Jointly published by Akadémiai Kiadó, Budapest
and Kluwer Academic Publishers, Dordrecht

Scientometrics,
Vol. 58, No. 3 (2003) 587–608

Informetric studies using databases: Opportunities and challenges

WILLIAM W. HOOD, CONCEPCIÓN S. WILSON

School of Information Systems, Technology and Management,
The University of New South Wales, Sydney (Australia)

Since their arrival in the 1960s, electronic databases have been an invaluable tool for informetricians. Databases and their delivery mechanism have provided both the source of raw data, as well as the analytical tools for many informetric studies. In particular, the citation databases produced by the Institute for Scientific Information have been the key source of data for a whole range of citation-based research. However, there are also many problems and challenges associated with the use of online databases. Most of the problems arise because databases are designed primarily for information retrieval purposes, and informetric studies represent only a secondary use of the systems. The sorts of problems encountered by informetricians include: errors or inconsistency in the data itself; problems with the coverage, overlap and changeability of the databases; as well as problems and limitations in the tools provided by the database hosts such as DIALOG. For some informetric studies, the only viable solution to these problems is to download the data and perform offline correction and data analysis.

Introduction

The creation of large electronic databases of bibliographic information has been a very significant step in the development of the discipline of informetrics. A whole range of informetric studies have been made feasible by access to electronic information that would have been quite infeasible prior to the electronic database era. WHITE & MCCAIN (1989) regard these databases as censuses of publications, which provide the basic data for informetric studies in the same way as population censuses provide the raw data for demographic studies. More recently, WILSON (1999) provides analyses of a range of informetric research, most of which are also based on data from online or other electronic databases. The large-scale development and use of these databases has occurred over the last 35 years or so. As useful and even essential as these databases are, many problems arise when using databases for informetric purposes. These will be addressed in this article.

Received July 15, 2003

Address for correspondence:
WILLIAM W. HOOD
School of Information Systems, Technology and Management,
The University of New South Wales, Sydney 2052, Australia
E-mail: W.Hood@unsw.edu.au

0138–9130/2003/US $ 20.00

Growth and history of databases

Much information has been transferred into electronic format and these databases can now be accessed via online database hosts, computer tapes, CD-ROMs or via web-based front ends. One of the largest of the online database hosts, DIALOG Information System, now has over 450 databases available that cover almost every imaginable discipline (DIALOG BLUESHEETS, 2002). The type of information contained in these databases consists of, *inter alia*, newspaper articles, surrogates of journal articles, full text of journal articles, statistical information, etc.

WILLIAMS (2002), in an introductory article to the *Gale directory of databases* provides an informative and useful introduction to the state of databases. A number of tables and graphs are provided tracing the growth and development of electronic databases. The number of databases in 2001 is given as over 12,900* (with the number of records in these databases at over 16,800 million. BURTON (1988, p. 43) summarises the progress made in electronic databases. This includes "extensive increases in data base coverage, rapid development of new data bases, and release of a wide variety of 'user friendly' tools to improve and facilitate access to existing services."

Classification of databases

WILLIAMS (2002) also provides a classification of the databases using a variety of different classification methods including the form of data representation, region or country of origin, subject category and media for distribution and access. HIBBS et al. (1984) classifies databases according to whether they are aimed at the general public, the academic community or business. LANCASTER & LEE (1985) provide a classification based on whether the database is pure science, applied science, popular press or congressional testimony as a means of tracking a topic through these four stages of development.

The contribution of databases to informetric studies

Databases have contributed to informetric studies in two distinct ways:

* The current figure as reported in File 230 of Dialog, Gale Directory of Online, Portable, and Internet Databases is as follows: "Gale Directory of Online, Portable, and Internet Databases covers more than 15,300 databases and database products of all types in all subject areas produced worldwide in English and other languages by more than 3,600 database producers. These databases are offered by some 2,000 on line services and database vendors and distributors." (*Gale Directory of Online, Portable, and Internet Databases,* 2002).

- Databases provide the data sources for informetric studies.
- The delivery mechanism or platform of the databases provide the analytical tools for informetric studies.

We will examine each of these in turn.

Databases as data sources

The first (and fairly obvious) major contribution that databases make to informetric studies is that they provide a source of data for research projects. More basic, PERSSON (1988) views searching of databases as a 'paper counting' activity and is in one sense an informetric study.

Each different database has a different set of fields, many of which are useful for different types of informetric analysis. STEFANIAK (1987) has a useful classification and listing of these fields. The following is an adaptation and modification of Stefaniak's list with some additions from DEOGAN (1987):

- Subject oriented fields (eg. classification codes, descriptors, identifiers, keywords, words in the title, words in the abstract, words in the full text).
- Type of publication (eg. journal paper, conference paper, book, patent, report, etc.).
- Source (eg. journal title, CODEN, ISSN number, ISBN number, patent number, year of publication, volume, number of issue, pages, name of publisher, place of publication).
- Responsibility (eg. name of authors, editors, translators).
- Geographical and institutional information (eg. country of its editor, name and corporate affiliation of the authors - name of organisation, city, country).
- Language(s) of publication.
- Secondary source (eg. year, volume and number of the abstract).
- Citations or references (eg. in the three ISI citation databases).

Further, MCGRATH (1996) discusses the 'objects of study' in informetrics, accompanied by concerns of a perceived inattention to basic units of analysis by informetricians. WILSON (1999, p. 117) enumerates various classifications of informetric research: "… by the types of data studied (eg., citations, authors, indexing terms), by the methods of analysis used on the data obtained (eg., frequency statistics, cluster analysis, multidimensional scaling), or by the types of goals sought and outcomes achieved (eg., performance measures, structure and mapping, descriptive statistics)". According to Wilson (1999), the basic unit of analysis in informetric studies

is a collection of publications (or more commonly, their surrogates). Each publication is a repository of properties or bibliographic fields with variable values, such as language, publication year, containing-journal, authors, and title. Each of these fields in turn also has properties, such as the language's number of printed works, the journal's editor, the author's institutional affiliation, and the institution's address.

Analytical tools for informetric studies

In addition to providing a source of data, many informetric studies using electronic databases utilise facilities, procedures and functions of the database host or software accompanying the database to perform statistical analyses. The tools available to the informetrician have increased and improved markedly over the last few years. INGWERSEN & CHRISTENSEN (1997, p. 206) strongly advocate, for example, the use of online data and list the following five advantages of online analysis:

1. it is fast;
2. it is inexpensive;*
3. instant results are provided using advanced processing tools;
4. both domain dependent and citation databases can be combined in the one analysis;
5. results are reproducible.**

One of the frequently used database hosts for informetric studies is DIALOG (DIALOG, 2002b). Some studies that use DIALOG as a data source include HAWKINS (1977), TENOPIR (1982), BYLER & RAVENHALL (1988), EGGHE (1988), WALKER (1990), WOLFRAM et al. (1990), REID (1992), and HOOD & WILSON (2001). Other database hosts are available; for example, PERSSON (1988) used the ESA/IRS (European Space Agency/-Information Retrieval Service) to illustrate similar advanced functionalities to that found in DIALOG for measuring scientific output through a

* This may or may not be the case since institutions have varying means of accessing online databases and varying costs depending on agreements negotiated with database producers. Additionally, processing time (and cost) will vary depending on the extent of experimentation undertaken by researchers before obtaining the desired results.

** The reproducibility is not absolutely guaranteed. Databases are constantly changing both by records being added and the structure and facilities of the database being altered. Databases are reloaded and reindexed, and databases are also removed from hosts from time to time. Despite techniques such as fixing the set through using either the accession number field or the update field (CHRISTENSEN & INGWERSEN, 1996), the results obtained by one researcher might not be replicable by any other researcher at a later date. CD-ROM databases, however, are fixed and if using the same version of the CD-ROM, should produce the same results.

variety of 'paper counting' techniques. INGWERSEN & CHRISTENSEN (1997, p. 207) list the features that are desirable in an online host:

- host has files relevant to the domain of interest;
- search results can be distributed across these files;
- removal of duplicate documents is possible;
- frequency analysis tools are available.

These features, which are available on the DIALOG system, are outlined below.

The OneSearch system allows a searcher simultaneously to search a number of databases (DIALOG, 2002d). This is a very useful tool in a multi-database environment. Selection of databases to search can either be by a list of individual databases, or by using the OneSearch/DIALINDEX (see below) categories, eg. INFOSCI for the Information Science subset of databases. The main problem with OneSearch is the lack of consistency between different databases (in field structure, indexing, output formats, etc.).[*] Also, there is a limit of 60 databases that can be searched at one time using OneSearch.

DIALINDEX is a facility whereby a search statement can be run against any subset of DIALOG databases or against the total set of databases available on DIALOG. This tool is useful for ranking of the databases for specific topics, as well as producing counts of records per database against year of publication. Printouts of the bibliographic records are not available. The major limitation of this type of analysis is that the topic of interest needs to be encapsulated in a single search statement of no more than about 240 characters so will typically be a brief statement of a topic consisting of keywords or phrases with appropriate connecting boolean operators; DIALINDEX is incapable of creating sets, and so cannot perform more complicated searches (DIALOG, 2002a). LANCASTER & LEE (1985) discuss the use of DIALINDEX as a source of data for research purposes, and in particular, for tracking the growth and movement of a topic through various databases – in their case, the topic of 'acid rain'. More recently, JACSÓ (1999) provides a brief review of the advantages and limitations of DIALINDEX as a database selection tool.

There are three commands on DIALOG (2002c) for handling duplicate records: Remove Duplicates (RD), Identify Duplicates (ID) and Identify Duplicates Only (IDO). It is not clear what algorithm DIALOG uses to match the duplicates, although MILLER (1990) provides some speculations. Duplicate detection works best between or among

[*] Some examples of the use of OneSearch in informetric studies include: VAN CAMP (1991), OJALA (1992), DE STRICKER et al. (1997), HOOD & WILSON (2001). Some references on multi-file searching in general include WANGER (1977), HAWKINS (1978), EPSTEIN & ANGIER (1980), and EVANS (1980).

databases that have similar record structures, eg. the three citation databases, *Science Citation Index* (SCI), *Social Science Citation Index* (SSCI) and *Arts & Humanities Citation Index* (AHCI). There is also a limit of 5000 records in a set that can be tested for duplicates.[*]

The RANK command is a very useful command for performing trend or statistical analysis on a set of records (DIALOG, 2002e). It was introduced by DIALOG for public use in 1993.[**] This command can be used for example to determine the top journals in a particular search result. Terms can be extracted from most phrase-indexed fields and then ranked in decreasing frequency order. The publication year field can be ranked in chronological order. Multiple fields containing similar data (eg. Descriptor (DE) and Identifier (ID) fields) can be grouped and ranked. Partial fields (eg. the first four characters of the International Patent Classification Code) can be ranked as well. There are a number of studies using the RANK command and a few recent publications are listed to provide the range to which this technique is used. WHITE (2001) used descriptors (DEs) from two databases, INSPEC and ERIC, to model the curricula for Drexel University's graduate programs in information systems and library and information science. WHITE (1996) demonstrates the use of the RANK command as a means of revealing interdisciplinarity in any field. OSAREH & WILSON (2002) used the *Science Citation Index* to study international collaboration among Iranian scientists. After the initial data set for Iranian scientific publications over some years was obtained, the RANK command was used on the geographical location (GL) field to identify international or cross-country collaboration. More generally the authors RANKed a number of fields (including the GL field) to obtain a comprehensive picture of science and research in Iran from 1975 to 2002 (WILSON & OSAREH, 2003).

Citation based informetric studies

Since GARFIELD's (1955) seminal paper, "Citation indexes for science", numerous studies have been done based on the ISI citation databases' unique cited reference (CR) field and its three subfields: cited author (CA), cited year (CY) and cited work (CW). WILSON's (1999, pp.125-150) review of Informetrics devotes a substantial section on

[*] If a search has more than 5000 records, this can sometimes be broken down by publication year, duplicates removed and then combined back into a full set.

[**] Other online services have a similar command to DIALOG's RANK command and some were introduced at a much earlier date. The GET command was introduced to Pergamon's Online ORBIT system in 1988, and the ZOOM feature was available on the European ESA-IRS system from at least 1982. (WHITE & MCCAIN, 1989).

citation analysis with examples of citation-based informetric studies. The festschrift in honor of Garfield edited by CRONIN & ATKINS (2000) includes 26 chapters, eight of which deal with various aspects of evaluative bibliometrics and more specifically, citation analysis. MARX et al. (2001) discuss similar DIALOG features for searching the *Science Citation Index* using the STN International host in Germany. A few years ago, the STN retrieval language MESSENGER added functions for carrying out statistical investigations; these functions provide opportunities for measuring the impact of scientific activities through the scientific literature.

The citation databases have been made available in at least five different formats: online, CD-ROM, computer tape, web-based access and paper. Each of these delivery mechanisms, despite being based on essentially the same data, provide quite different tools for the informetrician, and the types of studies that are feasible using one format, are not possible using another. Thus, anyone using this data for informetric studies has to consider not only the data that is required, but the tools available to manipulate the data to produce the required analyses.

Difficulties with electronic information for informetric studies

The use of databases is beset with many problems that may need to be overcome or at least managed, in order to extract useful information for informetric analyses. There are many references in the literature to the types of difficulties and problems that are encountered using these databases; these are categorised and listed below. The first group of problems relate to errors or lack of consistency in the data (at the micro level); the second group relate to other types of problems and difficulties in utilising databases for informetric purposes (at the macro level), and the third relate to problems with the tools that are available in the various platforms and delivery mechanisms. Some errors in the databases result from errors in the primary literature itself. These may be author-induced, for example, erroneous reference, etc. or production caused, for example, typographical misprints, etc. (COILE, 1977).

Many of these problems arise from the fact that most databases are created for information retrieval purposes; informetric studies are a secondary use of these databases. In many cases, the policies and procedures of the database producers and hosts are aimed at the primary purpose of information retrieval and are not ideal for informetricians.

Errors or lack of consistency of electronic data (micro level)

Spelling differences and errors: BOURNE (1977) examines the occurrence and frequency of spelling errors in 11 different bibliographic databases and determines the impact that these errors will have. Bourne found that index terms were misspelled in as many as 23% of the terms in one database, and in as few as 0.5% of the terms in another. FEDOROWICZ (1982) reported that over 30% of the sample terms examined were misspelled. Also reported was a study of the MEDLINE database by the National Library of Medicine which pointed to 80% of all terms occurring with a frequency of one or two, were misspellings. Misspellings have a very large impact when counting low frequency terms, as misspellings are usually unique. HOOD & WILSON (1992; 1994) also noted numerous spelling problems in their study.

Related to spelling errors are other differences including:

- abbreviation standards,
- differences in US versus UK spelling, and
- transliteration differences (BRAUN et al., 1995).

Subject indexing consistency: Most databases have indexing terms or keywords that have been assigned to records by indexers or authors. These may vary in quality and consistency. WHITE & GRIFFITH (1987) examined the quality of indexing in a number of medical databases. Different databases use different indexing vocabularies and also, indexing vocabularies change over time (INGWERSEN & CHRISTENSEN, 1997). HOOD & WILSON (1992; 1994) found major problems with the consistency of indexing in the LISA database. More recently, SAARTI (2001) looked at the consistency of subject indexing of novels using a Finnish fictional thesaurus by two groups: library professional and library patrons. He found the indexing very 'inconsistent' between the different group of indexers.

Names: There are many problems with the representation of author names in bibliographic databases. Some of these problems include: the abbreviated form "Surname, Initials" may represent two different authors; authors may change their preferred form of their names (e.g., D. J. Price later preferred Derek de Solla Price); citing authors may spell out full names or use various combinations of initials; and journals have different policies with regard to representation of author names (COILE, 1977).

PAO (1989) found many errors in the author field; these were categorised into nine headings: additions, omissions, transpositions, misspellings, spacings, punctuations, capitalisations, compound names, or combinations of the above. Examples are given for

each type of error. In Pao's study, errors were manually detected and corrected based on evidence to support a correction. The three steps involved included:

- spotting possible errors,
- examining the original bibliographic entry to find evidence to support a correction,
- making the actual correction on the data file.

If in doubt, no correction was made.

Journal titles: Journal names are also a great source of difficulty. Over time, a journal may change its name, split into two journals, two journals may merge, and the publisher or country of origin may change. The frequency of issue may change, special issues may be produced or an issue may not be produced. The cover date may differ from the actual date of production. Two journals may have the same or very similar names – e.g., *Journal of Education* published in Boston MA and *Journal of Education* published in London England; or *Library Science and Documentation* published in New York and *Library Science, with a Slant to Documentation* published in Bangalore, India – (COILE, 1977).

WILLIAMS & LANNOM (1981) reported a lack of standardisation of the journal title data element within and across databases. Four measures were developed to show the extent or lack of journal standardisation: number of different forms of journal name in a database, percent retrievable by each different form, number of non-contiguous entry points in a sorted list, and percent retrievable by best form of an element. In a response to this article, PITERNICK (1982) suggests the increased use of CODEN or ISSN as a possible solution to many of the problems caused by this lack of journal name standardisation.[*] DEOGAN (1987), PAO (1989) and STEFANIAK (1987) also note problems in standardisation of journal titles between and within databases. Problems also arise in trying to determine how to deal with journals that are translated into other languages. Should these be counted as two distinct journals or as one?

Dates: PAO (1989) found a few problems with dates. Inconsistencies were found such as journal issues listed under both "1983–1984" and "1983–84" as well as months not always being included nor standardised when used. These errors did not cause too many problems. WOLFRAM et al. (1990) also noted the problem of publication dates spanning across two years.

Corporate sources: Stefaniak (1987) notes problems with corporate sources arising from optional translation of foreign names and inconsistency in producing abbreviations

[*] Also suggested by INGWERSEN & CHRISTENSEN (1997).

for institutional names. Institutions may also change their name, merge, or split, making trend analysis difficult. Institutional unification is a significant problem for many informetric studies and many national studies report the need to 'clean-up' databases in order to reflect institutional counts more accurately (BOURKE & BUTLER, 1996b). In addition, the practice (intentional or otherwise) of omitting institutional affiliations is not trivial, especially in publications from databases for the 1980s and early 1990s. The case for Russia and the former USSR is a classic example (GARFIELD, 1990; WILSON & MARKUSOVA, in prep.) as well as for other developing countries such as Iran (WILSON & OSAREH, 2003).

Field structure and field delineation: In DIALOG, it is possible to output data from most databases in a tagged format to allow for easier post-processing or up-loading to an information retrieval system. However, depending upon the particular host, many databases have one field to represent a variety of pieces of information (MOED, 1988). The Source (SO) field in LISA on DIALOG, is a good example.[*] This field contains the volume, the issue number, the date of publication, the page numbers as well as notes. Due to inconsistencies in the formatting of this field, it is not easy to automatically parse this field into its constituent parts. MOED (1988, p. 135) notes that: "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." Inaccessibility of data is also noted in INGWERSEN & CHRISTENSEN (1997).

Other problems with information in databases (macro level)

Databases vary enormously in a variety of ways that affect their usefulness for informetric studies. According to BRAUN et al. (1995), variations can be either extrinsic (such as language, availability, price) or intrinsic (such as database content). JACSÓ (1997) evaluates the quality of the content of databases, vis-à-vis: scope, composition, source coverage, journal coverage, geographic coverage, language coverage, time period coverage, currency, accuracy, consistency and completeness.

Overlap: Errors in databases have the effect of masking duplicate records and thus reducing the calculated overlap statistic. These errors must be corrected as far as possible to obtain accurate calculations for overlap studies (PROVOST & NIEUWENHUYSEN, 1992). An early review of overlap in databases is given by STERN

[*] DIALOG File 61 – discontinued in 2002.

(1977); more recently, HOOD & WILSON (2003) provide a comprehensive study of overlap in databases on the subject of fuzzy sets.

Coverage: Coverage in a database of a particular discipline may be limited in many different ways (STEFANIAK, 1987). A particular database may be limited in the languages it covers or the geographical region of the publications. Even databases covering world literature are often biased towards the country or region of origin, which is often the US. Databases may also be limited in the types of literature covered; some may cover conference proceedings or monographs whereas others may not. STEFANIAK (1987) also lists the particular limitations of the *Science Citation Index*, which is often used for international comparisons.* INGWERSEN & CHRISTENSEN (1997) indicate the need to ensure that any dataset is unbiased, particularly when using the data to compare, for example, institutions; the citation databases may need to be complemented with data from discipline specific databases.

Time span: The time span of a particular database may be limited, particularly for retrospective searches (STEFANIAK, 1987; PERSSON, 1986). Most databases have not undertaken retrospective conversion of the pre-computer era data, thus including only records from the time of computerisation of the database.

Time delay in abstracting: If studying the state-of-the-art or recent developments in a field, the delay in processing of publications into a database may be significant (STEFANIAK, 1987). For example, ERNEST et al. (1988) found the main indexing lag for *Library Literature* to be on in the order of 4.3 months, for ERIC 7.6 months, and for LISA 10.3 months. Of course, this is in addition to the delay caused by the whole publishing process.

Missing data fields: STEFANIAK (1987) also points to the lack of certain important fields in some databases. These include the ISSN or CODEN field for journals or the corporate source field for authors. These omissions can either make some types of analysis very difficult or impossible. INGWERSEN & CHRISTENSEN (1997) note the obvious need that if doing an analysis by country, institution or journal that these data fields must actually exist in the database and be searchable. Also, problems may arise when, for example, the corporate source is only given for the first author or when there are fewer corporate sources than authors given. In the latter case, it may be difficult to match the correct corporate source to each of the authors.

* More detail about the limitations of SCI can be found in STEFANIAK (1987). See also CARPENTER & NARIN (1981), LANCASTER et al. (1984), SANDISON (1989), SMITH (1981), LUUKKONEN (1989), MOED (1989), SEGLEN (1989), and THORNE (1977).

Change in database policy or practice: Database producers may (and often do) change their policies and practices. WOLFRAM et al. (1990) note that this is a particular problem when doing any time-based analysis. "Sharp growth rate could be the result of increased coverage of the indexing service and not caused by an increase in the actual literature. Likewise, low growth rates may be an indication of poor coverage by an indexing service, and not a reflection of the state of the discipline." PERSSON (1986) notes the changing journal coverage of databases over time and JACSÓ (1997) provides a review of the length, width and depth of journal coverage in databases.

INGWERSEN & CHRISTENSEN (1997) note the sudden appearance of a full-text field in some databases; this can greatly distort the retrieval of records compared with those retrieved only via the other indexed fields.

BURTON (1988, p. 42) indicates that even in a single database, formats may change over time or a vendor's treatment of the data may vary even within a single database. "One example of a simple change that has major ramifications is the inclusion after 1975 of an explicit date of publication field in most data bases. Time series analysis on pre-1975 is not feasible until a special program can be written to scan the data and create a date field."

Isolating data types: Many analyses are predominantly interested in the primary literature represented by journal articles. Some databases allow this restriction to be carried out easily using the Document Type (or DT) field. However, different databases use this field differently and many databases don't have an equivalent field at all. This problem is noted by INGWERSEN & CHRISTENSEN (1997, p. 207-208, 214). Studies limited to journal articles include BRAUN et al. (1987), WOLFRAM et al. (1990), LANCASTER (1991), BRAUN et al. (1995) and HOOD (1998).

Type of inversion: Virtually all information retrieval systems have inverted files to allow searching on various fields. DIALOG has both word indexed and phrase indexed fields. Each field may have either, both or neither of these types of inversion. The type of inversion will affect the type of processing that is possible with that field (eg. use of the RANK command or searchability), as well as the retrievability of the records.

Field names: When doing searches on multiple databases using for example the OneSearch facility of DIALOG, the field names used in different databases may cause problems. The same field label may be used for different fields in different databases; alternatively, the same field may be given different names in different databases (INGWERSEN & CHRISTENSEN, 1997). Even the format for output varies across different databases, making it difficult to download records in a consistent format from a multi-file search. The same field label may be used for two different types of data. For

example, the AU field in INSPEC is used for both authors of papers and editors of conference proceedings (INGWERSEN & CHRISTENSEN, 1997).

Relevant information: In most cases, informetric studies of databases are not interested in the whole database; HOOD & WILSON (1992; 1994) are atypical in their analysis of the entire LISA database on CD-ROM. Relevant records must be selected using subject content criteria, or other criteria such as excluding meeting abstracts (BRAUN et al., 1995).

Data standardisation: Most studies involve some standardisation of the data before results can be calculated. BRAUN et al. (1995, p. 134) use an Information Science and Scientometrics Research Unit (ISSRU) standard, "which extends to a unified correction of name components such as van, de, di, d', del', etc., to the handling of German umlaut letters and to irregularities in volume and first page numbers." BOURKE & BUTLER (1996b) unified Australian institutional addresses from ISI tapes prior to any bibliometric analyses.

Difficulties with the database hosts for informetric studies

As well as the benefits and tools that the various hosts provide, which assist in informetric studies, there are also problems and pitfalls which need to be overcome and addressed. We will first discuss some of the available hosts and their characteristics.

Hosts for electronic databases

Online databases: Online databases have provided the data source for numerous informetric studies. A large number of databases are available from one of the online database hosts (such as DIALOG), and as discussed earlier, these hosts provide a considerable range of analytical tools for those doing informetric studies.

CD-ROM databases: As an alternative to online databases, CD-ROM databases sometimes provide access to large amounts of data in a cost effective manner. HOOD & WILSON (1992; 1994) used CD-ROM data effectively to examine the indexing terms used in the LISA database. However, INGWERSEN & CHRISTENSEN (1997) complain of a lack of robust processing facilities on current CD-ROM systems, as well as the difficulty of doing any multi-database analysis.

Web-based user-friendly front ends: There has been a proliferation of web-based user-friendly front ends to various online databases. Generally, these are designed for information retrieval purposes and are not very suitable for informetric studies. For example, the user-friendly version of the Citation indexes, *Web of Science* (WoS), is not very suitable as an informetric tool (ISI, 2002). In part this is due to the lack of some of the analytical (or manipulative) tools for informetric studies, for example, the facility to Rank various fields such as the author, journal or subject terms. However, we are beginning to see informetric studies using the WoS especially for citation analysis. TORRICELLA-MORALES et al. (2000) conducted a citation analysis of Cuban research publications cited in the WoS to describe various citation patterns. ROY et al. (2002) assessed the impact of scientific journals using citation data for otorhinolaryngology journals from the CD-ROM edition of SCI and from the WoS. A major advantage for using the WoS for citation analysis is the facility to obtain the number of times a researcher is cited regardless of his/her 'position' as first (or subsequent) author of research publications (WILSON & OSAREH, 2003). This is not the case with traditional hosts for electronic databases; for example, in DIALOG one can only search by the first author of a publication in the cited reference field or more specifically in the cited author field of the cited reference. DHYANI et al. (2002, p. 469) provide a survey of Web metrics and try to the answer the question: "Is Web informetrics any different, or is it just an application of classical informetrics to a new medium?"

Computer tapes: Another alternative to online studies, particularly citation-based studies using databases produced by ISI is to purchase the data on tape or disk. Local processing can then be performed on the data (INGWERSEN & CHRISTENSEN, 1997). An Australian example of using ISI tapes is provided in the work of BOURKE & BUTLER (1996a, 1996b). This has been the preferred delivery mechanism for many of the scientific indicators producers for many years.

Problems with the available electronic tools.

Some of the tools available on the database hosts have problems when used for informetric work. We highlight a few of these problems below.

Duplicate detection and removal: INGWERSEN & CHRISTENSEN (1997) note that the duplicate removal algorithm used by DIALOG is not completely safe. Records can either be wrongly identified as duplicates or duplicate records not identified.

MILLER (1990) indicates that the DIALOG algorithm checks for first author and title but not source or publication year. Errors in author or title fields can result in duplicates

being missed; articles with the same author and title, such as a conference paper later appearing as a journal article are wrongly tagged as duplicate.

In general, the large number of errors in the bibliographic databases will cause errors in the duplicate detection algorithm. The Identify Duplicates (ID) command can be used to check that all proposed duplicates are real duplicates; missed duplicates are much harder (or infeasible) to detect online. The success rate of DIALOG's command is claimed by DIALOG to be in excess of 90%, a claim supported by MILLER (1990).

Duplicate removal in DIALOG can be controlled by ordering the databases, and forcing DIALOG to remove a duplicate record from a particular database first. This technique can be used to advantage in the process called Reversed Duplicate Removal (RDR), where a duplicate record is removed from a less desirable database. INGWERSEN & CHRISTENSEN (1997) have a discussion of RDR as well as caveats in using DIALOG duplicate commands.

Ranking: There is a maximum of 50,000 terms that the RANK command on DIALOG can handle. HUDNUT (1993) lists some of the main weaknesses and limitations of the RANK command as implemented in 1993:

- Lack of standardisation of data across and within files means much merging of duplicate entries is required; this must be done manually.
- Some fields such as corporate source are only word-indexed, so don't provide quality of information possible with phrase-indexed data.
- Multilevel ranking not possible, eg. it is not possible to rank by author then by year.

In addition, we have also noticed the duplication of journal names in some databases: first by the full journal name and then repeated as the journal abbreviation. Unless the two versions of the journal name are merged in each instance, accurate counts for journals are not possible.

SNOW (1993) also points out that Ranking the DE field in medical databases produces many screens of general headings (e.g. human) before reaching the perhaps more interesting subject headings. Further, we have also noticed the double counting in the MEDLINE database for subheadings: the full version as well as the mnemonic two-character version (e.g. Adverse Effects and AE).

Offline informetric studies

As useful as the online and CD-ROM tools are for informetric studies, there are significant limitations as indicated above. For many informetric studies, the only way to

overcome these is to obtain the data on tape or to download the data for post-processing and offline analysis.

MOED (1988) argues that for several specific informetric applications, the software available on the host computers is not adequate. In these cases, the data should be downloaded into a local machine, and then software developed to perform the analysis properly. Moed was particularly interested in citation analysis, but the comments apply more broadly to many other parts of informetrics. It is certainly true for many informetric studies that the online systems do not provide the sorts of tools that are necessary.*

MIYAMOTO et al. (1989) and MIDORIKAWA et al. (1990) call for more methods and software tools for document analysis to be developed, due to the increasing amount of material available for analysis. INGWERSEN & CHRISTENSEN (1997) note that offline processing provides more flexibility for analysis compared with the online facilities. This has to be balanced against the cost and effort required downloading the records and possibly writing the software to carry out the processing. Tools such as The Bibliometrics Toolbox (BROOKS, 1998) provide a number of techniques for the analysis of downloaded informetric data. HOOD (1998), HOOD & WILSON (1999, 2002) out of necessity used offline processing of the data.

Difficulties of fixing problems: PAO (1989) reports that for small data sets of up to several hundred records, it is relatively easy to find and correct data errors. However, for larger data sets, it is much more difficult. Algorithmic methods may not be available and manual methods are time-consuming and costly. Many projects don't anticipate the time, effort and cost needed to clean up the data. In some studies, the errors in the data are not sufficiently large to significantly alter the overall results of the research. The cost of cleaning up the data may not be warranted compared with the small increase in accuracy obtained from the clean data.

However, as noted by INGWERSEN (1998), clean data may produce quite different results from online dirty data. As discussed in HOOD & WILSON (1999), Ingwersen speculated that the more 'flawed' the data set used in this study, the more its distribution over databases is similar to a Bradford-type hyperbolic distribution. Conversely, the more 'clean' the data set, the less 'Bradford-like' are the distribution curves (INGWERSEN, 1998).

* Of course, the criticisms made in these references are time dependent and relate to the facilities of the online systems at the time. More facilities and commands are being added to the online hosts, which may overcome some of the limitations as reported in the literature.

Using online data for offline studies: BURTON (1988) outlines the basic steps necessary to extract data from online databases into a form suitable for informetric analysis:

- refine the search strategy until satisfactory retrieval results are obtained;
- download the citation in the fullest and most explicitly tagged available format (e.g. format 4 in DIALOG, or by using the keyword tag);
- repeat steps above for each relevant database, creating discrete citation files for each database;
- add missing fields as necessary for analysis (eg. year of publication, country of origin, language);
- translate records to common format;
- identify duplicate citations. Duplicates can occur within the same databases as well as across databases (see for example, HOOD & WILSON, 1999; 2003);
- eliminate less-preferred form(s) of duplicates.

Conclusions

The difficulties outlined above need to be addressed in one way or another, either by the producers of the databases, or by those using the data for informetric purposes. If not addressed, then problems will most certainly occur. For example, PAO (1989) shows that the sorts or errors outlined above, if not corrected, can affect the fitting of the data to informetric models (Lotka's law in this case). The extent of these problems will depend on the type of study being performed. For some informetric studies, particularly those analysing the data at the macro level, many of the errors and limitations listed above will not be significant. But for others, such as performance measurement, the need to standardise corporate names for example will be crucial.

A number of authors in examining some of the problems outlined above call for more standardisation by the database producers or vendors. These include, *inter alia*, PAO (1989), STEFANIAK (1987), NORTON (1981), PITERNICK (1982), HAAS & CLARK (1992). COILE (1977) calls for errors to be corrected in the electronic databases by semi-automatic means. HAWKINS (1977, p. 17) notes: "Standardisation … is still very much needed. It will greatly improve the quality of bibliometric searches. It requires a considerable effort on the part of the database producer, as well as care by authors and editors in adopting a uniform style. Whether this standardisation will be achieved is still an open question." Some years later, HAWKINS (1981, p. 256) notes that "more conformity to common standards by database producers would be enormously helpful."

These problems do not appear to have been satisfactorily addressed in the 23 years since this paper was written.

The main problem is that most electronic databases are designed as Information Retrieval tools, and not as Informetric tools. Electronic databases can and are used as data sources for informetric studies, but the data usually requires significant manipulation or cleaning up. There is no such thing as clean data in electronic databases.

As the title of our paper implies, we have tried to outline the advantages ('opportunities') and disadvantages ('challenges') of using electronic databases for informetric studies. For the novice researcher entering the literature measurement field, we have provided a guide as to what is possible, what isn't, and perhaps how one can overcome the disadvantages of electronic databases discussed above. For the seasoned researcher, it may serve as a review (and a reminder) of how far we have come in the intersection of the subdisciplines of information retrieval and informetrics, bibliometrics and scientometrics since the 1970s. More importantly, it should continue alerting retrievalists and informetricians alike to push database producers and database vendors (hosts) towards meeting our need for 'cleaner' (more accurate) data. Despite all the disadvantages, we have found certain database hosts (eg. DIALOG or similar information retrieval systems with advanced functionalities, and the web-based version of the ISI's citation indexes, *Web of Science*) becoming more and more hospitable for quantitative studies of scholarly publications.

References

BOURKE, P., BUTLER, L. (1996a), Publication types, citation rates and evaluation. *Scientometrics*, 37 : 473–494.

BOURKE, P., BUTLER, L. (1996b), Standards issues in a national bibliometric database: the Australian case. *Scientometrics*, 35 : 199–207.

BOURNE, C. P. (1977), Frequency and impact of spelling errors in bibliographic databases. *Information Processing & Management*, 13 : 1–12.

BRAUN, T., BROCKEN, M., GLÄNZEL, W., RINIA, E., SCHUBERT, A. (1995), Hyphenation of databases in building scientometric indicators: Physics briefs, SCI based indicators of 13 European countries, 1980–1989. *Scientometrics*, 33 : 131–148.

BRAUN, T., BUJDOSO, E., SCHUBERT, A. (1987), *Literature of analytical chemistry: A scientometric evaluation*. Boca Raton: CRC Press, Inc.

BROOKS, T. A. (1998), The Bibliometrics Toolbox.
[Available at ftp://ftp.u.washington.edu/public/tabrooks/toolbox]

BURTON, H. D. (1988), Use of a virtual information system for bibliometric analysis. *Information Processing & Management*, 24 : 39–44.

BYLER, A. M., RAVENHALL, M. (1988), Using Dialindex for the identification of online databases relevant to urban and regional-planning. *Online Review*, 12 : 119–133.

CARPENTER, M. P., NARIN, F. (1981), The adequacy of the Science Citation Index (SCI) as an indicator of international scientific activity. *Journal of the American Society for Information Science*, 32 : 430–439.

CHRISTENSEN, F. H., INGWERSEN, P. (1996), Online citation analysis: a methodological approach. *Scientometrics*, 37 : 39–62.

COILE, R. C. (1977), *Error Detection in Computerized Information Retrieval Data Bases.* Arlington, VA: Center for Naval Analyses.

CRONIN, B., ATKINS, H. B. (Eds) (2000), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*. Medford, NJ, Information Today.

DE STRICKER, U., SERIO, S., CASEY, V. (1997), Information resources in Canada. *Database*, 20 : 18–32.

DEOGAN, M. S. (1987), On-line bibliometrics. *Lucknow Librarian*, 19 : 43–48.

DHYANI, D., NG, W. K., BHOWMICK, S. S. (2002), A survey of Web metrics. *ACM Computing Surveys*, 34 (4) : 469–503.

DIALOG (2002a), DIALINDEX. http://library.dialog.com/pocketguide/pktgde.pdf. pp. 46–48. 8th July, 2002.

DIALOG (2002b), DIALOG Home Page. http://www.dialog.com. 8th July, 2002.

DIALOG (2002c), Duplicate Detection. http://library.dialog.com/pocketguide/pktgde.pdf. p. 30. 8th July, 2002.

DIALOG (2002d), OneSearch. http://library.dialog.com/pocketguide/pktgde.pdf. pp. 27–30. 8th July, 2002.

DIALOG (2002e), RANK Command. http://library.dialog.com/pocketguide/pktgde.pdf. pp. 17–21. 8th July, 2002.

DIALOG Bluesheets (2002), Databases in Alphabetical Order. http://library.dialog.com/bluesheets/html/blf.html. 17th July, 2002.

EGGHE, L. (1988), Concentration places, concentration evolutions, and online information retrieval techniques for calculating them. *Information Processing & Management*, 24 : 109–121.

EPSTEIN, B. A., ANGIER, J. J. (1980), Multi-database searching in the behavioral sciences. Part 1: basic techniques and core databases. *Database*, 3 : 9–15.

ERNEST, D. J., LANGE, H. R., HERRING, D. (1988), An online comparison of three library science databases. *RQ*, 28 : 185–194.

EVANS, J. E. (1980), Database selection in an academic library: are those big multi-file searches really necessary? *Online*, 4 : 35–43.

FEDOROWICZ, J. (1982), A Zipfian model of an automatic bibliographic system: an application to MEDLINE. *Journal of the American Society for Information Science*, 33 : 223–232.

Gale Directory of Online, Portable, and Internet Databases. (2002), http://library.dialog.com/bluesheets/html/bl0230.html. 28th August, 2002.

GARFIELD, E. (1955), Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159), 108–111.

GARFIELD, E. (1990), The Russians are coming! Part 1. The red-hot 100 Soviet scientists, 1973–1988. In: *Essays of an information scientist: Journalology, KeyWords Plus, and other Essays*. Vol. 13, 202–215. Also available from: *Current Contents*. (24), 5–18, June 11, 1990.

HAAS, S., CLARK, M. (1992), Research journals and databases covering the field of agrochemicals and water pollution. *Science and Technology Libraries*, 13 : 57–64.

HAWKINS, D. T. (1977), Unconventional uses of on-line information retrieval systems: on-line bibliometric studies. *Journal of the American Society for Information Science*, 28 : 13–18

HAWKINS, D. T. (1978), Multiple database searching: techniques and pitfalls. *Online*, 2 : 9–15.

HAWKINS, D. T. (1981), Machine-readable output from online searches. *Journal of the American Society for Information Science*, 32 : 253–256.

HIBBS, J. E., BOBNER, R. R., NEWMAN, I., DYE, C. M., BENZ, C. R. (1984), How to use online databases to perform trend analysis in research. *Online*, 8 : 59–64.

HOOD, W. W. (1998), *An Informetric Study of the Distribution of Bibliographic Records in Online Databases: A Case Study Using the Literature of Fuzzy Set Theory (1965–1993),* PhD dissertation. Sydney, The University of New South Wales.
http://www.library.unsw.edu.au/~thesis/adt-NUN/public/adt-NUN1999.0033.

HOOD, W. W., WILSON, C. S. (1992), *An Analysis of the Indexing Used in the LISA Database.* (Ed.), Kensington, Australia: The School of Information, Library and Archive Studies, University of New South Wales.

HOOD, W. W., WILSON, C. S. (1994), Indexing terms in the LISA database on CD-ROM. *Information Processing & Management*, 30 : 327–342.

HOOD, W. W., WILSON, C. S. (1999), The distribution of bibliographic records in databases using different counting methods for duplicate records. *Scientometrics*, 46 : 473–486.

HOOD, W. W., WILSON, C. S. (2001), The scatter of documents over databases in different subject domains: How many databases are needed. *Journal of the American Society for Information Science and Technology*, 54 : 1242–1254.

HOOD, W. W., WILSON, C. S. (2002), Analysis of the fuzzy set literature using phrases. *Scientometrics*, 54 : 103–118.

HOOD, W. W., WILSON, C. S. (2003), Overlap in bibliographic databases. *Journal of the American Society for Information Science and Technology*, (in press).

HUDNUT, S. K. (1993), Finding answers by the numbers: statistical analysis of online search results. In: M. E. WILLIAMS (Ed.), *Proceedings of the 14th National Online Meeting,* (pp. 209–219), Medford, NJ, Learned Information.

INGWERSEN, P. (1998), Personal Communication.

INGWERSEN, P., CHRISTENSEN, F. H. (1997), Data set isolation for bibliometric online analyses of research publications: fundamental methodological issues. *Journal of the American Society for Information Science*, 48 : 205–217.

ISI (2002), Web of Science, http://www.isinet.com/isi/products/citation/wos/. 28th August, 2002.

JACSÓ, P. (1997), Content evaluation of databases. In: WILLIAMS, M. E. (Ed.) *Annual Review of Information Science and Technology*, Vol. 32. (pp. 231–267), Medford, NJ, Information Today.

JACSÓ, P. (1999), Database section tools. *Online & CD-ROM Review*. 23 : 227–229.

LANCASTER, F. (1991), *Bibliometric Methods in Assessing Productivity and Impact of Research.* (Ed.), Bangalore, Sarada Ranganathan Endowment for Library Science.

LANCASTER, F. W., LEE, J.-L. (1985), Bibliometric techniques applied to issues management: A case study. *Journal of the American Society for Information Science*, 36 : 389–397.

LANCASTER, F. W., MEHROTRA, R., OTSU, K. (1984), Some publication patterns in Indian and Japanese science: a bibliometric comparison. *International Forum on Information and Documentation*, 9 : 11–16.

LUUKKONEN, T. (1989), Publish in a visible journal or perish? Assessing citation performance of Nordic cancer research. *Scientometrics*, 15 : 349–367.

MARX, W., SCHIER, H., WANITSCHEK, M. (2001), Citation analysis using online databases: Feasibilities and shortcomings. *Scientometrics*, 52 : 59–82.

MCGRATH, W. E. (1996), The unit of analysis (objects of study) in bibliometrics and scientometrics. *Scientometrics*, 35 : 257–264.

MIDORIKAWA, N., MIYAMOTO, S., NAKAYAMA, K. (1990) A view of studies on bibliometrics and related subjects in Japan. In: BORGMAN, C. L. (Ed.), *Scholarly Communication and Bibliometrics*. (pp. 73–83), Newbury Park, SAGE Publications.

MILLER, C. (1990), Detecting duplicates: a searcher's dream come true. *Online*, 14 : 27–34.

MIYAMOTO, S., MIDORIKAWA, N., NAKAYAMA, K. (1989), A view of studies on bibliometrics and related subjects in Japan. *Communication Research*, 16 : 629–641.

MOED, H. F. (1988), The use of on-line databases in bibliometric analysis. In: L. EGGHE, R. ROUSSEAU (Eds), *Informetrics 87/88. Select Proceedings of the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval.* (pp.133–146), Netherlands, Elsevier.

MOED, H. F. (1989), Bibliometric measurement of research performance and Price's theory of differences among sciences. *Scientometrics*, 15 : 473–483.

NORTON, N. P. (1981), Dirty data – a call for quality control. *Online*, 5 : 40–41.

OJALA, M. (1992), Quality online and online quality. (The Dollar $ign), *Online*, 16 : 73–75.

OSAREH, F., WILSON, C. S. (2002), Collaboration in Iranian scientific publications. *Libri*, 52 : 25–35.

PAO, M. L. (1989), Importance of quality data for bibliometric research. In: C. NIXON, L. PADGETT (Eds), *National Online Meeting. Proceedings*. (pp.321–327), Medford, NJ, Learned Information.

PERSSON, O. (1986), Online bibliometrics. A research tool for everyman. *Scientometrics*, 10 : 69–75.

PERSSON, O. (1988), Measuring scientific output by online techniques. In: VAN RAAN, A. F. J. (Ed.), *Handbook of Quantitative Studies of Science and Technology*. (pp.229–252), Amsterdam, Elsevier Science.

PITERNICK, A. B. (1982), Standardization of journal titles in databases (letter to the editor), *Journal of the American Society for Information Science*, 33 : 105.

PROVOST, F., NIEUWENHUYSEN, P. (1992), Measuring overlap of databases in water supply and sanitation using sampling and the binomial probability distribution. *Scientometrics*, 25 : 201–208.

REID, E. O. F. (1992), Using online databases to analyze the development of a specialty: case study of terrorism. In: WILLIAMS, M. E. (Ed.), *13th National Online Meeting*. (pp. 279–291), Medford, NJ, Learned Information.

ROY, D., HUGHES, J. P., JONES, A. S., Fenton, J. E. (2002), Citation analysis of otorhinolaryngology journals. *Journal of Laryngology and Otology*. 116(5) : 363–366.

SANDISON, A. (1989), Thinking about citation analysis. *Journal of Documentation*, 45 : 59–64.

SAARTI, J. (2001), Consistency of subject indexing of novels by public library professionals and patrons. *Journal of Documentation*. 58(1) : 49–65.

SEGLEN, P. O. (1989), Evaluering av forskningskvalitet ved hjaelp af siteringsanalyse og andre bibliometriske metoder. In Norwegian. [Evaluation of research quality by means of citation analysis and other bibliometric methods]. *Nordisk Medicin*, 104, 331–335; 341–342.

SMITH, L. C. (1981), Citation analysis. *Library Trends*, 30 : 83–106.

SNOW, B. (1993), RANK: A new tool for analyzing search results on DIALOG. *Database*, 16 : 111–119.

STEFANIAK, B. (1987), Use of bibliographic data bases for scientometric studies. *Scientometrics*, 12 : 149–161.

STERN, B. T. (1977), Evaluation and design of bibliographic data bases. In: M. E. WILLIAMS (Ed.), *Annual Review of Information Science and Technology*. (pp. 3–30), New York, Knowledge Industry Publications for American Society for Information Science.

TENOPIR, C. (1982), Distributions of citations in databases in a multidisciplinary field. *Online Review*, 6 : 399–419.

THORNE, F. C. (1977), The citation index: another case of spurious validity. *Journal of Clinical Psychology*, 33 : 1157–1161.

TORRICELLA-MORALES, R. G., VAN HOODYDONK, G., ARAUJO-RUIZ, J. A. (2000), Citation analysis of Cuban research. Part 1. A case study: the Cuban Journal of Agricultural Science. *Scientometrics*, 47 : 413–426.

VAN CAMP, A. J. (1991), StarSearch for the health sciences. (Caduceus), *Database*, 14 : 99–101.

WALKER, G. (1990), Searching the humanities – subject overlap and search vocabulary. *Database*, 13 : 37–46.

WANGER, J. (1977), Multiple database use. *Online*, 1 : 35–41.

WHITE, H. D. (1996), Literature retrieval for interdisciplinary synthesis. *Library Trends*, 45 : 239–264.

WHITE, H. D. (2001), Computing a curriculum: Descriptor-based domain analysis for educations. *Information Processing & Management*, 37 : 91–117.

WHITE, H. D., GRIFFITH, B. C. (1987), Quality of indexing in online data bases. *Information Processing & Management*, 23 : 211–224.

WHITE, H. D., MCCAIN, K. W. (1989), Bibliometrics. In: WILLIAMS, M. E. (Ed.), *Annual Review of Information Science and Technology*, Vol. 24. (pp. 119–186), Amsterdam, The Netherlands, Elsevier Science Publishers B.V. for the American Society for Information Science.

WILLIAMS, M. E. (2002), The state of databases today: 2002. In: E. NAGEL (Ed.), *Gale Directory of Databases*. (pp. xvii-xxx) Detroit, Gale Group, Inc.

WILLIAMS, M. E., LANNOM, L. (1981), Lack of standardization of the journal title data element in databases. *Journal of the American Society for Information Science*, 32 : 229–233.

WILSON, C. S. (1999), Informetrics. In: WILLIAMS, M. E. (Ed.), *Annual Review of Information Science and Technology*, Vol. 34. (pp. 107–247), Medford, NJ, Information Today.

WILSON, C. S., MARKUSOVA, V. A. (in prep.), The effect of politico-economic changes in Russia from 1980 to 2000 on its scientific output as reflected in the *Science Citation Index*.

WILSON, C. S., OSAREH, F. (2003), Science and research in Iran: A scientometric Study. *Interdisciplinary Science Reviews*, 28(1) : 26–37.

WOLFRAM, D., CHU, C. M., LU, X. (1990), Growth of knowledge: bibliometric analysis using online database data. In: L. EGGHE, R. ROUSSEAU (Eds), *Informetrics 89/90: Selection of Papers Submitted for the 2nd International Conference on Bibliometrics, Scientometrics and Informetrics*, London, Ontario, Canada. (pp. 355–372), Amsterdam, The Netherlands, Elsevier.