*Jointly published by Elsevier Science Ltd, Oxford
and Akadémiai Kiadó, Budapest*

*Scientometrics,
Vol. 35, No. 2 (1996) 247-255*

EXPERIENCES WITH THE NATIONAL CITATION REPORTS DATABASE FOR MEASURING NATIONAL PERFORMANCE: THE CASE OF MEXICO

R. ARVANITIS,* J. M. RUSSELL,** A. MA. ROSAS**

* *ORSTOM-Mexique, Calle Cicerón 609, Col. Los Morales, 11530 México, DF (México)*
** *Centro de Informacion Científica y Humanística (CICH), Universidad Nacional Autónoma de México (UNAM), Ciudad Universitaria, 04510 México, DF ((México)*

(Received May 22, 1995)

The National Citation Report (NCR) is an integrated citation file supplied by the Institute for Scientific Information (ISI), of an individual country's articles in science and social sciences. Our experience with the NCR database for Mexico suggests that this is an important addition to the tools available for carrying out bibliometric analysis of research performance. However, in order to generate reliable and accurate indicators using these datafiles we recommend that these be handled by specialists well acquainted with the ISI information products and with the scientific setup of the country concerned.

Introduction

One of the main concerns expressed by *Glänzel* and *Schoepflin*[1] in their paper on the future of the fields of Scientometrics, Informetrics and Bibliometrics, and by other contributors to this special issue of *Scientometrics*,[2] was the need for a consensus with respect to methodologies used for the development of science and technology indicators. With tools such as the Science Citation Index CD-ROMs reaching larger audiences than ever before, the potential group of users carrying out bibliometric studies using these databases is growing.

If the validity of indicators is often questioned when trained bibliometricians carry out the process, then the prospect of non-specialists performing these types of studies introduces a new dimension to the problem. An ideal answer would be for the configuration and standardization of the datafiles to be such that results generated would not depend to any large extent on the expertise and knowledge of the person carrying out the inquiries. In present-day terms this goal is, perhaps, unrealistic given that the market is orientated primarily towards users wishing to carry out

0138-9130/96/US $ 15.00

bibliographic searches. However, sharing experiences on the pitfalls encountered in the creation of value-added datafiles for bibliometric purposes is an important step towards achieving greater reliability and replication of results.

Present study

In 1993 the Institute for Scientific Information (ISI) launched as a tool for relational bibliometrics, its service to provide users with National Citation Report (NCR), an electronic database of an individual country's articles in science and social sciences. The Science and Humanities Information Centre (CICH) of the National University of Mexico (UNAM) decided to purchase the NCR files for Mexico for use in scientometric projects related to Mexican research performance. The first use we put the NCR files to was as the data source on chemical research carried out in Mexico, and more particularly the role of the four main institutions in this field.[3] NCR turned out to be a good choice in that we found Mexican chemical research to be well-represented. However, we faced several methodological difficulties when carrying out this study, many of which were related to the correct identification of Mexican institutions.

Although it was our first experience with this database we had previously used the ISI Science Citation Index (SCI) and related files to measure different aspects of Mexican research, such as production of papers in different disciplines, levels of international collaboration, geographic distribution of the research effort, impact factors of journals where Mexican scientists publish.[4-6] So many of the difficulties encountered were not unfamiliar to us. However, the relational nature of the NCR datafiles presented us with a "new" organizational format for bibliographic databases.

In the present paper we evaluate the following general aspects of the NCR datafiles: relative coverage of Mexican research; identification and coding of institutions; field lengths; author identification; classification of journals; and perspectives for carrying out comparative studies between countries.

General considerations

The total number of NCR records for Mexico from 1981 to 1993 is approximately 23,500 giving an annual production of around 1,800 Mexican papers. These numbers are extremely small compared to the production rates of a country like the USA. When dealing with smaller total counts, the existence of errors will have a greater impact on results. Nonetheless, policy decisions are likely to be made on the basis of indicators regardless of the magnitude of the figures.

There is no doubt that access to NCR provides important additional information to that obtained by downloading records from the Citation Indexes CD-ROMs. The wider coverage of NCR with the addition of 1,300 from the Current Contents series and the provision of annual citation counts makes NCR a more comprehensive evaluation tool. Taking journals in Chemistry and related fields for the period 1981-1993 as an example of the added coverage of NCR, Fig. 1 shows the relative distribution of journals between SCI and CC in the NCR datafiles as a whole, and in the journals containing Mexican papers, together with the total numbers of Mexican papers in both ISI products. Although the majority of journals are covered by both SCI and CC, the addition of CC titles increased coverage by between 22-28%.

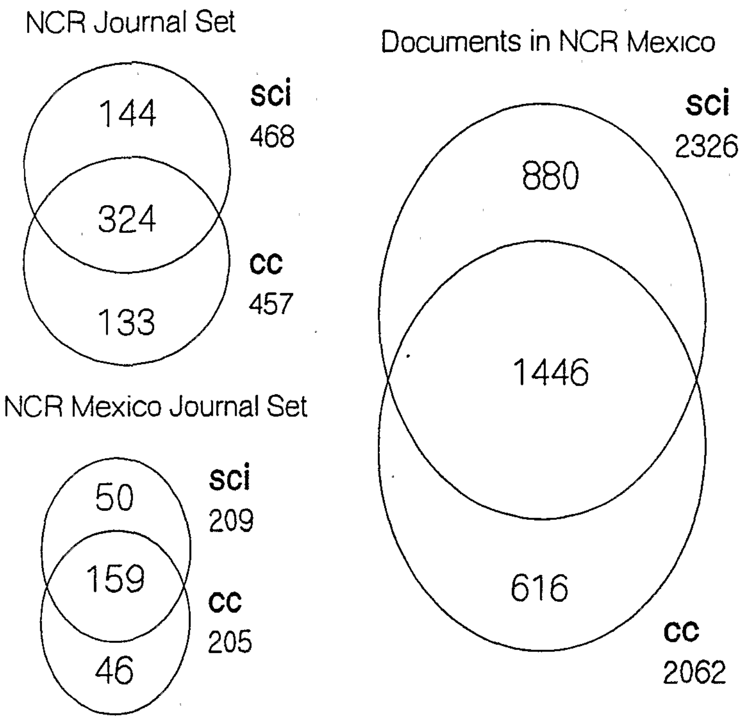

Fig. 1. Distribution in SCI and CC of journals and documents in Chemistry and related fields

The main restraint is the lack of citation data for the additional titles included in Current Contents making calculations of national production and citation rates impossible for instance in different disciplines, using a simple point and click method using software, such as Microsoft ACCESS. The ability to easily separate those records which include citation data (i.e. the journals included in Science Citation Index) from those without citation data (the additional journals in the different series of Current Contents) would greatly facilitate this kind of analysis and permit national comparisons. A comparative study, for instance, of the performance and citation rates of Latin American countries would be of great interest to the scientists and science administrators of the region.

Table 1

Journals in the NCR datafiles for Mexico in all science fields

Year	Total NCR	SCI	% SCI/Total
81	511	394	77.1
82	571	443	77.6
83	581	453	78.0
84	594	469	79.0
85	621	500	80.5
86	666	521	78.2
87	695	572	82.3
88	713	564	79.1
89	745	613	82.3
90	799	630	78.9
91	863	680	78.8
92	955	738	77.3
93	930	724	77.9

Table 1 illustrates the relative coverage of SCI journals in all fields of science where Mexican papers were published from 1981 to 1993. Approximately 80% of journals in the NCR datafiles are covered by SCI indicating that papers published in these titles will contain citation data. The percentage coverage of Mexican papers in SCI journals was a few percent lower. We identified 88 journal titles without a product or classification code corresponding to 261 Mexican papers (1.2% of total production). However these papers were concentrated in five Mexican journals which are apparently not covered either by the ISI citation indexes or by CC. One of these, *Veterinaria-México,* we know is indexed in only one ISI product, namely Focus On: Veterinary Science and Medicine. The number of journal titles without any coding increased considerably from 1991 onwards.

Our ability to correctly analyze trends in the production and citation rates of Mexican institutes would be considerably improved with access to general statistics on the ISI products, such as number of journals indexed annually, new journal titles and ceased titles, for example. Although this type of information is published periodically by ISI in the SCI printed version and through Journal Citation Reports, providing the NCR customers with these data would be extremely useful for users, such as ourselves, wishing to carry out in-depth analyses of the productivity and visibility of Mexican scientific research.

Identification of institutions

One of the main difficulties when attempting to identify documents published by a particular institution, using a commercial bibliographic database is the lack of normalization of institutional names. This is a particularly acute problem in the case of institutions from non-English speaking countries, as names are found in both the original language and in their translated form. Taking the UNAM as an example, we found 19 different combinations of its name, in both Spanish and English, in the institutional level of the corporate source in National Citation Reports for 1992 only (Table 2). These corresponded to 98% of the 1118 records pertaining to UNAM studies, the remaining 2% of UNAM documents were identified as a result of a knowledge of the individual institutes, faculties, etc., belonging to the UNAM, or by recognizing a certain element, such as a PO Box number, which had found its way into the institutional field. Identification of the three most common forms of the UNAM name; the abbreviated name in Spanish, in English, and its acronym, accounted for 94.5% of all occurrences. A preliminary study of the complete database (1981 to 1993) suggests that the number of variations is more in the order of 160.

As institutional coding is both time-consuming and tedious we have been looking into the possibility of an expert system which would do this job for us, as least as far as the main Mexican institutions are concerned. Considering that just four Mexican research institutions are responsible for around 70% of all Mexican papers published at international level, automatic indexing of these corporate sources would save a considerable amount of time, although it would not offset the cost of an expert system unless this could be shared by other institutions.

Table 2
Frequency of different elements occurring in the organization field of NCR by which records pertaining to the UNAM were retrieved for 1992

1) Variations and misspellings of the institutional name	
AUTONOMOUS NATL UNIV MEXICO	1
NACL UNIV MEXICO	1
NAT UNIV MEXICO	2
NATL UNIV MEXICO IZTACALA	1
NATL AUTONOMOUS UNIV MEXICO	707
NATL UNIV AUTONOMOUS MEXICO	2
UNAM	66
UNIV NAC AUTONOMA MEXICO	1
UNIV NACL AUTONOMA MADRID	6
UNIV NACL AUTONOMA MEXICO	283
UNIV NACL MEXICO	5
UNIV NATL AUTONOMA MEXICO	1
UNIV AUTONOMA MEXICO	11
UNIV NATL AUTONOMA MEXICANA	1
UNIV AUTONOMA NACL MEXICO	2
UNIV AUTONOMA MEXICO CITY	1
UNIV MEXICO	3
UNIV NATL AUTONOME MEXICO	1
UNIV NATL AUTONOMA MEXICO CIRC	1
Subtotal	1096 (98%)
2) By name of institutes, faculties, centres, etc.	18
3) By elements of the address	4
Total	1118

Our coding of Mexican institutions for the chemistry study was carried out by assigning each institution a unique ten letter code. We further identified departments, faculties, and institutes wherever possible for the three major institutions, UNAM, UAM (Universidad Autonoma Metropolitana) and IPN (Instituto Politecnico Nacional) included in our analysis.

However this additional, more specific coding was generally feasible only for the larger institutions, like the UNAM, and could be guaranteed only at faculty or institutional level. For instance, we could identify CINVESTAV (Centro de Investigaciön y de Estudios Avanzados), a large research institute belonging to the IPN, but not the departments making up the CINVESTAV. In the case of the UAM, we could rarely identify the specific departments involved and in many instances were not able to assign papers to one or more of the three main campuses of UAM due to insufficient information being provided by the database records. Great care was taken to correctly assign records to the UNAM or the UAM, a task complicated by the presence of the word "Autonomous" in both university titles. In many cases correct coding was the result of the authors' intimate knowledge of the three institutions and was sometimes achieved through familiar elements in the address field, such as a postal code.

Field lengths

One of the difficulties we had with NCR files were their XBase format with fixed length fields. Many addresses, complete in the main ISI files, are truncated in the NCR database where addresses are displayed in four hierarchical levels in the Corporate Source field: namely, Organization, Department, Laboratory and Section. We found that an important part of the institutional addresses is lost and that, in some cases, we were unable to correctly identify institutional affiliations due to the truncated fields. A possible solution to this problem is the provision of an overlay file where the excess information could be stored and called on only when needed.

Author identification

Another difficulty has to do with the fact that, although all author addresses are reported, it is not possible to identify the specific address of a given author. In some cases, a single institution is repeated, so that one supposes each of the authors in position one, two, three and so on, are members of that same institution. In other cases, many authors' names are reported but only one institution. A third situation is when one author reports more than one institutional affiliation. Assuming that the address to be noted in a scientific paper is that of the institution where the work was carried out, then it is unlikely that two institutions should be correctly included when referring to the work of a single author. These irregularities which are a source of confusion when carrying out bibliometric analysis, often stem from a lack of standardization on the part of journals where the articles were originally published. However, the large majority of scientific periodicals have some form of associating authors with their addresses, a facility which we believe would give any datafile an important added bonus when exploited for bibliometric purposes. NCR assigns an author position code to each author in the author file which could be linked with a particular institutional address in the address file. This facility would also make it possible to carry out in-depth studies of co-authorship patterns and institutional collaborations both at national and international level.

We have not attempted yet to measure the performance of individual Mexican scientists. However, due to the lack of any kind of standardization with respect to author names either by journal editors, and even by the authors themselves, this is a task which would require an considerable amount of time and effort even when aided by some kind of standardized list of researchers. This situation is made more difficult in developing countries by the fact that few scientists, particularly in applied areas, stay very long in research activities[7] and in the Spanish-speaking world by the custom of using both parents surnames.[8] For this reason reliable results could be expected only when evaluating the performance of a small elite of scientists easily identifiable by the constant use of a particular name form, or by other characteristics, such as a well-defined research interest, co-authorships, affiliations, etc..

Journal classification

NCR uses different classifications of journals according to which of the different ISI products these belong. The CC journal classification is comprised of a series of Category Codes plus a product code which indicates the particular series of Current Contents to which the title belongs. The SCI journals are assigned the subject classification used in the Journal Citation Reports plus the SCI product code. Individual journals are assigned one or more of these codes according to the range of fields covered, and the ISI products in which they are indexed. This system of subject coding for journals allows easy identification of the areas in which the Mexican papers are published. However, in analyses when it is necessary to determine the main thrust of the individual journals, it is often difficult to decide to which of the various disciplinary codes it is most accurate to assign the journal, especially when the different codes are not assigned to the same general disciplinary area.

Conclusions

NCR is an important addition to the tools available for the carrying out bibliometric studies related to national performance. It is a more comprehensive source for production data than the Citation Indexes alone. The inclusion of citation data for the great majority of journals makes this a unique tool for evaluating this particular aspect of a country s scientific impact. However, the absence of information on the citing papers prevents it from acting as a dual purpose database which could also be used for carrying out citation analyses required by individual researchers, or as a research tool with respect to informetric studies such as co-citation analysis. However, this facility would add considerably to the size and cost of acquiring the NCR datafiles even in the case of small countries, such as Mexico.

We are of the opinion that these datafiles should be handled by specialists with an intimate knowledge of how the ISI databases are structured, and of the scientific setups of the countries concerned, particularly for the correct identification of local authors and institutions. Knowledge of the scope of the different ISI products is also important. For this reason it would be extremely helpful for users to have easy access to statistical information on the different ISI products, such as relative journal coverage, which could perhaps be included within a user's manual to be provided with the NCR datafiles.

References

1. W. GLÄNZEL, U. SCHOEPFLIN, Little scientometrics, big scientometrics .... and beyond? *Scientometrics,* 30 (1994) 375-384.
2. J. M. RUSSELL, Back to the future for Informetrics, *Scientometrics,* 30 (1994) 407-410.
3. J.M. RUSSELL, R. ARVANITIS, A.MA. ROSAS, Institutional production cutting across disciplinary boundaries: an assessment of chemical research in Mexico. Proceedings of the Fifth International Conference on Scientometrics and Informetrics. 7-10 June 1995. River Forest, Illinois. Learned Information, Medford, NJ, in press.
4. H. DELGADO, J.M. RUSSELL, Impact of studies published in the international literature by scientists at the National University of Mexico, *Scientometrics,* 23 (1992) 75-90.
5. H. DELGADO, J.M. RUSSELL, Bibliometrical analysis of medical articles published in the international literature during the eighties by research institutes in the Mexican Republic, *Informetrics 91.* Selected papers from the Third International Conference on Informetrics. 9-12 August 1991, Bangalore, India. Sarada Ranganathan Endowment for Library Science, Bangalore, 1992, 130-147.
6. J.M. RUSSELL, The increasing role of international cooperation in science and technology research in Mexico. *Scientometrics.* Special issue on Latin America, 1995 in press.
7. J.M. RUSSELL, C.S. GALINA, Productivity of authors publishing in tropical bovine reproduction. *Interciencia,* 13 (1988) 311-311.
8. M.T. FERNANDEZ, A. CABRERO, M.A. ZULUETA, I. GOMEZ, Constructing a relational database for bibliometric analysis. *Research Evaluation,* 3 (1993) 55-62.