ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2022, Vol. 56, No. 4, pp. 187–196. © Allerton Press, Inc., 2022.
Russian Text © The Author(s), 2022, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2: Informatsionnye Protsessy i Sistemy, 2022, No. 7, pp. 25–35.

INFORMATION ANALYSIS

A Study of References in Russian Citing Publications (Sources) on Data from the Web of Science Core Collection

R. S. Gilyarevskii[a, b, *], A. N. Libkind[a, **], and I. A. Libkind[c, *]**

[a] *All-Russian Institute of Scientific and Technical Information, Russian Academy of Sciences, Moscow, 125315 Russia*
[b] *Lomonosov Moscow State University, Moscow, 119991 Russia*
[c] *OOO Servisnoe Byuro VIP, Moscow, 115580 Russia*
**e-mail: ruggero29@gmail.com*
***e-mail: anliberty@mail.ru*
****e-mail: ilya libkind@hotmail.com*

Received May 10, 2022

Abstract—The article presents the results of the first statistical analysis of an array of references (about 30 million) from Russian research articles, reviews, conference proceedings, and monograph chapters published in 1980–2020. The references are processed (restructured and identified) to make them machine-readable and to match them to cited publications. It is found that the content of the bibliographic description fields is not uniform. Methods for eliminating this ambiguity are proposed. Changes in the length of reference lists over time, the completeness of data in fields containing bibliographic descriptions of references, the dependence of successful identification of references on the database and the time of publication of the citing publication, and cross-identification of source publications and references are discussed.

Keywords: references, Russian source publications, restructuring references, identification of references, statistical analysis

DOI: 10.3103/S0005105522040033

INTRODUCTION

Objects of bibliometrics—books, articles, journals, and bibliographic references—have different social natures and require different approaches. Books and articles are definite and immediate—once published, they do not change their properties. By contrast, articles are linked to the journal in which they are published, and a journal is a living organism that can change its scope, periodicity, or title; it can be divided into separate series; or it can even be entirely discontinued. Citation indexes divide journals into topic categories, assigning the corresponding index to all articles of a given journal. However, most researchers neglect to consider the fact that in reality articles on the same topic may be dispersed across journals of different scopes, which means that journals publish articles on topics other than the journal's category, while most of the articles on the topic corresponding to that category end up in journals of other categories.

Bibliographic references are another matter. They connect two publications—the publication whose reference list it is from, i.e., the citing (source) publication, and the publication whose bibliographic description it contains, i.e., the cited publication. These references can be found in both journals of the same title and journals with completely different titles that are indexed in different categories. The citing articles, reviews, reports, and book chapters (sources of references) are dynamic: their number, as well as the number of journals, conference materials, and books, increases every year. The same reference, i.e., a bibliographic description that represents a publication, often appears in journals of different scopes (in physics, chemistry, or biology) that citation indexes assign to different categories or do not index at all. When authors of articles cite an article by another author, they may not be familiar with that other author and may not be aware of their possible connections with other authors and with the authors of other articles in the scientific citation network via various relationships, such as bibliographic coupling or co-citation.

Citation indexes such as the Web of Science (WoS), Scopus, and eLibrary process hundreds of millions of references every year and give researchers access to their published data. This study presents the first, to our knowledge, analysis of an array of about 30 million references from citing (source) publications by Russian authors published in 1980–2020 registered in the Web of Science Core Collection (WoS CC). The index makes references data available for research use. The citation reference (CR) field of WoS CC contains truncated bibliographic descriptions as provided by

the citing authors; they have not been substantially processed or used by anyone before this study. A reference in the CR field usually contains the last name and initials of the first author, the title of the journal (conference, book), the publication year, the volume or issue of the journal, the beginning page number, and the DOI (Digital Object Identifier) of the cited publication. Unfortunately, this field does not include the title of the referenced article (review, report, chapter) itself, and the title of the journal (conference, book) is usually given in an abbreviated and/or truncated form, and sometimes just as an abbreviation. Variants of the full name of the journal and its abbreviations often do not have a uniform, standard version.

So far, we have only processed bibliographic descriptions of citing (source) publications, the number of which is naturally smaller compared to cited publications, as they are only available in journals indexed in the Web of Science Core Collection, whereas cited articles do not have this limitation. This information was used to identify authors that reference research literature in Russian and the context of these references. We hope that the new array of references will allow us to study what Russian authors reference and in what context they do it, as well as how they use Russian and foreign scientific achievements. Enabling this required processing the array, which involved two operations: restructuring and identification. The first converts the texts of references into data in formats required for search and comparison procedures, the second eliminates different spellings of the same referents and establishes uniqueness (identity) of references to the same publications (articles, reviews, reports, book chapters).

SOURCE DATA. CITING PUBLICATIONS—SOURCES OF REFERENCES

The source data for this study was taken from the following seven databases of the international citation index WoS: the Science Citation Index-Expanded (SCI-E), the Social Science Citation Index (SSCI), the Art & Humanity Citation Index (A&HCI), the Book Citation Index-Science (BKCI-S), the Book Citation Index-Social Sciences & Humanities (BKCI-SSH), the Conference Proceedings Citation Index-Science (CPCI-S), and the Conference Proceedings Citation Index-Social Science & Humanities (CPCI-SSH). The most detailed analysis was performed on the data from the first three databases, the SCI-E, the SSCI, and the A&HCI. These databases were chosen as the focus of the study for the following reasons: first, each of them represents the corresponding academic field: the SCI-E database represents journals on the natural, exact, and applied sciences, the SSCI database features those on the social sciences, and the A&HCI database contains those on the arts and humanities; second, these three databases generally contain journal articles with the most substantiated scientific results.

The actual source data were articles, reviews, and non-journal publications—conference reports, book chapters—and bibliographic references contained in them.

When in 1962 E. Garfield first began publishing the SCI on thin paper, it consisted of two volumes: a voluminous index of references (the Citation Index) and a thin index of sources (Source Index), which were the only sources of references at the time, as the SCI was published monthly with quarterly, semi-annual, and annual cumulative publications, and due to the time lag, it could not contain any other references, as databases were not yet available to users [1]. The situation is different now, but for the sake of tradition we continue to refer to citing articles as sources. The considered citing (source) publications shall include Russian articles registered in at least one of the above seven WoS databases. The articles will be represented by their bibliographic descriptions and/or DOIs. These attributes should not be confused with bibliographic descriptions and often DOIs as well that represent references. Additional attributes of source publications may also include other metadata related to these publications, in particular:

- the DOI number;
- the identifiers assigned to publications registered in WoS, recorded in the UT (universally traceable identifier) field;
- the affiliation and country of residence of the author of the source publication, if available.

A publication is here considered Russian if at least one of its authors listed a Russian organization as their affiliation and/or Russia is indicated as a country of affiliation in the corresponding WoS database. The studied source publications from the SCI-E database covered the period from 1993 to 2020, i.e., 28 years. In the SSCI and the A&HCI, this period is significantly expanded and covers 1980–2020 (41 years). However, for the BKCI-S, BKCI-S, and CPCI-S the coverage periods are minimal—16 years (2005–2020), for the CPCI-SSH—31 years (1990–2020). If the source publication contains a non-empty reference list (the vast majority—91.9%—of publications from the analyzed array meet this condition), then instead of the phrase "source publication" we use the phrase "citing publication."

RESTRUCTURING REFERENCES: METHODS, TECHNIQUES, AND QUALITY ASSESSMENT

As noted, articles that have been cited by Russian source publications, i.e., included in reference lists of these citing publications, are to be represented by their bibliographic descriptions and/or DOIs (digital object identifiers). Sometimes, when the meaning is clear in context, these publications will just be called refer-

ences. As for reference lists, WoS databases store them in the CR text field. This field contains the entire reference list of each publication, which consists of a list of truncated bibliographic descriptions and some metadata of other documents referenced in this publication.

It is important to emphasize that the CR field is a text field. Text data can be used for the purposes of bibliometric analysis only after their restructuring. This restructuring is performed at the preliminary stages of the presented study. It involved separating bibliographic descriptions of references, as well as their DOI numbers, into individual elements. The separated and corrected (where possible) data were stored in the appropriate fields of a specially created table of our working database. The table also included two additional fields: one for the UT (universally traceable) identifier of the source publication (assigned to it in WoS) and one for the publication year.

The UT identifier, together with the working identifier assigned to each reference during the study, was used to establish the connections between references and the corresponding citing publications in the subsequent process of identification of references.

The second of the two additional fields records the publication year of the citing publication. We have limited the analysis to only references to cited publications published no earlier than 1700 that have undergone restructuring. The number of such references in Russian source publications across all seven analyzed WoS databases amounted to over 29 million (29 168 725 out of the total of 29 577 254, i.e., including references to before 1700).

The bibliographic descriptions of most of the references to publications from before 1700 were found to have incomplete elements. For example, the field Publication Source was empty in 93.8% of these publications. Some references have other elements in the Year Published field (beginning page, volume, etc.). All this can significantly reduce the quality of data; therefore, it is quite reasonable to exclude such references from consideration. Their share is slightly over 0.01% of the total studied array of references. Thus, the coverage period of the studied cited publications is 321 years (1700–2020). However, in many cases, to increase the reliability of data as well as their representability in graphs and tables, we limit the source arrays to only references to publications from no earlier than 1900.

Before proceeding to a more detailed analysis of the array of references in Russian publications, let us determine their total number. First we shall discuss the methods that can be used for that purpose.

NR method (number of references). The NR field of WoS databases records the number of references in the reference list for each source publication. Their sum is the total number of references.

CR method. As noted above, this field of WoS databases contains truncated bibliographic descriptions of all referenced publications for each source publication. Their sum is the total number of references.

The source data for the NR and CR methods are of different natures: the first method uses statistical data, presumably obtained in WoS by direct, rather uncomplicated processing of citing publications within this system and by its means. Meanwhile, the source data for the CR method include the results of the restructuring of this text field, which was a series of rather complex, sometimes iterative procedures developed and implemented outside of the WoS. Certain errors could occur at this stage.

The question is: is the rate of such errors high and are the data obtained as a result of the restructuring of the CR field reliable? We shall attempt to answer these questions by the comparing results obtained using the CR method with results obtained using the NR method. For these purposes, we assume that the NR field of the WoS, which has full access to the source data, accurately indicates the number of real references in the reference list of each source publication. Therefore, the coincidence rate of data obtained with the CR method and data obtained with the NR method can be used to assess the reliability of the performed restructuring.

To do this, we first compile an array of Russian source publications registered in the SCI-E database and published in 1993–2020. Next, we determine the number of references in these publications using the CR and NR methods. The CR method counted 22 674 974 references, the NR method—22 702 091 references, i.e., the number of references obtained with the CR method is only 0.12% less than the number obtained using the NR method. Thus, the above fact that the numbers of references obtained in two different ways are almost completely the same can be considered an indication of a high degree of reliability of the performed restructuring.

CHANGES IN THE NUMBER OF REFERENCES IN REFERENCE LISTS OVER TIME

To simplify the procedure for obtaining source data for calculating the average number of references and to increase the reliability of these source data, we used the *NR* approach. Let us denote the total number of references in the considered array with the symbols nr_i (Number of References), where the subscript i is some publication from the considered set of source publications I, $i \in I$. Let us sum the values of nr_i over the entire set I and then divide the resulting sum by the number of source publications in set I. As a result, we obtain the average number of references Avr_nr_I in reference lists of source publications from set I, i.e.:

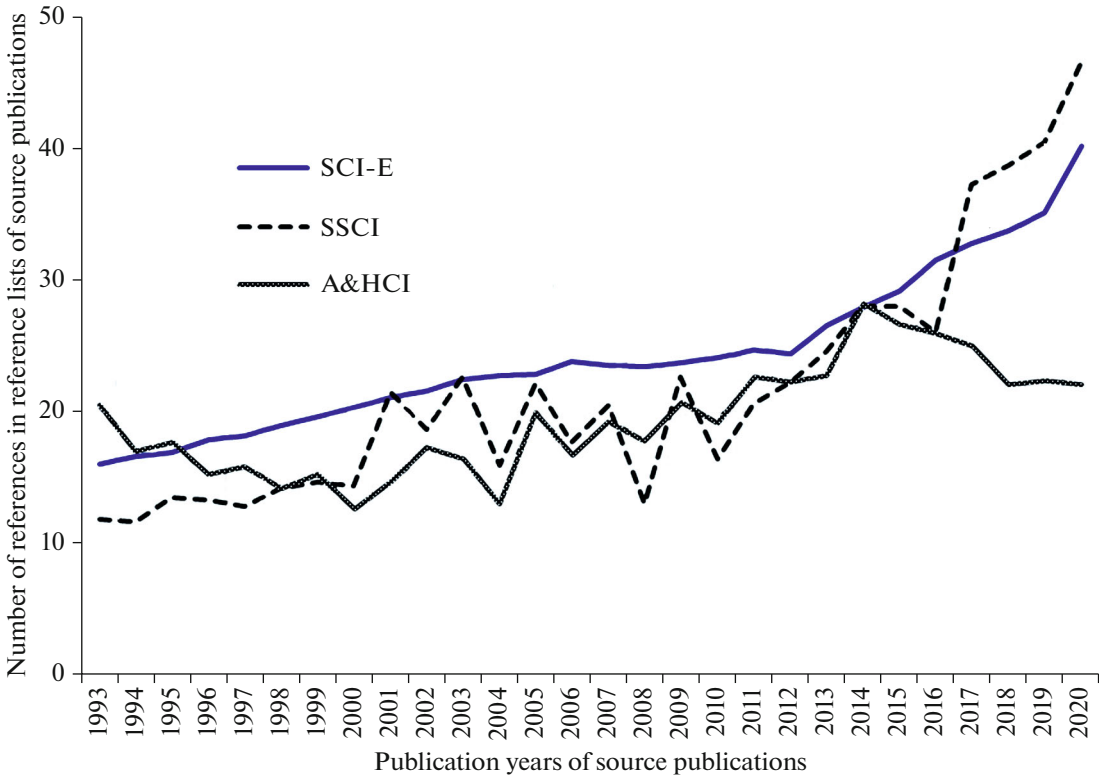

Fig. 1. Number of references in reference lists of Russian citing publications in the WoS Core Collection over time.

$$Avr\_nr_I = \frac{\sum_1^N nr_i}{N_I}. \quad (1)$$

It should be emphasized once again that these values have been obtained directly from the numeric field NR that only indicates the number of references, not their truncated bibliographic descriptions. It follows from Table 1 that the average number of references in reference lists of source publications on social sciences (the SSCI database) generally increases over time. Over the 41-year observation period (1980–2020), the average number of references in reference lists of publications in this academic field has increased more than three-fold: in 1980 it was 13.7 references, and in 2020 it was 46.6 references. However, in the period 1982–1992, this indicator showed a rather significant decrease. As for publications on arts and humanities (the A&HCI database), that field, similarly to social sciences, showed a decrease in the average number of references at the beginning of the study period that then reversed by 2000, when there was a rather unstable rise until 2016, followed by another decrease. For the natural, exact, and applied sciences (the SCI-E database), values were only available for the years after 1993, i.e., only for 28 years. During this period, the average number of references per publication increased dramatically: 16.0 (1993) compared to 40.2 (2020), i.e., a more than 2.5-fold increase. The sharpest increase in the number of references for both the SCI-E and the SSCI occurred after 2012 (Fig. 1), apparently in connection with the intensification of state support for publication activity of researchers.

IDENTIFICATION OF BIBILOGRAPHIC REFERENCES: APPROACHES, METHODS, TECHNIQUES

A detailed study of the obtained arrays of references and their corresponding citing (source) publications requires identification of references with each other, i.e., identification of their coincidence (identity), meaning that they reference the same publication. Let us clarify that the identification procedure is a looped iterative process of comparing the spelling variant of each attribute (bibliographic description and/or DOI value) of this reference with all the spelling variants of the attributes of the other references in the studied array. In other words, all references are compared to each other one by one. If two (or more) bibliographic descriptions and/or DOIs coincide, these references are assigned the same identifier and considered to be references to the same publication.

In WoS databases, reference lists with bibliographic descriptions of references in source publications are stored in the *CR* field. Bibliographic descriptions stored in this field have the following structure: the last

Table 1. The average number of references in reference lists of Russian citing publications in the WoS CC (1980–2020)

Year	SCI-E			SSCI			A&HCI		
	total number of references	total number of cited publications	number of references in one publication	total number of references	total number of cited publications	number of references in one publication	total number of references	total number of cited publications	number of references in one publication
1980				8476	619	13.7	6713	258	26.0
1981				8332	589	14.1	5836	259	22.5
1982				8655	547	15.8	7724	315	24.5
1983				10473	704	14.9	8214	396	20.7
1984				8839	644	13.7	9220	324	28.5
1985				9254	705	13.1	10121	336	30.1
1986				9052	638	14.2	8633	320	27.0
1987				7406	679	10.9	5508	302	18.2
1988				10067	940	10.7	6764	376	18.0
1989				9664	980	9.9	5954	349	17.1
1990				10503	1055	10.0	6163	319	19.3
1991				10277	1053	9.8	5400	236	22.9
1992				9449	907	10.4	5554	314	17.7
1993	400504	24980	16.0	10104	858	11.8	5977	292	20.5
1994	430738	25901	16.6	8409	721	11.7	4878	288	16.9
1995	464709	27489	16.9	10200	756	13.5	4821	274	17.6
1996	505383	28233	17.9	9729	735	13.2	5518	363	15.2
1997	527694	28998	18.2	9778	766	12.8	5565	352	15.8
1998	539163	28546	18.9	11315	802	14.1	5594	396	14.1
1999	557479	28463	19.6	11972	819	14.6	3803	250	15.2
2000	571654	28206	20.3	13483	940	14.3	2914	232	12.6
2001	555119	26332	21.1	12116	564	21.5	3004	206	14.6
2002	587882	27292	21.5	13378	717	18.7	3804	221	17.2
2003	587806	26186	22.4	15201	672	22.6	2648	162	16.4
2004	607955	26727	22.7	11512	724	15.9	1823	141	12.9
2005	612956	26811	22.9	12442	562	22.1	4997	251	19.9
2006	625409	26278	23.8	11649	661	17.6	4436	266	16.7
2007	651281	27637	23.6	12115	590	20.5	4572	238	19.2
2008	732677	31272	23.4	12880	991	13.0	4923	278	17.7
2009	718882	30263	23.8	18217	804	22.7	9031	437	20.7
2010	718708	29797	24.1	16935	1034	16.4	9275	485	19.1
2011	772740	31248	24.7	19299	937	20.6	10011	443	22.6
2012	818065	33545	24.4	26504	1189	22.3	12072	542	22.3
2013	872477	32869	26.5	26825	1089	24.6	10051	443	22.7
2014	945035	33820	27.9	30742	1098	28.0	15553	552	28.2
2015	1114868	38202	29.2	36296	1296	28.0	22887	861	26.6
2016	1251951	39726	31.5	47028	1808	26.0	25511	982	26.0
2017	1367060	41681	32.8	70901	1904	37.2	30198	1211	24.9
2018	1518764	44986	33.8	79046	2040	38.7	28726	1304	22.0
2019	1692568	48163	35.1	94752	2340	40.5	35581	1595	22.3
2020	1952564	48534	40.2	126258	2708	46.6	37601	1704	22.1

Table 2. Completeness of bibliographic description fields*

Field or set of fields of the structure	Number of references	Share in the total array of references (23449271), %
First author	11450901	48.8
Year	23450837	100.0
Source (Journal etc.)	23423011	99.9
Volume and/or Issue	16430245	70.1
Beginning Page	17239494	73.5
Year *AND* Source *AND* Volume and/or Issue *AND* Beginning Page	13853616	59.1
DOI (digital object identifier)	13955210	59.5
Year *AND* Source *AND* Volume and/or Issue *AND* Beginning Page *AND* DOI	10693285	45.6
(Year *AND* Source *AND* Volume and/or Issue *AND* Beginning Page) *OR* DOI	17115540	73.0

*Total array of references cited by Russian publications registered in at least one of the following WoS databases: the SCI-E, the SSCI, and the A&HCI; the citing publications were published in 1993–2020, the references—in 1700–2020, *AND* and *OR* are Boolean addition and multiplication operations.

name and initials of the first author; the publication year of the source; the source in which the reference is published (title of journal or book, name of conference); the volume or issue of journal; the beginning page number. In addition, many, but unfortunately not all, references have a digital object identifier (DOI). We shall consider two complementary approaches to solving the problem of identifying references: the first one uses a set of elements of the bibliographic description of the reference and the second one relies on the DOI of the publication.

Let us consider the capabilities of each of these approaches. To do this, we need to select references that meet the following conditions:

• the references are from Russian source publications from the period 1993–2020.;

• each of these source publications is registered in one of the three WoS databases: the SCI-E, the SSCI, or the A&HCI;

• the referenced publications were published in the period 1700–2020.

Note that the resulting array fully meets the representativity requirement—it contains over 23 million (23450837) references, which is 79.3% of the total number (29577254).

As can be seen in Table 2, using bibliographic descriptions for identification purposes (the first approach) makes it possible to cover only 48.8% of the entire array of references if the process uses data from strictly all fields. If it is allowed to exclude the First Author field from consideration (which somewhat reduces the reliability of identification), the coverage with the first approach reaches 59.1%. As for the second approach (using only DOI numbers), it can identify 59.5% of references (Fig. 1). Naturally, the share of references with a filled DOI field increases with time from past to present. However, even some references published a hundred years ago or earlier have DOIs.

The maximum possible coverage of references is achieved by joint application of the two approaches: in this way, almost three quarters of the array of references—73.0%—can be identified (see the last line of Table 2). This combined approach was therefore adopted for the study. As a result of the identification of the general array of more than 29 million (29577254) references for a subarray of 20.1 million (specifically 20133453) references, at least one identical reference to a cited publication was found for each reference (let us call this subarray of 20.1 million references the *core subarray*). Thus, the share of the core subarray in the total array is 68.1%, and the number of unique (excluding repeated) references in the core subarray turned out to be only slightly over 3.3 million (3310859). We believe that these data indicate a fairly high degree of identification in general.

Considering the results of the identification procedure in more detail reveals a number of trends. As a rule, earlier publication years of citing publications correlate with lower the degree of identification of references, i.e., publications that these citing publications reference (Table 3). Possible explanations are the following: first, DOIs have only been used since 2000. Consequently, publications published before 2000 could only receive these identifiers retroactively, which was not always done. Moreover, the practice of assigning DOIs did not spread immediately after the introduction of the identifier: it took time for it to reach to an increasingly wide range of editorial offices and publishers. Second, our preliminary analysis showed that earlier publication years correlate to less standardized bibliographic descriptions.

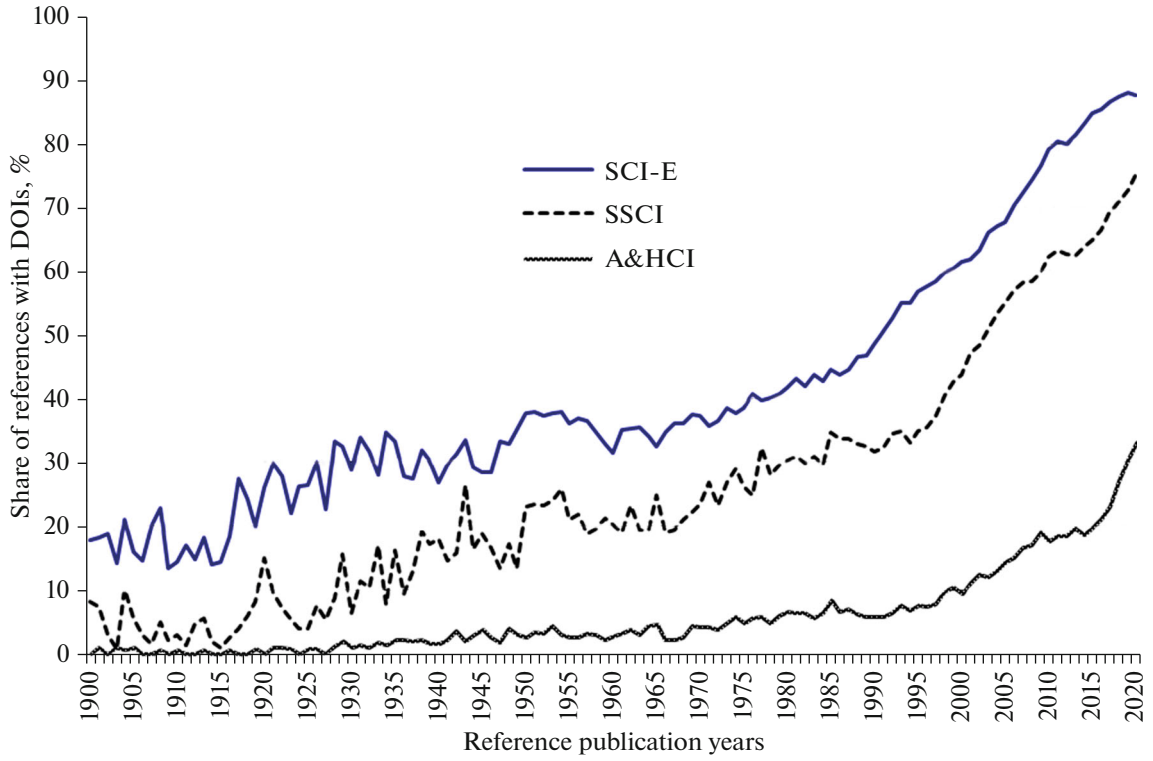

Fig. 2. Dependence of the number of references with DOIs in Russian citing publications on publication time within the 1993–2020 period and in the WoS database.

It can also be seen from Table 3 and Figure 2 that the achievable degree of identification of references is different across different WoS databases. Thus, the share of the core subarray in the array of SCI-E references is significantly higher (74.3%) than that for the SSCI (55.9%) and more than four times higher than for the A&HCI (16.9%). The share of references with a DOI is the highest in the SCI-E—60.5%, followed by the SSCI—48.9%, and then the A&HCI, very far behind—9.4%. The situation is similar in terms of completeness of the four bibliographic description fields (Source, Year Published, Volume or Issue, Beginning Page): 60.1, 45.3, and 14.6%, respectively. Obviously, the nature and traditions of citation in a particular academic field can significantly impact the share of core subarrays. Nevertheless, the provided data on the share of publications with DOIs and completeness of bibliographic description fields largely explain the differences in the share of core subarrays for different academic fields.

DISTRIBUTION OF REFERENCES ACROSS THEIR PUBLICATION YEARS

The data obtained after the restructuring of the CR field can be used, in particular, to construct distributions of references across their publication years. Such distributions are more meaningful and correct for shorter time periods of publication of the citing publications containing the analyzed references. In this case, along with the requirement of topic homogeneity (*topic closedness*, according to the terminology pro-

Table 3. Dependence of the degree of identification of references on the *WoS* database and the time of publication of citing publications

WoS database	Over the entire period (1993–2020)		1993–1999		2000–2006		2007–2013		2014–2020	
	all references	core subarray, %	all references	core subarray, %	all references	core subarray, %	all references	core subarray, %	all references	core subarray, %
SCI-E	22405680	74.3	3379105	66.9	4088312	70.3	5170891	68.8	9767372	81.4
SSCI	748798	55.9	60257	28.4	88869	37.4	119228	46.2	480444	65.1
A&HCI	296359	16.9	31280	10.5	21807	12.9	57419	16.2	185853	18.7

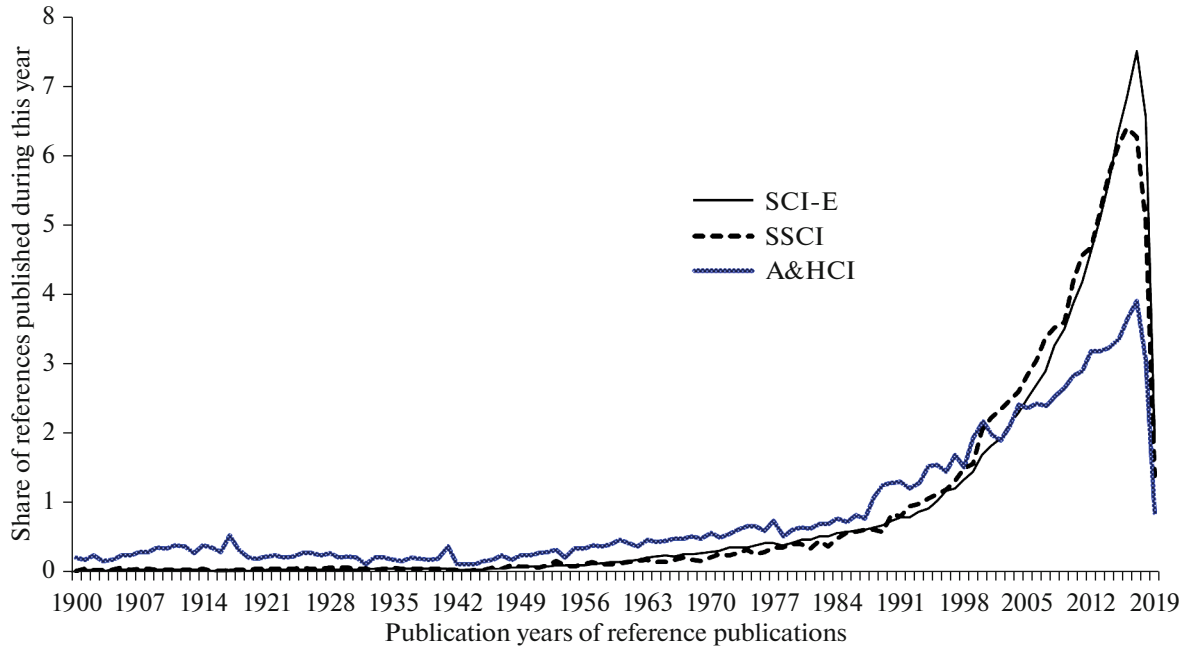

Fig. 3. Distribution of references from Russian source publications from journals published in 2019 across publication years of the cited publications from journals published in 1900–2019 excluding publications registered in several WoS databases at once.

posed in [2, 3]), another important requirement concerns a certain optimum simultaneity of citing publications.

The caveat of a certain optimum is relevant: on the one hand, correctly distributing references across publication years requires an instantaneous snapshot of the distribution corresponding to citing publications published at the same time. However, it would be absurd to make this requirement overly strict, reducing simultaneity to months or even days. We believe that, for the purposes of constructing such distributions, it would be reasonable to limit this requirement to publication times within the same year. Let us take 2019 as an example year and compile an array of references included in source publications from this year. Let us construct a distribution of such references separately for each of the following databases: the SSCI, the SCI-E, and the A&HCI. During the construction, we proceed from the repeatedly confirmed position that citation characteristics largely depend on the academic field of the citing publication.[1] Therefore, for the sake of testing integrity, we exclude from consideration borderline source publications that are included in any other databases besides the considered one. Thus, we exclude source publications from the SSCI if they are also registered in either the SCI-E or the A&HCI. This can be done using the UT identifier. Similarly, we also exclude source publications from the A&HCI database if they are also registered in either the SCI-E or the SSCI. The same applies to the SCI-E.

Fig. 3 presents a graph of the distribution of references across their publication years. The curves on the graph correspond to references published in 1900–2019 and cited by Russian source publications published in 2019 and registered in the corresponding WoS database. Noticeable differences in the distribution of references from different databases can be observed in the graph. The maximum value (7.5%) of the share of references published in a certain year (in their total number) corresponds to references in source publications on natural, exact, and applied sciences (SCI-E). The next maximum share of references is behind by a little over 1% (6.4%) and corresponds to publications on social sciences (SSCI). Note that the maximum for the SCI-E is observed in 2017, corresponding to a 2.5-year lag in relation to the citing publications, while for the SSCI that lag is 3.5 years. The maximum number of references in publications on arts and humanities (A&HCI) is almost twice (3.9%) smaller than the corresponding value for the SCI-E. Note that for the A&HCI, the share of references published 100 years ago or earlier is quite pronounced and in each of the considered years is within 0.2–0.3% of

[1] There is a certain amount of duplication between WoS databases, with the same journal with all its articles being simultaneously stored (registered) in several databases. In accordance with Bradford's law, not all articles of a journal correspond to its scope, and most of the articles on this topic are published in journals with other scopes. However, regardless of which WoS database(s) and how many of them an article is included in, it is assigned all the topic categories of these databases, but only one (single, identical) UT identifier [4]. Therefore, it is misleading to judge the number of publications in a particular academic field or discipline based on the codes and identifiers of articles in this or any other citation index [5].

Table 4. Results of cross-identification of source publications from 1993–2020 with references to publications from 1700–2020

Array of all Russian source publications from 1993–2020, including those that also act as references				Array of all references to Russian and foreign publications from 1700–2020 made in Russian publications in 1993–2020		Array of core references					
						core references that are also Russian source publications			all core references		
total number of SPs	number of publications that have become references	reference frequency	share of SPs that have become PR, %	total number of references	reference frequency	number of unique references	number of NON-unique	reference frequency	number of NON-unique core references	number of unique references	reference frequency
892185	430550	4.8	48.3	22404318	1.8	311124	2103535	6.8	12609504	2690206	4.7
30125	11041	2.1	36.7	748606	1.2	4123	16322	4.0	161070	54383	3.0
14769	2088	1.7	14.1	296342	1.0	641	2309	3.6	11555	4067	2.8

the total number of references. Thus, the share of old references in the academic field Arts and Humanities (A&HCI) exceeds the corresponding values for the field Social Sciences (SSCI)—0.01–0.03%—by an order of magnitude or more and the values for the field Natural, Exact, and Technical Sciences (SCI-E)—0.007–0.01% —by almost two orders of magnitude.

CROSS-IDENTIFICATION OF SOURCE PUBLICATIONS AND REFERENCES

The source data and the data obtained after processing can be used to take another step towards identifying publications. That step is to search the data obtained in the previous section to identify Russian publications that after their publication were cited by other Russian publications, i.e., publications that act both as source publications and as references. That would make it possible to pose and consider a number of interesting problems in the future.

Our approach to solving this problem is based on DOI numbers. Recall that in the WoS databases more than half (59.5%) of the bibliographic descriptions of cited publications published in 1700–2020 that were referenced by source publications in 1993–2020 have DOIs. More than two-thirds (71.7%) of Russian source publications published in that period and registered in one of the studied WoS databases also have DOIs. Consequently, source publications and references can be cross-identified by comparing their DOIs. Items with the same DOI number shall be considered to be the same publication that acts as a source publication in some cases and as a reference in other.

As can be seen from Table 4, almost half (48.3%) of Russian publications registered in the SCI-E were cited by other Russian publications. The corresponding value for the SSCI is significantly lower, at 36.7%, and that for the A&HCI it does not even reach 15% (14.1%). The objective of analyzing these differences, which can be assumed to be largely due to features, traditions, and citation ethics of the relevant academic fields [6], is outside of the scope of this study. However, one significant reason for these differences is, in our opinion, obvious. This results from the fact that we cross-identified the source publications and references based on coincidence of their DOIs. Naturally, the more source publications and references that have DOIs, the higher the probability that a source publication will be found among the references, i.e., will be identified as the same publication as one of the references. Reference to the data presented in the graph in Figure 2 clearly shows that they significantly correlate with the data in Table 4. Therefore, one of the reasons for the differences between the three studied WoS databases in terms of the share of source publications that have also become references is the difference in the share of publications that have DOIs. Moreover, it can be argued a priori that if all source publications and all references had DOI, then the corresponding shares of Russian publications referenced by other Russian publications would be significantly higher. That means that the real figures of citation of Russian publications by other Russian publications may be higher than those presented in Table 4.

Let us perform a brief test to estimate whether this assumption is correct, purely as a first approximation. Let us take the rather highly cited review "Unified cosmic history in modified gravity: From F(R) theory to Lorentz non-invariant models," whose authors include Russian researchers. This review was published in 2011 in the journal *Physics Reports-Review, Section of Physics Letters*. A WoS search in the array of global publications for the period 2011–2019 returns 1,571 publications that had referenced this review (as of 08.03.2022, there are 2361). Using the Refine option of the WoS interface, we select only the references from publications with a Russian affiliation. This returns 317 references. Meanwhile, the applica-

tion of our cross-identification procedure results in 310 references for this period. Thus, only 7 references, i.e., a little over 2% (2.2%), have been lost. Of course, the statistics for a single publication is not a reliable means of assessment. Nevertheless, it still gives some idea of lost data, and it can be assumed that these losses are generally not large.

CONCLUSIONS

Among the numerous resources of WoS databases, the text field CR, that contains reference lists from source publications, has not been statistically investigated before. Working with an array of about 30 million references of Russian authors made over a period of 41 years and stored in these databases, we have processed them to obtain answers to some questions to these data [7]. First, we were interested in how Russian researchers use their own and foreign scientific achievements, which works of their predecessors do they reference and in what way. Other questions of interest include mutual influence of various branches of science on each other and a number of other similar issues.

To find answers to these questions, we needed to process the array of references. The issues that arose during the processing shed light on the nature and features of the citation process itself, which create difficulties for extracting reliable and unambiguous data. Eliminating these problems required research, the process and results of which are presented in this article, including the results of statistical analysis and restructuring and identification of the array of references in order to translate their text data to a machine-readable format and match them to cited publications. In addition, changes in the average length of reference lists in citing publications were analyzed: over the study period, it has more than doubled. In addition, the research has confirmed the assumption of significant differences in the citation traditions of different academic fields—for example, the number of references in research on humanities and social sciences is several times smaller than that number in natural sciences, and their citation window is much longer.

Thus, the performed study has created an important foundation for further research deeper into the psychologically complex and understudied process of the so-called scientific citation.

FUNDING

The study was performed as part of the state assignment to the All-Russian Institute of Scientific and Technical Information, Russian Academy of Sciences no. 0003-2022-0001 and supported by the Russian Foundation for Basic Research (project nos. 20-07-00014 and 20-010-00179).

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Garfield, E., Citation indexes for science: A new dimension in documentation through association of ideas, *Science*, 1955, vol. 222, no. 3159, pp. 108–111. https://doi.org/10.1126/science.122.3159.108
2. Arapov, M.V. and Libkind, A.N., The concept of the closed information flow, *Autom. Doc. Math. Linguist.*, 1977, vol. 11, no. 2, pp. 77–94.
3. Libkind, A.N., One approach to study communication in science, *Scientometrics*, 1985, vol. 8, nos. 3–4, pp. 217–223. https://doi.org/10.1007/bf02016937
4. Akoev, M.A., Markusova, V.A., Moskaleva, O.V., and Pislyakov, V.V., *Rukovodstvo po naukometrii. Indikatory razvitiya nauki i tekhnologii* (Guide to Scientometry: Development Indicators of Science and Technology), Akoev, M.A., Ed., Yekaterinburg: Izd-vo Ural. Univ., 2021, 2nd ed.
5. Gilyarevskii, R.S., On the incorrect use of citation indices for assessment by comparison of divisions of science, *Autom. Doc. Math. Linguist.*, 2022, vol. 56, no. 1, pp. 26–29. https://doi.org/10.3103/S000510552201006X
6. Soos, S., Age-sensitive bibliographic coupling reflecting the history of science: The case of the Species Problem, *Scientometrics*, 2014, vol. 98, no. 1, pp. 23–51. https://doi.org/10.1007/s11192-013-1080-y
7. Boyack, K.W. and Klavans, R., Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front most accurately?, *J. Am. Soc. Inf. Sci. Technol.*, 2010, vol. 61, no. 12, pp. 2389–2404. https://doi.org/10.1002/asi.21419

Translated by A. Ovchinnikova