Research Evaluation. volume 5. number 3, December 1995. pages 177-181. Pre-refereeing paper

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

QUONIAM L.*, ROSTAING H.*, BOUTIN E.**,DOU H.*
*C.R.R.M. Fac. St. Jérôme. 13397 Marseille CEDEX 20. France. E-mail: crrm@crrm.univ-mrs.fr
**Centre de recherche LEPONT, BP 132. F-83957. La Garde CEDEX

Abstract:
Nowadays, with computer supported analysis of databases, constructing bibliometric or scientometric indicators may be considered as easy. In bibliography, many of those indicators may be found. The problem seems to us more to verify the accuracy of the global analysis, including the sampling of data. The global coherence of an analysis depends of the adequacy of all the steps of the analysis. We built and used an experimentation with on-line databases to demonstrate this through an example. Keeping the same protocol for data collections, we will use the same indicators over the various samples.

I. Introduction

The scientific production of a laboratory, abilities in collaboration building ([1], [2], [3]) are used as indicators of the activity of this laboratory even if they are only quantitative indicators and not quality indicators ([4]). We will use this indicator to show the accuracy of these indicators through various on-line bibliographic databases. As we are going to deal with searchers' work in detail, we will only look over our laboratory, according to the code of practice for information brokers ([5]).

II. Data Collection

One of our laboratory activity is bibliometry, scientometry and informetry. Those three words are nowadays well recognized and defined ([6]). We used those keywords with the Dialog Dialindex [1] to get the databases with the best coverage in this area. Our purpose is not to built the best data collection in each database. Our purpose is more to show the individual perception of each database, using key words recognized by the scientists of the area. Table 1 gives the main responses.

Table 1 Data collection strategy with DIALOG « DIALINDEX ».

N°	Base Collection in 01/1995	Bibliometr?	Scientometr?	Informetr?	1 OR 2 OR 3
61	LISA(Library&infoSci)-	2005	203	59	2071
144	Pascal-	1380	457	93	1606
202	Information Science Abs.-	884	298	52	931
7	Social SciSearch-	388	158	53	566
440	Current Contents Search-	317	155	63	479
434	SciSearch-	270	147	10	410

Then, considering this result, due to cost consideration, we used a maximum of CD-ROM for internal database constitution. LISA and PASCAL were collected out of CD-ROM. LISA was collected in the CD-ROM of the year 1994, and PASCAL in the CD-ROM from the years 1987 to 1994. Due to the great amount of duplicates between SCISEARCH, SOCIAL SCISEARCH and CURRENT CONTENTS SEARCH, the last database was not considered. An internal database has been downloaded from DIALOG from the two other databases without duplicates. Due to variability between CD-ROM and databases the internal databases do not have exactly the same amount of references. For example, PASCAL does not exist in CD-ROM before

[1] DIALOG Information Services. 3460 Hillview Av.. P.O. Box 10010. Palo Alto, CA 94303, USA

1987. The producers of those three internal databases are different (English for LISA, French for PASCAL, and American for SCISERACH). Two of these databases are multidisciplinary (SCISEARCH and PASCAL), and one thematic (LISA). The amount of bibliographic notices collected is reported in Table 2

Table 2 Number of bibliographic notices in our internal databases

Internal database	*Number of bibliographic notices*
SCISEARCH + SOCIAL SCISEARCH 01/95 Dialog	803
PASCAL from 1987-1994 CD-ROM	1191
LISA from 1994 CD-ROM	2229

III. Treatments

Over those databases, several treatments were achieved for internal reasons. We are now presenting few of them. Instead of building a common database without duplicates, we kept the separate databases to demonstrate the particularity of each one. Each database has his indexation practice, sources and point of view. Applying the same treatment over each separate database will outline the specificities.

A. Specificity in fields constitution and indexation practice

Outlining the databases' specificity could seem obvious but is very important. The SCISEARCH is the only database including author's citations, but includes a weak keyword index (240/803 references) which is better in PASCAL and LISA. SCISEARCH is also the only database who gives a description of the activity of the publication journal. PASCAL is the only database that provides the country of publication Our next treatment will attempt to emphasize one kind of specificity, the difference in data collection for those databases.

B. Specificity in sources and abstracting practice

To outline this point we determined the co-authors' groups. We do elaborate a software able to build groups (for example authors' groups) with a propagation algorithm ([7]). This algorithm is not at all based on a classic clustering technique, but use the natural structure of co-authoring ([8]). This algorithm only uses the co-presence of items in references. This means that the algorithm determines which are all the co-authors of an author. Then the algorithm determines which are the co-authors who work with the co-authors, and continue 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 all the studied database. When a group is finished, other groups are built up to the last author is entering a group or determined as an individual author. This algorithm may be applied over any field in the studied database. We choose authors' groups, but we could build keyword groups. The linked fields to the group may also be chosen depending of the interest of the analysis. This algorithm works very well with authors because the relations between authors are finite. Other authors published the same algorithm, but without linked fields to each built group ([9], [10]).The thresholds used for this analysis are explained in Table 4. The Table 3 represents our laboratory in the SCISEARCH database with those thresholds.

Table 4 Analysis parameters

Author Minimum frequency:	2
Co-authors minimum frequency:	1
Relations with other fields minimum frequency:	2

Other fields in SCISEARCH: affiliation, publication year, source, Journal subject category, author keywords, keywords plus.
Other fields in PASCAL: affiliation, publication year, source, English descriptors.
Other fiels in LISA: source, publication year, key words

Table 3 Our laboratory group in the SCISEARCH database

Numbers between parenthesis are frequencies

AUTHORS :
DOU H (3); QUONIAM L (2); HASSANALY P (2)
COAUTHORS RELATIONS :
DOU H [hassanaly p (2), quoniam l (2)]
HASSANALY P [dou h (2), quoniam l (2)]
QUONIAM L [dou h (2), hassanaly p (2)]
AFFILIATIONS :
DOU H [marseille; france (3)]
HASSANALY P [marseille; france (2)]
QUONIAM L [marseille; france (2)]
JOURNAL SUBJECT CATEGORY :
DOU H [information science & library science (2)]

With such parameters, our laboratory, through SCISEARCH, is perceived as a small laboratory (3 persons), without collaborations and which publish few (a maximum of 3 articles) in journals of the area of information science & library science.
With exactly the same parameters, we treated the LISA databases. First, it is important to notice the impossibility to extract affiliations out of the LISA CD-ROM. The author field is also not homogeneous. A bibliometric treatment of this field requires a manual control of the database. Table 5 shows the output of the treatment.

Table 5 Our laboratory group in the LISA database

Numbers between parenthesis are frequencies

AUTHORS :
Dou (8); Hassanaly (7); Quoniam (6); La Tela (2)
COAUTHORS RELATIONS :
DOU [hassanaly (7), la tela (2), quoniam (6)]
HASSANALY [dou (7)]
HASSANALY [la tela (2)]
HASSANALY [quoniam (5)]
LA TELA [dou (2)]
LA TELA [hassanaly (2)]
LA TELA [quoniam (1)]
QUONIAM [dou (6), hassanaly (5), la tela (1)]
SOURCES :
DOU [scientometrics (3)]
HASSANALY [scientometrics (2)]
QUONIAM [scientometrics (3)]
PUBLICATION YEAR :
DOU [1989 (2), 1988 (2)]
HASSANALY [1989 (2), 1988 (2)]
QUONIAM [1989 (2)]
KEY WORDS :
DOU [chemistry (2), bibliometrics (6), library materials (6), stock (6)]
HASSANALY [chemistry (2), bibliometrics (6), library materials (6), stock (6)]
QUONIAM [bibliometrics (5), library materials (5), stock (5)]

The perception of our laboratory begins to be different. Our laboratory appears to be constituted by four authors that still publish few (a maximum of height papers in the last six years). The initial specificity of our laboratory appears (chemistry).
With exactly the same parameters we treated the PASCAL database. Table 6 shows the output of the treatment.

Table 6 Our working group in the PASCAL database

Numbers between parenthesis are frequencies

AUTHORS :

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); HUOT 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)

COAUTHORS RELATIONS :

BEDECARRAX C [huot c (2)];
BILLARD P [dionne mc (1), longevialle c (2), paoli c (2)];
DIONNE MC [billard p (1), dou h (1), la tela a (1), longevialle c (1), paoli c (2)];
DKAKI T [dousset b (2), hilaire a (1), koussoube s (2), longevialle c (1)];
DOU H [dionne mc (1), hassanaly p (15), huot c (1), kister j (2), la tela a (7), paoli c (1), quoniam l (10)];
DOUSSET B [dkaki t (2), hilaire a (2), koussoube s (2), longevialle c (1)];
HASSANALY P [dou h (15), kister j (2), la tela a (6), quoniam l (8)];
HILAIRE A [dkaki t (1), dousset b (2), koussoube s (1), longevialle c (1)];
HUOT C [bedecarrax c (2), dou h (1), quoniam l (1)];
KISTER J [dou h (2), hassanaly p (2)];
KOUSSOUBE S [dkaki t (2), dousset b (2), hilaire a (1), longevialle c (1)];
LA TELA A [dionne mc (1), dou h (7), hassanaly p (6), paoli c (1), quoniam l (2)];
LONGEVIALLE C [billard p (2), dionne mc (1), dkaki t (1), dousset b (1), hilaire a (1), koussoube s (1), paoli c (2)];
NIVOL W [quoniam l (1), rostaing h (1)];
PAOLI C [billard p (2), dionne mc (2), dou h (1), la tela a (1), longevialle c (2), rostaing h (1)];
QUONIAM L [dou h (10), hassanaly p (8), huot c (1), la tela a (2), nivol w (1), rostaing h (1)];
ROSTAING H [nivol w (1), paoli c (1), quoniam l (1)]

AFFILIATIONS

BEDECARRAX C [paris, fra (2)]; BILLARD P [paris, fra (2)]; DKAKI T [toulouse, fra (2)]; DOU H [marseille, fra (18)]; DOUSSET B [toulouse, fra (4)]; HASSANALY P [marseille, fra (14)]; HILAIRE A [toulouse, fra (2)]; HUOT C [paris, fra (3)]; KISTER J [marseille, fra (2)]; KOUSSOUBE S [toulouse, 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)]

SOURCES :

DOU H [Cahiers de la documentation(2), Scientometrics (3)];
QUONIAM L [Cahiers de la documentation (2), Scientometrics (3)]

PUBLICATION DATE :

DKAKI T [1991 (2)]; DOU H [1990 (2), 1992 (4), 1991 (4), 1987 (4), 1989 (4)]; DOUSSET B [1991 (2)]; HASSANALY P [1991 (2), 1990 (2), 1987 (4), 1989 (5)]; KOUSSOUBE S [1991 (2)]; LA TELA A [1990 (2), 1987 (2)]; PAOLI C [1991 (2)]; QUONIAM L [1992 (2), 1990 (2), 1991 (4), 1989 (5)]; ROSTAING H [1993 (2)]

KEY WORDS :

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), coword 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 (8), 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), coword 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)];
NIVOL W [bibliometrics (2), bibliometric analysis (2), method (2)];
PAOLI C [bibliographic data (2), database (3), bibliometric analysis (4)];
QUONIAM L [informetrics (2), classification (2), code (2), bibliographic data (2), information processing (2), example (2), scientific literature (2), data analysis (2), patent document (2), methodology (2), congress (2), published document (2), automated processing (2), evaluation (2), bibliometry (3), statistical analysis (3), on line processing (3), downloading (4), data processing (4), europe (4), france (4), scientific research (4), bibliometrics (5), scientometrics (5), database (7), chemistry (8), bibliometric analysis (10)];
ROSTAING H [bibliometrics (2), scientific literature (2), bibliometric analysis (2), technological awareness (2), information processing (2)]

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 of 18 authors which publish normally (a maximum of 20 publications for a 7 year period). Collaborations appear with the IBM CEMAP Center of Paris (Huot and Bedecarrax), with the IRIT in Toulouse (Dousset, Dkaki, Koussoube), 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 in the graph is a function of the amount of co-publications between authors. Keeping in mind just the thick links, the structure of collaboration appears. Thick links between members of the same affiliations, thin links between members of different affiliations. This automatically

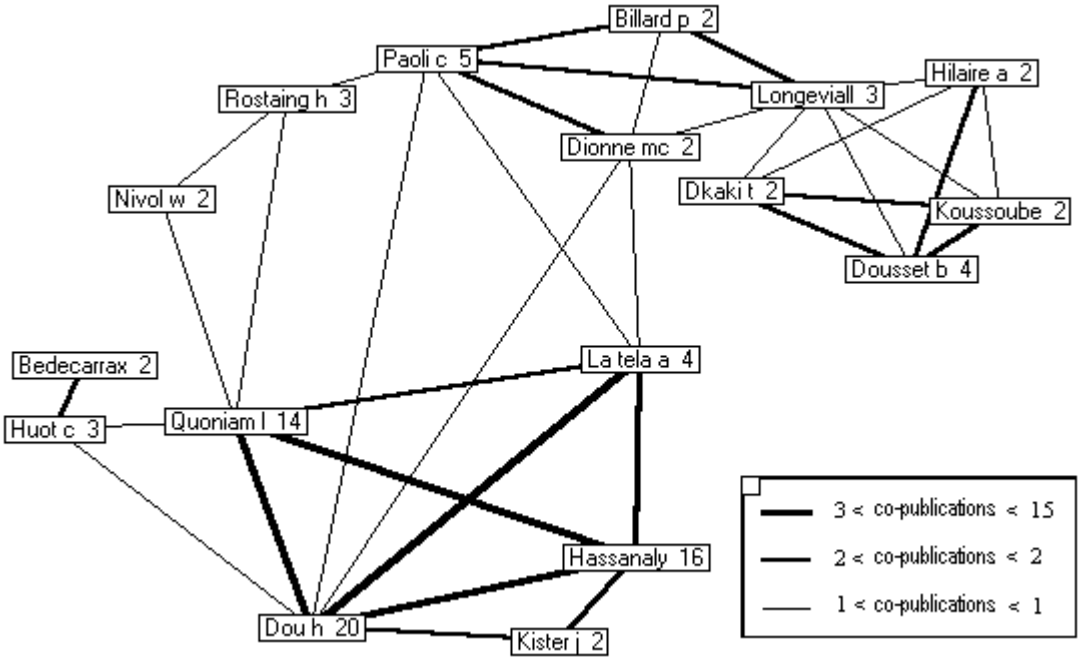

Figure 1 Co-authors relations in the PASCAL database.

generated graph was obtained using an algorithm built in our laboratory ([11]).

IV. Discussion

Depending on indicators, macro or micro level bibliometry ([12], [13]) and the application country, the incidence of the chosen database may be more important than the chosen indicator. An indicator may be considered efficient when he does not change the perception of the reality. What appends if the reality of the database is wrong? We did show that an American multidisciplinary database and an English specific database may deeply under estimate the activity of a French laboratory. Any one can imagine that the French multidisplinary database over-estimates the activity of this laboratory. We are not convinced of this point because this multidisciplinary database seems exhaustive in information science and other bibliometrics teams (which do not publish in French) are also under estimated in the American and English database in comparison with the French database. On the other hand, it is easy to imagine that other teams are over estimate in both American and English databases. But most of the teams keep the same importance from one database to the other. A consideration of both estimations (under and over) is the minimum for an « objective » evaluation. This leads us also to keep separate databases from various

producers to emphasize the contrast between databases. In that case, it is possible to try to explain the different point of view of the different databases. For place consideration, and to minimize the political problems of searchers' evaluation (5), we just use the sample of our working group. The bibliometrician, that uses on-line databases, has no incidence over the data collection of the producer or the server of the database. He as well cannot control the accuracy of indexation. He must very often control misspellings' author names or affiliations. A real estimation of the database must be performed before bibliometric indicators building. An analysis considering several databases is very often better.

Acknowledgment:
We would like to thank both Dialog and INIST [2] for providing their precious help.

V. Bibliography

[1] GLANZEL W; WINTERHAGER M
International collaboration of three east European countries with Germany in the sciences, Scientometrics, 1992; Vol. 25; N°. 2; p. 219-227

[2] KRETSCHMER H
Coauthorship networks of invisible colleges and institutionalized communities
Scientometrics; 1994; Vol. 30; N°. 1; p. 363-369

[3] PETERS HPF; VAN RAAN AFJ
Structuring scientific activities by co-author analysis: an exercise on a university faculty level
Scientometrics; 1991; Vol. 20; N°. 1; p. 235-255

[4] LIMING L; LIHUA L
Scientific publication activities of 32 countries: Zipf-Pareto distribution
Scientometrics, 1993; Vol. 26; N°. 2; p. 263-273

[5] EIIA, EUSIDIC, EIRENE
Code of practice for information brokers,
Information Services & Use 14 (1994), p. 115-121.

[6] EGGHE L.
Bridging the gap: conceptual discussions on infometrics,
Scientometrics, may 1994, Vol.30, N°1, p35.

[7] HAON H., PAOLI C., ROSTAING H.
Perception d'un programme de R.&D. à travers l'analyse bibliométrique des banques de données d'origine Japonnaises.
Acte du colloque I.D.T.-93. 22-24/06/1993. p. 63-70.

[8] LOGAN EL; SHAW WM JR
A bibliometric analysis of collaboration in a medical specialty
Scientometrics; 1991; Vol. 20; N°. 3; p. 417-426

[9] BORDONS M., ZULUETA M.A., CABRERO A., BARRIGON S.
Identifying research teams with bibliometrics tools. Proceedings of the fifth biennal conference of the international society for scientometrics and informetrics. Chicago USA. Learned information. Juin 1995. p. 83-92.

[10] ABD-EL-KADER M., MIQUEL J.F., DORE J.C.
Méthode d'analyse des thèmes et réseaux de la coopération scientifique nationale et internationale sur les céramiques dans les bases SCI. Colloque Les systèmes d'information élaborée. Juin 1995. Ile Rousse France.

[11] BOUTIN E., QUONIAM L., ROSTAING H., DOU H.
A new approach to display real Co-authorship and co-topicship through network mapping. Proceedings of the fifth biennal conference of the international society for scientometrics and informetrics. Chicago USA. Learned information. Juin 1995. p. 676.

[12] GRUPP H
On the supplementary functions of science and technology indicators : the case of West German telecommunications R & D
Scientometrics; 1990; Vol.19; N°.5-6; p. 447-472

[13] MIQUEL JF; OKUBO Y
Structure of international collaboration in science. II: Comparisons of profilesin countries using a link indicator
Scientometrics; 1994; Vol. 29; N°. 2; p. 271-297

[2] I.N.I.S.T. 2 allée du parc de brabois. 54514 Vandoeuvre-les-Nancy. CEDEX. France