JIS

Brief Communication Institutional addresses in the Web of Science: the effects on scientific evaluation

Carlos García-Zorita, Carmen Martín-Moreno, M. Luisa Lascurain-Sánchez and Elías Sanz-Casado

Laboratory of Information Metric Studies (LEMI), Department of Library Science and Documentation, Carlos III University of Madrid, Spain

Received 14 October 2005
Revised 12 December 2005

Abstract.

The effectiveness of the analytical tools used for the evaluation of scientific activity has been enhanced by the availability of bibliographic databases, in particular the standard-setting Institute for Scientific Information databases, whose operating rules are widely accepted by the scientific community. One of these rules is the availability in a single field of the institutional affiliations of all the authors of a paper. In practice this rule has been replaced by another, resulting from the inclusion of a new option, whereby records can be retrieved by the author's reprint address (Reprint Address field). The outcome is diversity in the information on affiliation that may generate some degree of uncertainty in connection with institutional attribution when discrepancies arise between the information contained in the two fields, mainly when the only option available is the reprint address. The present study found a high degree of uncertainty, however, essentially for the period prior to Web of Science, in particular for scientific evaluation in peripheral countries such as Spain.

Keywords: bibliometric studies; citation indexes; affiliation addresses; Web of Science

Correspondence to: Elías Sanz-Casado, Department of Library Science and Documentation, Carlos III University of Madrid, C/ Madrid 126, 28903, Getafe, Madrid, Spain. E-mail: elias@bib.uc3m.es

1. Introduction

The growing availability of bibliographic databases on electronic media, prompted by the surge of information and communications technologies in the area of information science, has intensified the development of bibliometrics, scientometrics and informetrics in recent years [1]. Any number of studies have drawn from these new disciplines to evaluate scientific activity by country, area of research and institution [2].

Hood and Wilson [1] drew attention to the importance of databases for such studies, as a source both of data and of the analytical tools needed to conduct them. These authors argue that on-line bibliographic products are nevertheless beset with a number of problems: on the one hand, constantly updated data are necessarily less stable, constituting an obstacle to the reproducibility of results; and on the other, the analytical tools available are more limited than in optical media and are therefore poorly suited to such studies.

Foremost among the databases, for their great quantitative and qualitative significance to bibliometrics, scientometrics and informetrics, are the well known

multidisciplinary citation indexes published by the Institute for Scientific Information (ISI). The ability to locate all the papers listed in the first citation index developed by Garfield [3] and marketed by ISI in 1964 was a huge step forward for research evaluation studies, for the insight afforded into a given paper's or author's influence and impact on the scientific community. As this citation index, which was and still is published on hard copy, evolved with information and communications technologies, the new formats appearing – on-line, computer tape or CD-ROM – rendered the search and retrieval of information more convenient.

These databases have come to play a relevant role in scientific policy in many countries, where they are used as tools to evaluate the research conducted by authors and institutions. The information they contain is structured into a number of fields that prove to be highly useful for different types of analysis, such as the scientific productivity of institutions or countries – for which criteria such as authorship are also considered; author, document or journal impact or visibility; and inter-institutional scientific co-operation.

The appearance of internet databases and the relative ease of retrieving on-line information, along with the evolution of citation indexes, had a catalytic effect on bibliometric research [4], and led the ISI to offer a new web-based service called the Web of Science (WoS), which has become the backbone of the Thomson ISI Web of Knowledge (WoK) platform [5]. It includes the three main ISI databases and an expanded version (SCIE) of the Science Citation Index (SCI) covering a larger number of journals and featuring the possibility of retrieving the papers cited using any of the authors as a search criterion.

Given the growing interest in WoS, this paper aims to analyse one of the characteristics of this new format, whose limitations and misuse have received less attention in the literature than others, in connection with the quality of the information provided and its effectiveness as a research evaluation tool. The limitations identified in the past were generally related to the bias observed in terms of the language in which papers are written [6]; the country of origin of the journals, as well as the representative value of the selected journals included in the databases [7–11]; erroneous publication of impact factor [12]; typographical errors in authors' names; the fact that papers are cited under the first author's name only [1] and so on. However, very few studies have focused on institutional affiliation, where the problems encountered are related to the lack of standardization in the addresses listed. This may be attributable less to database indexation policy than to a lack of standardization among journals, which do not always list all authors' addresses.

There was, in this regard, a rule that had become a standard for bibliometric analyses, which consisted in the assumption, by the researchers involved, that the address field in the ISI citation indexes contained the institutional affiliation of all the authors signing a paper. This rule was regarded as valid for both the hardcopy and the optical media electronic versions of products, where there was only one address field, a feature that has been maintained in the on-line manual for version 7.0 of WoS [13]. It must be borne in mind, however, that the database can only be held accountable for entering the information contained in the journals themselves. Such an address field could be used for studies of different types, since records were retrievable by country, city or institution. Furthermore, it had the added value of containing the addresses of all the authors signing the paper, so studies could be conducted on co-operation between institutions or countries, an issue of enormous interest when evaluating scientific activity.

In the WoS, however, institutional affiliation has been separated into two fields: one that assigns an institutional address to the author to whom requests for reprints (RP) are to be sent, and another for research addresses (C1), consequently generating a new query rule over which the researcher has no control. Since the system retrieves information from both fields when the institutional affiliation is requested, in the records finally obtained, field C1 may possibly be empty.

This new rule ensures accuracy in matching authors with their respective institutions when the reprint address is the criterion, but if the RP address differs from the research addresses recorded, some uncertainty is created around the information that appeared in previous formats. This uncertainty is even greater where discrepancies appear between the two affiliation fields or where only one affiliation is listed. How valid is the new rule when the sole address available is the RP? In the event of multiple authorship, is the affiliation shown for a specific author in the RP field attributable to all the other authors? There were no such uncertainties under the old rule, nor would they exist under the new one if all the research addresses were always available.

In light of this situation, the present quantitative study was conducted to analyse the uncertainty generated by the differentiation of institutional addresses in the WoS, in particular in a country such as Spain, where scientific policy has been and continues to be

largely based on ISI databases. If the number of records with an empty research address field (C1) were small in comparison to total productivity, the uncertainty generated would admittedly be of minor significance and its influence on scientific activity evaluation studies minimal: but how reliable would the results be otherwise?

2. Methodology

Spanish scientific production as listed in the Science Citation Index Expanded (SCIE) from 1985 to 2004 was analysed in this study. The review was based on a sample consisting of around 20% of the papers appearing annually in the WoS platform. The raw data consisted of the information retrieved when 'SPAIN' was entered in the CU (country) field, an operation performed in January 2005.

The documents retrieved were processed with a personal bibliographic manager – ProCite® v.5.03 – to which the data were transferred with ad hoc methodology [14]. Managing the data in this way made it possible to differentiate two institutional addresses for each of the records analysed: the authors' research addresses contained in the field labelled C1 and the address for requesting reprints in the field labelled RP. This latter address also includes the name of the author to whom the request for reprints should be sent.

Table 1
Spanish scientific production listed in SCIE/WoS (1985–2004)

Year	Spanish records	Records retrieved		Records analysed	
	A	B	%	C	%
1985	6,976	1,400	20.1%	1,398	20.0%
1986	8,073	1,600	19.8%	1,598	19.8%
1987	8,717	1,700	19.5%	1,697	19.5%
1988	9,139	1,800	19.7%	1,815	19.9%
1989	9,685	1,900	19.6%	1,911	19.7%
1990	10,547	2,100	19.9%	2,083	19.7%
1991	11,616	2,300	19.8%	2,301	19.8%
1992	13,970	2,800	20.0%	2,801	20.1%
1993	14,746	3,000	20.3%	3,006	20.4%
1994	16,057	3,200	19.9%	3,199	19.9%
1995	17,874	3,600	20.1%	3,600	20.1%
1996	19,547	3,900	20.0%	3,887	19.9%
1997	21,439	4,300	20.1%	4,314	20.1%
1998	24,026	4,800	20.0%	4,813	20.0%
1999	24,939	5,000	20.0%	4,919	19.7%
2000	24,518	4,900	20.0%	4,909	20.0%
2001	26,653	5,300	19.9%	5,187	19.5%
2002	26,910	5,400	20.1%	5,638	21.0%
2003	30,962	6,200	20.0%	5,931	19.2%
2004	30,804	6,200	20.1%	6,242	20.3%
total	357,198	71,400	20.0%	71,249	19.9%

3. Results

Spanish scientific output in the period studied according to the SCIE/WoS included 357,198 papers, broken down by year as shown in Table 1, column A. Based on the figures in that column, Spanish scientific production in the SCIE grew by a factor of 4.42 in the period analysed.

Consequently, the mean cumulative growth rate amounted to 8.13%, with a rise from 6976 in 1985 to 30,804 in 2004. Column B of the table gives the number of papers downloaded yearly from the WoS which, with a total of 71,400, comes to 20% of production. Column C shows the number of papers by year of publication obtained with the methodology described. The analysis was conducted on this sample of the total production, or 19.9% of the total retrieved from the WoS. The discrepancy observed between the number of papers retrieved for a given year (column B) and the number analysed for that year (column C) is due to the fact that in the Web of Science the reference year is the year processed, which may differ from the year of publication [13].

Given the purpose of this study, the presence of research addresses was analysed in the 71,249 records in the sample. The results are given in Table 2, which shows the number, year by year, of papers containing such addresses, both in absolute numbers and percentages.

These results suggest that there are two clearly differentiated patterns of behaviour in the period studied (K-S statistic, 2.133; $p < 0.01$): one prior to 1998 and the other from that date on. From 1985 to 1997, 65% of the records contained a research address, whereas in the period between 1998 and 2004 the percentage climbed to 99.8%.

In light of the large percentage of papers with no research address (C1), this study focused on the ones with an institutional affiliation in the RP field only, to ascertain how many were signed by a single author and

Table 2
Presence of research addresses in papers

Year	No. of papers with research addresses	%
1985	666	47.6%
1986	840	52.6%
1987	843	49.7%
1988	876	48.3%
1989	1,038	54.3%
1990	1,234	59.2%
1991	1,396	60.7%
1992	1,834	65.5%
1993	2,142	71.3%
1994	2,162	67.6%
1995	2,595	72.1%
1996	2,915	75.0%
1997	3,300	76.5%
1985–97	21,841	65.0%
1998	4,798	99.7%
1999	4,904	99.7%
2000	4,896	99.7%
2001	5,172	99.7%
2002	5,626	99.8%
2003	5,920	99.8%
2004	6,239	100.0%
1998–2004	37,555	99.8%

how many by two or more authors. The results are given in Table 3.

Column A shows the number of papers analysed for each year. Column B gives the number of papers having the reprint request address only and column C the percentage. Further to the values in these two columns, while during the earlier years studied approximately 50% of the records had only one address – the reprint address – these values gradually declined, with percentages of under 0.5% after 1998.

The table likewise shows the number of papers having an RP field only that were signed by a single author (column D) and the number signed by two or more authors (column E). It may be deduced from these two columns that during most of the period analysed (1985–97), only a small number of papers were signed by a single author (with a mean of 14.21%), while the bulk (85.79%) of the studies were co-authored.

The percentage values in the final column (F) are the key to measuring the degree of uncertainty in the WoS. The percentage of uncertainty around institutional affiliation in this type of record tends to decline, ranging from 46.1% in 1985 to 18.5% in 1997, and dropping to negligible levels after the latter year.

4. Discussion and conclusions

The use in peripheral countries of ISI databases as virtually the sole tool for evaluating research in many areas, particularly science and technology, and the availability of such databases on web platforms (WoS) is being encouraged by scientific policy in many countries, such as Spain, with the purchase of a licence for use by the Spanish research community as a whole. This ease of access leads to more intensive use of ISI databases, with regional and national agencies adopting the respective production analysis and journal ranking criteria in their own evaluations, on which researcher accreditation and promotion are based. This has meant that Spanish scientists are publishing their research more and more frequently in journals listed in such databases, as may be inferred from the results of the present study, which show that the number of papers published by Spanish institutions grew more than fourfold in the period analysed.

With regard to institutional affiliation, the former rule seems to have been modified. This rule, which consisted in assuming that the address field in the ISI citation indexes contains the institutional address of all the authors signing a paper, had become a standard in studies evaluating scientific activity. But a new rule has appeared in the WoS, the platform increasingly used in bibliometric studies, consisting of differentiating between two fields for authors' institutional affiliations. The new rule generates a degree of uncertainty that had not existed hitherto, in connection with institutional affiliation, particularly as regards co-authored documents in which the ISI indexes only provide the address of the author to whom reprint requests should be sent. This uncertainty could be ignored if the volume of records involved, i.e. co-authored documents with no research addresses, were negligible with respect to total production. This is not entirely the case, however, inasmuch as two patterns of behaviour can be clearly distinguished in the sample analysed with respect to the presence of research addresses (C1). Only 65% of the records for papers listed between 1985 and 1997 contain a research address, whereas this percentage rises to 99.8% for papers listed from 1998 to 2004. Moreover, the data vary considerably in the earlier period: hence, around 50% of the records for the first four years have research addresses, whereas the mean for the last four (1994–97) rises to 72.8%.

Table 3
Degree of uncertainty of Web of Science

Year	Reprint author only					
	Papers analysed			One author	Co-authored papers	
Year	A	B	C B/A	D	E	F E/A
1985	1,398	732	52.4%	87	645	46.1%
1986	1,598	758	47.4%	92	666	41.7%
1987	1,697	854	50.3%	120	734	43.3%
1988	1,815	939	51.7%	139	800	44.1%
1989	1,911	873	45.7%	109	764	40.0%
1990	2,083	849	40.8%	88	761	36.5%
1991	2,301	905	39.3%	114	791	34.4%
1992	2,801	967	34.5%	155	812	29.0%
1993	3,006	864	28.7%	110	754	25.1%
1994	3,199	1,037	32.4%	157	880	27.5%
1995	3,600	1,005	27.9%	156	849	23.6%
1996	3,887	972	25.0%	153	819	21.1%
1997	4,314	1,014	23.5%	215	799	18.5%
1998	4,813	15	0.3%	6	9	0.2%
1999	4,919	15	0.3%	8	7	0.1%
2000	4,909	13	0.3%	3	10	0.2%
2001	5,187	15	0.3%	5	10	0.2%
2002	5,638	12	0.2%	7	5	0.1%
2003	5,931	11	0.2%	6	5	0.1%
2004	6,242	3	0.0%	0	3	0.0%
	71,249	11,853		1,730	10,123	

The records containing no institutional address in field C1 give cause for concern with respect to the high rate of co-authorship, particularly in the earlier portion of the period studied. Although the percentage of uncertainty about institutional affiliation in such records tends to decline, it ranges from 46.1% in 1985 to 18.5% in 1997, after which the rates plummet to negligible levels. Such sharp differences in behaviour patterns over time generate additional uncertainties. This new rule must be taken into account by the authors of bibliometric analyses when evaluating both the scientific production of countries and institutions, and inter-institutional and international co-operation.

References

[1] W.W. Hood and C.S. Wilson, Informetric studies using databases: opportunities and challenges, *Scientometrics* 58(3) (2003) 587–608.

[2] K.C. Garg, An overview of cross-national, national, and institutional assessment as reflected in the international journal Scientometrics, *Scientometrics* 56(2) (2003) 169–99.

[3] E. Garfield, Citation indexes for science: a new dimension in documentation through association of ideas, *Science* 122(3159) (1955) 108–11.

[4] B. Cronin, Bibliometrics and beyond: some thoughts on web-based citation analysis, *Journal of Information Science* 27(1) (2001) 1–7.

[5] P. Jacsó, *Péter's Digital Reference Shelf: Web of Science Citation Indexes*. Available at www.galegroup.com/servlet/HTMLFileServlet?imprint=9999&region=7&fileName=reference/archive/200408/webscience.html (accessed 14 November 2005).

[6] T.N. van Leeuwen, H.F. Moed, R.J.W. Tijssen, M.S. Visser and A.F.J. van Raan, Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance, *Scientometrics* 51(1) (2001) 335–46.

[7] J. Rey-Rocha, M. Martin-Sempere, F. Lopez-Vera and

J. Martinez-Frias, English versus Spanish in science evaluation, *Nature* 397(6714) (1999) 14.
[8] G. Paris, G. De Leo, P. Menozzi and M. Gatto, Region-based citation bias in science, *Nature* 396 (1998) 210.
[9] A. Fernández-Cano and A. Bueno, Multivariate evaluation of Spanish educational research journals, *Scientometrics* 55(1) (2002) 87–102.
[10] H. Andersen, Influence and reputation in the Social Sciences: how much do researchers agree? *Journal of Documentation* 56(6) (2000) 674–92.
[11] A. Sidiropoulos and Y. Manolopoulos, A new perspective to automatically rank scientific conferences using digital libraries, *Information Processing and Management* 41 (2005) 289–312.
[12] L.L. Lange, The impact factor as a phantom. Is there a self-fulfilling prophecy effect of impact? *Journal of Documentation* 58(2) (2001) 175–84.
[13] ISI, *Web of Science 7.0. Workshop Manual* (Thomson, Philadelphia, 2004).
[14] E. Sanz, C. Suárez-Balseiro, C. García-Zorita, C. Martín-Moreno, M.L. Lascurain-Sánchez, Metric studies of information: an approach towards a practical teaching method, *Education for Information* 20(2) (2002) 133–44.