Research on the Applicability of Author Identifiers in Scientific Paper Databases

Authors: Dongbo Shi, Deng Hui, Yang Zhijian, Liu Ningjie, Liu Yuxiu, Mao Yufei, Shi Dongbo

Date: 2024-06-06T00:00:00+00:00

Abstract

Objective: To examine the coverage and accuracy of author identifiers in major scientific paper databases, and to verify whether they can be directly utilized for empirical research in scientometrics and science & technology policy.

Methods: Using publications from 825 Chinese scientists as a standard dataset, we retrieved and collected scientist identifiers and their publication information from scientific paper databases, calculated data coverage, accuracy, and robustness, and employed a difference-in-differences method for experimental replication to test database applicability.

Results: First, the WOS, Scopus, AMiner, and OpenAlex databases can retrieve over 90% of Chinese scientist identifiers, while ORCID coverage is less than 50%. Second, Scopus demonstrates the highest accuracy at 85.2%, whereas OpenAlex shows the lowest at only 51.2%. Third, directly employing database author identifier data for empirical research introduces non-negligible errors.

Limitations: The accuracy dataset primarily comprises young scientists and does not cover social sciences and humanities at the disciplinary level, representing certain limitations.

Conclusion: Author identifiers in current major databases cannot yet be directly applied to large-scale empirical research; establishing a standardized scientist achievement certification information platform could enhance the accuracy of Chinese author name identification.

Full Text

A Study on the Applicability of Author Identification Numbers in Scientific and Technical Paper Databases

Dongbo Shi[1], Hui Deng[2], Zhijian Yang[3], Ningjie Liu[1], Yuxiu Liu[4], Yufei Mao[5]

[1]School of International and Public Affairs, Shanghai Jiao Tong University, Shanghai, 200030, China
[2]School of Economics and Management, Nanjing University of Technology, Jiangsu, 210000, China
[3]School of Management, Zhejiang University, Zhejiang, 310058, China
[4]School of Juris Masters, China University of Political Science and Law, Beijing, 100091, China
[5]School of Economics and Management, Tongji University, Shanghai, 200092, China

This work is supported by the National Natural Science Foundation of China (Grant No. 72374140).

Abstract

[**Purpose**] To evaluate the coverage and accuracy of author identification numbers (author IDs) in major bibliographic databases and assess whether they can be directly used in empirical research in the field of science of science and science policy.

[**Methods**] The ground truth dataset consists of publications from 825 Chinese scientists. We retrieve and collect scientist identifiers and their publication information from bibliographic databases to calculate coverage, accuracy, and robustness metrics. We further assess the validity of these author IDs for empirical research by replicating a top-tier journal article using data collected through the author IDs.

[**Results**] First, WOS, Scopus, AMiner, and OpenAlex can retrieve more than 90% of Chinese scientists' identifiers, while ORCID' s coverage is less than 50%. Second, Scopus achieves the highest accuracy at 85.2%, whereas OpenAlex' s accuracy is only 51.2%. Third, directly using publication data collected through author IDs for empirical research introduces non-negligible bias.

[**Limitations**] The ground truth dataset is primarily composed of young scientists and does not cover social sciences and humanities, which presents certain limitations.

[**Conclusion**] Currently, author identification numbers from major databases cannot be directly applied to large-scale empirical research. Establishing a standardized information platform for scientist publication certification could improve the accuracy of Chinese author name disambiguation.

Keywords: Bibliographic databases; Author identification number; Author-name disambiguation

1. Introduction

As the world undergoes unprecedented changes unseen in a century, the global economy has entered a downward cycle, and competition among major powers for technological supremacy has intensified like never before. Scientific and technological innovation has become the primary battlefield for international strategic competition. The decisive force in this competition lies in talent, making the implementation of a talent-strong nation strategy a major and urgent task for the Party and the country. At the Central Talent Work Conference, General Secretary Xi Jinping pointed out that although China "already possesses a large, high-quality, and increasingly optimized talent pool that plays an increasingly prominent role," the reform of the talent development system and mechanism has been insufficient in both "breaking old patterns"and "establishing new ones," and a talent development system with both Chinese characteristics and international competitive advantages has not yet been truly established [1]. Since the 18th National Congress, China' s scientific research has achieved new historic accomplishments, with China' s high-quality papers ranking first in the world for the first time [2]. China stands at a critical juncture of transitioning from quantity to quality, from catching up to leading. Against the backdrop of sustained R&D investment and significant progress in higher education, building a talent system that conforms to the laws of scientific research and supports original innovation is key to becoming a technological powerhouse. Among these, incentive systems including compensation design [3] and talent evaluation constitute the foundation of the talent development system [4], affecting the efficiency of transforming China' s scientific and technological resources into outputs and requiring long-term in-depth research.

Research on talent mechanisms and systems cannot proceed without the support of science of science theory and empirical studies. Talent evaluation, talent programs, and incentive system reforms all require precise policy evaluation, which demands that research units and data shift from the regional and institutional level to the individual scientist and team level. Bibliometric data is an indispensable foundation for such research, with databases like Web of Science and Scopus commonly used to study scientist evaluation [5,6], mobility [7,8], and incentives [9]. However, many scientists share identical surnames and given names (or initials), making it challenging to distinguish authors with the same name in databases as distinct scientists in reality—a problem known as author name disambiguation. This phenomenon is particularly severe among Chinese scientists [10]. Without solving this problem, accurate scientist-level empirical research becomes impossible, let alone theoretical and policy research.

Therefore, this study uses the ground truth dataset of 825 Chinese scientists

compiled by Shi et al. [11] to examine the coverage and accuracy of author identification numbers in major bibliographic databases—Web of Science, Scopus, OpenAlex, ORCID, and AMiner—and tests whether these database identifiers can be directly used in empirical research through replication experiments.

2. Related Research

Current approaches to author name disambiguation fall into two categories: algorithmic automatic generation and author self-reporting (claiming). The former offers more comprehensive coverage, while the latter provides higher accuracy. Author name disambiguation algorithms use analytical or pattern recognition methods to cluster papers with similar author names, automatically generating author identification numbers [12,13]. The most notable example is the algorithm developed by Torvik and Smalheiser in 2009 for the MEDLINE database, which was eventually integrated into PubMed [14] and laid the foundation for research on scientific collaboration [15], research direction selection [16], gender issues [17], and peer review [18]. Unfortunately, Torvik and Smalheiser' s data only supports research in medicine and life sciences and cannot be applied more broadly across disciplines [14]. OpenAlex uses machine learning algorithms to perform name disambiguation for all authors and has open-sourced both its algorithm source code and data, injecting new momentum into science of science research. However, most other disambiguation algorithm developers have not provided open-source algorithms and data, and the computational resources required to replicate these algorithms often exceed the capabilities of science of science researchers.

To address name ambiguity issues in databases, major bibliographic database operators and non-profit organizations have chosen an alternative technical route. In 2008, the Web of Science (WOS) launched ResearcherID, a unique identifier that allows scientists to register and claim their papers within the WOS database. In 2012, the non-profit organization Open Researcher and Contributor Identifier (ORCID) released user identifiers, enabling authors to register and maintain personal education and work histories as well as publication records. Today, many international journals require authors to provide their ORCID upon manuscript submission [19]. Scopus' s Scopus Author Identifier combines automatic generation algorithms with scientist feedback [20].

These database author identification numbers provide new high-quality research data for science of science studies at the individual or team level. For example, Moed et al. used Scopus AuthorID to study immigrant scientists [20], while Khurana and Sharma combined ResearcherID, AuthorID, and ORCID to study how the h-index can be used for scientist evaluation [21]. Such data have recently been applied to research on Chinese scientists, such as Zhao et al.'s use of ORCID data to demonstrate that returnee scientists do not exhibit stronger publication capabilities than domestic scientists [22], a conclusion that contradicts common perceptions in academia.

The accuracy and coverage of author identification numbers in bibliographic databases directly affect the reliability and validity of empirical papers using these data. Conclusions drawn from inaccurate data can be misleading, while conclusions from accurate but incomplete data often lack representativeness. Therefore, it is essential to examine the applicability of author identification numbers. Using data from 193 German Leibniz laureates, Aman demonstrated that Scopus AuthorID achieves recall and precision rates as high as 97% and 100%, respectively [23], and confirmed that Scopus AuthorID can be used to track scientists' international mobility. Kawashima and Tomizawa used Japan' s KAKEN funding database to show that Scopus AuthorID' s recall and precision rates are 98% and 99%, respectively [24]. Boudry and Durand-Barthez found that both ORCID and ResearcherID cover less than 20% of a sample of French scientists, with many IDs not including complete publication records [25]. Evidently, the accuracy and coverage of author identification numbers vary significantly across different populations. Notably, no existing research has examined these issues specifically for Chinese scientists, which limits the application of relevant author identifiers in China' s science of science and science policy research.

3. Data and Methods

3.1 Standard Dataset

This study uses the ground truth dataset of publications from 825 Chinese scientists compiled by Shi et al. [11] as the standard dataset (Table 1). This dataset includes Chinese scientists who earned their PhDs between 1997 and 2014, with an average graduation year of 2007. The sample comprises 14% women, 18% who earned PhDs in mainland China, and 65% who earned PhDs in the United States. By 2019, 49% of the scientists worked at academic institutions in mainland China, 42% in the United States, and the remainder primarily in Europe, Japan, and Hong Kong. The dataset covers all natural science fields, with the largest representation from engineering and materials science (22%) and medicine (21%), and the smallest from earth sciences (9%). Thus, this dataset is representative as a ground truth.

Shi et al. [11] collected these scientists' complete publication records from PhD graduation through 2019 in SCI/SSCI-indexed journals from personal homepages (41%), Google Scholar (39%), ResearchGate (12%), and other sources (Table 2). The dataset includes an average of 56 papers per scientist. Notably, 60 scientists have missing data for certain years; for these cases, publications from missing years were excluded from subsequent calculations. Importantly, using the 765 scientists with complete publication records does not change the study's conclusions. Additionally, the dataset includes three high-energy physics scientists with 850, 919, and 1,174 publications, respectively; removing these three scientists also does not alter the conclusions.

3.2 Author Identification Numbers and Publications in Bibliographic Databases

Based on the scientists' work histories and research fields in the standard dataset, we retrieved corresponding author identification numbers from bibliographic databases. We selected four commonly used bibliographic indexing databases in science of science and science policy research—Web of Science, Scopus, OpenAlex, and AMiner. The first three provide personal author identification numbers (AuthorID), while AMiner provides scientist profile pages containing publication lists.

ORCID (Open Researcher and Contributor Identifier) is a user identifier launched on October 16, 2012, by the non-profit organization ORCID. It assigns a unique 16-digit identifier to each registered scientist, providing researchers with a unique identity. Scientists can link their ORCID accounts to publications in WOS and Scopus [26]. In 2012, ORCID was integrated into WOS ResearcherID. It is important to note that this study examines ORCID as a single data source, focusing on scientists whose ORCID records are complete but whose standard dataset does not include Google Scholar or ResearchGate data.

Web of Science (WOS), established in 1964, covers academic journals, conference papers, and citation data globally across natural sciences, social sciences, arts, and humanities. Its Science Citation Index (SCI) and Social Sciences Citation Index (SSCI) are authoritative datasets for science of science and science policy research. As of December 2023, SCI includes over 9,500 journals and 61 million papers, while SSCI includes over 3,500 journals and 10 million papers. Since 2008, WOS has offered the unique identifier ResearcherID. Initially, ResearcherID required user registration and manual linking to WOS papers; later, it introduced automatic generation algorithms to create author identifiers for unclaimed papers [27].

Scopus, launched by Elsevier in 2004, is an abstract and citation database dating back to 1966, covering life sciences, social sciences, natural sciences, and medicine. Using information on authors and publications—such as affiliations, research fields, article titles, citations, and co-authors—Scopus employs advanced algorithms to assign a unique Scopus Author Identifier to each author, automatically distinguishing homonymous authors and matching name variations [20].

AMiner, launched in March 2006, is a next-generation scientific intelligence analysis and mining platform developed by Professor Jie Tang' s team at Tsinghua University' s Department of Computer Science and Technology. It aggregates global scholar profiles, institutional profiles, and journal profiles across all disciplines, including natural sciences, social sciences, and humanities [28]. AMiner extracts and integrates academic data from distributed networks, creates semantic-based profiles for each researcher, uses generative probabilistic models to model papers, authors, and venues, discovers interesting patterns in researcher social networks, and provides search services such as expertise search

and relation search, offering researchers a profile dataset [29].

OpenAlex, launched by OurResearch in January 2022, is a free and open global academic research database indexing open-access journals and research outputs across disciplines [30]. OpenAlex comprises five entity types: works, authors, institutions, venues, and concepts. It inherits data from Microsoft Academic Graph and performs name disambiguation for all authors using machine learning algorithms [31].

We searched for scientists from the standard dataset in these five databases and recorded their corresponding identifiers and publication information. ORCID, WOS, Scopus, and AMiner offer web search functionality. The retrieval process is illustrated in Figure 1 [Figure 1: see original paper]. First, we input scientist information (name, institution, etc.) into the database search interface. Second, based on education background, work history, research field, and start time from the standard dataset, we selected matching identifiers with publication records from the search results. If multiple matching identifiers existed, we recorded the top three most relevant ones (for AMiner, we recorded the profile URL), using the most accurate ID for subsequent analysis. Finally, we collected publication information including DOI, title, journal, publication year, accession number, and authors. OpenAlex provides PostgreSQL data; we matched author IDs in the authors table through full-name search and then matched these IDs with the author_{id} column in the works table to obtain work IDs and institutional information (raw_{{{affiliation}}_{{string}}}) for each paper. We then retrieved publication details including DOI, title, journal, publication year, accession number, and authors. The appendix describes the retrieval process and data collection for scientist Chen Jianing across different databases.

3.3 Matching Database Publications with the Standard Dataset

For WOS and ORCID datasets, we obtained ORCID for 402 scientists (24,181 papers) and 1,115 WOS ResearcherIDs for 777 scientists (45,485 papers). We connected ORCID and WOS datasets with the standard dataset using WOS accession numbers.

For Scopus, AMiner, and OpenAlex, we followed these steps:

Step 1: Restrict database publication scope. Since the standard dataset only includes papers published in SCI/SSCI-indexed journals, we first limited database publications to those in SCI/SSCI-indexed journals based on the Journal Citation Report' s annual journal lists. We also restricted publications to the same years covered in the standard dataset for each scientist. As shown in Table 3 , the proportions of Scopus, AMiner, and OpenAlex data covered by SCI/SSCI are 72.9%, 67.1%, and 61.9%, respectively.

Step 2: Match papers by Digital Object Identifier (DOI). Based on the previous step, we directly matched papers with identical DOIs. The matching rates for Scopus, AMiner, and OpenAlex are 77.2%, 65.5%, and 4.7%, respec-

tively. OpenAlex' s proportion is significantly smaller because its number of candidate IDs and papers is two orders of magnitude larger than other databases.

Step 3: Match by journal, publication year, and exact title. For papers missing DOI information (in either database or standard dataset), we matched by exact journal, publication year, and title. The matching rates for the three databases in this step are 5.1%, 4.2%, and 0.2%, respectively.

Step 4: Fuzzy title matching with manual verification. For papers with exact journal and year matches but non-exact title matches, we calculated title similarity (defined as the proportion of overlapping words after removing symbols) and manually verified pairs with similarity exceeding 80%. This step matched 0.9% of Scopus papers and 0.7% of AMiner papers. OpenAlex's volume was too large for manual verification, potentially underestimating its accuracy by approximately 1%. However, as we will see, this proportion has negligible impact on the final accuracy evaluation.

Through these four steps, we identified intersections between each database' s publication set and the standard dataset for evaluating author identification performance.

3.4 Evaluation Metrics

We assess database author identification numbers using coverage, accuracy, and robustness metrics. For scientists with multiple retrieved identifiers, we use average metrics as final measures.

Coverage (CV): The proportion of scientists for whom author identification numbers can be retrieved under search conditions. This metric determines the database' s applicability scope.

B3 Precision (BP), B3 Recall (BR), and B3 F1-score (BF1) are defined as follows:

$$BP = \frac{|D_i \cap G_i|}{|G_i|}$$

$$BR = \frac{|D_i \cap G_i|}{|D_i|}$$

$$BF1 = \frac{2 \times BP \times BR}{BP + BR}$$

where D_i represents the publication dataset under a scientist' s database identifier, G_i represents scientist i' s standard publication dataset, and N represents the number of scientists retrieved from the database. The B3 accuracy metric is commonly used in literature to measure algorithmic performance [32]. BP describes the proportion of identified papers that truly belong to the scientist,

while BR describes the proportion of the scientist' s actual publications that are correctly identified. Since BP and BR involve a trade-off (e.g., a high-precision algorithm may miss more papers), we use their harmonic mean to represent overall performance.

To measure robustness, we introduce standard deviations of precision and recall:

$$SD_{BP} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{|D_i \cap G_i|}{|G_i|} - BP\right)^2}$$

$$SD_{BR} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{|D_i \cap G_i|}{|D_i|} - BR\right)^2}$$

3.5 Empirical Experiment

To further validate whether author identification numbers can be used in science of science and science policy empirical research, we replicate Shi et al.' s [11] study using datasets identified from different databases to test whether consistent conclusions can be drawn. Shi et al. [11] investigate whether young Chinese scientists have more successful careers after returning to China compared to their peers who remain overseas. Using a standard difference-in-differences approach with publication count as the dependent variable, the study employs the interaction between a return-to-China indicator and time dummies as the key independent variable, controlling for individual and year fixed effects. The regression equation is:

$$Y_{i,t} = \alpha + \beta Post_{i,t} \times Treat_i + \gamma Post_{i,t} + \eta_i + \mu_t + \varepsilon_{i,t}$$

where $Y_{i,t}$ represents scientist i' s number of publications in year t, $Treat_i$ indicates whether scientist i returned to China, and $Post_{i,t}$ indicates the post-return period. The study uses a matching strategy, only including scientists with similar age, educational background, and research capabilities in the regression.

4. Results

4.1 Coverage and Accuracy

As shown in Table 4 , the coverage rates of identification numbers in WOS, Scopus, AMiner, and OpenAlex all exceed 91%, with Scopus highest at 98.5% and OpenAlex second at 96.7%. Notably, ORCID' s coverage is merely 48.7%, far below expectations. Although this exceeds Boudry and Durand-Barthez' s [25] findings, it still means over half of Chinese scientists have not registered ORCID or maintained their profiles, making retrieval impossible. Since the standard dataset comprises young scientists, coverage might be even lower for more senior Chinese scientists. From a coverage perspective, WOS, Scopus,

AMiner, and OpenAlex can locate the vast majority of Chinese scientist identifiers for empirical research, but ORCID' s insufficient coverage may introduce non-negligible bias.

Precision varies significantly across databases. Scopus achieves the highest precision at 83.1%, but this is substantially lower than results reported by Aman [23] and Boudry and Durand-Barthez [25], indicating Scopus' s algorithm performs worse for Chinese scientists than for other populations. Surprisingly, ORCID' s precision, though higher than WOS, AMiner, and OpenAlex, is only 82.6%, far from the expected near-100% (theoretically, ORCID data should be highly accurate as it is personally maintained). Two factors likely contribute to this: first, retrieval of incorrect ORCIDs yields zero precision, accounting for 2.7% of all identifiers; second, ORCID allows authors to use third-party platforms (e.g., Scopus, Crossref) to manage personal data, which automatically imports platform data into ORCID, reducing precision.

WOS and AMiner achieve precision rates of 63.5% and 73.6%, respectively, while OpenAlex' s precision is only 39.7%, meaning these databases assign non-authored papers to scientists.

Recall rates are relatively similar across databases. Scopus leads at 87.4%, while the other four databases range between 72.4% and 73.8%. WOS, Scopus, and OpenAlex have higher recall than precision, causing their author IDs to overestimate scientists' publication counts. OpenAlex shows the largest bias, overestimating by nearly 200%. Conversely, ORCID and AMiner underestimate publication counts, with ORCID underestimating by an average of 3.1 papers per scientist.

Overall, Scopus demonstrates the highest accuracy with an F1-score of 85.2%, at least 7% higher than other databases, likely due to Scopus' s emphasis on continuous improvement of its name disambiguation algorithm and attention to Chinese scientist datasets. Additionally, Scopus shows significantly higher robustness than the other four databases. As Figure 2 [Figure 2: see original paper] illustrates, Scopus performs best overall, while OpenAlex performs worst with an F1-score of only 51.2%, possibly because: first, OpenAlex lacks scientist authentication and verification mechanisms; second, OpenAlex has not used high-quality Chinese scientist datasets to train its algorithm.

4.2 Heterogeneity

Author name disambiguation essentially identifies a scientist' s true papers (accurate papers) from a set of papers with identical author names (candidate papers). The closer the information between candidate and accurate papers, the greater the challenge. When more individuals with the same name work in a scientist' s institution and field, filtering accurate papers becomes more difficult. Since workplace (including region) and field are important variables in empirical research, data bias directly distorts research conclusions.

We divide each scientist' s work experience into mainland China and overseas components to examine database accuracy across regions (Table 5). Contrary to expectations, most databases (except ORCID) show higher accuracy for papers published while scientists work in mainland China. Additionally, except for AMiner, all databases overestimate scientists' publication counts, particularly for work conducted in mainland China.

We further examine disciplinary differences in database accuracy (F1-score), categorizing scientists into six fields: chemistry, earth and environmental sciences, engineering and materials science, information science, life sciences, and mathematics and physics. As Figure 3 [Figure 3: see original paper] shows, all databases (except OpenAlex) demonstrate substantially lower accuracy in information science than in other fields. Scopus shows smaller disciplinary variation and outperforms other databases' maximum accuracy across all fields.

Finally, we regress database accuracy (F1-score) on scientists' personal characteristics. Table 6 shows that ORCID, WOS, and Scopus achieve higher accuracy for younger scientists, while ORCID and AMiner show higher accuracy for scientists working in mainland China, though the magnitude is limited. OpenAlex demonstrates lower accuracy for female scientists than for male scientists, suggesting that using OpenAlex may incorrectly estimate gender differences in research productivity.

4.3 Replication Results

Table 7 presents replication results of Shi et al. [11] using datasets from different databases. Column (1) reports results from the standard dataset, while columns (2)-(6) use data from each database. The standard dataset yields a coefficient estimate of 0.210 ($p<0.01$). Estimates using ORCID and OpenAlex are both non-significant, likely due to ORCID' s small sample size and OpenAlex' s low accuracy. Although WOS, Scopus, and AMiner produce positive and significant coefficients, all three models overestimate the true coefficient by 55%, 99%, and 85%, respectively—biases that cannot be ignored. Despite Scopus' s superior performance in coverage, accuracy, and robustness, its 99% overestimation substantially limits its applicability. Therefore, based on this evidence, current database author identification numbers cannot be directly applied to empirical research.

5. Conclusion and Discussion

This study uses the ground truth dataset of 825 Chinese scientists compiled by Shi et al. [11] to examine the coverage and accuracy of author identification numbers in Web of Science, Scopus, OpenAlex, ORCID, and AMiner. We find substantial variation in database accuracy, ranging from 51.2% to 85.2%, with Scopus achieving the highest accuracy and OpenAlex the lowest. While WOS, Scopus, AMiner, and OpenAlex can locate the vast majority of Chinese scientist identifiers for empirical research, ORCID' s coverage of less than half may intro-

duce non-negligible bias. Replication experiments further reveal that database accuracy is affected by scientists' work region and discipline, confirming that current database identification numbers cannot be directly applied to empirical research.

How should researchers use scientist publication data for empirical studies? The answer depends on the research sample and unit of analysis. When the analysis unit is the individual scientist and the sample size is not large, we recommend collecting scientists' complete personal profiles and using name disambiguation algorithms based on career experience and citation networks developed by Liu [33] and Shi [11]. These algorithms achieve significantly higher accuracy than database author identification numbers, operate efficiently, and have been recognized by top international journals. For large-scale data analysis where these algorithms are not feasible, researchers should first test database author identification numbers using a small ground truth dataset and report the robustness of their results. Furthermore, we call on domestic institutions to establish a standardized scientist publication certification platform that assigns unique identifiers to all scientists working in China and incentivizes them to actively maintain their publication records. This would not only solve the author name disambiguation problem for Chinese scientists at its source but also provide valuable training datasets for improving name disambiguation algorithms.

References

[1] Xi Jinping. Deepening the Implementation of the Strategy for Strengthening the Nation with Talents in the New Era to Accelerate the Construction of a World-Class Talent Hub and Innovation Highland[J]. Qiushi, 2021(24):4-15.

[2] Woolston C. Nature Index Annual Tables 2023: China tops natural-science table[J]. Nature, 2023.

[3] Cao Yifan, Sheng Chuangxin, Tong Feng. Problems and Countermeasures of Introducing Foreign Strategic Scientists to the Guangdong-Hong Kong-Macao Greater Bay Area[J]. Science and Technology Management Research, 2023,43(14):78-84.

[4] Wei Haiyong, Li Zuchao. The Establishment and Application of Knowledge-based Talent Incentive Model: Based on the Perspective of Achievement Needs Theory[J]. Science & Technology Progress and Policy, 2008(6): 169-171.

[5] Wang Jiaxun, Qiu Junping. An Analysis of Chemistry Academic Paper: A Case Study of the First Five Batches of "Recruitment Program for Global Young Experts" [J]. Journal of Modern Information, 2019, 39(2):8-16.

[6] Zhang Lihua, Ji Lu, Chen Xin. A study of the difference of researchers' academic performance during their professional career[J]. Science Research Management, 2021,42(5):182-190.

[7] CHEN Kaihua, YANG Yifan, CHEN Guang, ZHANG Ruhao. Research on

the Characteristics of Global Talent Flow at Different Levels: Based on Paper Data of a Hundred-year from Scopus[J]. Science of Science and Management of S.& T, 2023,44(04):3-20.

[8] Wei Licai, Huang Yi. On the Influence of Academic Mobility on the Scientific Research Productivity of Returned Young Science & Engineering Talents[J]. Research in Higher Education of Engineering, 2020,(01):67-73.

[9] Teng Guangqing, Lyu Jing, Jiang Yao, Tuo Rui, Peng Jie. Impact of Research Funding on Research Topics Based on STM[J]. Journal of Modern Information, 2022,42(05):58-68.

[10] Kim J, Kim J, Kim J. Effect of Chinese characters on machine learning for Chinese author name disambiguation: A counterfactual evaluation[J]. Journal of Information Science, 2023, 49(3): 711-725.

[11] Shi D, Liu W, Wang Y. Has China' s Young Thousand Talents program been successful in recruiting and nurturing top-caliber scientists?[J]. Science, 2023, 379(6627): 62-65.

[12] Elliott S. Survey of author name disambiguation: 2004 to 2010[J]. Library Philosophy and Practice, 2010, 473: 1-11.

[13] Ferreira A A, Gonçalves M A, Laender A H F. A brief survey of automatic methods for author name disambiguation[J]. Acm Sigmod Record, 2012, 41(2): 15-26.

[14] Torvik V I, Smalheiser N R. Author name disambiguation in MEDLINE[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2009, 3(3): 1-29.

[15] Freeman R B, Huang W. Collaborating with people like me: Ethnic coauthorship within the United States[J]. Journal of Labor Economics, 2015, 33(S1): S289-S318.

[16] Myers K. The elasticity of science[J]. American Economic Journal: Applied Economics, 2020, 12(4): 103-134.

[17] Zhou S, Chai S, Freeman R B. Gender homophily: In-group citation preferences and the gender disadvantage[J]. Research Policy, 2024, 53(1): 104895.

[18] Krieger J L, Myers K R, Stern A D. How Important is Editorial Gatekeeping? Evidence from Top Biomedical Journals[J]. The Review of Economics and Statistics, 2021: 1-33.

[19] Carter C B, Blanford C F. All authors must now supply ORCID identifiers[J]. Journal of Materials Science, 2017, 52(11): 6147-6149.

[20] Moed H F, Aisati M, Plume A. Studying scientific migration in Scopus[J]. Scientometrics, 2013, 94: 929-942.

[21] Khurana P, Sharma K. Impact of h-index on author' s rankings: an improvement to the h-index for lower-ranked authors[J]. Scientometrics, 2022, 127(8):

4483-4498.

[22] Zhao Z, Bu Y, Kang L, et al. An investigation of the relationship between scientists' mobility to/from China and their research performance[J]. Journal of Informetrics, 2020, 14(2): 101037.

[23] Aman V. Does the Scopus author ID suffice to track scientific international mobility? A case study based on Leibniz laureates[J]. Scientometrics, 2018, 117(2): 705-720.

[24] Kawashima H, Tomizawa H. Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan[J]. Scientometrics, 2015, 103(3): 1061-1071.

[25] Boudry C, Durand-Barthez M. Use of author identifier services (ORCID, ResearcherID) and academic social networks (Academia.edu, ResearchGate) by the researchers of the University of Caen Normandy (France): A case study[J]. Plos one, 2020, 15(9): e0238583.

[26] Akers K G, Sarkozy A, Wu W, et al. ORCID author identifiers: A primer for librarians[J]. Medical Reference Services Quarterly, 2016, 35(2): 135-144.

[27] Web of Science [EB/OL].[2023-12-1].https://webofscience.help.clarivate.com/en-us/Content/wos-researcher-id.htm.

[28] AMiner [EB/OL].[2023-12-1].https://www.aminer.cn/introduction/.

[29] Song Y, Situ F, Zhu H, et al. To be the Prince to wake up Sleeping Beauty: The rediscovery of the delayed recognition studies[J]. Scientometrics, 2018, 117: 9-24.

[30] OpenAlex [EB/OL].[2023-12-1].https://openalex.org/about/.

[31] Priem J, Piwowar H, Orr R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts[J]. arXiv preprint arXiv:2205.01833, 2022.

[32] Levin K, Cashore B, Bernstein S, et al. Overcoming the tragedy of super wicked problems: constraining our future selves to ameliorate global climate change[J]. Policy sciences, 2012, 45(2): 123-152.

[33] Liu Weichen, Shi Dongbo, Li Jiang. Name Disambiguation for Chinese Authors Using Their Career Experience and Citation Networks[J]. Journal of Information Resources Management, 2020,10(06):82-89+100.

Corresponding author: Dongbo Shi, ORCID: 0000-0003-1191-9103, Email: shidongbo@sjtu.edu.cn.

Author Contributions: Dongbo Shi: conceptualized the study, designed the methodology, collected and analyzed data, wrote and revised the manuscript; Hui Deng: collected and verified data, wrote and revised the manuscript; Zhijian Yang: collected and verified data; Ningjie Liu: processed data; Yuxiu Liu: collected data; Yufei Mao: collected data.

Conflict of Interest Statement: All authors declare no competing interests.

Supporting Data: [1] Dongbo Shi. Scientist dataset. scientist.csv. [2] Dongbo Shi, Hui Deng, Zhijian Yang, Yuxiu Liu, Yufei Mao. Database author identification numbers and publication information data. (File) iddata.

Appendix: Database Retrieval Procedures

This study uses a ground truth dataset of young scientists to search for these scientists in Web of Science, ORCID, Scopus, AMiner, and OpenAlex databases, collecting their identifiers and publication data. The process is as follows: First, input scientist information (name, institution, etc.) into database search interfaces. Second, verify whether search results match the target scientist by comparing name, education background, work institution, research field, and start time with the ground truth dataset. Third, record the scientist' s identifier; if multiple matching identifiers exist, record the top three most relevant ones (for AMiner, record the profile URL). Fourth, after confirming the scientist' s identity, collect and save publication information to a table named "author name_{unique} ID," including DOI, title, journal, year, accession number, and authors. Figure 1 [Figure 1: see original paper] illustrates this basic retrieval process.

Using scientist Chen Jianing as an example (Figure 2 [Figure 2: see original paper]), we document the retrieval and data collection process across five databases.

1. Web of Science

1. Access the search interface at https://www.webofscience.com/wos/author/search
2. Input the scientist' s name (Figure 3 [Figure 3: see original paper])
3. Refine search by work institution (Figure 4 [Figure 4: see original paper])
4. Collect publication information (DOI, title, journal, year, accession number, authors, Researcher ID) into "2_{chen} jianing_{HPU}-2037-2023"

2. ORCID

1. Access https://orcid.org
2. Input information: first name, last name, institution name (Figure 5 [Figure 5: see original paper])
3. Verify key information consistency (institution, start year, end year) (Figure 6 [Figure 6: see original paper])
4. Collect publication information (DOI, title, journal, year, accession number, authors, ORCID ID, Researcher ID) into "chen jianing_{0000}-0002-7525-1424"

3. Scopus

1. Access https://www.scopus.com/search/form.uri?display=authorLookup#author
2. Input information: first name, last name, institution name (Figure 7 [Figure 7: see original paper])
3. Verify education, work institution, research background, and start time; record ID (Figure 8 [Figure 8: see original paper])
4. Collect publication information (DOI, title, journal, year, accession number, authors, Scopus ID) into "Chen Jianing_{55864209500}.csv"

4. AMiner

1. Access https://www.AMiner.org/
2. Input information: first name, last name, institution (Figure 9 [Figure 9: see original paper])
3. Review basic information: education background, work institution, start time (Figure 10 [Figure 10: see original paper])
4. Collect publication information (DOI, title, journal, year, authors) into "chen jianing_{AMiner}.cn/profile/jianing-chen/54056041dabfae91d3fdb590"

Publication data for Chen Jianing collected from WoS, ORCID, Scopus, and AMiner are saved in tables named "author name_{unique} ID"(Figure 11 [Figure 11: see original paper]), with AMiner scientist IDs being profile URLs.

Note: Figure translations are in progress. See original paper for figures.