DSI DATA SCIENCE AND INFORMETRICS
VOL.3, NO.1, FEB. 2023

DATA PAPER

Most productive Chinese authors between Web of Science and VIP database

Fei Shu

Hangzhou Dianzi University, Hangzhou, China

ABSTRACT

The purpose of this dataset is to compare Web of Science and Chinese bibliometric database in terms of authors and their performance. This dataset includes top 100 most productive authors in over 100 disciplines from two databases. The overlap or difference between two sub-datasets from two databases implies the differences between Web of Science and Chinese bibliometric databases in terms of measuring Chinese research performance.

Background & Summary

China' s research activities, as measured by the number of publications or research and development (R&D) investments, have been experiencing a period of rapid growth over the past quarter-century. As a result, more and more bibliometric studies focus on China and try to evaluate China' s contribution to the world' s scientific activity. However, previous studies find inconsistent results when measuring Chinese research activities using different data sources (Liang, 2003; Meho & Yang, 2007; Shu et al., 2019). The objectives of this dataset are to compare an international bibliometric database (i.e., WoS) with a Chinese bibliometric database in terms of authors and their output, to demonstrate the extent of the overlap between the two groups of Chinese scientific elites in both international and Chinese bibliometric databases, and to determine the effect of disciplines. The results of this study will indicate the extent to which international bibliometric databases can be used to evaluate Chinese national research production as a whole and in individual research disciplines.

Methods

For this dataset, the Web of Science (WoS) and the Chinese Science and Technology Periodical Citation Database (VIP) are used as data sources because of their coverage and representation.

WoS and VIP use different disciplinary classification systems. WoS assigns journals to 232 subject categories while the VIP classifies Chinese literature into 35 fields and 457 subfields. Equivalences between the WoS and VIP disciplinary classification systems were first established based on the descriptions of each subject category. This produced 116 obvious one-to-one matches. Dance was removed from the list since no Chinese publication was found in this WoS category. Therefore, 115 disciplines with equivalent classes across WoS and VIP were compared in this study, which account for 66.08 % of Chinese publications

(959,728 of 1,452,380) in WoS and 65.15% of literature (19,472,497 of 29,889,566) in VIP. This list includes 83, 21 and 12 disciplines in Natural Sciences, Social Sciences, and Arts and Humanities respectively.

Some inconsistencies between the WoS and VIP classification systems were also found. WoS adopts the journal classification system assigning indexed journals to roughly 250 WoS categories while VIP classifies the discipline at the paper level (the paper classification system) using the Chinese Library Classification Scheme. An inclusive classification is applied to both databases; in other words, journals or papers may be assigned to one or multiple disciplines, which produces 1,240,677 and 22,727,318 assignments in WoS and VIP respectively.

All papers with a Chinese address (CU = Peoples R China) published between 2008 and 2015 (n=1,452,380) as well as their bibliographic information were retrieved from WoS and assigned to relevant disciplines. In the 115 selected disciplines, Chinese authors contributed the most papers in *Chemistry, Physics* (92,342), followed by *Engineering, Electrical & Electronic* (70,318) and *Optics* (49,038) while they only contributed 2, 5 and 6 papers in *Folklore, Literary Theory & Criticism* and *Film, Radio & Television* respectively. On the other hand, 29,940,090 Chinese papers published between 2008 and 2015 were indexed by VIP under 457 subfields (disciplines), ranging from 1,667 papers in *Physics, Condensed Matter* to 4,223,457 papers in *Education & Educational Research* in the 115 selected disciplines. No correlation was found between WoS and VIP in terms of the number of publications among these 115 disciplines.

In each discipline, Chinese authors were ranked by their number of published papers during the period of 2008-2015 in both WoS and VIP dataset. The top 100 (and tied) authors in the 115 disciplines were retrieved and formed 115 pairs of author groups, for a total of 26,969 records in the two databases.

Although WoS indexes the complete first name of the authors from 2008 onwards, author name ambiguity remains an issue in WoS, especially since different Chinese names can be transliterated to a single English name. The issue of author name ambiguity is less important in the VIP data, as full author names are recorded using Chinese characters. However, there remain cases where Chinese authors share the same Chinese name.

Both automatic and manual validation were performed to disambiguate author names in the WoS and the VIP data. A combination of the author' s full name and her/his primary institutional affiliation was used for automatic validation. A pilot test with fully manual validation was conducted based on data from 10 selected disciplines (Shu et al., 2016), and the results indicated that the automatic validation allows to disambiguate about 97% of WoS data and almost all VIP data. Exceptional cases were caused by two or more Chinese authors that share the same (Chinese or English) name, and who were active within the same institution or the same discipline. In addition to the automatic validation, a thorough manual validation (that lasted about 6 months) was performed to disambiguate these exceptions. In each discipline, the same name affiliated to different institutions was validated as either an author having multiple affiliations or different authors sharing the same name. Incomplete entries and inconsistent formats were also corrected. The manual validation disambiguated 120,953 ambiguous records regarding Chinese author names.

Meanwhile, in addition to typos and incomplete entries, serious institutional name ambiguity was also found in WoS data. For example, JINAN-UNIV refers to Jinan University located

[1] History is classified as discipline under both Social Science and Arts and Humanities.

at city of Guangzhou in the province of Guangdong while UNIV-JINAN refers University of Jinan located in the city of Jinan in the province of Shandong; BEIJING-UNIV-TECHNOL (Beijing University of Technology) and BEIJING-INST-TECHNOLOGY (Beijing Institute of Technology) are two different institutions while both BEIJING-INST-CHEM-TECHNOL and BEIJING-UNIV-CHEM-TECHNOL refer to the same Beijing University of Chemical Technology (formerly Beijing Institute of Chemical Technology). Both CHINESE-ACAD-MED-SCI and PEKING-UNION-MED-COLL refer to the same institution with two different names (Chinese Academy of Medical Science and Beijing Union Medical College). Institution name disambiguation was conducted manually at the same time as the author name disambiguation was performed, and clarified 1,398 ambiguous records regarding Chinese institution names.

Among the 26,969 records retrieved from WoS and VIP (14,911 records from WoS and 12,058 records from VIP), 12,270 and 11,066 Chinese elite researchers as well as their primary affiliated institutions were identified from WoS and VIP, respectively, across the 115 selected disciplines. As noted above, Chinese scientific elites in multiple disciplines tied for the top 100 ranking. In addition, the total numbers of Chinese scientific elites in 7 disciplines in WoS and 3 disciplines in VIP totalled fewer than 100 because fewer than 100 Chinese authors published papers in these disciplines between 2008 and 2015.

Technical Validation

In order to compare the WoS and VIP in terms of the most productive authors in each of the 115 disciplines chosen, the number of papers per author was compiled in order to produce ranked lists of top Chinese authors in WoS and VIP. The top 100 (and tied) authors in terms of the number of publications produced between 2008 and 2015 in the 115 identified disciplines formed 115 pairs of Chinese scientific elite researchers. The amount of overlap between each of these 115 sets of researchers indicated whether the Chinese scientific elites found in the WoS is the same as the one found in the VIP. For each discipline, the overlap between those researchers who are among the top 100 in WoS and the top 100 in VIP (hereafter referred to as the overlap rate) was calculated based on the formula,

$$\text{overlap}=\frac{\text{number of shared Chinese authprs}}{(\text{numberof top 100 and tied VIP authors}+\text{number of top 100 and tied Wos authors})/2}$$

For example, the overlap rate is 20% when 22 shared authors are found between 105 authors in WoS and 115 authors in VIP (considering that the number of top 100 authors may equal more than 100 when ties are included).

The publication counts presented in this paper were based on the number of articles, notes, and review articles but exclude editorials, book reviews, letters to the editor and meeting abstracts that are not generally considered original contributions to scholarly knowledge (Moed, 1996). In China, not all co-authorship credits are assigned based on an individual' s scientific contribution but on the basis of seniority (Shen, 2016). However, Chinese bibliometric databases, including VIP, give full credit to all co-authors when counting the number of publications. This study applied the same approach regardless of the argument on whether a full count or divided count is better to measure the co-authorship.

In addition to the overlap rate, eight indicators were also compiled for each discipline for the purpose of data analysis, as shown in Table 1.

Table 1 List of Indicators Used in Data Analysis

Indicator	Description
The Overlap Rate	The share of Chinese scientific elites found in both databases
The number of VIP papers	The number of papers that were published between 2008 and 2015 and indexed by VIP
The number of VIP authors	The number of Chinese scholars who published at least one paper indexed by VIP between 2008 and 2015
The number of Chinese WoS papers	The number of papers that were published by Chinese scholars between 2008 and 2015 and indexed by WoS
The number of Chinese WoS authors	The number of Chinese scholars who published at least one paper indexed by WoS between 2008 and 2015
The number of WoS papers	The number of papers that were published between 2008 and 2015 and indexed by WoS
The ratio of Chinese WoS papers to all WoS papers (Ratioc2w)	The share of Chinese WoS papers to all WoS papers
The ratio of Chinese WoS papers to all Chinese papers (Ratiow2c)	The share of Chinese WoS papers to all Chinese papers including both WoS papers and VIP papers
The ratio of Chinese WoS authors to VIP authors (Ratiow2v)	The ratio of the number of Chinese WoS authors to the number of VIP authors

Data Records

The dataset is stored in a Microsoft Access file "WoS-CSI", which is in Figshare with a DOI as 10.6084/m9.figshare.21979967. The "WoS-CSI" file consists of 15 tables; and the descriptions of the major three tables are as below:

- CSI: Top 100 most productive authors as well as their affiliated institutions in all disciplines.
- SCI: Top 100 most productive authors as well as their affiliated institutions in all natural science disciplines.
- SSCI&AHCI: Top 100 most productive authors as well as their affiliated institutions in all Social Sciences and Humanities disciplines.

Reference

Liang, L. (2003). Evaluating China´s research performance: how do SCI and Chinese indexes compare?. *Inter–disciplinary Science Reviews, 28* (1), 38–43.

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *ASI Journal of the American Society for Information Science and Technology, 58* (13), 2105–2125.

Moed, H. (1996). Differences in the construction of SCI based bibliometric indicators among various producers: A first over view. *Scientometrics, 35* (2), 177–191.

Shen, S. X. (2016). Negotiating authorship in Chinese universities : How organizations shape cycles of credit in science. *Science, Technology, & Human Values, 41* (4), 660–685.

Shu, F., Julien, C.–A., & Larivière, V. (2019). Does the Web of Science accurately represent Chinese scientific performance. *Journal of the Association for Information Science and Technology, 70* (10), 1138–1152. doi: 10.1002/asi.24184

Shu, F., Larivière, V., & Julien, C. (2016). *National and international scientific elites: An analysis of Chinese scholars.* Paper presented at the 21st International Conference on Science and Technology Indicators, Va–lencia, Spain.