over 96 % cluster purity – demonstrating strong performance on multi-faceted datasets.

Recent research emphasizes graph-based methods leveraging relationships among entities. Wang et al. (2025) developed a heterogeneous graph neural network (H-GNN) that incorporates semantic and topological features, improving the representation of ambiguous author nodes. Their model achieved an F1 score of 0.834 on the AMiner dataset. Similarly, Santini et al. (2022) introduced a knowledge graph embedding-based approach using literal metadata (e.g., names, titles, venues), clustering author entities via hierarchical agglomerative clustering without needing labeled training data.

Approaches using contrastive learning and variational autoencoders have been explored for capturing both relational and non-relational aspects of documents. Pooja et al. (2021) proposed an unsupervised method that embeds co-authorship and metadata separately using variational graph autoencoders and fuses these representations through neural techniques. The method improved generalization across heterogeneous datasets compared to traditional global embedding methods.

Tools like ANDez (Kim and Kim 2024) integrate top-performing ML models in an open-source Python environment. These platforms aim to enhance accessibility and reproducibility in AND tasks by bundling feature extraction, similarity scoring, and model evaluation modules.

2.4 The Role and Limitations of Digital Identifiers

Persistent digital identifiers such as the Open Researcher and Contributor ID (ORCID) play a crucial role in the accurate disambiguation of author names, improving metadata quality, and enhancing the discoverability of scholarly outputs. ORCID was introduced as a community-driven initiative in 2012 and has since become one of the most widely adopted systems for uniquely identifying researchers across disciplines (Fenner et al. 2014).

ORCID provides a unique 16-digit identifier (ORCID iD) that links researchers to their scholarly work, affiliations, and funding data. It serves not only as a tool for attribution but also as a bridge connecting multiple systems, including Crossref, DataCite, Scopus, ISNI, and institutional repositories (Fenner et al. 2014; Panigabutra-Roberts 2025). This interoperability supports researcher identification across platforms and facilitates metadata synchronization.

The ORCID metadata model includes fields for employment, education, works, funding, and more. However, Morgan and Eichenlaub (2018) note that many ORCID records remain sparsely populated – so-called “orphan records” – due to the optional nature of metadata entry and the burden of self-maintenance. Their longitudinal API-based analysis showed that a significant number of ORCID profiles lacked basic metadata such as institutional affiliations or publications, limiting their effectiveness for name disambiguation.

Adoption of ORCID varies significantly across countries and institutions. For example, Bouchard and Boudry (2025) found in a national survey that ORCID adoption among researchers in France remains uneven, with higher uptake in natural sciences and lower awareness in the humanities. Similarly, a Spanish study by Bordons et al. (2024) revealed macro- and micro-level inconsistencies in ORCID implementation across research fields and institutions, with better integration in larger universities and among more senior researchers.

Institutional systems also play a critical role in adoption. The University of Tennessee, Knoxville (UTK) provides a compelling case study: integration of ORCID into the faculty activity reporting system (Elements) led to higher adoption among faculty compared to graduate students (Panigabutra-Roberts 2025). Nevertheless, incomplete requirements for electronic theses and dissertations (ETDs) limited ORCID uptake among early-career researchers.

Moreover, many systems lack full integration with ORCID. Although ORCID data can be leveraged in library cataloging and linked data initiatives such as WorldCat and VIAF, these implementations are still partial (Panigabutra-Roberts 2025). The metadata is often underutilized in discovery systems, and the lack of mandatory data fields results in minimal record utility for authority control and interoperability.

While ORCID's technical architecture supports integration through APIs and data synchronization, actual adoption and implementation across repositories and workflows remain inconsistent. The ODIN project highlighted that lack of interoperability between identifier systems (e.g., ORCID, DOI, and DataCite) can impede data linking and reuse, emphasizing the need for a cohesive infrastructure strategy (Fenner et al. 2014).

Further, limitations arise from privacy settings and the self-managed nature of ORCID profiles. Users can choose to make their data public, private, or accessible only to trusted parties, which affects the completeness and visibility of metadata across systems (Panigabutra-Roberts 2025). Without institutional mandates or automated workflows, many researchers do not regularly update their records, undermining ORCID's potential as a reliable disambiguation tool.

2.5 Technological Tools and Services in Practice

Author name disambiguation (AND) has seen significant technological advancements in recent years, especially in terms of tools, services, and infrastructure supporting large-scale metadata processing and disambiguation workflows. Among the most widely discussed technologies are Web APIs and RESTful services, which provide scalable and interoperable interfaces for integrating disambiguation systems into digital library ecosystems.

A notable example is the ReCiter system developed by the Weill Cornell Medicine Library. ReCiter is an open-source, identity-driven authorship prediction algorithm that leverages institutional data such as HR records and past publication histories. Its design prioritizes integration with institutional workflows through RESTful APIs, enabling real-time author-publication association at scale. ReCiter combines rule-based logic with machine learning components and is optimized for deployment within academic institutions (Albert et al. 2021).

Another major contribution to the AND landscape is the web service developed by Ferreira et al. (2012), which provides a RESTful API for scholarly databases. This system incorporates supervised machine learning algorithms and enables real-time processing of metadata across various bibliographic sources. The tool was particularly evaluated using DBLP data and demonstrated strong precision and recall performance. The emphasis on offering a public web API highlights the push toward accessible, platform-agnostic disambiguation services.

However, access limitations remain a challenge for implementing such technologies uniformly across regions and institutions. Many APIs rely on proprietary data (e.g., Scopus, Web of Science), limiting their applicability in open science contexts. Additionally, some tools are closely coupled with specific platforms, reducing cross-platform portability.

Beyond API services, newer approaches emphasize institutional name normalization and disambiguation as a complementary task to AND. The AffilGood project (Duran-Silva et al. 2024) focuses on standardizing noisy and multilingual institutional affiliation strings using annotated datasets and entity-linking tools. This work highlights how institution-level disambiguation supports cleaner author identification and enhances metadata quality. The integration of resources such as the Research Organization Registry (ROR) is central to this task.

Moreover, domain-specific variations in metadata completeness have been noted as a recurring issue. Disciplinary differences affect the quality of affiliation data, author names, and co-authorship metadata. Delgado et al. (2018) point out that metadata consistency and structure are highly variable across fields, especially on the web, where multilingualism and informal web content complicate extraction and clustering tasks. These disparities introduce challenges for cross-disciplinary AND systems, which must adapt to heterogeneous data environments.

Finally, there is increasing interest in hybrid architectures that blend RESTful APIs with attention-based deep learning models. For example, a hybrid attention-based system proposed by Zhou et al. (2024) integrates multi-head attention mechanisms with author context modeling, showing promise for scalable AND tasks when paired with large knowledge graphs. Such systems are often modular, enabling integration via APIs for downstream applications such as recommendation systems or digital repositories.

APIs and services have advanced author disambiguation, limitations related to platform dependence, metadata variation, and disciplinary coverage persist. Future directions must emphasize open interoperability, multilingual support, and adaptive frameworks that address the practical needs of libraries, researchers, and institutions across diverse regions and scholarly ecosystems.

3 Research Methodology

3.1 Research Design

This study employed a bibliometric analysis approach integrated with elements of systematic review to analyze scholarly literature on Author Name Disambiguation (AND) published between 2005 and 2024. Bibliometric analysis was used to map the research landscape, trends, and collaborations, while a systematic approach guided the formulation of inclusion criteria, screening procedures, and reproducible data extraction. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework was adapted to ensure transparency in document selection and refinement processes.

3.2 Data Source and Search Strategy

The Scopus database was selected as the primary source for data retrieval due to its comprehensive coverage of multidisciplinary and high-impact publications. A Boolean search query was designed to capture the breadth of literature on author name disambiguation using the Title-Abstract-Keyword (TITLE-ABS-KEY) field to avoid missing relevant records. The final search string was as follows:

> TITLE-ABS-KEY ((“author name disambiguation” OR “researcher name disambiguation” OR “name ambiguity” OR “name confusion” OR “author identification” OR “author attribution” OR “author recognition” OR “name matching” OR “author meta-data” OR “author profiling” OR “name variant resolution” OR “name disambiguation” OR “author entity resolution”)) AND PUBYEAR > 2004 AND PUBYEAR < 2025 AND (LIMIT-TO (DOCTYPE , “cp”) OR LIMIT-TO (DOCTYPE , “ar”)) AND (LIMIT-TO (LANGUAGE , “English”)). This search yielded an initial corpus of 2,580 documents.

3.3 Screening and Refinement Process

Following retrieval, documents were screened using predefined inclusion and exclusion criteria. Inclusion criteria were: (a) peer-reviewed journal articles or conference papers; (b) English language; (c) focus on author name disambiguation or related concepts. Exclusion criteria included: editorials, reviews, book chapters, short surveys, errata, and documents with incomplete metadata.

A total of 2,333 documents met the basic criteria after applying publication year and document type filters. Next, documents were subjected to quality screening for relevance, metadata completeness, and removal of duplicates. 171 documents were identified as duplicates or containing insufficient metadata and were excluded. Finally, the top 2,004 documents were selected based on citation count and keyword relevance for in-depth bibliometric analysis. Figure 1 presents the PRISMA flow diagram detailing the document selection process.

3.4 Data Analysis Tools

Data preprocessing and analysis were conducted using a combination of specialized bibliometric software tools:

- Bibliometrix R package (including the Biblioshiny web interface) for data import, refinement, and statistical analysis;
- VOSviewer for network visualization of co-authorship, keyword co-occurrence, and country collaborations;
- Microsoft Excel for tabulation, graphing, and trend calculations;
- Scopus for data export and citation metrics.

The bibliometric indicators analyzed include publication trends, authorship patterns, citation impact, core sources, productive institutions, and thematic evolution.

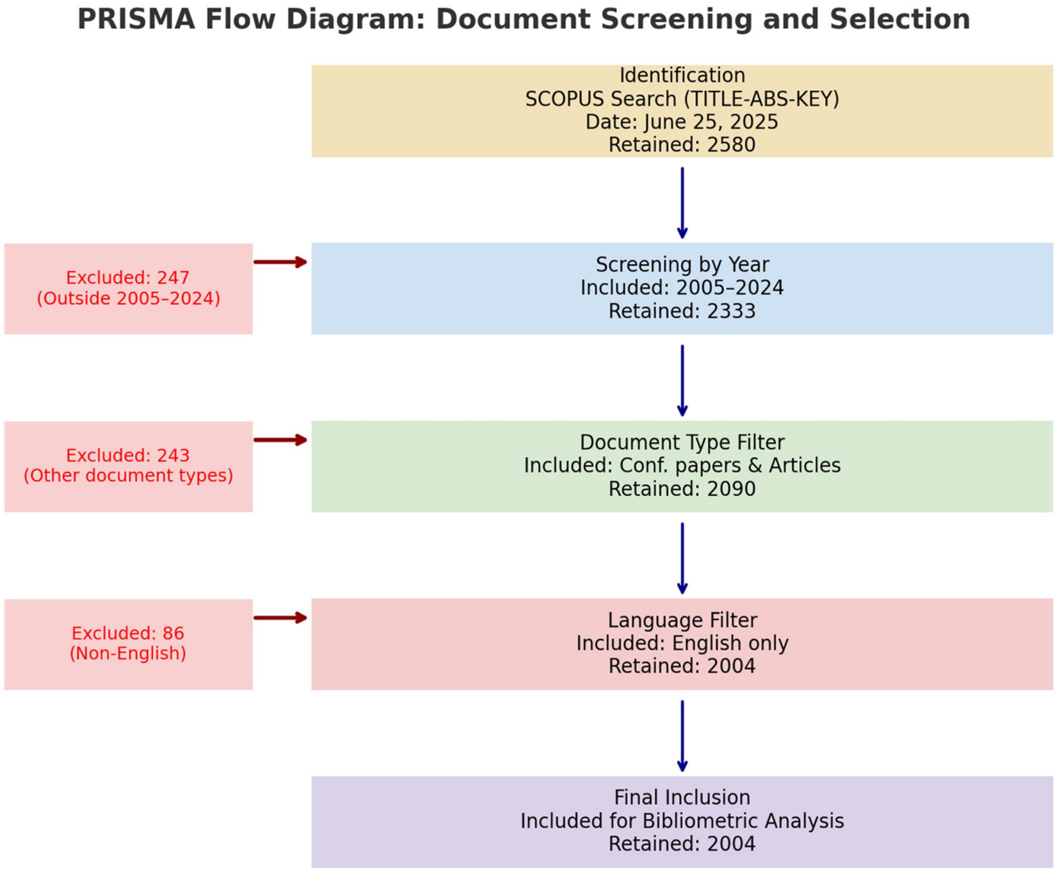

Figure 1: PRISMA flow diagram of document screening and selection for bibliometric analysis.

3.5 Validation and Reproducibility

To ensure methodological rigor, all data collection and analysis steps were documented and can be replicated using the same search string and tools. Thresholds for inclusion in visualizations (e.g., minimum number of documents or citations for network maps) were selected based on standard bibliometric practices to avoid skewing the analysis.

4 Results

4.1 Annual Publication Output and Citation Trends (2005–2024) (Related to Table 1, Table 2, and Figure 2 – Describes Scholarly Output Growth and Citation Performance Over Time)

Table 1 summarizes key bibliometric indicators for 2004 documents on author name disambiguation (AND) published between 2005 and 2024. The dataset spans 895 distinct sources and includes 4,608 contributing authors, with an average of 3.52 co-authors per document, indicating strong collaboration. A notable 19.51 % of publications involved international co-authorship. Conference papers (61.5 %) significantly outnumber journal articles (38.5 %), underlining the technical and emerging nature of this research domain.

Table 1: Descriptive summary of the bibliometric dataset on author name disambiguation (2005–2024).

Description	Results
Main Information About Data	
Timespan	2005–2024
Sources (Journals, Books, etc.)	895
Documents	2004
Annual Growth Rate (%)	7.94
Document Average Age	8.42
Average Citations per Document	17.23
References	46,974
Document Contents	
Keywords Plus (ID)	7,401
Author's Keywords (DE)	3,587
Authors	
Authors	4,608
Authors of Single-Authored Documents	168
Authors Collaboration	
Single-Authored Documents	189
Co-Authors per Document	3.52
International Co-Authorships (%)	19.51
Document Types	
Article	772
Conference Paper	1,232

Figure 2 shows the yearly output of scholarly articles related to Author Name Disambiguation (AND) from 2005 to 2024, based on 2,004 selected publications. The data reveals a steady and consistent increase in research activity over the two-decade period.

- Initial Phase (2005–2012): The field witnessed a slow but steady rise in publications, from 26 articles in 2005 to 76 in 2012. This phase likely reflects foundational research and early methodological explorations in AND.
- Growth Phase (2013–2019): A significant growth trajectory begins in 2013 (107 articles) and peaks in 2019 with 168 articles. This surge coincides with increased attention to machine learning and semantic analysis in digital libraries, as well as the rise of platforms like ORCID and Scopus as major bibliometric databases.
- Maturity Phase (2020–2024): A slight decline in publication volume is observed post-2019, with numbers decreasing from 153 in 2020 to 111 in 2024. This plateau may be attributed to topic saturation, greater reliance on established tools, or shifts toward broader themes like researcher profiling and AI-driven metadata cleaning.

The consistent growth up to 2019, followed by a slight tapering, suggests that the domain has entered a phase of consolidation, with increased specialization and diversification in research themes and technologies applied to AND.

Table 2 shows citation performance by publication year. Articles from 2008 and 2005 recorded the highest mean citations per article (67.26 and 52.04 respectively), demonstrating the lasting impact of early foundational studies. Recent years show declining citation averages due to recency, though a gradual increase in 2021 indicates renewed scholarly attention.

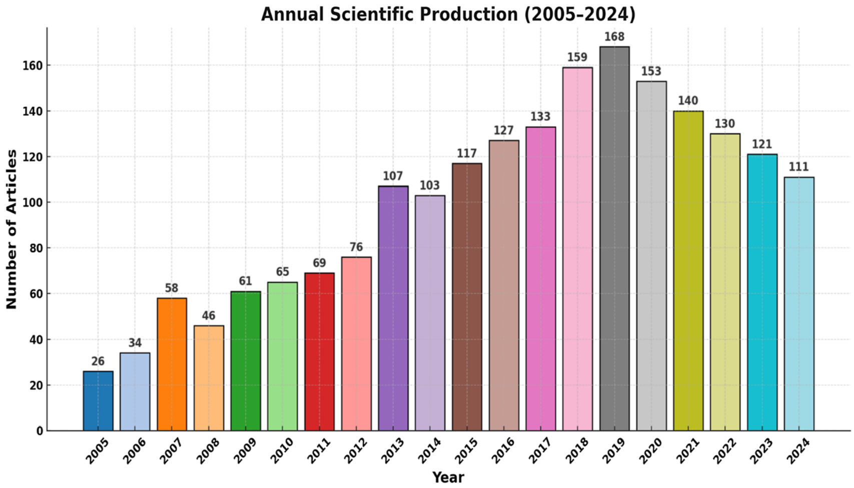

Figure 2: Annual growth of scholarly output in author name disambiguation (2005–2024).

Table 2: Year-wise citation trends and average citations per article (2005–2024).

Year	MeanTCperArt	N	MeanTCperYear	CitableYears
2005	52.04	26	2.48	21
2006	38.32	34	1.92	20
2007	46	58	2.42	19
2008	67.26	46	3.74	18
2009	26.89	61	1.58	17
2010	16.48	65	1.03	16
2011	24.52	69	1.63	15
2012	23.63	76	1.69	14
2013	10.86	107	0.84	13
2014	18.28	103	1.52	12
2015	11.97	117	1.09	11
2016	15.46	127	1.55	10
2017	20.43	133	2.27	9
2018	13.44	159	1.68	8
2019	8.71	168	1.24	7
2020	10.65	153	1.78	6
2021	28.99	140	5.8	5
2022	5.88	130	1.47	4
2023	4.67	121	1.56	3
2024	1.47	111	0.74	2

4.2 Core Publication Sources and Their Impact (Related to Table 3 and Table 4 – Discusses Key Journals, Conference Proceedings, and Their Bibliometric Indicators)

Table 3 applies Bradford's Law to classify core sources. Zone 1 includes seven venues, led by CEUR Workshop Proceedings and Lecture Notes in Computer Science. These venues dominate AND literature due to their recurring publication of technical innovations and conference proceedings focused on digital libraries, machine learning, and data mining.

Table 4 compares the impact of top sources. CEUR Workshop Proceedings achieved the highest total citations (TC = 2,617), while Lecture Notes in Computer Science leads in early contributions (PY_start = 2005).

Table 3: Core journals and conference proceedings in author name disambiguation: Bradford's law distribution.

Source	Rank	Freq	cumFreq	Zone
CEUR Workshop Proceedings	1	359	359	Zone 1
Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	2	173	532	Zone 1
Scientometrics	3	45	577	Zone 1
ACM International Conference Proceeding Series	4	29	606	Zone 1
Proceedings of the ACM/IEEE Joint Conference on Digital Libraries	5	26	632	Zone 1
Advances in Intelligent Systems and Computing	6	25	657	Zone 1
Communications in Computer and Information Science	7	23	680	Zone 1
International Conference on Information and Knowledge Management, Proceedings	8	18	698	Zone 2
IEEE Access	9	14	712	Zone 2
Journal of the Association for Information Science and Technology	10	13	725	Zone 2

Table 4: Impact metrics of key publication sources on author name disambiguation research.

Element	h_index	g_index	m_index	TC	NP	PY_start
Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	24	36	1.143	1914	173	2005
CEUR Workshop Proceedings	22	40	1.1	2,617	359	2006
Scientometrics	21	33	1.313	1,159	45	2010
Proceedings of the ACM/IEEE Joint Conference on Digital Libraries	13	26	0.619	798	26	2005
International Conference on Information and Knowledge Management, Proceedings	11	18	0.579	543	18	2007
Journal of the Association for Information Science and Technology	11	13	0.917	330	13	2014
Information Processing and Management	9	11	0.5	551	11	2008
Behavior Research Methods	7	9	0.389	482	9	2008
IEEE Transactions on Knowledge and Data Engineering	7	7	0.368	1869	7	2007
Proceedings – IEEE International Conference on Data Mining (ICDM)	7	7	0.35	465	7	2006

Scientometrics, although lower in volume, shows strong normalized impact (m_index = 1.313), reflecting the scholarly depth and citation longevity of articles it published.

4.3 Author Productivity and Institutional Contributions (Related to Figure 3, Table 5, and Table 6 – Highlights Leading Authors, Productivity Patterns Using Lotka's Law, and Top Institutions)

Figure 3 presents the most prolific authors in AND research. Paolo Rosso and Francisco Rangel emerge as key contributors, with strong interconnections as later confirmed by the co-authorship map (Figure 4). Their leadership in stylometry and author profiling has significantly shaped methodological directions in the field.

Table 5 presents the distribution of author productivity in Author Name Disambiguation (AND) research, analyzed using Lotka's Law. The data shows a classic long-tail distribution where the majority of authors contribute minimally, while a few demonstrate high productivity.

- Single-publication authors dominate the landscape, with 3,616 individuals (78.5 %) publishing only one document. This reflects either peripheral interest in AND or early-career contributions without follow-up studies.

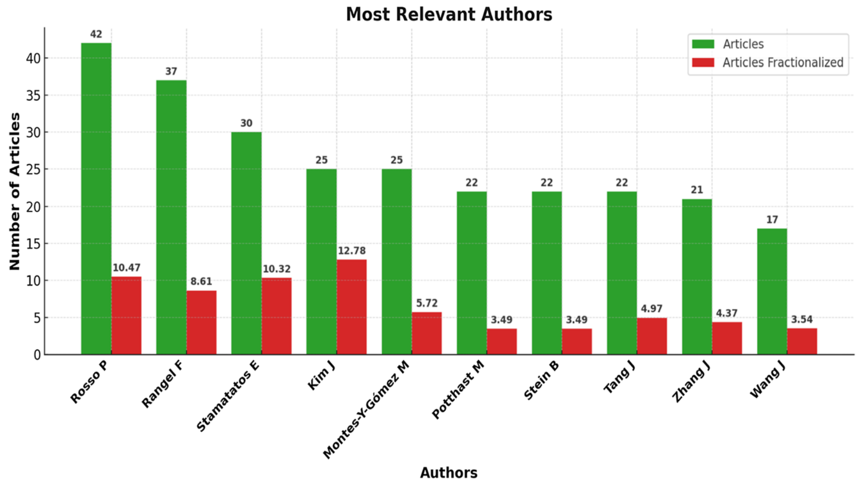

Figure 3: Top contributing authors by publication count in name disambiguation studies.

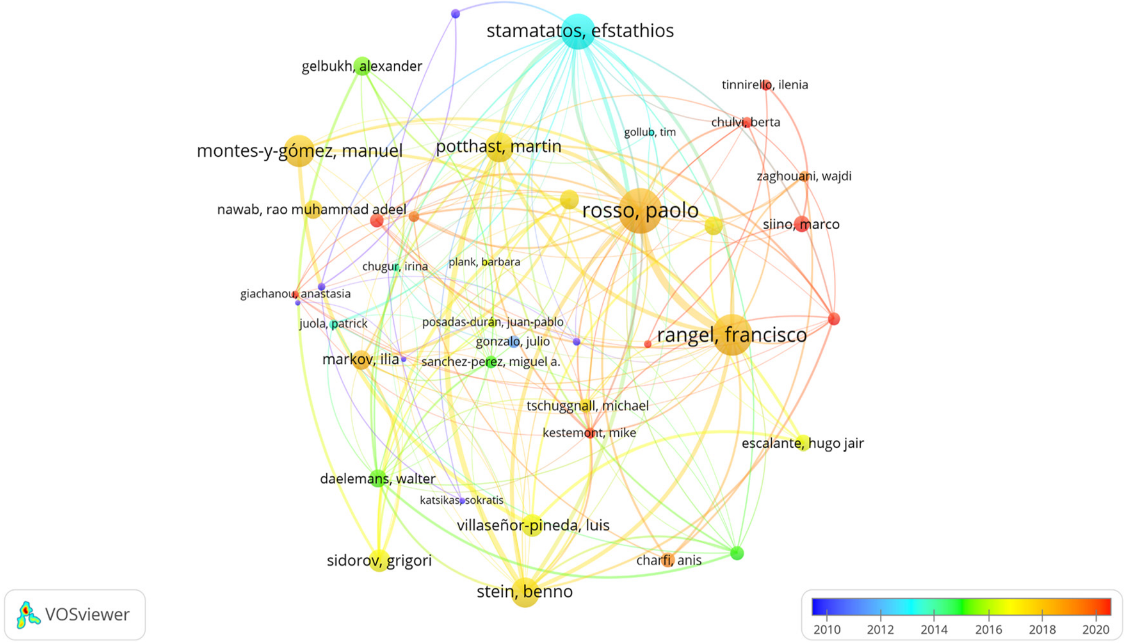

Figure 4: Co-authorship network of highly cited authors in name disambiguation research (2005–2024).

- Two- and three-time contributors account for 11.2 % and 4.2 %, respectively, indicating a slightly more engaged group of researchers but still not sustained contributors.
- Only 50 authors (1.1 %) published five papers, and fewer than 0.5 % of authors published six or more, emphasizing the presence of a small core community consistently producing AND-related scholarship.
- The highest observed productivity comes from a single author with 42 publications, underscoring their central role in the field's development. A handful of others contributed 20 or more papers, suggesting a stable elite driving research trends.

This distribution strongly conforms to Lotka's Law, which posits that the number of authors publishing n papers is inversely proportional to n^2. The observed skew highlights the fragmentation and early-stage maturity of the field, where sustained engagement by a broader base of authors is still developing.

Table 5: Distribution of author productivity in name disambiguation literature: a Lotka's law analysis.

Documents written	N. of authors	Proportion of authors
1	3,616	0.785
2	514	0.112
3	192	0.042
4	128	0.028
5	50	0.011
6	32	0.007
7	20	0.004
8	13	0.003
9	8	0.002
10	6	0.001
11	5	0.001
12	5	0.001
13	3	0.001
14	1	0
15	3	0.001
16	2	0
17	1	0
21	1	0
22	3	0.001
25	2	0
30	1	0
37	1	0
42	1	0

Table 6 shows top institutions include Tsinghua University (70 articles) and Universitat Politècnica de València (41), reinforcing the central role of China and Europe in AND research. Notably, Latin American and Southeast Asian universities such as Universitas Sriwijaya and Federal University of Minas Gerais show emerging engagement.

Table 6: Top contributing institutions in author name disambiguation research.

Affiliation	Articles
Tsinghua University	70
Universitat Politècnica de València	41
Instituto Nacional de Astrofísica	33
University of the Aegean	29
Beihang University	26
National University of Defense Technology	26
Foshan University	25
Bauhaus-Universität Weimar	24
Federal University of Minas Gerais	24
Universitas Sriwijaya	24

4.4 Co-authorship and Collaboration Patterns (Related to Figure 4 and Figure 5, Table 7 – Examines Author-Level and Country-Level Research Collaborations and Network Structures)

Table 7 quantifies collaboration at the national level. China and the USA lead in publication volume, while Italy and Germany show higher Multiple Country Publication (MCP) ratios, indicating active international research engagement.

Figure 4 shows a co-authorship network visualization of 251 authors in the domain of author name disambiguation research, created using VOSviewer. The authors included in this map each have at least one publication and a minimum of 100 citations. These thresholds, applied to a total of 5,010 authors, resulted in 251 authors selected for mapping based on their total link strength in co-authorship networks.

Each node in the network represents an individual author. The size of the node reflects the number of documents authored, while the thickness of the connecting lines (edges) indicates the strength of co-authorship links. Clustering resolution was set to 5.0, and 19 clusters emerged, each representing a collaborative research community. The color overlay reflects the average publication year, with a gradient ranging from blue (older publications) to yellow/orange (more recent activity).

Notably, Paolo Rosso, Francisco Rangel, Martin Potthast, Efstathios Stamatatos, and Benno Stein appear as central figures, indicated by their large node sizes and dense connections. For instance, Rosso (42 documents, 1,554 citations, 123 link strength) and Rangel (37 documents, 1,494 citations, 121 link strength) are heavily connected, often collaborating within the same cluster. Other prominent contributors include Montes-y-Gómez, Manuel (25 documents, 381 citations), Gómez-Adorno, Helena, and Strotmann, Andreas.

The network also displays semi-peripheral and emerging scholars like Ilia Markov, Luis Villaseñor-Pineda, and Zaghouani, Wajdi, who have strong intra-cluster ties but fewer inter-cluster connections. Isolated or weakly linked authors may represent specialists or independent contributors within the domain.

This visualization demonstrates how collaborative relationships shape the development of author name disambiguation research. The presence of tightly knit clusters highlights sustained partnerships, while cross-cluster connections suggest interdisciplinary collaboration, especially in areas intersecting with computational linguistics, information retrieval, and machine learning.

Figure 5 shows a country-level co-authorship network visualization in the field of author name disambiguation research, based on bibliometric data extracted from the Scopus database. The map was generated using VOSviewer, where countries were analyzed as the unit of analysis under the co-authorship method. Countries with at least 5 publications and 100 citations were included, resulting in a final selection of 42 out of 128 countries that met these thresholds.

Each node represents a country, with the size of the node reflecting the number of documents produced. The thickness of the connecting lines indicates the strength of collaborative ties between countries, measured as

Table 7: Geographic distribution and collaboration patterns of corresponding authors.

Country	Articles	SCP	MCP	Freq	MCP_Ratio
China	198	154	44	0.099	0.222
USA	196	169	27	0.098	0.138
India	80	75	5	0.04	0.063
Germany	65	49	16	0.032	0.246
Spain	65	52	13	0.032	0.2
Mexico	41	31	10	0.02	0.244
United Kingdom	37	30	7	0.018	0.189
Japan	35	33	2	0.017	0.057
Brazil	33	31	2	0.016	0.061
Italy	32	20	12	0.016	0.375

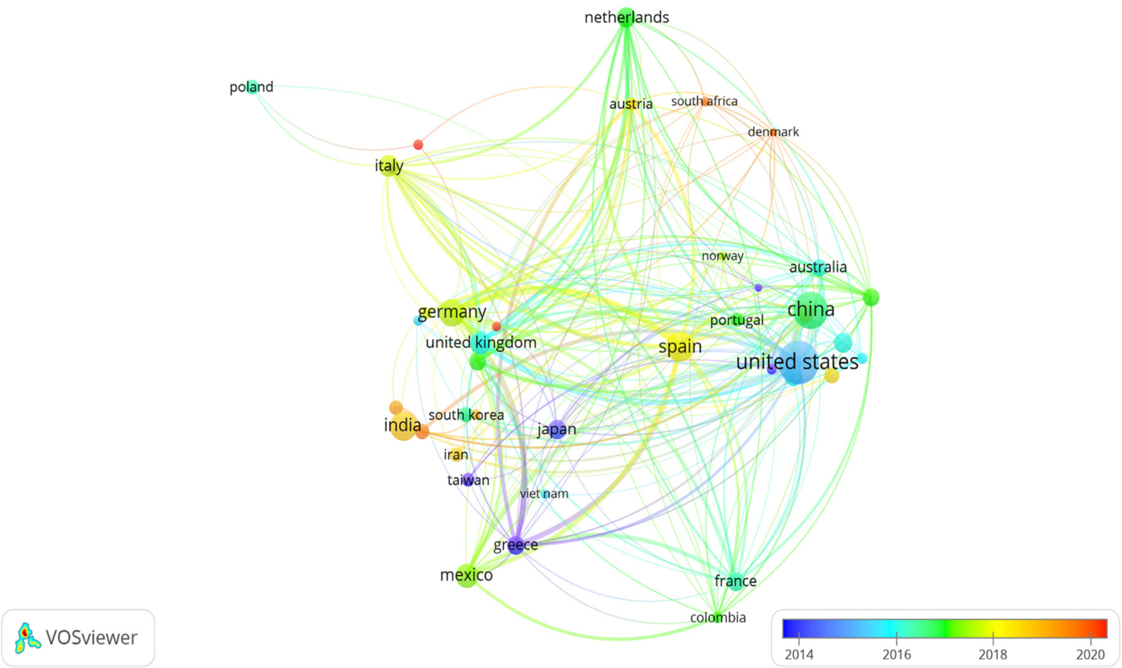

Figure 5: Country-level co-authorship network in name disambiguation research (2005–2024).

co-authorship link strength. The color scale represents the average year of publication, ranging from earlier (blue) to more recent (yellow/orange) research activity. Clusters were identified using a resolution of 5.0, resulting in 6 distinct clusters, suggesting geographic and institutional collaboration trends.

Notably, the United States, China, Spain, and Germany emerge as the most dominant contributors, both in terms of publication volume and international collaboration. The United States leads with 434 documents, 12,997 citations, and the highest total link strength of 190, followed by China (278 docs, 88 link strength), Spain (169 docs, 140 link strength), and Germany (136 docs, 131 link strength). These countries form the core hubs in the network, engaging in widespread co-authorship with other nations.

Countries like India, Mexico, and Greece also show significant collaboration, although with lower centrality. Interestingly, despite its high publication volume (152 documents), India shows relatively weaker co-authorship ties (link strength = 9), indicating a more nationally concentrated research output. In contrast, Belgium, Netherlands, and Switzerland exhibit moderate document counts but high link strength, suggesting strong international engagement.

This map demonstrates that global research on name disambiguation is predominantly driven by collaborations among Western, East Asian, and a few emerging research economies. The overlay of publication years further indicates an increasing trend of international co-authorship in recent years, especially among European and Asian countries.

4.5 Thematic Structure and Keyword Co-occurrence Analysis (Related to Figure 6)

Figure 6 illustrates a co-occurrence network of 345 keywords that appeared at least ten times across the dataset, extracted using full counting from all keywords. The visualization reveals several prominent clusters that represent thematic concentrations in the literature related to author name disambiguation and its associated domains.

The most dominant keywords in terms of occurrence and centrality include “author identification”, “author profiling”, and “classification (of information)”, signifying core research focuses. These terms are tightly

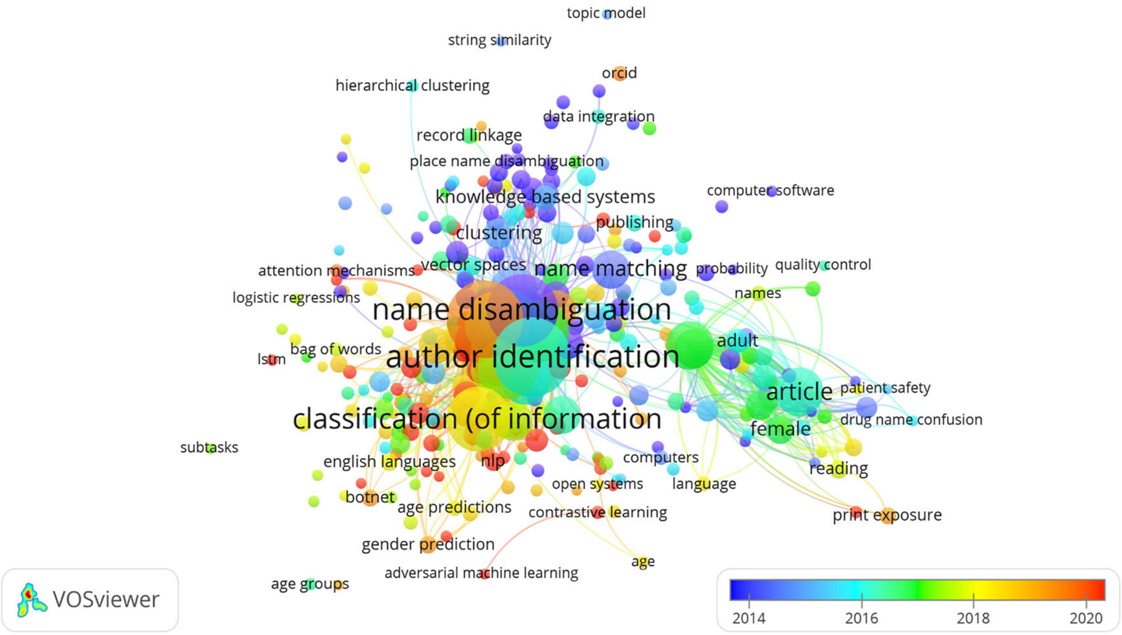

Figure 6: Co-occurrence network visualization of author keywords (2014–2020).

interconnected, forming the central hub of the network. Surrounding these central nodes are clusters indicating subdomains and emerging research interests such as “gender prediction,” “logistic regression,” “record linkage,” “clustering,” “hierarchical clustering,” and “ORCID.”

The color-coded clusters demonstrate topical groupings, with some clusters oriented toward natural language processing (NLP) and machine learning, while others lean toward bibliometric techniques, information retrieval, and privacy concerns in author data. The time-overlay (from 2014 to 2020) further shows the progression and recent intensification of interest in areas like adversarial machine learning, contrastive learning, and author name disambiguation in medical and scientific publishing platforms.

This keyword co-occurrence map highlights the interdisciplinary nature of research on author name disambiguation, bridging computer science, library science, and scientometrics.

5 Discussion

This study aimed to analyze the global research landscape of Author Name Disambiguation (AND) using bibliometric techniques, with the objectives of: (1) mapping publication trends; (2) identifying key sources and authors; (3) analyzing collaboration patterns; (4) evaluating institutional and national contributions; and (5) exploring the thematic and methodological evolution of the field. Drawing from 2,004 records indexed in Scopus from 2005 to 2024, the analysis presents critical reflections grounded in both quantitative output and conceptual development.

5.1 Objective-Driven Growth Patterns and Temporal Shifts

As shown in the annual scientific production data, there has been a marked increase in research output on AND over the past two decades, particularly from 2013 to 2019. This pattern mirrors the growing need to manage the integrity of scholarly metadata amid a global surge in digital publishing and bibliometric assessments. Studies such as Hussain and Asghar (2017) and Cappelli et al. (2025) emphasize the rising complexity of name ambiguity in large-scale repositories like Scopus and Web of Science, which this study confirms through a temporal spike in publication and citation activity. The post-2019 plateau, while notable, suggests that the field may be entering a consolidation phase, wherein foundational frameworks are being adapted rather than re-invented.

5.2 Convergence with Technology-Driven Paradigms

The literature reviewed demonstrates a strong convergence of AND research with machine learning, natural language processing, and semantic web technologies (Wang et al. 2025; Zhou et al. 2024). This is supported by keyword co-occurrence analyses in this study, which show dominant themes such as "author profiling," "classification," "semantic disambiguation," and "ORCID integration." These thematic patterns align with the ongoing trend toward hybrid, context-aware disambiguation models that combine co-authorship networks, publication metadata, and linguistic features.

5.3 Fragmentation in Author Productivity and Limited Continuity

The Lotka's Law distribution of author productivity reveals a high degree of fragmentation in the field. Approximately 78.5 % of contributors authored only a single publication, while fewer than 0.5 % have sustained publication records. This finding aligns with earlier claims by Liu et al. (2014) and Tekles and Bornmann (2020) about the ad hoc nature of AND research participation. The limited long-term engagement by most researchers suggests either the technical barrier to entry or the perception of AND as a secondary research theme within broader areas like digital libraries or information retrieval.

5.4 Collaborative Concentration and Geographic Inequities

A clear clustering of institutional and national collaboration was observed, with leading contributions from China, the United States, Italy, and Germany. The analysis of co-authorship patterns reveals strong regional networks, but relatively low international link strength in countries like India and Brazil. These findings echo concerns in the literature regarding the lack of global standardization and insufficient attention to linguistic and cultural diversity in AND systems (Bouchard and Boudry 2025; Xu and Hu 2024). The literature stresses the necessity for collaboration across geopolitical regions, particularly when building multilingual or culturally adaptive AND systems.

5.5 Core Sources and Disciplinary Silos

This study finds that conferences (e.g., CEUR Workshop Proceedings, LNCS) dominate AND publication venues, reaffirming the literature's portrayal of AND as a predominantly computer science-driven field. Journals such as Scientometrics appear less frequently but exhibit higher normalized citation impact, suggesting that theoretical contributions often emerge from LIS or scientometric communities (Robinson-Garcia et al. 2025). This disciplinary divide poses both a challenge and an opportunity for cross-pollination between technical developers and metadata policy advocates.

5.6 Emerging Gaps and Future-Ready Innovations

Although progress is evident, several gaps remain. The reviewed literature and findings point to the underutilization of blockchain technology, decentralized identifiers, and open interoperability standards in AND frameworks. Only a few initiatives (e.g., ODIN, ReCiter) propose holistic, multilingual, and scalable solutions

(Albert et al. 2021; Fenner et al. 2014). The limited inclusion of regional naming conventions – particularly in South Asia, the Middle East, and Sub-Saharan Africa – risks algorithmic bias and undermines the goal of equitable research evaluation (Donathan et al. 2025).

5.7 Implications

The findings substantiate the claims in prior literature that AND is both a technical and policy issue. While algorithmic improvements continue to dominate the research agenda, broader institutional engagement, metadata curation standards, and persistent identifier integration remain critical for long-term solutions. For librarians, bibliometricians, and platform developers, this study underlines the need for cross-sector collaboration and alignment between machine-readable author identifiers and human-readable authority records.

6 Conclusions

This study offers a comprehensive bibliometric assessment of Author Name Disambiguation (AND) research using 2,004 publications indexed in Scopus from 2005 to 2024. Guided by five objectives, it mapped publication trends, source distribution, author productivity, collaborative structures, and thematic patterns. Findings show steady growth in scholarly output, especially after 2013, reflecting increasing attention to accurate author attribution in digital bibliographic databases. The dominance of conference proceedings and a small group of prolific authors suggests that the field is technologically driven but lacks widespread and sustained participation.

Co-authorship networks and country-level collaborations revealed a concentration of research activity in China, the United States, and parts of Europe. In contrast, underrepresentation from regions such as South Asia, Africa, and Latin America points to persistent global imbalances in metadata research. Thematic analysis highlights the shift toward machine learning–enabled disambiguation methods, with frequent use of terms such as “semantic disambiguation,” “author profiling,” and “classification,” underscoring a clear movement toward more sophisticated, automated frameworks.

Despite these contributions, the study is subject to several limitations. It relied solely on Scopus, potentially excluding relevant literature indexed in other databases or published in non-English languages. The keyword-based analysis, while effective for identifying broad themes, does not capture full-text or methodological details. Additionally, advanced techniques for tracking thematic evolution over time, such as those offered in Biblioshiny, were not applied.

Building on these findings, several directions are recommended. Future studies should expand their data sources and languages to capture a more representative global view. The adoption of advanced bibliometric tools, including thematic evolution mapping and author trajectory analysis, would offer deeper conceptual insights. Targeted efforts to include underrepresented regions are essential for building equitable global disambiguation systems. Interdisciplinary collaboration – particularly between computer scientists, librarians, and policy makers – can help bridge technical development with practical metadata governance. Although not directly examined in this study, the integration of persistent identifiers like ORCID remains a promising avenue for further empirical investigation.

This study provides a structured understanding of the growth, contributors, and knowledge dynamics of AND research. It serves as a foundation for future work that aspires to develop more inclusive, accurate, and interoperable systems for author identification in scholarly communication.

Author agreement/declaration: This statement certifies that all authors have seen and approved the final version of the submitted manuscript. This manuscript has not received prior publication and is not being considered for publication elsewhere.
Funding information: The authors state no funding involved.

Author contributions: The authors contributed equally to the conceptualization, intellectual discussion underlying this study, literature exploration, writing, reviews and editing, and accepted responsibility for the content and interpretation.
Conflict of interest: The authors state no conflict of interest.

References

Albert, P.J., Dutta, S., Lin, J., Zhu, Z., Bales, M., Johnson, S.B., Mansour, M., Wright, D., Wheeler, T.R., and Cole, C.L. (2021). ReCiter: an open source, identity-driven, authorship prediction algorithm optimized for academic institutions. *PLoS One* 16: e0244641.

Bordons, M., Moreno-Solano, L., and González-Albo, B. (2024). ORCID identifier adoption in Spanish scholarly communication: a macro and micro level perspective. *Learn. Publ.* 37: e1606.

Bouchard, A. and Boudry, C. (2025). Knowledge and use of the ORCID author identifier in France: a national survey. *Learn. Publ.* 38: e2004.

Cappelli, F., Colavizza, G., and Peroni, S. (2025). Recent developments in deep learning-based author name disambiguation. *Proc. 21st Conf. Inf. Res. Sci. Connect. Digital Libr. Sci. (IRCDL 2025)* 3937: 15.

Delgado, A.D., Montalvo, S., Martínez Unanue, R., and Fresno, V. (2018). A survey of person name disambiguation on the web. *IEEE Access* 6: 59496–59514.

Donathan, D.I., Nason, M., Tullney, M., Shi, J., and Alperin, J.P. (2025). *Evaluating multilingual metadata quality in crossref* (No. arXiv:2503.11853). arXiv, https://doi.org/10.48550/arXiv.2503.11853.

Duran-Silva, N., Accuosto, P., Przybyła, P., and Saggion, H. (2024). AffilGood: building reliable institution name disambiguation tools to improve scientific literature analysis. In: Ghosal, T., Singh, A., Waard, A., Mayr, P., Naik, A., Weller, O., Lee, Y., Shen, S., and Qin, Y. (Eds.). *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*. Association for Computational Linguistics, Bangkok, Thailand, pp. 135–144, Available at: https://aclanthology.org/2024.sdp-1.13/.

Fenner, M., Haak, L.L., Thorisson, G.A., Ruiz, S., Vision, T.J., and Brase, J. (2014). ODIN: the ORCID and DataCite interoperability network. *Int. J. Knowl. Learn.* 9: 305–325.

Ferreira, A.A., Gonçalves, M.A., and Laender, A.H. (2012). A brief survey of automatic methods for author name disambiguation. *ACM Sigmod Record* 41: 15–26.

Firdaus, F., Fahreza, I., Nurmaini, S., Darmawahyuni, A., Sapitri, A.I., Rachmatullah, M.N., Lestari, S.D., Fachrurrozi, M., Afrina, M., and Putra, B.W. (2022). Identification of Indonesian authors using deep neural networks. *Comput. Eng. Appl. J.* 11: 15–24.

Hussain, I. and Asghar, S. (2017). LUCID: author name disambiguation using graph structural clustering. In: *2017 Intelligent Systems Conference (IntelliSys)*, London, UK, pp. 406–413.

Islamaj Dogan, R., Murray, G.C., Névéol, A., and Lu, Z. 2009. Understanding PubMed® user search behavior through log analysis. *Database*, 2009: bap018.

Kim, J. (2018). Evaluating author name disambiguation for digital libraries: a case of DBLP. *Scientometrics* 116: 1867–1886.

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

Kim, J. and Kim, J. (2024). ANDez: an open-source tool for author name disambiguation using machine learning. *SoftwareX* 26: 101719.

Liu, W., Islamaj Doğan, R., Kim, S., Comeau, D.C., Kim, W., Yeganova, L., Lu, Z., and Wilbur, W.J. (2014). Author name disambiguation for PubMed. *J. Assoc. Inf. Sci. Technol.* 65: 765–781.

Manzoor, A., Asghar, S., and Amjad, T. (2022). Toward a new paradigm for author name disambiguation. *IEEE Access* 10: 76055–76068.

Morgan, M. and Eichenlaub, N. (2018). ORCID iDs in the open knowledge era. In: Méndez, E., Crestani, F., Ribeiro, C., David, G., and Lopes, J.C. (Eds.). *Digital libraries for open knowledge*, Vol. 11057. Springer International Publishing, Cham, Switzerland, pp. 308–311.

Mryglod, O., Nazarovets, S., and Kozmenko, S. (2023). Peculiarities of gender disambiguation and ordering of non-English authors' names for Economic papers beyond core databases①. *J. Data Inf. Sci.* 8: 72–89.

Panigabutra-Roberts, A. (2025). Representing researchers in the library linked data environment: a case study of ORCID users at the University of Tennessee, Knoxville. *J. Librariansh. Sch. Commun.* 13: eP18195.

Pooja, Km., Mondal, S., and Chandra, J. (2021). Exploiting similarities across multiple dimensions for author name disambiguation. *Scientometrics* 126: 7525–7560.

Protasiewicz, J. and Dadas, S. (2016). A hybrid knowledge-based framework for author name disambiguation. In: *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Budapest, Hungary, pp. 000594–000600.

Rehs, A. (2021). A supervised machine learning approach to author disambiguation in the web of science. *J. Informetr.* 15: 101166.

Robinson-Garcia, N., Corona-Sobrino, C., Chinchilla-Rodríguez, Z., Torres-Salinas, D., and Costas, R. (2025). The use of informetric methods to study diversity in the scientific workforce: a literature review. *Quant. Sci. Stud.* 6: 652–685.

Rodrigues, N.S. and Ralha, C.G., 2024. A hybrid machine learning method to author name disambiguation. In: *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*. SBC, Brasil, pp. 108–117.

Santini, C., Gesese, G.A., Peroni, S., Gangemi, A., Sack, H., and Alam, M. (2022). A knowledge graph embeddings based approach for author name disambiguation using literals. *Scientometrics* 127: 4887–4912.

Sebo, P., De Lucia, S., and Vernaz, N. (2021). Accuracy of PubMed-based author lists of publications and use of author identifiers to address author name ambiguity: a cross-sectional study. *Scientometrics* 126: 4121–4135.

Shi, J., Nason, M., Tullney, M., and Alperin, J. (2025). Identifying metadata quality issues across cultures. *Coll. Res. Libr.* 86: 101.

Tekles, A. and Bornmann, L. (2020). Author name disambiguation of bibliometric data: a comparison of several unsupervised approaches. *Quant. Sci. Stud.* 1: 1510–1528.

Wang, G., Sun, Z., Hu, W., and Cai, M. (2025). Author name disambiguation based on heterogeneous graph neural network. *PLoS One* 20: e0310992.

Xu, S.B. and Hu, G. (2024). Rethinking the author name ambiguity problem and beyond: the case of the Chinese context. *Account. Res.* 32: 1–24.

Zhou, Q., Chen, W., Zhao, P.-P., Liu, A., Xu, J.-J., Qu, J.-F., and Zhao, L. (2024). Towards effective author name disambiguation by hybrid attention. *J. Comput. Sci. Technol.* 39: 929–950.