INFORMETRICS 87/88
L. Egghe and R. Rousseau (Editors)

THE USE OF ON-LINE DATABASES FOR BIBLIOMETRIC ANALYSIS

H.F. Moed

Science Studies Unit, Lisbon Institute
University of Leiden, Stationsplein 242
2312 AR Leiden, The Netherlands

Abstract

Databases containing bibliometric information on published scientific literature play an important role in the field of quantitative studies of science and in the development and application of Science and Technology indicators. For these purposes, perhaps the most important and probably the most frequently used database is the Science Citation Index, produced by the Institute for Scientific Information. SCISEARCH, the on-line version of the Science Citation Index (SCI), is included in several host computers. However, other databases are used as well, such as Physics Abstracts or Chemical Abstracts.
In this contribution, potentialities and limitations of several online databases as sources of bibliometric data in a number of host computers will be discussed. The discussion will focuss on the on-line version of the Science Citation Index, and on citation analysis.
It will be argued that for several specific bibliometric applications, on-line databases and software implemented in the host computer do not provide appropriate facilities. In fact, for these specific applications, one should first download the primary data from the host into a local computer (PC, Mainframe). Next, dedicated software should be developed on a local level, in order to perform the bibliometric analyses properly. This will be illustrated by presenting a number of applications, related to citation analysis ('impact measurement') and co-citation analysis ('mapping fields of science').

1. INTRODUCTION

The subject of this paper is the exploration of potentialities and limitations of the use of 'online' scientific literature databases for bibliometric purposes.
Nowadays a large number of databases containing bibliographic information on published scientific literature is being produced. Some of these focus on specialized subjects, others may have a much broader scope and claim to cover the main sources of information of an entire scientific discipline, or even of science as a whole (multidisciplinary literature databases). The content of these databases is usually published in a hard copy form, enabling users to do manual searches. However, during the last decade a number of organizations have arised that store the information in computer-readable form into computers, use modern telecommunication techniques to enable users to have access to the computer system, and offer software to search the databases for relevant scientific material.
Probably many new developments are awaiting us in the near future, both with respect to the number of databases brought online, the number of host organizations offering online facilities, the telecommunication techniques applied, the logical structure of the databases, the hardware, and - last but not least - the software facilities offered to search the databases. therefore, results of an exploration of the potentialities of these online scientific databases for

bibliometrical purposes should be considered as highly time-dependent: what is not possible today, might become possible tomorrow. Actually, the experiences described below refer to the 'state of the art' anno 1987.
In this contribution, the use of online literature databases is considered mainly from the point of view of a 'bibliometrician', i.e. as the person collecting data on quantitative aspects of science, using the formal scientific literature as data source, and focusing on bibliographic information such as author names, institutional (corporate) addresses, lists of cited references, titles, abstracts, indexing terms, and keywords. One can make a distinction between two general perspectives from which a bibliometrician collects his quantitative bibliographic data: science policy and science studies. Operating from the perspective of science policy, a bibliometrician seeks to collect data and to calculate indicators of aspects of science that are considered as policy relevant, such as scientific activity or impact of (groups of) researchers in a specific field of country. The other perspective - science studies - is rather broad, and embraces a large number of topics, such as the study of the development of scientific specialties or the study of the communication system within science.
Researchers at the Science Studies Unit of the LISBON-Institute at the University of Leiden have experienced with online literature databases for almost two years. In fact, in several projects these databases are used extensively. The most important database in these projects is SCISEARCH, the online version of the Science Citation Index, produced by the Institute for Scientific Information at Philadelphia (USA). SCISEARCH is implemented in several host computers. In our projects, we used the host of the Deutsches Institut für Medizinische Documentation und Information (DIMDI) at Cologne (Federal Republic of Germany). In this paper, an overview is given of many techniques that were developed. Four bibliometric analyses will be presented in which data were gathered from online scientific literature databases. The specific questions and problems studied in these analyses arise primarily from the perspective of science policy. One must keep in mind that the analyses presented here have not been completed yet, and therefore no definitive conclusions can be drawn. However, they provide a clear illustration of the type of questions or problems a bibliometrician has to deal with, and of the specific bibliometric data that are needed to tackle the various problems.
The structure of this paper in as follows. First, the main limitations of online literature databases are discussed briefly. Next, a general overview is given of the techniques that were developed to collect relevant data from online databases and to handle these data. Four cases will be presented in which specific bibliometric data, combined with other types of information are used to study certain problems or phenomena. Finally, conclusions are drawn with respect to potentialities and limitations of the online databases.
As will become clear in the next sections, downloading of bibliographic data from online literature database and storing these data on computer diskettes, constitutes a main element of our method. I emphasize that these data will only be used for the bibliometrical research projects indicated in the paper and will not be used for any other purposes.

2. MAIN LIMITATIONS OF ONLINE LITERATURE DATABASES

2.1 Unification problem

Probably the most important problem that one encounters in using literature databases is the so called unification problem: the same 'object' can have different names. To give a simple example: The University of Leiden appears in the corporate source field (indicating the institution of which the authors of a publication are working) of the SCISEARCH database at least in the following variations:

Leiden State Univ
State Univ Leiden
Univ Leiden
Leiden Univ
Rijksuniv Leiden
Sterrewacht Leiden
Gorlaeus Labs
State Univ Leyden

To an analyst having some background knowledge on the University of Leiden it is evident that all names listed above indicate the same university. However, for a computer comparing character strings the variations are completely different.
It should be noted that the unification problem does not merely occur in online scientific literature databases, but, for instance, also in databases containing information on clients of firms or banks, and regardless of whether these database are computerized or not.
Literature databases often contain data fields that are not unified, although such fields may contain information that is relevant for bibliometrical purposes. These fields should be unified first, before correct bibliometric numbers can be determined. Doing the unification online is in many cases almost impossible and would be very expensive.
The phenomenon of several names indicating the same object does not only occur within one database, but arises also if two databases are merged, such that for each individual publication information from both databases is combined. Since differences may exist between the formats of author names or journal titles in the two databases, the problem arises: how can identical scientific publications be identified ?

2.2 Not all bibliographic information is stored in separate, searchable fields

A typical example is the fact that in some databases the year of publication of scientific publications is not a separate field. As a consequence, it is not possible to select publications from some specific year using a simple host command (of course, selecting articles from a specific 'entry date' is always possible). According to our experiences, of much greater importance is the fact that in none of the databases the volume number and starting page number of articles are separate, searchable fields. Of course, certain types of scientific publications do not have a volume - or page number (for instance books, reports, theses), but journal articles do have this information. And journal articles constitute an important object in bibliometric analyses. Publication year, volume number and starting page number are most useful for the identification of identical scientific journal articles in the case that variations exist in author names or journal titles.

2.3 Host software may not be very powerful

The host of the European Space Agency (ESA) at Frascati (Italy) has implemented a so-called zoom command, enabling users to produce simple frequency tables of the occurence of words or phrases in some specific data fields for a selected set of articles. The host STN offers a similar facility, though it seems to be less powerful then ESA's zoom command since it can handle only a limited amount of data. Other hosts, such as DIMDI, do at present not have such a facility at all. Although in simple frequency tables much of the original information may be lost (particularly relationships between the words of phrases), these tables can still be most useful for bibliometrical purposes.

2.4 Additional information is needed.

According to our experiences, in most bibliometric applications the bibliometric data collected from a literature database can produce significant results only if these data are combined with other types of information. Actually, one should

construct a compound database in which to each scientific publication other data are added, either bibliometric or non-bibliometric. In particular, a database containing article-by-article information from different scientific literature databases seems to be a promising tool for bibliometric studies.

3. GENERAL OUTLINE OF THE METHODS AND TECHNIQUES

The main elements are shown in Figure 1. An IBM-compatible Personal Computer (PC) acts as an interface between host computers containing literature databases on one side and the IBM Mainframe-computer of the Leiden University Computer Centre on the other. The software packages used for the PC are: an operating system (DOS), an editor and two terminal emulating programs: one for communication with the host computers, and one for interaction with the IBM Mainframe. By loading the emulation software into the memory, the PC can act as a terminal. In addition, it provides facilities for 'file transfer' from PC to host computer or to IBM Mainframe.

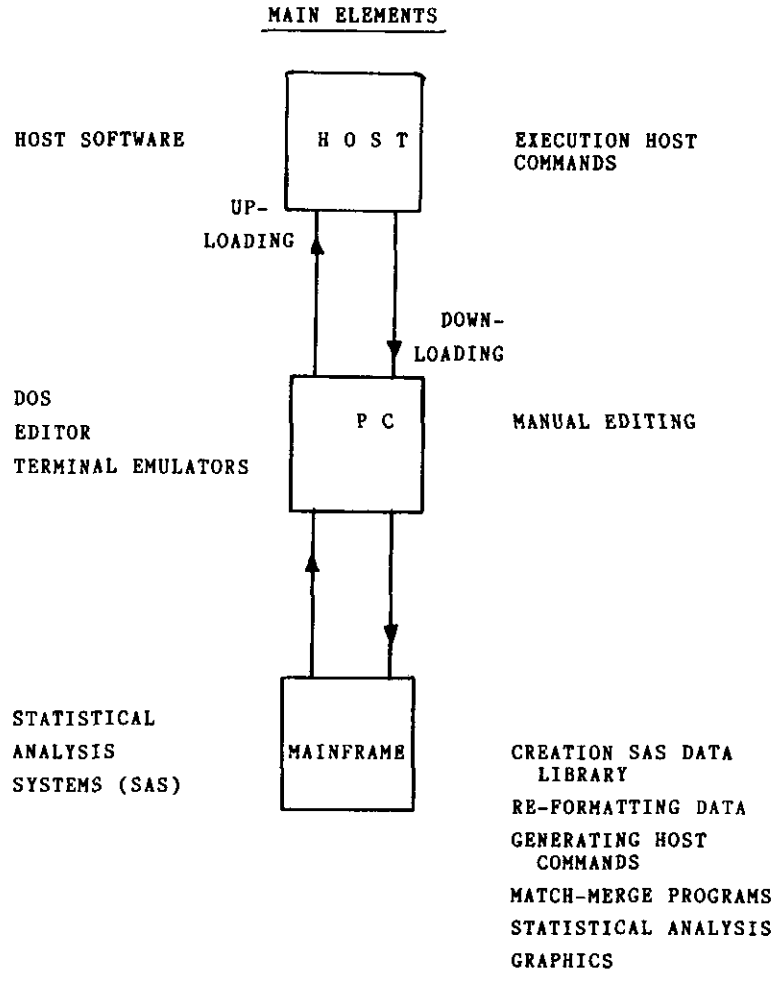

Fig. 1: Main elements of the methods and techniques applied

File transfer between PC and host computer involves both uploading and downloading. Downloading means that information send by the host to the PC is stored on PC diskette or hard disk. The uploading facility is used in the following manner. First, a file containing host commands is created either by manual editing or by computer programming. Next, the file is opened and the first command is sent to the host. After execution of the command, the host sends a 'host prompt' - for DIMDI the prompt is a question mark - to the PC. This is the sign for the file transfer software to send the next command to the host, and so on.
In the present configuration, a dial-up modem is used to connect the PC-terminal to a node of a national data-communication network of the Netherlands Post, Telegraph, and Telephone Company. This network in its turn is connected to several international networks, along which most host computers can be accessed. Selected data from host computers are downloaded to the PC, and sent to the IBM Mainframe. In the Mainframe computer, we use the 'Statistical Analysis System' (SAS), a software package developed by SAS Institute Inc., North Carolina (USA). The SAS system provides tools for information storage and retrieval, data-modification and programming, report writing, file handling, statistical analysis and graphical display (SAS-Graph). The main task is to create a SAS Data Library, containing the bibliographic information that was downloaded from the hosts. Essentially, to each publication a unique sequence number is assigned. Several datasets are created, each containing specific types of information on these publications that correspond roughly to the various data fields in the online literature databases: primary bibliographic data (i.e. author names, journal title, volume number, starting page number, and publication year), data on corporate addresses, on reference lists, on citations (i.e. on articles citing a specific publication), and on indexing terms, classification codes and keywords. Within each dataset, the data are splitted into subfields or variables. For instance, separate variables exist for the volume number, starting page number and publication year, both of source articles and of the cited references. The general structure and content of such a SAS Data Library is given in Table 1. For several projects in our research program, data libraries have been created. These libraries contain data on articles from different scientific research fields, and are extracted from different online literature databases. In order to create these libraries,the data from the hosts must be re-formatted. The format of the data varies from host to host and from database to database. Computer programs were written, that re-format data from SCISEARCH (host: DIMDI), Chemical Abstracts (host: STN), Mathematical Abstracts (STN), BIOSIS (STN), Excerpta Medica (DIMDI), and INSPEC (ESA). In addition, a program was written that re-formats data, downloaded from the host ESA, in the so called Format X.
Once a SAS Data Library is created, all SAS procedures can be applied easily. For instance, sorting procedures, match-merge programs, procedures to generate all kinds of frequency tables, advanced statistical procedures and facilities for graphical display. In addition, data fields can be unified.

4. FOUR CASES

4.1 Scientific Activity in Belgium: publication counts by subfield using SCISEARCH (DIMDI host).

Recently, two data fields containing significant information for bibliometric analyses were added to the SCISEARCH database at the host DIMDI. The first is the *corporate country field* (CCO Field) that contains the name of the country from which the article originated. These corporate country names are unified by DIMDI. The second new data field contains information on the journal category to which the articles belong (SH or SC field). These categories correspond roughly to scientific subfields, and are based on a classification of scientific journals. Using these new data fields, one can obtain an indication of the degree of scientific activity of a specific country within the various scientific subfields. To generate the relevant data, we proceeded as follows. As the time period for

the analysis we chose the total time span of the most recent segment of SCISEARCH at DIMDI, that is the period from 1-1-1983 tot 31-5-1987. We decided to perform an analysis on articles from Belgium. First, we downloaded data on the total number of articles published in each category, by displaying the part of the Master Index that relates to the category codes (SC field).

Table 1: General structure of a SAS Data Library containing bibliographic information

Dataset	Variables/fields
Primary dataset	sequence number publication first author second author ... journal title year of publication volume number starting page number ending page number language document type number of references
Corporate source data	sequence number publication publishing institute publishing city publishing country
Title data	sequence number publication title
Dataset classification codes	sequence number publication classification code
Dataset indexing terms	sequence number publication indexing term
Dataset references	sequence number (source) publication cited first author cited journal title cited volume number cited starting page number cited publication year
Dataset citations	sequence number publication number of citations

The number of different categories appeared to be 196. Next, we generated on our PC by manual editing a host command of the type:

F CCO = Belgium
F SC = AA and 1
F SC = AB and 1

The first command selects all articles with Belgium as a corporate country name. The host creates a set of these articles and assigns the number 1 to this set. The next command selects articles that have a specific category code (for instance, 'AA') and that are included in set number 1. All commands were sent to the host and were executed one after another. At the end 197 sets were created. By displaying the search history the relevant data were downloaded. We were connected to the host for 30 minutes. Total costs (both host and database fees) amounted to some 60 DM. Thus two files were created, containing for each subfield the total number of articles and the number of Belgian articles respectively.

The files were sent to the IBM Mainframe, were re-formatted and merged together, using SAS software. Finally, simple indicators were calculated. We calculated an activity index for each subfield, dividing the percentage of Belgian articles in a subfield relative to the total number of articles from all countries in that subfield, by the percentage of Belgian paper in all subfields relative to the total number of articles in the SCISEARCH segment. Thus, one can indicate subfields in which Belgian researchers are active, compared to the overall Belgian activity in science as a whole. The subfields in which Belgium is very active in this sense are listed in Table 2. Only subfields are selected for which the Belgian Activity Index exceeds 1.5, and that contain a total number of articles from all countries that exceeds 1000.

Table 2: Subfields in which Belgian researchers are most active compared to the overall Belgian activity in Science

Subfield	Number of Publications	Activity Index
Physiology	1441	3.5
Tropical Medicine	207	3.4
Mathematical Physics	135	2.8
Endocrinology and Metabolism	833	2.4
Nuclear Physics	345	2.3
Biochemistry & Molecular Biology	2626	2.2
Respiratory System	483	2.1
Gastroenterology	758	2.1
Mineralogy	85	1.8
Spectroscopy	289	1.7
Dermatology & Veneral Diseases	377	1.7
Mathematical Methods, Physical Sciences	167	1.7
Pharmacology & Pharmacy	1577	1.7
Analytical chemistry	577	1.6
Agriculture	546	1.6
Rheumatology	125	1.6
Virology	167	1.5
Pure Mathematics	188	1.5
Urology & Nephrology	414	1.5
Physics, Particles & Fields	146	1.5

With respect to the interpretation of Table 2 three comments should be made. First, in the calculation of the Activity Indexes no input measures are involved, such as the number of researchers working in each subfield. Second, only articles are counted that are published in journals processed for the SCISEARCH database. Finally, the classification of journals into categories is a rather global one, and is probably mainly based on an inspection of the journal titles.

4.2 Citation analysis of scientific journals

Bibliometric data were collected for a number of scientific journals. For each article in these journals, information was gathered on publishing authors and institutions. The names of the institutions were unified. In addition, it has been determined how many times each article has been cited during the first three (calender) years after publication date, the year of publication included (short term impact).

First, publication data were collected from SCISEARCH at the DIMDI host. Only data fields containing information on authors, source, corporate source and article type were downloaded. Next, citation data were collected, using host commands of the following type:

F RF = EUR J PHARM AND RF = 1983 AND PY = 1985

Executing this command, the host selects all source articles in the database published in 1985 that contain in the reference lists at least one reference in which the string 'EUR J PHARM ' appears as a journal title, and '1983' as the publication year of a cited article. 'Eur J PHARM ' indicates the European Journal of Pharmacology. Journal titles in the reference field are not unified. As a consequence, we had to determine first in which variations the titles of our journals appeared in the reference lists, by displaying the index of cited journal titles. Thus, a large set of articles was created, that contained all references to the selected journals that we were able to find in the database. Of these source articles, we downloaded only the relevant references, i.e. references to our journals and to the specified publication years. A typical example of a downloaded reference is:

Pieters JGM; EUR J PHARM; 1983; vol. 86; pg. 519

The numbers 86 and 519 indicate the volume number and the starting page number of the cited reference, respectively.
Publication and citation file were sent to the IBM Mainframe. Data were re-formatted using computer programs in SAS, and a SAS Data Library was created. Finally, the number of citations to each article was counted by match-merging publication- and citation-file. Since variations occur in author names and journal titles, the problem arises as to how identical publications can be identified, and by which fields or variables the publication- and citation-file should be matched. We used a match-key containing the following variables :
- the first four characters of the last name of the first author
- his first initial
- the first character of the journal title
- the volume number
- the starting page number (at most four digits)
- the year of publication.

The problem of defining an appropriate match-key is very interesting indeed, a detailed discussion goes beyond the scope of this paper.
It should be noted that the method of citation analysis described above was developed at the beginning of 1987, and is based on the facilities that DIMDI offered in that period. Until June 1987, it was almost impossible to select citations to a given article using one (or few) commands, since only three subfields could be searched: the author name, the journal title and the publication year of a cited reference. Recently, DIMDI changed the structure of the reference (RF) field. Now, the publication year of a cited reference is not a separate searchable subfield anymore. However, the possibility exists to add in a citation search information on the volume number and starting page number of a cited reference. At present we are developing a new method for online citation analysis, using the new facilities offered by DIMDI. Results will be published in the near future.
The costs of the method described above depend on the number of articles involved, and on the number of times these articles are cited. For the journals in our pilot project the online costs were approximately 1 DM per article.

Table 3: The most highly cited articles from three journals in the field of Pharmacology

Rank	Publishing Institute	Number of citations
1	Schering Plough Corp, Bloomfield, USA	86
2	Merck Sharp & Dohme Res Labs, USA	85
3	Reckitt & Colman Ltd, Hull, Great Brittain	76
4	H Lundbeck & Co AS, Copenhagen, Denmark	74
5	Upjohn Co, Kalamazoo, USA	70
6	St Thomas Hosp & Med Sch, London, Great Brittain	69
7	Ctr Rech Merrell Int, Strasbourg, France	65
8	Karolinska Inst, Stockholm, Sweden	59
9	Vet Adm Ctr, Philadelpia, USA Univ Penn, Philadelphia, USA	57
10	ICI Ltd, Macclesfield, Great Brittain	52

The result of two analyses are presented in Table 3 and Figure 2, respectively. The first is an analysis of highly cited articles from three journals in the field of pharmacology. The following three journals were included: the European journal of Pharmacology, the British Journal of Pharmacology and the journal Pharmacology and Experimental Therapeutics. Articles were selected from volumes, published in 1983 and 1984. As has been stated before, for each article the number of citations was counted, received during the first three years after publication date, the year of publication included. The ten most highly cited articles are listed in Table 3. Review articles are not included.

In Table 3 one observes that the five most highly cited articles are published by pharmaceutical industries. The overall percentage of articles from pharmaceutical industries is only 13%. The question arises as to why so many highly cited articles are published by pharmaceutical industries. One can pose at least three explanations for this outcome:

(i) Research in pharmaceutical industries has the highest quality, compared to universities or governmental agencies.
(ii) Pharmaceutical industries publish more 'methodological' articles in which a new practical technique or method is introduced.
(iii) These industries produce the most controversial papers.

Further research into this problem is needed.

The second analysis relates to another journal, the Journal of Nuclear Materials, and to 'impact factors'. The impact factor of a journal calculated by ISI, measures the average number of citations in a specific year (say, 1986) to articles, published in that journal in the two previous years (1985 and 1984). We calculated a measure that is very similar to the impact factor: the average number of times articles published in a journal in, say, 1984, are cited during the period 1984-1986. Our impact factor is slightly more than two times the 'official' ISI impact factor. However, we calculated our impact factor for each individual volume in the journal. We took into account the month in which each volume was published. Moreover, we included additional information on the type of volume, whether it is a proceedings volume or a 'regular' volume. For proceedings volumes we determined whether these were produced by 'typeset' or by 'camera ready' method. The results are presented in Figure 2.

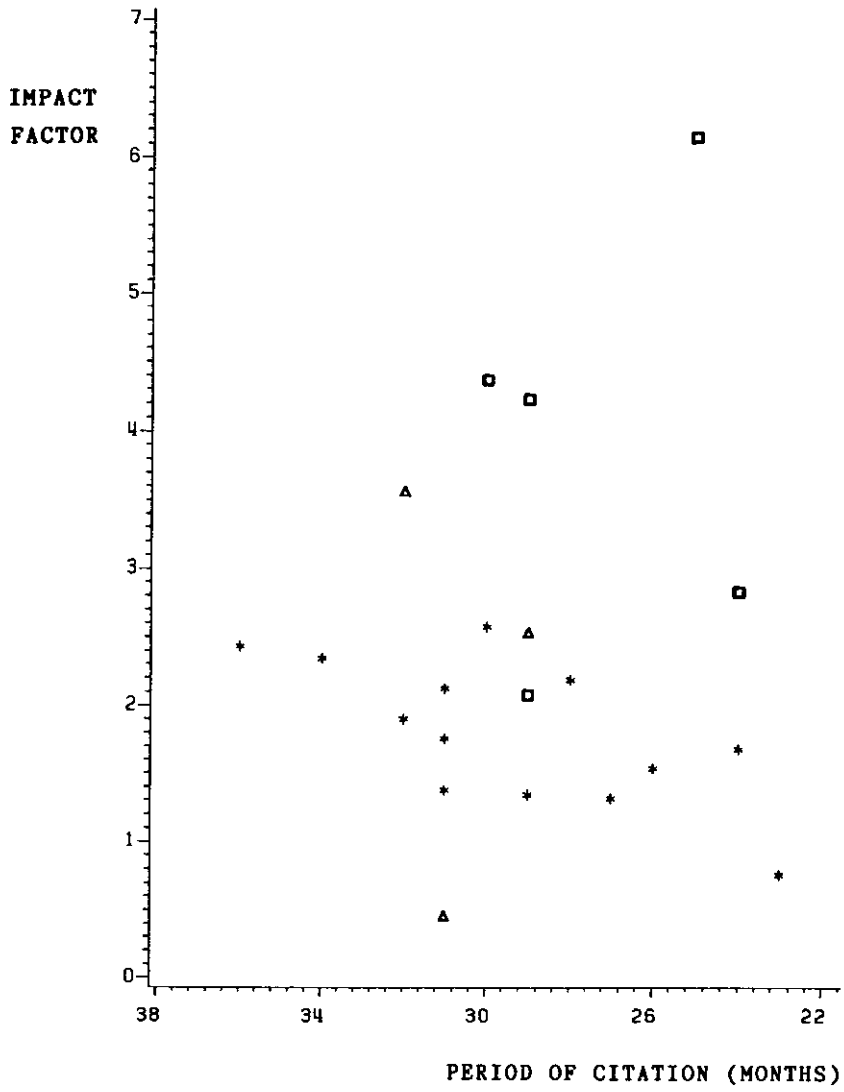

Fig. 2: Citation analysis of a scientific journal: impact factor by volume.

Figure 2 shows, first of all, that large differences exist between the scores of the various volumes. These differences remain completely invisible if one calculates an impact factor by aggregating all articles published in a specific year. Particularly for the regular volumes one observes a decrease of the average impact, as the period, during which citations were counted, decreases. This is an artefact of the method of counting citations: volumes that are published in January in a specific year have three years (36 months) to be cited, while volumes published at the end of December in that year have only two years (24 months).
Finally, it appears that - for this journal - most of the proceedings volumes are highly cited compared to regular volumes, and that the 'typeset' proceedings are cited more frequently than the 'camera ready' proceedings.

4.3 Combining citation data from SCISEARCH and information from other Abstracting/indexing services

At present two projects are conducted in which citation data collected from SCISEARCH online are combined with information from several other databases: Chemical Abstracts, Physics Abstracts, Mathematical Abstracts, BIOSIS, and Excerpta Medica. One project is financed partly by the Netherlands Organization for the Advancement of Scientific Research (ZWO) and examines differences between publication and citation characteristics in a number of scientific subfields. The second project is related to co-citation analysis, and is financed partly by Ministry of Education and Science (through the Advisory Council for Science Policy in the Netherlands, RAWB). This latter project is discussed extensively by R.R. Braam, H.F. Moed and A.F.J. van Raan (1988). The main purpose is to perform citation- or co-citation analyses based on the reference lists in articles from a number of specific scientific subfields. The subfields are defined by using some combination of classification codes, indexing terms, keywords, assigned to articles by other abstracting services such as Chemical Abstracts. In addition, in one case journal titles were used as well to define a scientific subfield.
I will briefly sketch how we proceeded in creating a SAS Data Library containing all the relevant information. The principal elements are summarized in Figure 3. First a scientific subfield is defined, using a combination of classification codes, indexing terms, keywords, and journal titles, in abstracting/indexing services such as Chemical Abstracts or Excerpta Medica. All information on the articles (except the abstract) was downloaded to the PC and sent to the IBM Mainframe, were a SAS Data Library was created using SAS software. Next, we had to add the reference lists of these articles to the Library. For articles published in journals processed by ISI, these reference lists can be found in the SCISEARCH database (host: DIMDI). We decided to focus only on articles published in these ISI journals. The problem arises as to how a given publication can be found as quick as possible (if present) in the SCISEARCH database. In order to solve this problem, we re-formatted the names of publishing authors into the format used in SCISEARCH, by means of relatively simple SAS programs. It appeared that selection of articles based on year of publication, author names, and words from the title provided very good results. By computer programming a file of host commands was generated containing the appropriate information. The commands were uploaded to SCISEARCH. As a result, a set of articles was created for which we found a match to our original publication file. For each publication, we downloaded information on authors, source, corporate source, and the complete reference lists. The data were reformatted and a second Data Library was created. Finally, the two libraries were merged into one large SAS Data Library, using the same matchkey as the one described in section 4.2.

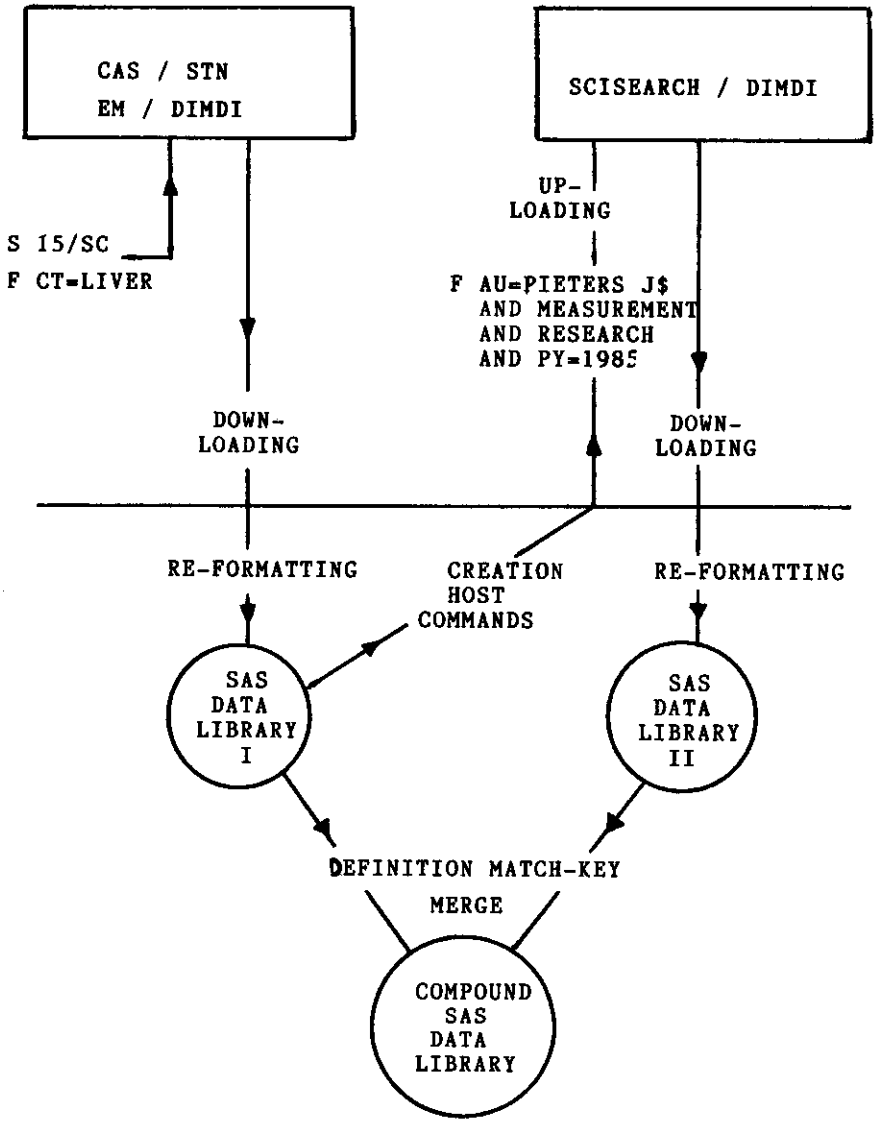

Fig. 3: Combining two databases: general scheme

The compound SAS Data Library offers the opportunity of performing bibliometric analyses in which data from various literature database are combined. As an example, I present a bibliometric matrix, in which results from a co-citation analysis are compared to biosystematic codes from the BIOSIS-database. The matrix is shown in Table 4. The analysis related to approximately 1000 articles from the field of Chemoreception research. First, the articles were clustered by means of a single linkage co-citation clustering program, based on the reference lists in these articles. For details on the co-citation clustering procedure, I refer to the contribution by R.R. Braam, H.F. Moed and A.F.J. van Raan (1988). The columns in the matrix refer to the various co-citation clusters. Next, for each cluster we counted the frequency of occurence of the biosystematic codes, assigned to the articles by BIOSIS. As a first approximation, we considered only codes that occurred at least three times in a cluster; in addition, we skipped articles that appeared in more than one co-citation cluster. Thus, a matrix was generated in which the frequency of occurrence of each biosystematic code in the various clusters can be analysed. The numbers in the cells indicate the percentage of articles in a cluster having a specific code, relative to the total number of articles in that cluster.

Table 4: Comparison of co-citation clusters and Biosystematic codes (BIOSIS)

Biosystematic code	co-citation cluster numbers								
	1	2	3	4	5	6	7	8	9
Muridae (mouses, rats)	55		82	50	100	73			
Diptera (flies)		20							
Lepidoptera (butterflies)		66					33		
Cricetidae (hamsters)			8						
Coleoptera (beetles)							67		
Osteichthyes (bonefish)								86	
Orthoptera (locusts)									50
Malacostraca									28

Table 4 shows that the biosystematic codes are not distributed uniformly amongst the various co-citation clusters. To be more specific, for each cluster a biosystematic code exists that account for 50% or more of all articles. For four clusters a code exist that is assigned to more than 80% of the articles. These results suggest that the use of biosystematic codes or other types of information such as indexing terms, might be useful in defining the scientific content of a co-citation cluster. We have planned further research on this subject.

5. SUMMARY AND CONCLUSIONS

In this contribution some limitations and potentialities of online scientific literature databases are discussed mainly from the point of view of a bibliometrician, collecting data on quantitative aspects of science. The discussion is based on the experiences made by researchers of the Science Studies Unit of the LISBON-Institute (University of Leiden) during the past two years. Four main limitations were encountered. Relevant data fields in a database may not be unified. Not all relevant bibliographic information is stored in separate, searchable fields. Host software may not be sufficiently powerful. Finally, in many cases significant outcomes can be produced only if the data from a literature database are combined with other types of information. Our general approach to deal with these limitations, is to download the relevant data and to create our own 'dedicated' database, using a software package that does not only provide tools for information storage and retrieval, but also for data-modification and programming, file handling, and advanced statistical analysis. This enables one to do the unification 'offline', based on detailed information on all variations that occur. Master-files can be created that contain both variations and unified names. All relevant data fields can be made searchable. Many data manipulations can be carried out rather easily. Other types of information can be added to the database. In particular, data from several literature databases can be combined. 'Match-keys' can be defined that contain the minimum amount of information that is necessary to identify a particular publication, and that may be assumed to appear in all correct bibliographic specifications of that publication.
I hope to have shown that this approach is able to generate significant bibliometrical results, that are worthwhile being examined to a greater detail.

ACKNOWLEDGEMENT

Part of the research described in this contribution was financed by Elsevier Science publishers, as a main element of a joint research project. I acknowledge also the Netherlands organisation for the Advancement of Scientific Research (ZWO), the Advisory Council for Science Policy in the Netherlands (RAWB) and the Ministry of Education and Sciences for financial support.

REFERENCES

Braam, R.R., Moed, H.F. and van Raan, A.F.J., Mapping of science: critical elaboration and new approaches, a case study in agricultural biochemistry. These proceedings (1988).