Digitalization and scientometrics in assessing the migration of scientists

Sudakova A.E.
Ural Federal University
Yekaterinburg, Russia
ae.sudakova@gmail.com

Tarasyev A.A.
Ural Federal University
Yekaterinburg, Russia
a.a.tarasyev@urfu.ru

***Abstract* — The article presents the development of bibliometrics as a tool for assessing scientific activity. The term “bibliometrics” has various formulations and applications: the development of a discipline, the citation of scientists, the study of affiliations of scientists. The transformation of this tool from analog to digital format is also presented. The bibliometric data was first used in the 1920s, however, the method gained popularity after the creation of the Science Citation Index. The controversy of using this method lies in the methodology of the studied object (for example, whether citation is an indicator of scientist productivity, whether it is possible to consider mobility as relocation, affiliate change, or co-authorship with scientists from other universities and research organizations). The study presents an algorithm for generating, processing and analyzing bibliometric data, the need for development of which is dictated by the current situation.**

***Keywords* — bibliometrics, scientometrics, scientist mobility, academic mobility, scientist migration.**

I. Introduction

Digitalization spreads across all areas of human activity, and its impact can only be assessed ambiguously. On the one hand, this process accelerates and simplifies human activities, on the other hand, it can have a destructive effect on the human personality.

The article discusses the issue of migration of scientists and the possibility to analyze it using digital technologies.

Scientists are direct participants of scientific and technological progress. The level of development of such progress is a priority in most countries. And while countries compete for the best minds in the world, scientists are looking for better conditions for their work. Therefore, the study of the scientific potential of the country and the migration of scientists is very relevant.

We clarify that this article considers mobility as a type of migration. We will discuss the term “mobility of scientists” in a separate article.

In the process of analyzing the migration of highly qualified specialists, it is especially important to study the quantitative and qualitative relationships arising from mobility. Unlike irrevocable migration, the scientist mobility has both positive and negative effects for all parties involved: individual’s increased knowledge, dissemination of knowledge (already existing and acquired), the flow of intellectual capital.

If we consider the scientist mobility on a national scale, then the boundaries between positive and negative effects on the intellectual capital increment are blurred. In this case, scientific potential and increment of knowledge are broadcasted domestically, which will be clearly manifested only within the receiving and sending organizations. But the situation is different with the cross-border migration (mobility). The outflow of intellectual capital is one of the negative effects (for the scientist’s country of origin).

Migration, with some adjustments, is generally accountable (migration data are presented on the Federal State Statistics Service website and are registered in The Federal Migration Service). However, migration of certain categories of the population (highly qualified specialists, including scientists) is difficult to assess. Firstly, mobility is characteristic of this category. Secondly, migration records for certain categories of the population are not kept, or at least not presented in the public domain. In this regard, the issue of collecting statistical data becomes relevant. In view of the above, we note that the assessment of the scientists’ migration has several specific features: how to assess the loss of intellectual capital (with double affiliation or with its complete change); how does migration and its types (mobility) affect the dissemination of knowledge and scientific effectiveness.

II. Theoretical frame

It is becoming possible to assess the interaction of scientists, analyze mobility, and evaluate migration due to the digitalization of scientific activities. Namely, it is happening due to the publication of articles in scientometric databases (SCOPUS, WoS, eLibrary, PubMed), the development of professional networks – research gate, the assignment of identification numbers of scientists – ORCID, and other tools.

Bibliometrics was first defined in 1969 by Alan Pritchard [1, p. 348] as the «application of mathematical and statistical methods to books and other means of communication». In fact, bibliometrics is a general term that includes several mathematical laws formulated to describe the creators and the creation of literary works. Bibliometrics is used not only for verbal analysis, but also for identification of important qualitative and quantitative characteristics of a scientist (publication activity, country, citation, co-authors, and other indicators). Bibliometric analysis is becoming increasingly popular nowadays. However, its origins go back to the 1920s.

Edward Wyndham Hulme first used the term in 1922 to describe the processes of science and technology by counting documents [2]. Hulme summarized the results of Cole and Eales [3] and prepared an original work on the growth of British patents (linking them to social processes in the UK) and the changes displayed in the International Catalogue of Scientific Literature (linking changes in the subject and country of literature production with international events).

After that, the term “bibliometrics” was not used for about 20 years. The term reappeared only in 1943 in the work of C. Gosnell [4]. The essence of the term remained unchanged: identification of something based on the analysis of articles. Then there was a break in the use of the term, which lasted until 1962, when it was used by Raisig [5]. After that, the term comes into use, and bibliometric analysis gains popularity. According to the Web of Science bibliometric database, the popularity of the method has been rapidly growing since the 1980s and until 2019.

The meaning and purpose of using the term “bibliometrics” are formulated differently:

- it is used for studying the progress of the discipline [6]. Among the fundamental works, we note the study by Eugene Garfield and his Science Citation Index [7-9], the works of Russian researchers G. E. Vladutz [10], V.A. Markusova [11], A.I. Mikhailov [12] and others;
- collection and interpretation of statistical data related to books and periodicals: to demonstrate historical movements, to determine the use of books and magazines at the national or universal level, and to determine the frequency of use of the material in public places [5] (considering the age of this work, we can clarify and modify this definition: to determine the level of article citation). One of the goals of this analysis is to evaluate the performance of a scientist. The first analysis of this type was presented by Gross in 1927 [13];
- analysis of affiliations from publications, which is used to identify qualitative and quantitative characteristics of researchers for the purpose of assessing migration (including mobility) of a scientist (the author's interpretation).

The general meaning of bibliometrics was presented in the work of Alan Pritchard in 1969, where it was defined as “the application of mathematics and statistical methods to books and other means of communication” [1]. The general wording of the term is relevant to this day, despite the age of the publication. This work also discussed the fact that the term “bibliometrics” had very close ties with the similar terms “biometrics”, “econometrics” and the Russian term “scientometrics”.

In 1988, J. Budd [14] formulated what can be analyzed using bibliometrics: “this unobtrusive or non-reactive aspect of methodology is attractive to researchers studying the process of communication in scientific disciplines”. This idea can be completed with the phrase “it also allows us to evaluate scientific performance”. This is evidenced by the founder of SCI [15]. However, the idea was formulated in 1927, long before the digitalization of bibliometrics. It was reflected in Lotka’s law [16], which is based on measuring the scientist’s productivity according to the citation of his works.

There are many works on bibliometrics and its active use as a tool for scientometrics. However, it is worth noting that there are researchers who do not support the use of this tool, so the expressed point of view is debatable. For example, F. Thorne [17] argues that the level of citation cannot always be interpreted as an advantage of the author: citations can be used to criticize the study and its results. Unlike Thorne, L. Smith argues that quality is an unfounded indicator, and journal quality in correlation with article citation may be the answer to the search for quality criteria [18].

It raises the question: if scientific literature has been used as a tool for analysis for a long time, then how does bibliometrics progress or develop? Prior to the creation of the SCI index and scientometric databases, information was processed manually. Due to the development of digitalization, a simplification of the process has occurred (or is still occurring). Standards for bibliometric references allow the quick search of information; the unity is created for each individual publication - as a result, it is possible to track the popularity (citation) of the article.

Three stages can be distinguished in the development of bibliometrics: manual data processing, digitalization of bibliometrics and creation of SCI, data generation and data mapping using software (VOSviewer, SPSS, Pajek).

D. Price's work is one of the first studies that covers the results of processing bibliometric data [19] and builds a citation network. Based on the results of this study, a mathematical model for the growth of citation networks was developed, and the Price’s law on the aging of scientific literature was formulated. From this point on, the importance of SCI increases, which makes it a valuable source for studying certain areas of science and networks of scientific communications, as well as a means of assessing the effectiveness of scientific research.

It should be noted that Russian scientometric studies are not covered in this article (in the USSR, computer programs were not used as widely as in the rest of the world, and this article specifically reviews the processing of bibliometric data using digitalization). However, scientometrics has been actively developing in Russia without the use of computer programs, and the works of Russian scientists are well-known in other countries [8].

How can we study migration, including mobility of scientists, using bibliometrics? Researchers have begun to fully use the indicators of scientometric databases, so let us review a number of migration studies which apply this data.

Traditionally, the migration and mobility of scientists are studied and evaluated using state statistics [20], [21], analysis of resumes and personal web pages [22], [23], [24], questionnaires and interviews with scientists [25], [26], as well as state and administrative databases [27].

In the context of digitalization, however, two more tools can be used to study migration (including the mobility of scientists): social networks (including professional ones, for example, researchgate), and bibliometric data from scientometric databases (SCOPUS, Web of Science and others). Among the studies devoted to the collection of bibliometric data, we note the works of Dubois P. [28], Moed H.F. [29], van Eck N.J. [30] and other authors, some of which are presented later in the article.

Analysis of bibliometric data allows conducting a comparative study of the publication activity of "mobile" and "non-mobile" authors [31], as well as studying the impact of migration on the development of various disciplines [32]. The affiliation method also allows studying the mobility of certain groups of (elite) scientists who are few in number, but who are important for the development of science [33]. These “digital footprints” can capture the movement of scientists between countries [34], as well as the number of representatives of various disciplines in certain countries or organizations [35]. They also allow conducting the analysis of migration flows [29].

G. Laudel [36] was one of the first to advocate the use of bibliometrics to build global indicators of scientific mobility. Laudel [33], who worked with the English-language text database of medical and biological publications (PubMed), collected data on the first authors, beginning with the publications from 1980. She supplemented her analysis with data on doctoral degrees, so the analysis was limited to a specific classification of elite scientists who published at least three articles in the field of science and nature between 1980 and 2002. This is a restrictive data set in terms of both disciplinary and country coverage. The article also analyzed how to determine an “elite” scientist. In some cases, this issue was solved using scientometric databases, but there were also other tools [37], [38].

Speaking of Russian studies, we note E. Dyachenko’s work [39], which compares the structure of internal migration networks for Russian and American physicists (in particular, for scientists working in the field of applied physics). Data collection was carried out by the so-called analog method (manually). Scientists' data were obtained from articles indexed in the Web of Science database. The resulting network was visualized and analyzed using UCINET software (UCINET 6 for Windows). As a result, hypotheses about the relationship between mobility and scientific productivity were put forward.

The study by Van Eck and Waltman presents tools for the aggregation and visualization of bibliometric data [30]. The article shows how to use VOSviewer for building and viewing bibliometric maps. Unlike programs like SPSS and Pajek, VOSviewer pays special attention to the graphical presentation of bibliometric maps. VOSviewer’s functionality is especially useful for displaying large bibliometric maps in an easily understandable format.

III. BIBLIOMETRICS AS A TOOL FOR ASSESSING MIGRATION

Given the weak structure and the presence of a significant unstructured component of the analyzed data, the non-relational database MongoDB, as well as the Hadoop server based on Cloudera CDH, is the primary data warehouse. Spark is the main tool for implementing distributed data processing. Interaction with Spark is carried out through specialized packages of the scripting language R. The analysis of scientific mobility is carried out using a data search and processing program created on the basis of original algorithms developed by the authors of the project. Ruby programming language was used to develop a database processing algorithm. Algorithmic data cleaning (removal and verification of homonyms, verification of authors with a unified publication, and other parameters) was carried out after collecting statistical data.

At the first stage, mining of articles affiliated with the selected organization is carried out using scientometric databases. The source data array for articles is presented as a column matrix $A_{s1}=(a_{s1})_{r\times1}$, s=1,...,r, where the column of the organization is the row with the article’s title; r is the total number of articles affiliated with the selected organization. At the second stage, the matrix A_{s1} is expanded by adding new columns of characteristics to the analyzed data set: author’s identifier, name, country of affiliation, additional affiliation and number of citations for the analyzed article. To do this, mining of personal profiles of authors is carried out using the author’s ID. At the third stage, the collected data is compared in order to obtain generalized information about all the articles affiliated with the selected organization and information about all the authors who worked on the articles.

A new matrix is formed to identify scientists involved in the processes of academic mobility. This matrix reflects information on the number of works the authors published with third-party organizations during the specified period of time. An i-parameter, which defines the author’s identifier, is introduced to accomplish this. Matrix B is formed according to the results of the authors search algorithm in the collected data array. In this matrix, the b_{ij} element is obtained by accumulating values in cases where the identification numbers of the authors coincide with the authors’ target numbers in the array.

As a result of these transformations, the matrix $B=(b_{ij})_{m\times l}$ is obtained. Rows of this matrix represent all authors of scientific papers, and columns i=1,…,m represent all identified academic organizations. This matrix is compiled for each analyzed year; therefore, each article has 3 main indexes that determine its position in the analyzed data array. It should be noted that the resulting matrix B is a sparse matrix with elements reflecting information on the number of articles b_{ij} by the author i from organization j.

In the compiled matrix, the first column represents the analyzed academic organization. If the affiliation of the main organization is determined, then the b_{ij} element is fixed in the first column with the indication of the number of articles. When a change in the author’s affiliation is detected, the b_{ij} element is assigned the value of the article, and then it is put in the column j, where j is the serial number of the column indicating the organization to which the author has moved. In order to output the “brain drain” data based on the resulting data array, elements of the matrix B should be checked for compliance with a number of conditions for the time interval k=2011, ..., 2017=1, ..., N

From a practical point of view, the implementation of the algorithm involves verifying the entire array of data on the articles affiliated with the selected organization and on the authors who worked on them. At the same time, the current affiliations of author profiles are verified using the API. In case they do not coincide with the profile of the original organization, the data on authors is transferred to a separate array for further processing and visualization. As a result, there are two data arrays. The first one reflects external academic partners and characterizes academic mobility. The second data array reflects the general situation of publication activity and academic mobility between the selected scientific organization and organizations whose employees participated in joint research with employees of the original organization.

This algorithm results in a matrix of scientists. It indicates their main individual characteristics, including the area of their work, the quantity and quality of scientific works, as well as current and previous places of work. The final matrix $X=(x_{ij})_{c\times d}$ reflects the number of articles written by the author i=1,...,c from the academic organization j=1,...,d.

In general, the solution to the problem can be represented as follows: data mining using the API involves the convolution of the database by articles, authors and scientific organizations; the resulting array is cleaned using the IDs of authors, articles and organizations; affiliations in articles and in the author's profiles are verified; after the verification, there are statistical arrays of academic activity and sparse matrix reflecting the interaction of authors and scientific organizations.

IV. Conclusion

Digitalization has a clear positive effect on bibliometrics. Firstly, it provides access to scientific literature to a wide audience. Secondly, it provides an opportunity to evaluate the scientific field (development of individual areas of science, productivity, migration). However, data collection is most often carried out using the analog method, and the computer programs have not yet been applied properly in this area.

In addition, the use of bibliometrics as a way to evaluate scientific activity has been widely introduced only recently. Therefore, it is necessary to define the problem to study migration and its types: (1) domestic databases (in Russia – eLibrary) cannot be ignored if the purpose is to study the interaction (mobility) of a scientist on a national scale, (2) analysis of foreign databases (SCOPUS, WoS, etc.) is a priority if the goal is to assess irrevocable migration or a cross-border mobility. Moreover, industry and professional databases (for example, PubMed) should be considered in any analysis.

Acknowledgment

Grant of the President of the Russian Federation "Migration of Russian scientists in cross-border and national space: problems, efficiency, assessment" (MK-1737.2019.6).

References

[1] A. Pritchard, "Statistical bibliography of bibliometrics?" *Journal of Documentation, vol.* 25(4), pp. 348–349, 1969.

[2] E. W. Hulme, "Statistical bibliography in relation to the growth of modern civilization". London, 1923, 64 p.

[3] Cole F. J. and Eales N. B. The history of comparative anatomy. Part I—A statistical analysis of the literature. *Science Progress* 11(44), April 1917, p. 578–96.

[4] Gosnell C. F. The rate of obsolescence in college library book collections as determined by an analysis of three select lists of books for college libraries. PhD thesis. New York University. Sept. 1943.

[5] Raisig L. M. Statistical bibliography in the health sciences. *Bull. Med. Lib. Assoc.*, 50(3), July 1962, p. 450–61.

[6] Pritchard A. *Statistical bibliography; an interim bibliography*. North-Western Polytechnic, School of Librarianship. May 1969, 60p. (SABS-5; PB 184 244)

[7] Garfield E. Citation Indexes for Science // Science. 1955. Vol. 122. № 3159. P. 108–111.

[8] Guide to scientometrics: indicators of the development of science and technology / M. A. Akoev, V. A. Markusova, O. V. Moskaleva [et al.]. – Yekaterinburg : Ural University Publishing House, 2014. – 250 p.

[9] Garfield E. Citation indexing: its theory and application in science, technology, and humanities. New York: Wiley, 1979. P. 274.

[10] Vladutz G. E., Nalimov V. V., Styazhkin N. I. Scientific and technical information as one of the tasks of cybernetics // Advances in Physical Sciences. 1959. Vol. 69. № 1. P. 13–56.

[11] Markusova V. A. The first Soviet index of bibliographic references in computer science // Scientific and technical information. Ser. 1. 1976. № 2. P. 30–32.

[12] Mikhalov A. I., Chernyi A. I., Gilyarevskyi R. S. Fundamentals of Scientific Information. M.: Science, 1965. P. 435.

[13] Gross P. L. K., Gross E. M. College Libraries and Chemical Education // *Science.* New Series, Vol. 66, No. 1713 (Oct. 28, 1927), pp. 385-389

[14] Budd J. M. A bibliometric analysis of higher education literature // Research in Higher Education, 1988, vol. 28, is. 2, pp 180–190

[15] Garfield E. Citation Index in Sociological and Historical research // Current Contents. 1969. № 9. August 26. P. 42–46

[16] Lotka A. J. The frequency distribution of scientific productivity // Journal of the Washington Academy of Sciences. Vol. 16, No. 12 (June 19, 1926), pp. 317-323

[17] Thorne F. C. The Citation Index: Another Case of Spurious Validity. Journal of Clinical Psychology 33(0ct. 1977):1157-61

[18] Smith L.C. Citation analysis // Library Trends, 1981. Vol. 30, iss.1. P. 83-106.

[19] Price D. J. de S. Networks of Scientific Papers // Science. 1965. Vol. 149. № 3683. P. 510–515.

[20] Arvizu DE, Bowen RM (eds) (2014) National Science Board. 2014. Science and Engineering Indicators 2014. National Science Foundation, Arlington, VA. http://www.nsf.gov/statistics/seind14/

[21] OECD (2013) OECD science, technology and industry scoreboard 2013: innovation for growth. OECD Publishing, Paris. doi:10.1787/sti_scoreboard-2013-en

[22] Cacibano C, Bozeman B (2009) Curriculum vitae method in science policy and research evaluation: the state-of-the-art. Res Eval 18(2):86–94. doi:10.3152/095820209X441754

[23] Sandstrцm U (2009) Combining curriculum vitae and bibliometric analysis: mobility, gender and research performance. Res Eval 18(2):135–142. doi:10.3152/095820209X441790

[24] Woolley R, Turpin T (2009) CV analysis as a complementary methodological approach: investigating the mobility of Australian scientists. Res Eval 18(2):143–151. doi:10.3152/095820209X441808

[25] Boring P, Flanagan K, Gagliardi D, Kaloudis A, Karakasidou A (2015) International mobility: findings from a survey of researchers in the EU. Sci Public Policy. doi:10.1093/scipol/scv006

[26] Flanagan K (2015) International mobility of scientists. In: Archibugi D, Filippetti A (eds) The handbook of global science, technology, and innovation. Wiley, Chichester, pp 364–381. doi:10.1002/9781118739044.ch17

[27] De Filippo D, Casado ES, Gymez I (2009) Quantitative and qualitative approaches, to the study of mobility and scientific performance: a case study of a Spanish university. Res Eval 18(3):191–200. doi:10.3152/095820209X451032

[28] Dubois P, Rochet JC, Schlenker JM (2014) Productivity and mobility in academic research: evidence from mathematicians. Scientometrics 98(3):1669–1701. doi:10.1007/s11192-013-1112-7

[29] Moed HF, Halevi G (2014) A bibliometric approach to tracking international scientific migration. Scientometrics 101(3):1–15. doi:10.1007/s11192-014-1307-6

[30] van Eck N.J., Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping // Scientometrics (2010). Vol. 84, iss. 2. P. 523-538. https://doi.org/10.1007/s11192-009-0146-3

[31] Pierson AS, Cotgreave P (2000) Citation figures suggest that the UK brain drain is a genuine problem. Nature 407(6800):13–13. doi:10.1038/35024218

[32] Borjas GJ, Doran KB (2012b) The collapse of the Soviet Union and the productivity of American mathematicians. Q J Econ 127(3):1143–1203. doi:10.1093/qje/qjs015

[33] Laudel G (2005) Migration currents among the scientific elite. Minerva 43(4):377–395. doi:10.1007/s11024-005-2474-7

[34] Furukawa T, Shirakawa N, Okuwada K, Sasaki K (2012) International mobility of researchers in robotics, computer vision and electron devices: a quantitative and comparative analysis. Scientometrics 91(1):185–202. doi:10.1007/s11192-011-0545-0

[35] Deville P, Wang D, Sinatra R, Song C, Blondel VD, Barabбsi AL (2014b) Career on the move: geography, stratification, and scientific impact. Sci Rep. doi:10.1038/srep04770

[36] Laudel, G. (2003). Studying the brain drain: Can bibliometric methods help? Scientometrics, 57(2), 215-237.

[37] Zuckerman, H. (1995). Scientific elite: Nobel Laureates in the United States. Transaction Publishers.

[38] Hunter, R.S., Oswald, A.J., & Charlton, B.G. (2009). The elite brain drain. The Economic Journal, 119(538), F231-F251.

[39] E. L. Dyachenko. Internal migration of scientists in Russia and the USA: the case of physicists // Scientometrics, 2017, Volume 113, Issue 1, pp 105–122. https://doi.org/10.1007/s11192-017-2478-8

.