Treating bibliometric indicators with caution: their dependence on the source database
Luc Quoniam, Hervé Rostaing, Eric Boutin, Henri Dou

To cite this version:

HAL Id: sic_00826944
https://archivesic.ccsd.cnrs.fr/sic_00826944
Submitted on 28 May 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Database comparisons

Treating bibliometric indicators with caution: their dependence on the source database

L Quoniam, H Rostaing, E Boutin, H Dou

Nowadays, with computer-supported analysis of databases, constructing bibliometric or scientometric indicators may be considered easy. The problem is more to verify the accuracy of the global analysis, including the sampling of data. The global coherence of an analysis depends on the adequacy of all the steps. Using on-line databases, an experiment was designed to demonstrate this. Keeping the same protocol for data collection, the same indicators are used over the various samples. The results from three separate databases are profoundly different.

Scientific production and collaborative ability are used as indicators of the activity of a laboratory even though they are quantitative not qualitative indicators. We have tested the accuracy of these indicators through various on-line bibliographic databases. Since we were dealing with researchers' work in detail, we looked only at our own laboratory, according to the code of practice for information brokers.

Data collection

Among our laboratory activities are bibliometry, scientometry and informetry. We used these three words, well recognised and defined nowadays, as keywords with the Dialog Dialindex to get the databases with the best coverage in this area. Our purpose was not to build the best data collection in each database, but more to show the individual perception of each database, using keywords recognised by the scientists of the area. gives the main responses.

Cost considerations led us to use as much information as possible from CD-ROM to constitute our internal database. Lisa was collected from CD-ROM for 1994, and Pascal for 1987–94. Because there are so many duplicates between SciSearch, Social SciSearch and Current Contents Search, the latter database was not considered. An internal database was downloaded from Dialog from the other two without duplicates.

There is some variability between the CD-ROM versions and the original databases, so the internal databases do not have exactly the same number of references as shown in Table 1. For example, Pascal...
Analysis

From these databases, we performed several analyses for internal purposes. We now present a few of them. Instead of building a common database without duplicates, we kept the databases separate to demonstrate the particularity of each one. Each has its own indexation practice, sources and point of view. Applying the same treatment over each separate database will outline the specificities.

Fields constitution and indexation practice

Outlining the databases’ specificity could seem obvious but is very important. SciSearch is the only database including author’s citations, but it has a weak keyword index (240/803 references) which is better in both Pascal and Lisa. SciSearch is also the only database which gives a description of the activity of the journal. Pascal is the only database that provides the country of publication. Our next analysis emphasises one specificity — the data collection practices.

Sources and abstracting practice

To outline this point we determined the groups of co-authors. We have devised software which builds groups (for example authors’ groups) using a propagation algorithm. It is not based on a classic clustering technique, but the natural structure of co-authoring, using only the co-presence of items in references. This means that the algorithm determines all the co-authors of an author. Then it determines the co-authors who work with the original co-authors, and continues until there are no new co-authors in the group.

Then, the algorithm builds a new database, linking for each group, the affiliations, keywords and papers used by this group in the whole of the studied database. When one group is finished, other groups are built up until the last author enters a group or is recognised as an individual author.

This algorithm may be applied over any field in the studied database. We chose authors’ groups, but we could build keyword groups. The fields linked to the group may also be chosen depending on the purpose of the analysis. The algorithm works very well with authors, because the relations between them are finite. Other researchers have published a similar algorithm, but without the linked fields to each built group.

The thresholds used for this analysis are explained in Table 3. The Table 4 represents our laboratory in the SciSearch database with those thresholds.

With these parameters, our laboratory, through SciSearch, is perceived as small (three persons), without collaborations and publishing few (a maximum of three articles) in journals in information science and library science.

With exactly the same parameters, we analysed the Lisa databases. First, it is important to note that it is impossible to extract affiliations from of the Lisa CD-ROM. The author field is also not homogeneous. A bibliometric treatment of this field requires manual control of the database. Table 5 shows the output of the treatment.

The perception of our laboratory begins to change. It appears to have four authors that still publish few (a maximum of eight papers in the last six years). The initial specificity of our laboratory appears (chemistry).

Table 2. Number of bibliographic notices in internal databases

<table>
<thead>
<tr>
<th>Internal database</th>
<th>Number of bibliographic notices</th>
</tr>
</thead>
<tbody>
<tr>
<td>SciSearch + Social SciSearch 01/95 Dialog</td>
<td>803</td>
</tr>
<tr>
<td>Pascal from 1987-1994 CD-ROM</td>
<td>1191</td>
</tr>
<tr>
<td>Lisa from 1994 CD-ROM</td>
<td>2229</td>
</tr>
</tbody>
</table>
With exactly the same parameters we analysed the Pascal database. Table 6 shows the result. The initial specificity of our laboratory appears (chemistry), but also a rather good definition of our activity in information science: automated processing, classification, code, data analysis, data processing, decision making, downloading, graphics, information layout, information processing, on line processing, patent document, relational analysis, research indicator, research program, scientific research, scientific technical information, software, statistical analysis, technological awareness, tool, user interface. Our publications in French journals appear too (Cahiers de la documentation).

Our working group is now made up of 18 authors which publish normally (a maximum of 20 publications over a seven-year period). Collaborations appear with the IBM CEMAP Center of Paris (Huot and Bedecarrax), with the IRIT in Toulouse (Dousse, Dkaki, Koussoubou), with the CEDOCAR (Centre de DOCUMENTATION des ARMées) (Paoli, Dionne, Hilaire, Longevialle) and the CNRS URA 1409 G.O.A.E. (Kister).

The graph of our team is shown in Figure 1. The thickness of links is a function of the number of co-publications between authors. Keeping in mind just the thick links, the structure of collaboration appears. There are thick links between members with the same affiliation, thin links between members with different affiliations. This graph was generated automatically using an algorithm created in our laboratory. 

Discussion

Depending on the indicators, macro- or micro-level bibliometrics and the application country, the choice of database may be more important than the choice of indicator. An indicator may be considered efficient when it does not change the perception of the reality. What happens if the reality of the database is wrong?

We showed that an American multidisciplinary database and an English specific database may
### Table 6. Our working group in the Pascal database

<table>
<thead>
<tr>
<th>Authors</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dou H (20); Hassanaly P (16); Quoniam L (14); Paoli C (5); La Tela A (7); Dousset B (4); Rostaing H (3); Longevialle C (3); Nivol W (2); Koussoube S (2); Kister J (2); Hilaire A (2); Dkaki T (2); Dionne Mc (2); Billard P (2); Bedecarrax C (2)</td>
<td>Cahiers de documentation (2), Scientometrics (3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Co-authors relations</th>
<th>Publication date</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Affiliations</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bedecarrax C [paris, fra (2)]; Billard P [paris, fra (2)]; Dkaki T [toulouse, fra (2)]; Dou H [marseille, fra (18)]; Dousset B [loulouse, fra (4)]; Hassanaly P [marseille, fra (14)]; Hilaire A [lousse, fra (2)]; Huot C [paris, fra (3)]; Kister J [marseille, fra (2)]; Koussoube S [loulouse, fra (2)]; La Tela A [marseille, fra (7)]; Longevialle C [paris, fra (2)]; Nivol W [marseille, fra (2)]; Paoli C [marseille, fra (2), paris, fra (3)]; Quoniam L [marseille, fra (11)]; Rostaing H [marseille, fra (3)]</td>
<td>Bedecarrax C [bibliometrics (2), bibliometric analysis (2), database (2), application (2), data processing (2), data analysis (2), patent document (2), relational analysis (2)]; Billard P [bibliometric analysis (2), database (2)]; Dionne MC [bibliometric analysis (2), database (2), bibliographic data (2)]; Dkaki T [bibliometrics (2), graphics (2), information layout (2), data processing (2), information processing (2)]; Dou H [classification (2), decision making (2), tool (2), research program (2), method (2), firm strategy (2), case study (2), research indicator (2), information processing (2), content analysis (2), frequency (2), information science (2), patent document (2), methodology (2), congress (2), published document (2), evaluation (2), code (3), bibliographic data (3), scientific technical information (3), data analysis (3), on line processing (3), data processing (4), statistical analysis (4), europe (4), france (4), bibliometry (5), bibliometrics (6), downloading (6), scientific research (6), scientometrics (6), chemistry (10), bibliometric analysis (10), database (12)]; Dousset B [bibliometrics (2), information layout (2), data analysis (2), bibliometric analysis (2), software (2), user interface (2), graphics (3), data processing (3), information processing (3)]; Hassanaly P [decision making (2), tool (2), research program (2), scientific technical information (2), research indicator (2), information processing (2), content analysis (2), frequency (2), code (2), information science (2), patent document (2), methodology (2), congress (2), automated processing (2), data processing (3), data analysis (3), europe (3), france (3), on line processing (3), statistical analysis (4), downloading (4), scientific research (5), bibliometry (5), scientometrics (6), chemistry (8), database (9), bibliometric analysis (10)]; Hilaire A [data processing (2)]; Huot C [bibliometric analysis (2), database (2), application (2), data processing (2), data analysis (2), patent document (2), relational analysis (2), bibliometrics (3)]; Kister J [scientific research (2), tool (2), research program (2)]; Koussoube S [bibliometrics (2), graphics (2), information layout (2), data processing (2), information processing (2)]; La Tela A [chemistry (2), scientometrics (3), data processing (2), statistical analysis (2), bibliometry (2), scientific research (2), bibliometric analysis (3), database (7), downloading (4)]; Longevialle C [data analysis (2), bibliometric analysis (2), database (2)].</td>
</tr>
</tbody>
</table>
An analysis involving the best solution. The most control underesti-
mate or affiliations. An accurate estimate of the validity of the database must be performed before building bibliometric indicators. An analysis involving several databases is very often the best solution.

References

1. W Glanzel and M Winterhager, "International collaboration of three east European countries with Germany in the sciences", 

2. H Kretschmer, "Coauthorship networks of invisible colleges and institutionalized communities", 

3. H P F Peters and A F J Van Raan, "Structuring scientific activities by co-author analysis: an exercise on a university faculty level", 


5. EIA, EUSDIC, EIRENE, "Code of practice for information brokers", 

6. L Egghe, "Bridging the gap: conceptual discussions on informetrics", 


9. M Bordons, M A Zulketa, A Cabrero and S Barrigon, "Identifying research teams with bibliometrics tools", 

10. M Abd-El-Kader, J F Miquel and J C Dore, "Méthode d'analyse des thèmes et réseaux de la coopération scientifique nationale et internationale sur les céramiques dans les bases SCI", 

11. E Boulin, L Quoniam, H Rostaing and H Dou, "A new approach to display real co-authorship and co-topisphie through network mapping", 

12. H Grupp, "On the supplementary functions of science and technology indicators: the case of West German telecommunications R & D", 

13. J F Miquel and Y Okubo, "Structure of international collaboration in science. II: Comparisons of profiles in countries using a link indicator", 

Note: Numbers in parenthesis are frequencies.
Research Evaluation

Special issue on S&T indicators: Part II
Guest editor Anthony F J van Raan

Editorial: Introduction to a special issue on the Fourth International Conference on Science and Technology Indicators

Treating bibliometric indicators with caution: their dependence on the source database
L Quoniam, H Rostaing, E Boutin, H Dou

Linkage between agency-supported research and patented industrial technology
Francis Narin, Kimberly S Hamilton and Dominic Olivastro

On the importance of regional innovation flows in the European Union: some methodological issues in the Italian case
Simona Iammarino, M Rosaria Prisco and Alberto Silvani

Measuring innovation in services
Rinaldo Evangelista and Giorgio Sirilli

Basic indicators for describing the stock of human resources in science and technology (HRST)
Werner Grünwald, George Paterson and Daniela Schackis

Evaluation of national biomedical research outputs through journal-based esteem measures
Grant Lewis

Roles of bibliometrics in scientific communication
Hans E Roosendaal

Plus

Book reviews on the challenge of scientometrics, regional innovation, and evolutionary economics

Index to volume 5, 1995

Article summaries, author affiliations:
back cover ISSN: 0958-2029
Research Evaluation

Special issue on S&T indicators: Part II

Treating bibliometric indicators with caution: their dependence on the source database
L Quoniam, H Rostaing, E Boutin, H Dou (CRRM Faculty St Jerôme, France) Pages 177-181

Linkage between agency-supported research and patented industrial technology
Francis Narin, Kimberly S Hamilton and Dominic Olivastro (CHI Research, USA) Pages 183-187

On the importance of regional innovation flows in the European Union: some methodological issues in the Italian case
Simona Iannarino, M Rosaria Prisco and Alberto Silvani (Irsds-Cnr, Italy) Pages 189-206

Measuring innovation in services
Rinaldo Evangelista and Giorgio Sirilli (National Research Council of Italy) Pages 207-215

Basic indicators for describing the stock of human resources in science and technology (HRST)
Werner Gräweil (Eurostat, Luxembourg), George Paterson and Daniela Schackus (Consultants to Eurostat) Pages 217-224

Evaluation of national biomedical research outputs through journal-based esteem measures
Grant Lewison (Wellcome Trust, UK) Pages 225-233

Roles of bibliometrics in scientific communication
Hans E Roosendaal (Elsevier Science, The Netherlands) Pages 237-241

Nowadays, with computer-supported analysis of databases, constructing bibliometric or scientometric indicators may be considered easy. The problem is more to verify the accuracy of the global analysis, including the sampling of data. The global coherence of an analysis depends on the adequacy of all the steps. Using on-line databases, an experiment was designed to demonstrate this. Keeping the same protocol for data collection, the same indicators are used over the various samples. The results from three separate databases are profoundly different.

CHI Research's early work on the citation linkage between patented technology in the USA, and the underlying research science base using 1987/88 US patents has been massively expanded to include citations from 1992/94 US patents, and an analysis of the cited US papers and the agencies supporting them. There is a very strong within-country component to the linkage: inventors in the US system cite their own country's papers approximately three times as often as would be expected, when adjusted for the size of the country's science. The linkage is strongest in the highly scientific areas of technology, and is quite subject specific. Over the six years separating the studies, there has been a remarkable three-fold increase in linkage. A large fraction of these papers cited in patents originate in the US university system, and are supported by US research support agencies.

Using an Italian questionnaire not specifically designed to study geographic phenomena, the potential and limitations of the regionalisation of innovation information is analysed. There were difficulties involved in introducing the territorial dimension into the existing questionnaire and proposals are made for incremental amendments to the survey method for collecting and presenting the data, aimed at capturing the regional interdependencies. These include a request for the location of the main source of information, the broad geographic area from which technology has been acquired or transferred, and the sector and region of origin of the innovation.

Some of the problematic issues in the measurement of innovation in the service sector are discussed, in particular the applicability of the definitions and methodologies set out in the OECD Oslo Manual on the statistical measurement of technological innovation, which was used in the recent European Community Innovation Survey of the manufacturing sector. The results of a pilot survey in Italy and other surveys are compared and some suggestions made for the design of a questionnaire for the service sector. It is suggested that the Oslo Manual framework can be used to collect innovation data in the service sector with some changes in the definition of innovation and in the list of expenditure items. The economic impact of technological and organisational innovation, though considered a crucial issue, is still not amenable to statistical measurement.

Science and technology indicators have been used since the 1960s, yet it is only within the last few years that more pressing policy needs for wider-ranging measures of human resources in science and technology (HRST) began to emerge. The measure of stocks of HRST, and the development of indicators which make international comparisons possible, will be a major step forward. The objective of this paper is to present the concepts of HRST and some indicators for the measurement of basic HRST populations.

A study was carried out to determine the volume and esteem of scientific publications in 16 subfields of clinical medicine and biomedical research for the UK and 11 other OECD nations for 1988-93 as part of the UK Government's Technology Foresight Programme. The subfields were defined by means of sets of specialist journals and, for multi-disciplinary and general journals, keywords in the titles of the papers. The esteem of the papers was based on subjective views of scientific administrative staff of the relative value of papers in different classes of journal. The weighted sum of the numbers of papers from each country as a percentage of the world total was compared with the percentage based on the simple sum, to give a 'journal esteem factor'. This provides a more judgement-based and timely assessment of research quality than the commonly used indicators based directly on numbers of citations.

In view of changes in scientific communication towards an electronic environment, this paper sketches the development of bibliometric tools for use in a scientific publishing environment. The main areas of application are: quality control of individual products; market analysis; strategic analysis for long-term planning; and product development. Each of these requires different indicators and tools, which are critically discussed. The influence of the type of journal and scientific discipline on the usefulness and interpretation of the impact factor is addressed briefly, and requirements for indicators are discussed against the background of possible future changes in the scientific communication process. This leads to speculation on the role of bibliometrics, and how to organise it, in an electronic communication environment.
In volume 5 of the journal

Research Evaluation:

Basic indicators for describing the stock of human resources in S&T, Werner Grünwald (Eurostat, Luxembourg) et al, December, 217-224.

Bibliometric study of highly-cited European physics papers in the 1980s, Wolfgang Glänzel (Hungarian Academy of Sciences) et al, August, 113-122.

Changes in Latvia's science indicators in the transition period, Janis Kristapsons et al (Latvian Academy of Sciences), August, 151-160.

Elements of evaluation of S&T policy in Austria, Michael Steiner (Graz University) et al, April, 98-108.

Evaluation of national biomedical research outputs through journal-based esteem measures, Grant Lewison (Wellcome Trust, UK), December, 225-235.

Evaluation of S&T policy in Ireland, Mark Boden (PREST, Manchester) and Michael Fitzgibbon (S&T Evaluation Unit, Dublin), April, 55-62.

France: the guarantor model and the institutionalization of evaluation, Philippe Larédo and Philippe Munstar (Ecole des Mines/CSI), April, 11-22.

Irregularity of Ukraine biological fields, A A Korenovoy et al (STEPS Centre, Kiev), August, 143-150.

Linkage between agency-supported research and patented industrial technology, Francis Narin et al (CHI Research, USA), December, 183-187.


On the importance of regional innovation flows in the EU: some methodological issues from Italy, Simona Lammarino (National Research Council, Rome), December, 188-206.

Patterns of science and technology policy evaluation in Germany, Stefan Kuhlmann (Fraunhofer Institute - ISI, Germany), April, 23-34.

Research actors and the state: research evaluation and the evaluation of S&T in Spain, Luis Sanz-Merendez (CSIC, Madrid), April, 79-88.

Research and technology evaluation in the UK, Philip Hills and Alison Dale (PREST, University of Manchester, UK), April, 35-44.

Research evaluation benefits from international exchange, Erkki Ormala (Science and Technology Council, Finland), April, 2.

Research evaluation in European national science and technology systems, Luke Georgiou (PREST, University of Manchester, UK), April, 3-10.

Research performance at the microlevel: analysis of structure and dynamics of pharmacological research teams, Maria Bordons et al (CINDOC, Madrid), August, 137-142.

Roles of bibliometrics in scientific communication, Hans E Roosendaal (Elsevier Science BV, Amsterdam), December, 237-241.

R&D and innovation at the firm level: improving the S&T policy information base, Morley S Lipsett et al (Simon Fraser University, Vancouver), August, 123-129.

R&D evaluation in Italy: a science and technology policy view, Alberto Silvani and Giorgio Sirilli (National Research Council, Rome), April, 69-78.

R&D evaluation in Portugal, C Marciano da Silva et al (New University, Lisbon), April, 89-97.


The patchwork of the Dutch evaluation system, Arie Rip and Barend van der Meulen (University of Twente, The Netherlands), April, 45-54.

Towards a national system of research evaluation in Greece, Nikolaos Kastrinos (PREST, University of Manchester), and Yannis Katsoulacos (Athens University), April, 63-68.

Transformation of the science system in East Germany: a comparative study, Hans-Jürgen Czerwon (Berlin, Germany), August, 131-136.

Treating bibliometric indicators with caution: their dependence on the source database, Luc Quoniam et al (Faculté des Sciences St Jerôme, Marseilles, France) December, 177-181.

Turn over for details and order form

Beech Tree Publishing
10 Watford Close, Guildford, Surrey GU1 2EP, UK
Telephone and fax +44 1483 567497
E-mail: page@scipol.demon.co.uk
Welcome to Research Evaluation the journal

It is refereed, international, and interdisciplinary. It started in 1991. It has two main aims:

- to bring together the growing literature on research appraisal, monitoring and evaluation (some of it reporting research and experimental results, some systematically drawing on experience, etc); and
- to provide a forum for communicating ideas and experiences, including book reviews, and the other traditional features of a journal.

The ‘Research’ in the title ranges from the individual research project to the national level. Thus it includes research projects and programmes, researchers and institutes, and types of research output—and the public and private sectors, natural and social sciences.

The ‘Evaluation’ in the title applies to evaluation.

Research Evaluation is relevant to you if you are a researcher, a sponsor, a user of research, a research policy analyst, or in any similar job. It is not committed to any specific approach or philosophy, quantitative, qualitative or otherwise.

Subscription information

Published April, August and December. A4 size, 184-208 pages per volume, acid-free paper. ISSN 0958-2029. All back issues available. Volume 6 (1996) will be published in early 1997. Air-speeded to addresses outside UK, Prices in pounds and US dollars:

<table>
<thead>
<tr>
<th>Period</th>
<th>1997</th>
<th>1996</th>
<th>1995</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>£74, $131</td>
<td>£71, $126</td>
<td>£68, $121</td>
</tr>
<tr>
<td>Third world</td>
<td>£57, $101</td>
<td>£55, $97</td>
<td>£53, $94</td>
</tr>
<tr>
<td>Personal</td>
<td>£35, $62</td>
<td>£34, $61</td>
<td>£33, $59</td>
</tr>
</tbody>
</table>

Single order or back-issue: £25 or US$41; two or more: one-third of the 1996 annual price each.

ORDER FORM

I wish to subscribe to Research Evaluation for __ year(s).
First issue: April/August/December 19...
Rate: normal __ third world/East Eur __ personal/student __ and/or I want single issues listed below __
I enclose £/$__ payable to Beech Tree Publishing.
Other negotiable currencies: add £5/US$9 per cheque.
or Charge to my Visa/Mastercard/EuroCard/Access card:
Card # ___________ Expires ___________
or Bank transfer: send me details __
or Please invoice us __ Send me a free sample copy __

Name and address ____________________________________________

To: RE, 10 Watford Close, Guildford, Surrey GU1 2EP, UK
or your usual library subscription agent.

LIBRARY REQUEST

Please consider subscribing to Research Evaluation. It would be useful for [list courses, etc].

It may be more useful than our existing subscription(s) to:

Please also talk about this to:

My name and dept:

0895