ISSN 1019-3316, Herald of the Russian Academy of Sciences, 2016, Vol. 86, No. 6, pp. 490–499. © Pleiades Publishing, Ltd., 2016.
Original Russian Text © D.V. Kosyakov, A.E. Gus'kov, E.S. Bykhovtsev, 2016, published in Vestnik Rossiiskoi Akademii Nauk, 2016, Vol. 86, No. 11, pp. 1015–1025.

Review

Russia's Academic Institutes as Mirrored by Webometrics

D. V. Kosyakov[a1], A. E. Gus'kov[b2], and E. S. Bykhovtsev[b3, *]

[a] *Trofimuk Institute of Petroleum Geology and Geophysics, Siberian Brach, Russian Academy of Sciences, Novosibirsk, Russia*
[b] *State Public Scientific Technological Library, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia*
e-mail: [1]kosykovDV@ipgg.sbras.ru; [2]Guskov@spsl.nsc.ru; [3]gor2991@ya.ru

Received October 16, 2015

Abstract—How adequately and fully do the indicators used in webometric research help assess the quality and size of scientific websites? This question is asked by the authors of this article. They analyze the sites of Russian academic institutes and their indicators. The degree of value stability of webometric indicators is demonstrated; examples of their discrepancy with assumed meanings are given; and alternative interpretations are offered. This research resulted in a monthly replenished free-for-all database of webometric indicators of academic sites, hosted at http://webometrix.ru, which could become a tool for planning measures to improve the representation of scientific organizations on the Internet and a basis for generating webometric site rankings.

Keywords: webometrics, site ranking, scientific organizations, impact, open access, Google Scholar, sites.

DOI: 10.1134/S1019331616050063

Most contemporary scientists obtain necessary data from scientific websites, and the concept of open network access to research results is being promoted across the world. In addition, the Internet plays an important role in the interaction between science and society, providing the opportunity to publicize the latest achievements and disseminate scientific knowledge. It is necessary to know well and use efficiently this powerful tool of the 21st century.

The principles that underlie the World Wide Web and the accumulated experience of bibliometric research have largely influenced the special features of perception and the methods of quantitative analysis of the academic web space (the totality of sites of research and educational organizations). Since the mid-1990s, attempts have been made to formalize the basics of research methods and the corresponding terms—*netmetrics, web metrics, Internet metrics, webometrics, cybermetrics,* and *web bibliometrics* [1]. The most popular term today is *webometrics*. The number of webometric studies in the world has increased noticeably, which is proved by the growth of the number of Scopus-indexed publications on this topic (Fig. 1).

A widespread method of studying the academic web space is to rank the sites/domains of academic organizations; it started with the works of the Cybermetrics Lab group headed by Isidro Aguillo. As a result, a constantly updated webometric ranking of universities, research centers, clinics, business schools, and open archives was formed (the Webometrics Ranking of World Universities project: www.webometrics.info). In our opinion, the ranking reflects sufficiently adequately the efficiency of the scientific activity of an organization, which is confirmed by comparison with authoritative ratings of higher educational institutions [2].

The approach developed by this group of Spanish researchers underlay a number of replicas at the regional level, including in Russia. At least six attempts were made to rank domestic scientific organizations and universities, including the following:

(1) a ranking of the sites of scientific institutions of the RAS Siberian Branch (Institute of Computational Technologies, RAS SB, http://www.ict.nsc.ru/ranking),

(2) a webometric ranking of Russia's scientific institutions (Institute of Applied Mathematical Research, RAS Karelian Research Center, http://webometrics-net.ru),

(3) a webometric research service of scientific sites (Far East Geological Institute, RAS FEB, http://fareastgeology.ru/webometrics),

(4) a webometric index of Russian higher educational institutions and research institutes (Institute of Scientific and Educational Information, Russian Academy of Education, http://ru-webometrics.info),

[*] Denis Viktorovich Kosyakov is Deputy Director for information technologies of the Trofimuk Institute of Petroleum Geology and Geophysics (IPGG), RAS Siberian Branch (SB). Andrei Evgen'evich Gus'kov, Cand. Sci. (Eng.), is Director of the State Public Scientific Technological Library (SPSTL), RAS SB. Egor Sergeevich Bykhovtsev is a 2nd-category librarian of the SPSTL, RAS SB.

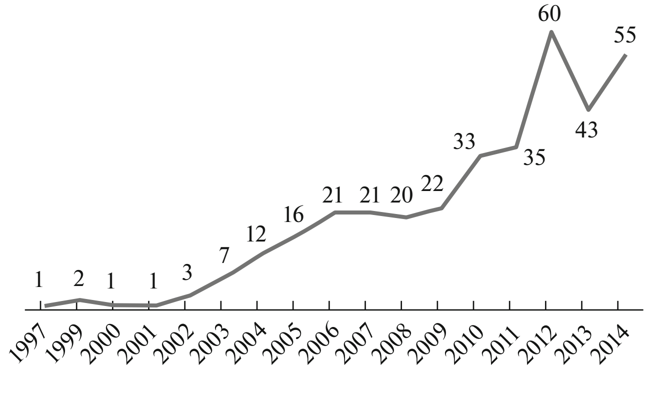

Fig. 1. Number of articles on webometrics indexed by the Scopus database from 1997 through 2014.

(5) a ranking of sites of higher educational institutions and institutes (Siberian Federal University, http://webometrics.sfu-kras.ru), and

(6) a ranking of sites of scientific organizations of the Russian Agricultural Academy.

All these studies are based on a general set of webometric site indicators and the use of commercial search engines to measure parameters. Let us consider the following characteristic indicators:

- the number of site pages indexed by a search engine (site size) is the indicator of the degree of "presence" of an organization on the Internet;
- the number of full-text documents (the file formats pdf, doc(x), ppt(x), and, less often, ps are considered) is the indicator of an organization's "openness," volumes of research results as academic articles, preprints, reports, and educational materials posted in public space;
- the number of "academic" materials indexed by the specialized search engine Google Scholar is the indicator of the quality of materials displayed with no access restrictions; and
- the number of references from other sites to the pages of a site under study (incoming links) is the indicator of the degree of acknowledgment of the level of an organization and its scientific results in society; an alternative indicator is also considered, the Yandex thematic citation index.

Ideally, the above indicators should reflect the results of all scientific and educational activities of an organization and the measure of acknowledgment of their importance. The actual picture, however, does not match the ideal. Despite the fact that the above ratings are based on the same indicators, the final results may differ noticeably owing not only to different calculation formulas and approaches to the choice of measurement object but also to the specifics of the indicators themselves and ways of measuring.

A separate problem is verification of the initial data of rankings and their comparison with one another, because the measured and adjusted values of the indicators open up only in the ranking of the Institute of Applied Mathematical Research, Karelian Research Center, RAS. Combined values open up also in the ranking of the Institute of Computational Technologies, RAS SB.

In order to determine the completeness and adequacy of widely used webometric indicators for finding the volume and quality of the scientific contributions of an organization, it is necessary to analyze the indicators themselves and the dependence of their values on the specific organization of the web space. We conduct such a case study of the sites of Russian academic institutes to examine simultaneously the academic segment of the scientific Runet.

The methodology of measuring webometric indicators. The purpose of this article is to overview the academic web space formed by the sites of the institutes and centers of the Russian Academy of Sciences, the Russian Academy of Agricultural Sciences, and the Russian Academy of Medical Sciences, which were reassigned during the reform of state-owned academies to the Federal Agency for Scientific Organizations of Russia (FASO Russia). Russian Government Executive Order No. 2591-r of December 30, 2013, approved a list of 826 federal state institutions and 181 federal state unitary enterprises subordinate to FASO Russia. Among them are 648 research institutes and observatories, 49 research centers, five museums, eight libraries, and four botanical gardens. The remaining organizations are not for science and research.

Out of 714 research organizations, we were able to find 612 institutions (86%) with operating official sites located at unique addresses. We did not consider organizations for which sites or pages were located in subsites of regional research centers.

Webometric analysis of sites uses specialized or commercial web crawlers—bots that help surf all pages of a site by links found on already opened pages, beginning with the homepage [3, 4]—or the data of search engines [5]. The justifiability of using search engines for webometric objectives has been repeatedly challenged due to the incomplete coverage of the web space, as well as due to the instability, poor accuracy, and the closed nature of algorithms. As the sizes of sites grow, navigation becomes more complicated, reducing the internal coherence of sites and making the use of web crawlers problematic.

The original study [6] used data of the Google, Yahoo, Live Search, and Exalead search engines to collect webometric indicators. However, serious changes have happened in the market of search engines since then. In 2009, Microsoft replaced Live Search with the Bing search engine, based on semantic search technology that was purchased in 2008 from Powerset (https://en.wikipedia.org/wiki/Bing). In 2009, an agreement between Microsoft and Yahoo was announced according to which Yahoo was to switch its search to the Bing database by 2012. Exalead focused on the development of corporate decisions, having an extremely negligible share in the market of search engines. By now the project Webometrics Ranking of World Universities uses only the Google data. The domestic options of rankings use the Yandex data and, less often, the Bing data. This choice appears justified: Yandex occupies over 50% of the Russian market, and Google, over 40%. In addition, Google dominates the world search market by a significant margin (over 68%), and Bing (jointly with Yahoo) takes second place with its 19%.[1] The project of the RAS Karelian Research Center also uses BeeCrawler software to correct data and analyze internal links [7].

All the three search engines—Google, Bing, and Yandex—support queries like "site:site.domain. name," the execution of which may result in assessing the quantity of pages (number of answers) indexed by a search engine on a site by an indicated domain name and on all sites placed on subdomains of the indicated domain. Moreover, Yandex ignores the prefix "www" in a domain name, and in order to obtain the number of pages on a specific site, one should use queries like "host:site domain name" in Yandex. It is impossible to obtain the number of pages of a site that is accessible by default by the domain name without a prefix (e.g., "www") in the Google and Bing search engines. Large values are rounded in the web interfaces of search engines. Web services of search engines can also be used to obtain data. However, unlike Bing and Yandex, the Google web service yields values substantially different from the standard search interface (Fig. 2). Further we will consider the results obtained with the Google standard interface and the Bing and Yandex program interfaces (web services).

The data on the number of documents on sites are obtained similarly with search engines using qualifications on the file format (file extension). Google Scholar also allows searches indicating the name of a site, making it possible to find the full texts of articles in periodicals and abstract collections located on one site or on all sites of a domain, as well as correctly formulated metadata about such publications without full texts. The number of links to a site can be obtained with specialized commercial services.

Objects of analysis. Among the 612 organizations for which we were able to find official sites in individual domains, most of them (513) are institutes and research centers (67); then follow libraries, museums, botanical gardens, observatories, etc. Most of the domains are located in the zone .ru (587 domains), but there are domains in zones .org (9), .com (5), .рф (4), .su (3), .info (3), and .net (1); 398 sites are located in second-level domains; 198, in third-level domains; and 16, in fourth-level domains.

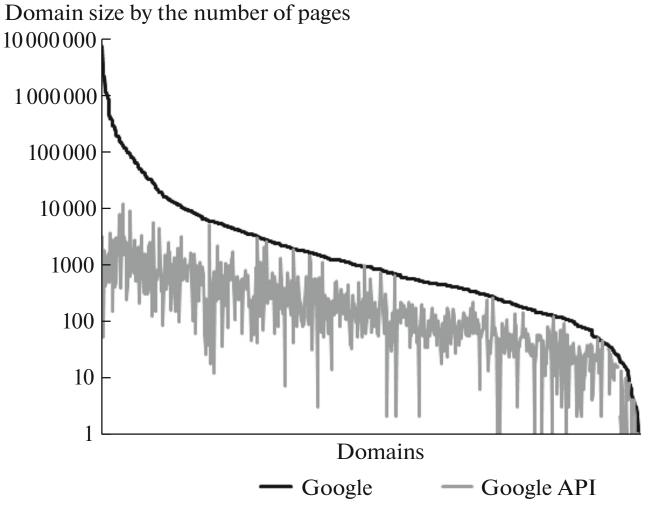

Fig. 2. The ratio of domain sizes of academic organizations measured in June 2015 with the Google standard interface and the programming interface Google API. The domains were sorted by size in the Google standard interface in descending order.

Most sites are located on the servers of institutes; about 80 organizations use hosting services, the most popular of which are Masterhost and Timeweb. The most popular web server is Apache (55%), and then come nginx (33%) and IIS (10%). The most used content management systems (CMS) are Wordpress, Bitrix, and Drupal (39, 36, and 35 sites, respectively). By our estimates, over 300 sites are constructed without CMS. Among analytical tools of web traffic processing, the most frequently used is Yandex Metrika (179 sites); then come Google Analytics (106 sites) and LiveInternet (75 sites). Several sites have two and more analytical counters, and 379 (62%) sites have none.

The official sites of 108 organizations are accessible by default with the host name that matches the name of the main domain; 504 organizations use prefixes (as a rule, "www"). The sites of many organizations are accessible by the address both with the "www" prefix and without it (mirrors), which creates problems when obtaining webometric data. Most organizations (53%) have only an official site—one web host in a domain or two mirrors—with the "www" prefix and without it; others have several subdomains (e.g., the site of the SPSTL, RAS SB, with the address www.spsl.nsc.ru has a subdomain, webirbis.spsl.nsc.ru); 16% of the organizations have more than 10 such individual sites, and 4%, more than 40.

The sizes of academic domains. The sizes of domains (the number of pages on domain sites) measured by search engines are distributed exponentially, except for "tails." The values obtained with the web version of Google exceed by about two times similar assessments obtained with API Yandex and Bing.

[1] According to http://www.gs.seo-auditor.ru/, https://www.netmarketshare.com/search-engine-market-share.aspx.

However, this is mainly achieved due to maximal assessments, because in over 40% of cases, the size of a site according to Bing turns out to be larger than according to Google (for Yandex, in 35% of cases). Finally, the Bing and Yandex assessments turn out to be larger by about equal proportions. "Small" sites up to 50 pages usually represent "business cards" and practically have no information about scientific activities, while "large" sites, on the contrary, contain databases of publications and/or scientific content.

The dynamics of the total sizes of academic sites can be accessed at http://webometrix.ru/link/h01.

Table 1. The results of intraday measurements of the sizes of randomly chosen sites in the Google standard interface, measured on June 6, 2015

Site www.ispras.ru		Site www.bakulev.ru	
site size, pages	data retrieval time	site size	data retrieval time
2610	12:14:09	134000	13:17:53
2600	12:38:53	134000	13:42:47
2600	13:03:41	152000	14:07:43
2600	13:28:33	152000	14:32:38
1780	13:53:27	134000	14:57:38
2760	14:18:22	134000	15:22:34
2610	14:43:22	134000	13:17:53

Table 2. Results of measuring the size (in pages) of individual sections of the site www.bakulev.ru, January 2015

Path	Google	Bing	Yandex	Pages
/	14600	8460	3364	
/about/	14100	4010	1835	
/about/news/	13900	42	452	600
/about/press/	898	224	193	1048
/about/reviews/	126	3	4	
/about/structure/	643	2570	339	
/cyclo/	153	113	119	
/education/	105	79	49	
/en/	195	152	90	
/publish/	453	4480	348	333
/publish/bcatalog/	263	207	239	196
/publish/jcatalog/	2	5370	0	1
/science/	792	349	251	
/science/dissertation/	204	103	70	
/science/dissertation/timing/	193	0	64	
/search/	1	2270	0	

Most likely, the noticeable fluctuations in indicators, especially in Google, are due to several factors:

- the reconstruction of search indices in general;
- the specific operation of algorithms that assess the number of search query results;
- the specifics of algorithms that include pages in search indices; and
- the correction or introduction of mistakes to indexed sites.

A perennial study, individual results of which are also available on the site http://www.worldwidewebsize.com, shows a significant fluctuation in the sizes of the search indices Google and Bing, which, no doubt, also affects the number of answers that result from specific queries.

Judging by intraday observations (Table 1), noticeable fluctuations in the number of results for one and the same search query are possible even in a short time interval. They are most noticeable for Google; the Yandex and Bing results are also subject to fluctuations, but less noticeably.

Observations conducted for more than a year over the indexing of the IPGG RAS SB site with the tools of the Google, Bing, and Yandex webmasters support the assumption that a significant difference exists in the principles of search index formation. Bing and Yandex are much stricter in placing pages into search indices, excluding pages with nonunique or irrelevant content in terms of these search engines. It is characteristic that, when the search bot detects pages, they are included into the index and their verification for compliance with the internal rules and exclusion from the index are done with a delay. Apparently, large site updates may cause a sharp growth in the number of pages measured by a search engine with a subsequent gradual decrease.

Several problems exist that are mainly related to webmaster errors, distorting the results. Here are some of them.

Error 404. Table 2 shows the results of a sampling analysis of the sizes of site parts of the Bakulev Scientific Center for Cardiovascular Surgery (www.bakulev.ru). The site's good structure allows for analysis and, in a number of cases, even for an authentic assessment of the actual number of pages.

The section News (route /about/news/) shows well differences in assessments by various search engines, primarily, Google's overstated assessment. The site forms the page of a specific piece of news by calling the PHP script at http://www.bakulev.ru/about/news.php with ID parameters = xxxxx (news number) and SID = 16. For absent news (wrongly formed ID), a warning page "Element not found" is generated, but the server generates this page with reply code 200 (ok) instead of 404 (not found). The search engine's bot may mistakenly consider such a page as existing, leading to an overstated assessment.

The problem of the absent file robots.txt. Exceptions in the file robots.txt help inform bots and crawlers of site sections that should not be processed and collected. This mainly relates to numerous technological sections of a site. The absence or wrong content of such a file may entail the inclusion of technological content into the index of a search engine. For example, Bing and Yandex indexed the embedded technical documentation of the Apache web server on the site of the Institute of Chemical Kinetics and Combustion, RAS SB (www.kinetics.nsc.ru).

The problem of various routes. Very often, one and the same site content can be accessed by various routes or with different parameters in the address line. For example, on the site of the RAS Institute of World Economy and International Relations (www.imemo.ru), the search index happened to include simultaneously the usual version of the page and the print version, which doubled the site's size according to the data of search engines.

The problem of alternative names and mirrors. The site of the RAS Institute of Geography (www.igras.ru) consists of two parts, which are accessible at www.igras.ru, as well as at igras.ru. In this respect, some pages were indexed with one host name and some, with another, while some pages got both host names. For example, the section "Staff," which contains personal pages (/staff), was indexed as 94 to 207 for Google and 19 to 270 for Bing in the space www.igras.ru and igras.ru, respectively.

The link http://webometrix.ru/link/h02 represents tables of leaders by the number of pages according to Google, Yandex, and Bing for each month, beginning with January 2015. It is easy to notice that the leaders are widely represented by the domains of libraries and regional research centers: the Central Scientific Library, RAS FEB (cnb.dvo.ru); the Irkutsk Research Center, RAS SB (isc.irk.ru); the SPSTL, RAS SB (spsl.nsc.ru); the Central Scientific Agricultural Library (cnshb.ru); the RAS Library for Natural Sciences (benran.ru); the Udmurt Research Center, RAS UB (udman.ru); the Central Scientific Library, RAS UB (cnb.uran.ru); and the Karelian Research Center, RAS (krc.karelia.ru).

The high standings of regional research centers depend on the fact that all or most sites of scientific institutions that belong to them are located on subdomains of the main domain of a research center, thus adding weight to the main domain. This also explains the invariably high standing of the RAS Siberian Branch in the Webometrics.info ranking: the domain nsc.ru, where most Novosibirsk institutes are located, is taken into account.

Libraries gain the leading positions because online catalogs are located on their sites. Thus, the main weight of the SPSTL, RAS SB, comes from the online catalog based on the Web Irbis (webirbis.spsl.nsc.ru) system, where Google in some months finds over half a million pages, many of which correspond to replies to various queries to the search subsystem. The electronic catalog of the Central Scientific Library, RAS FEB (libserver.cnb.dvo.ru), by Google's estimates, contains over 230 000 pages. Neither Yandex nor Bing find many pages in these sections of the site, because the contents of the catalogs are formed dynamically, and, as was noted above, these search engines treat more conservatively both the detection of such contents on the site and its inclusion into search indices. A different situation is observed on the site of the Central Scientific Agricultural Library, the "heaviest" section of which—encyclopedias, dictionaries, and directories (www.cnshb.ru/AKDiL)—is indexed equally well by all search engines (77 000, 63 000, and 83 000 pages according to Google, Yandex, and Bing, respectively).

The large size of the sites of scientific institutes with high standings is formed due to various factors. Let us consider them using examples.

In the domain of the RAS Mathematical Institute (mi.ras.ru), along with the official site and many sites of scientific groups and conferences (at least 80 in total), there is one of the numerous mirrors of the portal of the European Mathematical Information Service (emis.mi.ras.ru), included into the Google index in a volume of over 200 000 pages (http://webometrix.ru/link/h03). It is also interesting to note sharp changes in the size of the domain mi.ras.ru according to Google data. After reaching peak values in July (810 000 pages), the indicators dropped to 20 000–50 000 pages in early 2016.

In the domain of the Dorodnitsyn Computer Centre, RAS (http://webometrix.ru/link/h04), the main size (over 230 000 pages according to Google) is provided by the site of the Faculty of Informatics and Applied Mathematics of St. Tikhon's Orthodox University for the Humanities (www.pstbi.ccas.ru), where a mirror of the database "New Martyrs, Confessors Who Suffered for Christ during the Years of Persecutions of the Russian Orthodox Church in the 20th Century" is located.

The domain of the Siberian Supercomputer Center (sscc.ru, for further details, http://webometrix.ru/link/h05) has over 75 sites, including at least two versions of the official site (www.sscc.ru and www2.sscc.ru). One of the sites is supported by the Tsunami Laboratory of the Institute of Computational Mathematics and Mathematical Geophysics, RAS SB (tsun.ssdd.ru), and contains catalogs and databases of tsunamis, earthquakes, volcanoes, and meteorite craters. The total volume of the laboratory's site is 186 000 pages according to Google.

The domain of the Ershov Institute of Informatics Systems, RAS SB (iis.nsk.su, in further detail, http://webometrix.ru/link/h06), has over 60 sites, including the site of the Faculty of Mechanics and Mathematics of Novosibirsk State University and the archives of Academician A.P. Ershov, amounting to over 110 000 pages according to Google.

The domain of the Lebedev Physical Institute, RAS (http://webometrix.ru/link/h07), contains over 50 sites, including the sites of divisions, projects, and personal sites. One of the largest is the site of the TESIS project of the Laboratory of X-Ray Astronomy of the Sun (www.tesis.lebedev.ru). Chronological databases of images of the Sun, magnetic storms, solar flares, and sunspots are located on it. Yandex assesses the size of this site at 38 000 pages.

In the domain of the Ioffe Physicotechnical Institute, RAS (ioffe.ru, http://webometrix.ru/link/h08), the site of journals published by the institute has a large size (88 000, 48 000, and 63 000 pages according to Google, Yandex, and Bing, respectively); it contains both cards of individual issues and articles and a full-text archive.

Thus, several reasons for the large size of domains are as follows:

- online catalogs and databases, which can be the result of proprietary research or the full or partial replication of other scientific resources;
- full-text archives of periodicals;
- archives of outstanding scientists; and
- resources not related directly to the institute but developed with its participation or supported on its computational facilities.

Sometimes, the large size of a site determined by a search engine may be due to an error in indexing or counting the number of query results. Thus, the incredible result of the RAS Institute of Astronomy of more than 2 mln pages according to Bing in July–August 2015 (inasan.ru, http://webometrix.ru/link/h09) is (a) subject to significant fluctuations (from 700 000 to 2.88 mln pages in two weeks, both downwards and upwards) and (b) not proved true by other search engines. In addition, (c) it was impossible to locate the position of a large block of pages on the site by early August 2015. Bing assessed the total number of pages on the site as 2.1 mln, 1070 pages in the Russian version (www.inasan.ru/rus) and 324 in the English one (www.inasan.ru/eng). The site also has personal sections (e.g., www.inasan.rssi.ru/~kurbatov/) and conference sections, but it is impossible to detect a large section among them. Since search engines limit the size of search results (Yandex to 1000 results; Google and Bing to about 500–600), it is impossible to study this problem in detail.

Site sizes measured by Yandex and Bing do not change much in time (except for the unexplainable data on the RAS Institute of Astronomy, which almost doubled the total indicator). This is confirmed by high correlation ratios between monthly series, which do not drop below 0.9. Google measurements are less stable; the correlation ratio between the series was 0.49 in January–February, 0.3 in March–April, and 0.56 in June–July (2015 data). However, high values of rank correlation ratios (Spearman's coefficient, 0.99, and Kendall's coefficient, above 0.9) allow us to assume that the main reason for changes is the general reconstruction of the search index.

The dependence between measurements of different search engines is much worse. The correlation ratio between the Bing–Yandex series for July 2015 is 0.03; Bing–Google, 0.007; and Yandex–Google, 0.25. The rank coefficients appear much better, 0.89, 0.87, and 0.91 for Spearman's coefficient and 0.74, 0.70, 0.77 for Kendall's coefficient, respectively. The webometrics.info ranking uses the logarithmic normalization of values, which is justified because the correlation ratios between normalized values are more than 0.8 (http://webometrix.ru/link/h10).

The number of full-text documents. The search engines Google, Yandex, and Bing can also be used to assess the number of files posted on sites, taking into account their types. Analysis has shown predictably that the most popular format for documents is ps (over 55%). A large number of pdf (about 40%) and ps files on a site usually means archives of publications uploaded for public access (mostly one or several journals published by an organization). The next most popular is the Microsoft Word format (doc, docx, 10.5%), representing various documents. The largest number of files in the Microsoft PowerPoint format (ppt, pptx, fewer than 1%) belong to seminar materials; Microsoft Excel files (xls, xlsx, also fewer than 1%) usually represent application forms and various reporting tables with summary data.

The dynamics of the number of documents on academic sites that hit the indices of search engines can be accessed at http://webometrix.ru/link/h11 and http://webometrix.ru/link/h12. It is clearly seen that these indicators are much more stable than the total number of pages.

The leading positions in the number of documents are mainly ensured by online publications of conference materials and archives of periodicals. For example, full-text archives of four journals issued by the institute, as well as materials of numerous conferences, are published on the domain of the Ioffe Physicotechnical Institute, RAS (ioffe.ru). The site of the Central Scientific Library, RAS UB (cnb.uranr.ru), contains a large number of pdf files, which are scanned tables of contents of issues of periodicals from the library collections. The sites of the RAS Institute of Russian Literature (pushkinskijdom.ru) and the Peter the Great Museum of Anthropology and Ethnography (kunstkamera.ru) display rich electronic libraries. The sites of the RAS Zoological Museum (zin.ru); the RAS Institute of Sociology (isras.ru); and the Sobolev Institute of Mathematics, RAS SB (math.nsc.ru), have full-text archives of periodicals published by these institutions. The files on the sites of the Institute of Metal Physics, RAS UB, represent full-text materials (publications, the archive of the information bulletin *Perspektivnye Tekhnologii* (Promising Technologies)), as well as abstracts to the articles in the journal *Physics of Metals and Metallography (Fizika Metallov i Metallovedenie)* as pdf documents. The sites of the Budker Institute of Nuclear Physics, RAS SB (inp.nsk.su), contain a large number of files with various contents: news, teaching materials, extended abstracts of dissertations, reviews, and publications of the staff members. The bulk of files in the pdf format in the domain of the Pacific Oceanological Institute, RAS FEB (poi.dvo.ru), are the Japanese Meteorological Agency's perennial archives of weather maps for the territory of the Pacific Ocean, uploaded for anonymous access.

The assessment of the number of files by search engines is much more stable than the total number of pages, which is also confirmed by high correlation ratios from month to month (no less than 0.9, mainly about 0.98), as well as between the readings of various search engines (no less than 0.83).

The number of "academic" materials in the Google Scholar index. The group that supports the Google Scholar project is working intensively to improve the indexing mechanism and to extend content coverage. This is expressed by the constant growth of the index size and, consequently, the total number of documents indexed in the Russian academic web space (http://webometrix.ru/link/h13).

Very few files from the first ten by number present on sites hit leading positions according to Google Scholar (http://webometrix.ru/link/h14). Of the total number of over 1 mln pdf and ps documents indexed by Google on the sites of institutes in July 2015, only 128 000 hit the Google Scholar index. Google Scholar confidently indexes only articles in periodicals, extended abstracts of dissertations, technical reports, and conference materials (if they are represented as individual files and not as a collection in general). It is possible to noticeably improve indexing by providing the search bot with necessary information about a publication on a card page in metatags according to the Highwire Press, Eprints, BE Press, PRISM, and Dublin Core standards. Moreover, this is how it is possible to index not only full-text files but also publications in the HTML format, as well as publications without full texts but with abstracts. Therefore, the number of publications indexed by Google Scholar may exceed the number of files.

As in the case with files, the main content indexed by Google Scholar is archives of scientific journals, conference materials, and staff papers. Note the site of the Orekhovich Institute of Biomedical Chemistry (ibmc.msk.ru). On the site of the journal *Biomedical Chemistry*, published by this institute, only three pdf files are indexed (actually, there are many more, but the scans of articles were not processed sufficiently qualitatively; therefore, Google is unable to index them), Scholar finding 7950 full-fledged publications. This is ensured by providing all metainformation necessary for indexation on the pages of the site. Metadata for Scholar are available on the sites of the Institute of Computational Technologies, RAS SB (ict.nsc.ru); the Trofimuk Institute of Petroleum Geology and Geophysics, RAS SB (ipgg.sbras.ru); and the Keldysh Institute of Applied Mathematics, RAS (keldysh.ru).

Links. The latest versions of the webometrics.info ranking use the commercial services Ahrefs and Majestic SEO to determine the number of incoming links, which is related to the decreased importance of incoming links in the ranking of Google search results and the cessation of updating of the database of these links (http://mysiteauditor.com/blog/is-google-pagerank-dead/). Yandex has also reduced significantly the role of incoming links in its algorithms primarily because of the active use of incorrect methods of search engine optimization (SEO), related to the purchase of links at specialized link farms (https://en.wikipedia.org/wiki/Link_farm).

The number of links has been measured, starting from June 2015, by the services Ahrefs (http://ahrefs.com) and Majestic SEO (https://majestic.com), which use their own web crawlers for link search. As was announced in the Ahrefs service blog, the index size exceeded 1 bln pages in July, which is significantly smaller than the size of the Google index (over 45 bln pages); however, it appears to be the largest public database of links. The dynamics of the total number of links is given at http://webometrix.ru/link/h15. The leaders in the number of incoming links and referring domains are given at http://webometrix.ru /link/h16.

Let us analyze in more detail a few examples of leading-in-number incoming links of academic domains. The bulk of links to the VINITI site (viniti.ru) comes from the site worldwidescience.org (Global Scientific Gateway), a project that supports federative search in scientific databases and portals.

Ahrefs detected a large number of links (from several hundred, thousand, and even tens of thousands) to one of the sites of the Siberian Supercomputer Center—the above tsunami laboratory—from a large number of domains (over 250) mostly to currently nonexisting pages. The referring sites do not belong to scientific topics and seem to represent classical link farms. Traces of such search optimization are also seen at the site sat.isa.ru of the SAT@home project, dedicated to voluntary distributed computing on the BOINC platform in the domain of the RAS Institute for Systems Analysis.

A significant number of links to the site of the Ioffe Physicotechnical Institute, RAS (ioffe.ru), come from its three mirrors (Table 3). The bulk of links to the site of the RAS Institute for System Programming also comes from several domains, in 50% of cases, from the copyright symbols on each page. Mirrors of the official

Table 3. Top 10 hosts that refer to scientific sites by the number of references

Host	Number of references	Number of target sites	Commentaries
worldwidescience.org	3593589	2	"Global scientific gate," references to the VINITI database
library.gpntb.ru	214979	72	Russian cumulative catalog of Russia's SPSTL scientific–technical literature. References to document holders in library cards
ec-dejavu.ru	213597	6	Déjà vu culture encyclopedia. References from each page to one already inaccessible page on the site of the RAS Institute for System Programming
www.ioffe.rssi.ru	210367	45	RAS Ioffe Institute's site mirror. Navigation has several absolute links to the main site, www.ioffe.ru
www.mathnet.ru	189634	290	All-Russia mathematical portal, developed by the RAS Mathematical Institute. Links to site mi.ras.ru on each page in the copyright
variable-stars.ru	150623	290	The Astronet site's subsite of the journal *Peremennye Zvezdy* (Variable Stars) with search by astronomical sites. The Ahrefs bot has indexed numerous various versions of search answers with links to the sites of the RAS Special Astrophysical Observatory, RAS Space Research Institute, etc.
mitsa.dommod.ru	138392	6	Another mirror of the RAS Ioffe Institute
world-it-planet.org	112931	2	The IT Planet Olympiad's site. Each page has a list of partners and sponsors, the RAS Institute of System Programming among them
194.85.224.34	108110	6	Another site of the RAS Ioffe Institute
www.politstudies.ru	94497	2	*Polis. Political Studies* journal site; among the founders is the RAS Institute of Sociology. Navigation has a link to the institute's site

Table 4. Top 10 hosts that refer to scientific sites by the number of target domains

Host	Number of target domains	Number of references	Commentaries
ru.wikipedia.org	1050	17772	The Russian version of Wikipedia
webometrics-net.ru	910	15759	Ranking of the Karelian Research Center's sites
википедия.орг.рф	516	4052	A full proxy clone of Russian Wikipedia
elementy.ru	508	7112	Popular science project Elements. Contains catalogs of organizations, individual divisions, periodicals, etc.
gruzdoff.ru	482	3236	Partial copy of Wikipedia. A CC BY-SA 3.0 license is given at the bottom of each page, but no links to Russian Wikipedia
www.library.ru	478	70580	Information and reference portal. Contains a catalog of periodicals and references to texts
www.edu.ru	459	1238	Federal portal Russian Education. Contains a catalog of institutes
www.ivolga.ru	445	2561	Business portal of the Volga Federal District. Contains a catalog of sites
window.edu.ru	445	8409	Subsection of the portal Russia Education: a single access window to information resources. Links to materials on sites of institutes
dic.academic.ru	434	4410	Partial copy of Wikipedia's Russian content with no license

site and subsidiary projects also generate links to the site of the RAS Institute of Sociology (isras.ru).

The site of the TESIS project in the domain of the Lebedev Physical Institute, RAS, attracts a large number of links from the navigation area of several sites of the Our Planet project.

The above examples show that many external links are links from mirrors of the main site, links from navigation elements or copyright symbols of friendly or subordinate sites, and links from various catalogs and search engines and, therefore, are technical (Table 4).

Ahrefs and Majestic SEO update the base, excluding links that have disappeared in case of continuing inaccessibility of these pages; however, to all appearances, they do not check the accessibility of target pages.

Some ambiguities can be lifted by the following:

- counting only unique pairs of "referring domain–target page address," which, in particular, would exclude links from copyright and navigation elements and
- excluding or separating into an individual indicator links to the home page of a site.

This approach, being very resource intensive, still does not help filter most "garbage" links. Recognizing these problems, the authors of the webometrics.info ranking introduced a correction of this indicator in the latest versions: the exclusion from calculation of the ten domains with the largest number of links for each target domain. However, this by no means always excludes "suspicious" sources, and for organizations with quality links to their sites, this may lead to a painful decrease in the indicator. Nevertheless, correction of this indicator by formal rules appears to be problematic.

Yandex's thematic citation index. Several Russian rankings use Yandex's thematic citation index (TCI), an indicator calculated by a closed formula for ranking search results, taking into account the qualitative characteristic of links to a site from other sites [8, 9]. According to Yandex information, the algorithm of index calculation takes into account the thematic closeness of a resource and sites that refer to it. The value of TCI, according to our observations, largely depends on the lifetime of a site. The index is assigned to a site and not to a domain in general.

The TCI values of academic sites vary in a wide range, from 4500 to 0 (http://webometrix.ru/link/h17). More than half of the sites acquire a TCI value below 100; 25%, from 100 to 500; 13%, from 500 to 1000; and only 7%, more than 1000.

The main approaches in webometrics speculate on an idealistic concept of web space, a peculiar "spherical Internet in a vacuum." The basic elements of this concept are the following provisions.

- Information on the Internet resides in a single copy. Indeed, why would an ideal world need copies of already extant content?
- Materials are published exclusively by their authors or publishers. There is no reason to assume that someone may upload someone else's content.
- Hyperlink = logical link. It is assumed that any mention of an information object is accompanied by a link to it. On the contrary, the presence of a link means a real connection between information objects.
- A domain's content is relevant to the organization's activity. There is no reason for a scientific organization or university to support sites on its domain space that are not related to research and educational processes, or to publish irrelevant content on its website.

The analysis of webometric indicators of the academic web space shows that the real situation differs significantly from the ideal. The main problems are inaccurate measurements by the tool used and the presence of irrelevant content in the object measured. Thus, the sites and domains of scientific organizations often contain content that does not belong to their sphere of activity and mirrors of general or third-party resources; there are files that have no relation to scientific and educational activities and publications by associates of other organizations. The most problematic sphere appears to be the analysis of external links to the sites of organizations. Only an insignificant part of them can be seen as an analog of citation in publications. At the same time, many mentions of scientists, scientific projects, publications, and results on other websites manage without any links, or links are given to other versions of pages and documents placed on resources that have no relation to a particular scientific organization.

Such anomalous situations are typical of only some organizations that are among ranking leaders; the others have many fewer examples of irrelevant content. Therefore, except for individual cases, the indicators can serve as a measure of an institute's contribution to the development of the academic web space, open access to scientific information, and promotion and popularization of scientific research results. The publicly available tool of analysis and comparison of webometric indicators of institute sites will help them conduct a more conscious policy of improving sites, see their place in the academic web space, and finally lead to a qualitative and quantitative improvement of representation of Russian academic science on the Internet.

The monthly updated results of collected webometric indicators of the domains of Russia's academic organizations obtained by this study and the tools of their analysis are available at http://webometrix.ru. Further research trends will include comparison of various site rankings, more detailed analysis of indicators and rankings depending on the type of organizations and scientific trends, and expansion of the research area to nonacademic science and education.

REFERENCES

1. M. Thelwall, L. Vaughan, and L. Björneborn, "Webometrics," Annu. Rev. Inf. Sci. Technol. **39**, 81–135 (2005).
2. I. F. Aguillo, J. Bar-Ilan, M. Levene, and J. L. Ortega, "Comparing university rankings," Scientometrics **85**, 243–256 (2010).

3. M. Thelwall, "A web crawler design for data mining," J. Inf. Sci. **27**, 319–325 (2001).

4. V. Cothey, "Web-crawling reliability," J. Am. Soc. Inf. Sci. Technol. **55**, 1228–1238 (2004).

5. J. Bar-Ilan, "The use of web search engines in information science research," Annu. Rev. Inf. Sci. Technol. **38**, 231–288 (2004).

6. I. F. Aguillo and J. L. Ortega, "Fernandez webometric ranking of world universities," Higher Education in Europe **33**, 233–244 (2008).

7. A. A. Pechnikov, "Measurements of webometric indicators," Mezhdunar. Zh. Eksp. Obrazovaniya, No. 10, 400–403 (2013).

8. Yu. I. Shokin, O. A. Klimenko, E. V. Rychkova, and I. V. Shabal'nikov, "Ranking the sites of scientific organizations of RAS SB," Vychisl. Tekhnol. **13**, (3), 128–135 (2008).

9. A. I. Khanchuk and V. V. Naumova, "The information space of the RAS Far Eastern Branch," Vestn. DVO RAN, No. 4, 122–129 (2009).

Translated by B. Alekseev