# Data Set Isolation for Bibliometric Online Analyses of Research Publications: Fundamental Methodological Issues

**Peter Ingwersen* and Finn Hjortgaard Christensen**
*Centre for Informetric Studies, Royal School of Librarianship, Birketinget 6, DK 2300 Copenhagen S, Denmark. E-mail: pi@db.dk*

**The aim of the article is to emphasize and illustrate the retrieval dimensions of data collection activity online and their influence on the research evaluation outcome. The attempt is to reinforce the link between online retrieval and bibliometrics. Given that various forms of publication counts and citation analyses provide a valuable and revealing quantitative starting point for more qualitative indications and assessments of Science and Technology (S&T) performance, it is evident that their reliability and objectivity must be undisputed as far as possible. The article discusses the basic problems and limitations inherent in online bibliometric data collection and analyses, and points to possible solutions by means of illustrative case studies and examples. The reason for performing local publication analyses online often arises because of the increased use of external research assessments made by centralized bodies. For small institutions in small countries, like the North European one, such self-analyses may in addition provide valuable and inexpensive insights into novel S&T niches to explore. The major concern is the extent to which online bibliographic and domain dependent databases, as a supplement to the Institute for Scientific Information (ISI) citation files, are suitable for quantitative analysis and mapping of R&D outcome. By merging these two different types of databases into a single cluster, the method of duplicate removal becomes crucial. The article introduces a novel removal procedure by describing and exemplifying the principle of Reversed Duplicate Removal (RDR). RDR enables the analyst to take control of the location of the duplicates and to perform tailored analyses of the overlap of identical documents between files. It is well known that the databases themselves present obstacles directly associated with the process of performing online retrieval of the information necessary for further analysis. Problems encountered are, for instance, poor or inconsistent subject indexing within a single database or among several databases. Name form inconsistencies as to authors, institutions, and journals, the lack or inaccessibility of vital data in the database structures, etc., also present obstacles. On the other hand, comprehensive online bibliometric analyses are in many ways easier, faster, and less expensive to perform locally than those made using the independent CD-ROM versions of the relevant databases. In contrast to the online versions, the CD-ROM systems demonstrate a vital shortage of robust data processing and manipulation facilities. The downloading of records from a variety of CD-ROM files, the cleaning-up process, and the ensuing data processing activities become cumbersome and resource demanding. Regardless of database versioning, the degree of awareness of these retrieval and set isolation factors, such as the relevant search commands, syntax, and the analysis assumptions on the part of the analyst, plays an important role for the quality of the analysis outcome.**

## 1. Introduction

The domain-related bibliographic databases available online and, in particular, the citation databases produced by the Institute for Scientific Information (ISI), provide an invaluable resource for a wide range of informetric analyses. The term informetrics designates a recent extension of the traditional bibliometric analyses also to cover non-scholarly communities in which information is produced, communicated, and used. Diffusion analyses of, for instance, political issues in newspaper full-text databases become, hence, an integrated element of informetric research. The traditional bibliometric applications include the study of communication patterns, the identification of research fronts, historical studies of the development of a discipline or a domain, and the evaluation of the research activities of countries, institutions, or individuals (Cronin, 1984; Garfield, 1979).

The use of publication and citation data for the evaluation of research is an increasingly important tool in research policy, funding, and peer review procedures, where

* To whom all correspondence should be addressed.

Received July 31, 1995; revised November 28, 1995 and January 30, 1996; accepted March 27, 1996.

there is a noticeable lack of quantitative performance indicators that can be used to aid the assessment of, for example, promotion applications or resource allocation. In an increasing number of countries, the evaluation of publicly funded research is performed by an independent and centralized body, i.e., a state agency or a selected university department of science studies. Similarly on their own behalf, the individual research institutions or scholars wish to monitor their own performance or observe the international competition via comparative studies. Increasingly, the reason may be to check the external assessments made by centralized bodies. For small institutions in small countries, like the North European ones, such self-analyses may, in addition, provide valuable insights into novel Science & Technology (S&T) niches to explore. However, for both the assessors at the centralized R&D evaluation bodies and the individuals or single institutions, it is vital that the data sets, on which the variety of analyses are carried out, are valid and unbiased.

Data set production can be done in several ways. It can range from the purchase of extractions of the ISI files dealing with a particular country or institution, to using the CD-ROM versions of these databases. In our case, the data set isolation and analysis is carried out *online*. The online analyses have five advantages: They are 1) fast; 2) inexpensive; and they 3) provide instant results by means of advanced online processing tools. Perhaps more important, they allow for 4) the direct combination of domain-dependent and ISI databases; and 5) they are reproducible.

The disadvantage is that they are less flexible from the point of view of the analyst. Dou, Hassanaly, La Tela, & Quoniam (1990) point to and discuss how commercially available software may provide offline (and automatic) analyses on downloaded online material. However, the cost of and resource allocation for downloading large quantities of references followed by duplicate removal from several different files are heavy compared to the more direct online method. In between these two solutions to data set isolation, we find the CD-ROM application. In this case, commercial software is mandatory for any distinct bibliometric analysis.

Several studies have pointed to the benefits, as well as the problems associated with, the application of online databases and host software for publication and citation analyses. Persson (1986, 1988) has described the retrieval techniques necessary for performing instant analysis online, as has more recently, for instance, Wissmann (1993). The more general limitations in relation to the specific use of citation counts in any application, and the problems which must be taken into account when citation data are to be used as research indicators, are discussed in Sandison (1989) and Smith (1981). More critical discussions of their validity are provided by Luukkonen (1989), Moed (1989), Seglen (1989), and Thorne (1977). Obviously, publication counting, citation analyses, and scientific domain mapping cannot stand alone as science indicators. On the other hand, when such analyses are endeavored, they ought to be carried out as comprehensively as possible. Since specialized research, which often is carried out in small countries, often get published in specialized international journals not accepted by ISI, we strongly recommend the supplementary use of core discipline-oriented databases during data collection. The main reason is the high probability of such works being cited in journals admitted by ISI. Based on these ideas, Christensen and Ingwersen (1995) report on the results of an ongoing methodological project concerning online data set generation. Basically, it covers the initial stages of tuning data sets online as well as methods for duplicate removal and processing. It attempts to link online retrieval with bibliometrics.

In this article, we enhance the principle of Reversed Duplicate Removal (RDR). RDR enables the analyst to take control of the location of the duplicates and to perform tailored analyses of the overlap between files of identical documents. The discussion of RDR, and the intricate issue of duplicate processing, are illustrated by a typical case of countries contributing research in a selected topic and a Bradford distribution of core journals for the same topic. Further, the multi-dimensional usability and processing of the overlap of duplicates is demonstrated and discussed by samples of online data extractions related to the detailed topical mapping of a domain. The article incorporates the most recent developments of the host software mandatory for online analysis. At the same time, we attempt to demonstrate the most important limitations as well as the pitfalls and errors associated with this type of data isolation. By being aware of the problems often hidden from, ignored by, or completely unknown to analysts inexperienced in online retrieval, it is believed that the basic collection of data may be improved in qualitative terms.

Fundamentally, these problems are concerned with 1) the data structures, their contents and form in the various databases; and 2) the command language features and online data processing tools. When making a significant impact, these two dimensions will be dealt with in relation to the various types of data set isolation discussed below. In common to the range of bibliometric data set isolations online is the use of certain search features, e.g., database clustering and duplicate removal, data processing, e.g., frequency analysis, as well as the limitations in the database contents, and the name form problems for persons and institutions.

The article is organized in four major sections and conclusions. The first section discusses and demonstrates the issues of database selection as well as the clustering and tuning of the data set. The following section concentrates on duplicate removal techniques, including RDR, followed by a discussion and demonstration of the available online data set processing tools. The penultimate section analyses briefly the common and specific bibliographic database features influencing the process of data

set generation. The concluding section summarizes the parameters available for set isolation, and hence, for publication counting. It is assumed that the basic Dialog retrieval commands are known to the readers. Thus, examples and cases are only explained in detail with respect to the more advanced search and processing features.

## 2. Common Online Retrieval, Processing, and Database Issues

The following retrieval features are mandatory if comprehensive publication counts and analysis as well as citation analyses are to be carried out online on several files simultaneously: Cross-file searching, database clustering, and duplicate removal.

A cross-file searching facility enables the analyst to observe the number of documents held in different databases associated with a country, an institution, a topic or domain, or combinations of these search parameters. Other, more form-based search parameters may in principle be included, e.g., particular document types, time periods, or language. Based on the response from the cross-file facility, the analyst may select a cluster of all or of the dominating files to be fused into one set and further analyzed. This crude set should be tuned, i.e., directly disturbing items, e.g., editorials, letters to the editor, etc., or unwanted document types should be isolated from the set. This separation depends on the purpose of the analysis. If the aim is an analysis of journal articles on a certain topic, it is evident the tuning method will differ from that applied to book reviews for the same topic. We call this isolation of document type(s) ''tuning,'' and the process is carried out by means of the Document Type index (DT=) in the databases. Schubert, Glänzel, & Braun (1989) discuss a similar tuning process of citable items when analysing SciSearch. In our case we are in particular interested in the overlap between several databases, e.g., SciSearch and Inspec. Since editorials, notes, conference papers, etc. are types only found in one of the files, they are kept separate from the data set. This ''tuning'' process is demonstrated in Figure 2.

In order to avoid duplicate documents, these should be removed from the final data set to be analyzed. This operation is complicated and is demonstrated in section 2.2 below. Duplicate removal in addition implies a further reduction of the number of databases pertaining to the subsequent analyses.

When the duplicates have been removed, the final data set(s) should be analyzed as easily and inexpensively as possible. What is required is an online frequency analysis tool which allows the analyst to break down the set into, for instance, time series of production ratio, national or institutional distribution. The goal may also be to produce Top-20 lists of authors, institutions, countries, or journals in a field, and to observe the citation impact of journals or individual researchers. Such tools exist in the search software of various online vendors, such as the Rank and Zoom facilities at Dialog and ESA/IRS, respectively. They are, however, functionally different, as discussed below in section 3. A similar frequency tool has been installed in late 1994 at the STN Online Service under the name SmartSelect (Huber, 1995).

If the starting point is to isolate a country, an institution, or a particular journal, it presupposes that the database fields, i.e., the corporate source (CS=) and journal name (JN=) fields, actually exist, contain the names searched for, and are searchable. The way of inversion of the fields in question is crucial, as is the level of name form control. Similar conditions apply to the subsequent analyses of the isolated set in the form of distributions by country, institutions, authors, or journals.

A particular case concerns the isolation of a topic or a domain as the basic data set parameter. The problem is the nature of comprehensiveness of the final data set. If the interest is to isolate and analyze the total number of documents which also incorporate the topic as a minor aspect, the search should include the abstract and full-text fields. On the other hand, if the concern is publications dealing with the topic basically as a major aspect, the data set isolation might profit from searching the index term and title fields only. A help exists if the database(s) in question contain an exhaustive and searchable subject category field which covers the topic or domain in question, as demonstrated by the analyses by Persson (1988). These database specific issues are discussed briefly in section 4 below.

### *2.1. Cross-File Searching, Database Clustering, and Data Set Tuning*

Given a topic (asteroids) and a year span (e.g., 1992–1994) as an example of the starting parameters for the isolation of a data set, the purpose might be to analyze the worldwide breakdown by country, institutions, and authors, or to perform citation analyses of the topic.

Initially, the adequate online host is selected, i.e., an online vendor which a) provides the files relevant to the domain, i.e., astronomy; b) allows for the automatic distribution of search results across these files; c) makes duplicate removal possible; and d) provides access to frequency analysis tools. In this case, these mandatory conditions are fulfilled at least at Dialog, ESA/IRS, and STN. Although the latter service also provides the citation files, including SciSearch (SCI), we have selected Dialog, mainly because of its widespread use.

Via the Dialindex facility, which constitutes a database of the 400 files provided by Dialog and categorized by S&T disciplines, the cluster category ''Physics'' is chosen since it includes the domain ''astronomy.'' The entry of the search statement in the form of a string into the chosen cluster (ss asteroids and py = 1992:1994) results in a set of 1,772 isolated documents. (Note that these initial commands and online results are not illustrated in the figures). The command ''set detail on'' invokes the distri-

bution among the different files. In this case, one can observe that the documents adhere to six of the files in the cluster, among which are Inspec, Aerospace, Pascal, and SCI. An example of the cluster at a much later stage is illustrated in Figure 1. The order of the files in the cluster is important and can be altered by the command ''set files,'' as demonstrated in Figure 3 below. At this point, the set is comprehensive but also very crude. It is comprehensive because all topical fields in the basic indexes of the cluster files were actually searched, including the abstracts in which the topic ''asteroids'' may be dealt with as a minor aspect only. The crudeness does not refer to the simplicity of the search statement which, in principle, could have been a quite elaborate one. The lack of refinement refers to the fact that it certainly contains duplicates as well as all the document types available in the various files forming the cluster.

The initial tuning of the crude set, prior to duplicate removal, refers to the common fact that the SCI database includes short editorials, letters to editors, and corrections/errata published in the journals indexed by SCI. Such items would supposedly be superfluous in a publication count and a subsequent evaluation process. Consequently, SCI-Search is initially isolated from the cluster and limited to the relevant document types: ''articles, reviews, or reviews, bibliographies.'' When united with the remaining five files, the original data set is reduced by 42 insignificant items to 1,730 references to documents indexed in the six different files. One may note, however, that this new set (set 12, Fig. 1) still contains several different document types, since the other databases in the cluster have not been streamlined to one type only. In order to tune the set further to contain journal articles only, it becomes necessary to generate sets for the remaining files that are structured accordingly. Hence, it is vital to check how the different files denote their ''article-type,'' e.g., as ''journal paper'' in Inspec. An example is demonstrated below (Fig. 2). In many cases, the article type is not explicitly stated, exactly because the major portion of a file consists of articles.

In addition, we may observe that by not limiting the data set to SCI Search only, the initially tuned data set incorporates several relevant domain-dependent databases and thus provides a larger potential for more complete and interesting publication analyses: Of the 1,730 items on ''asteroids,'' 1,266 documents derive in fact from files other than SCI. These documents may be articles published in specialized journals, R&D reports, and conference papers. Evidently, these items are potentially citable.

### *2.2. Removal of Duplicates*

When dealing with a data set from a cluster of files, the removal of duplicates is mandatory for any subsequent analysis. Operationally the removal is easy. However, the fundamental question is: From which databases do we want to remove the duplicates, and in which file to keep them? Awareness of this issue, and thus of the so-called Reversed Duplicate Removal (RDR) technique, is crucial for the outcome of the subsequent analyses. A second, often overlooked fact exists: The set of the duplicates may, in addition, be of interest. This overlap of items across files may probably indicate that such documents or sources are of particular importance (Pao, 1993). Third, by removing the duplicates it becomes apparent which databases should really be analyzed. Figure 1 shows the basic operation, and Figure 3, data set B, demonstrates the consequence when a different removal sequence across files is applied, i.e., when RDR is introduced into a file cluster.

On Figure 1, set 12 consists of the initially tuned 1,730 documents from the cluster of six files. Note that the original file order has been kept, i.e., that the Inspec file is the leading database in the file sequence and consequently contains the duplicates between Inspec and the other databases.

We may observe that set 13 (Fig. 1) has been reduced from 1,730 items to 1,155. The approximately 600 (tri-) or duplicates have been removed primarily from the files 108 (Aerospace), 144 (Pascal), and 434 (SCI), previously displaying 479, 230, and 506 items, respectively. The Mathfile is reduced to zero items and may hence be excluded from further analysis. The duplicates are kept in the Inspec file. The display also shows that a publication analysis (for this topic/time period) may prudently be carried out in three of the files comprising a 95% sample of the total research output.

The significance of this sequential order of removing the duplicates is important, i.e., that the file order in the cluster begins with Inspec, file 2. A subsequent publication analysis will therefore primarily depend on the file structure and contents of this file. Below in section 3, Figure 4 demonstrates a biased result from this file order: Certain publication analyses will be distorted, if the analyst is unaware of the order of sequence.

Regardless of the method of duplicate removal, the resulting data set will still consist of several document types as demonstrated in the previous section. If the ISI files are included in the clustered set, as was the case in Figure 1, there might exist a bias towards journal articles in that set. A complete tuning prior to the removal of duplicates might thus be necessary, in order to isolate and control the variety of document types contained in the set of clustered files. Figure 2 demonstrates this principle of complete tuning of a data set to ''journal articles'' and the resulting distribution between the files. This is followed by duplicate removal in the original file order (Fig. 3, data set A), as well as in the reversed order of duplicate removal (RDR) (Fig. 3, data set B). In order to simplify the problems, the data set generation is carried out from the start and is performed on two databases only: Inspec and Sci Search. In Inspec, the denotation for journal article is ''Journal Paper'' or ''JP,'' as shown in set 8 (Fig. 2).

We observe that the completely tuned data set from

```
?rd s12                                   (remove duplicates from set 12= 1730 items)

SET       ITEMS     DESCRIPTION

          S13       1155      RD s12 (unique items)

?set detail on
?ds s13

File      Items
2         476                 (Inspec file)
62        27                  (SPIN(R) file)
108       298                 (Aerospace file)
144       65                  (Pascal file)
239       0                   (Mathfile)
434       285                 (SCI file)
```

FIG. 1. Duplicate removal from an initially tuned set (set 12) consisting of a cluster of six databases on Physics; duplicates are kept in file 2, Inspec.

the two files prior to duplicate removal holds 933 documents; after duplicate removal (Fig. 3), it contains 783 unique items and covers approximately 70% of the six-file data set on the identical topic and time period (1,155 items, Fig. 1). Note, however, that the final data set (A) is limited to ''journal articles'' only.

More importantly, the RDR provides us with two different sets, but containing the same items. Data set (A) actually contains 138 (overlapping) items in Inspec, which in data set (B) are incorporated in SCI. In the case of a comprehensive statistical sample made out of three core files, e.g., Inspec, Aerospace, and SCI (see Fig. 1), the RDR operation will produce six sets in total, in order to cover all the file order sequences. The choice of the

```
?b 2,434                                  (the original file order)
?ss asteroids and py=1992:1994

S1        5388      ASTEROIDS
S2        2550193   PY=1992 : PY=1994
S3        999       ASTEROIDS AND PY=1992:1994

?ss dt=(article or review or review, bibliography or jp)

S4        8716823   DT=ARTICLE
S5        87742     DT=REVIEW
S6        139950    DT=REVIEW, BIBLIOGRAPHY
S7        3907431   DT=JP
S8        12866723  DT=(ARTICLE OR REVIEW OR REVIEW, BIBLIOGRAPHY OR JP)

?ss s3 and s8

          999       S3
          12866723  S8

S9        933       S3 AND S8             (the set completely tuned to "articles")

set detail on
?ds s9

File      Items
2         464                             (no. of items in Inspec)
434       469                             (no. of items in SCI)
S9        933       S3 AND S8
```

FIG. 2. Complete tuning sequence of a data set to ''journal articles'' (set 9 = 933 items) distributed onto two files, Inspec and SCI. The set numbering has been restarted.

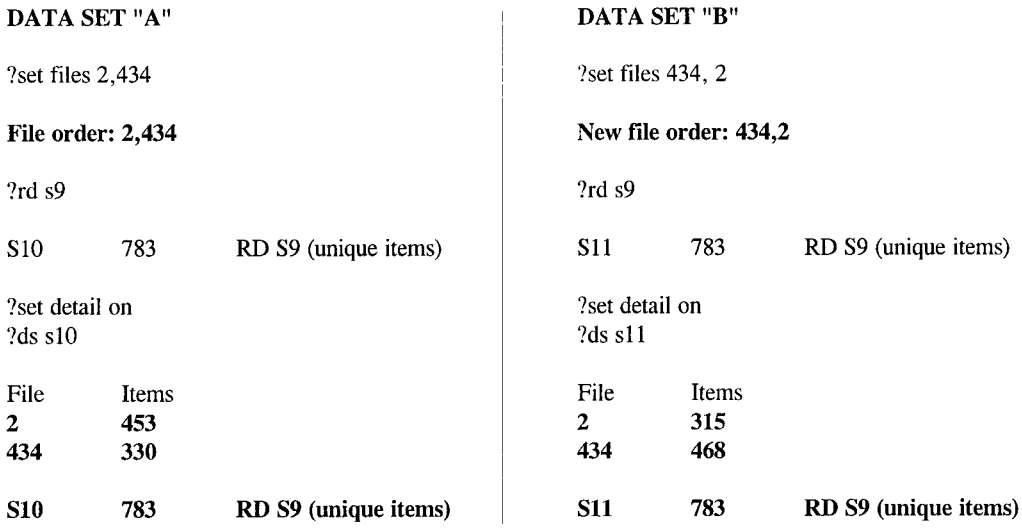


FIG. 3. Duplicate removal (Data Set ''A''): Duplicates are placed in Inspec. Reversed duplicate removal (RDR) from the same tuned set, placing duplicates in SCI (Data Set ''B'') by means of the command: Set files 434,2. The overlapping set consists of (453–315 = 138) items.

set(s) to be analyzed for publication characteristics will then depend on the purpose of the analysis.

An analysis carried out on data set (A) will depend on the data structure and contents of the first-order file, namely Inspec, for more than 60% of the original set of 783 items. An analysis by country or author institutions will not yield a satisfying result, because Inspec only provides the affiliation of the first author. In this respect, data set (B) is more valid since SCI covers all author affiliations. The overlapping 138 items, now contained in the leading SCI file of data set (B), will thus incorporate the affiliations and countries of the secondary authors. Still, the remaining 315 items from Inspec in set (B) will yield corporate names or countries only adhering to the first author. Figures 4 and 5 below demonstrate the analyses of the two data sets for contributing countries by the application of the RANK processing tool.

In contrast, a subject category or topical analysis carried out on data set (A) will yield more information than a set (B) analysis, because Inspec is indexed in a more specific and controlled manner than SCI. In our demonstration example on ''asteroids 1992–94'' from Inspec and SCI, we may thus see the data sets (A) and (B) as statistical samples, each covering 60% of the 783 items or, if you like, 40% of the initially tuned set of 1,155 items (Fig. 1). One may note that the set of 138 items overlapping the two files may be of particular interest since it may be isolated and explored according to all the parameters made available by both files (see section 3.1 below). This fact provides the main reason for also involving domain-related databases in addition to the ISI files in data set creation.

An analysis of the contributing authors will yield the same result regardless the data set type, since both sets contain all authors.

The problem is, however, that the algorithm for removing duplicates is not completely safe. The 11 items in data set (A) (set 10), which are removed from Inspec compared to the original crude set, i.e., the 464 items in set 9 (Fig. 2), seem dubious. Logically there should not have been any duplicates in the first order file in data set (A). A control [by means of the command: ''ss s9 from 2''; ''ido'' (identify duplicates only)] demonstrates that all the 11 items retrieved as duplicates from Inspec were actually different from one another. They ought to have been added to the data set (A). According to Miller (1990), the reason for the error may be that the algorithm checks for identical title and first author, but not for the source and publication year. This was the case in our set of duplicates. The sources and years were different. As it is, the algorithm functions in such a way that articles previously published as conference papers, with identical titles and authors, are removed. Miller (1990, p. 27) states: ''Unless you are foolishly trustworthy or just don't care, I would always recommend using IDO to see what the system has identified as duplicates before removing them with the RD command.'' This rule of thumb has not been followed in the subsequent examples and, therefore, there is an additional error potentiality.

## 3. The Advanced Online Data Processing Tools

The most simple online data processing tools consist of the term frequency analysis programs which are provided by the major online hosts: The RANK feature by Dialog, and the ZOOM device by ESA/IRS. Zoom came on the online market already in 1981/1982 (Ingwersen, 1984) whilst the Rank facility became publicly available in 1993 (Dialog, 1993). Fundamentally, both features provide the searcher with a list of terms, authors, or journal names from a selected set of items sorted by decreasing frequency of occurrence.

?rank s10 gl detail cont (detailed ranking of data set (A) on Geo-Location, GL= )

RANK: S10/1-783 Field: GL= File(s): 2,434
(Rank fields found in 330 records -- 43 unique terms)

| RANK No. | File | No.items in File | No.Items Ranked | %Items Ranked | Term |
|---|---|---|---|---|---|
| 1 | 434 | 4717240 | 159 | 48.2% | USA |
| 2 | 434 | 144477 | 39 | 11.8% | GERMANY |
| 3 | 434 | 648289 | 37 | 11.2% | FRANCE |
| 4 | 434 | 911994 | 24 | 07.3% | ENGLAND |
| 5 | 434 | 783524 | 21 | 06.4% | JAPAN |
| 6 | 434 | 533873 | 18 | 05.5% | CANADA |
| 7 | 434 | 311511 | 16 | 04.8% | ITALY |
| 8 | 434 | 253349 | 12 | 03.6% | AUSTRALIA |
| 9 | 434 | 59005 | 7 | 02.1% | BRAZIL |
| 10 | 434 | 218041 | 7 | 02.1% | NETHERLANDS |
| 11 | 434 | 44112 | 7 | 02.1% | RUSSIA |
| 12 | 434 | 193060 | 7 | 02.1% | SWEDEN |
| 13 | 434 | 551022 | 7 | 02.1% | UNION OF SOVIET SOCIALIST REPUBLICS |
| ... | ... | ... | ... | ... | ... |
| 20 | 434 | 97937 | 3 | 00.9% | DENMARK |

FIG. 4. Abbreviated list of countries producing the items in Data Set (A), ranked according to frequency of occurrence in the set (set 10, Fig. 3). Field GL = Geo Location; File order: Inspec prior to SCI; GL-field found in 330 of 783 items; 43 unique terms = 43 unique country names.

Zoom was originally designed to support the searcher during the retrieval process by providing conceptual feedback from the system. Therefore, it may be carried out on certain database fields only, mainly on the title, index term, author, corporate source, journal name, and classification code fields. However, the zooming can be done regardless of the mode of inversion of these fields. Zoom does not provide alphabetic sorting, e.g., by author name, nor does it present the frequency in percent.

The Rank command (Fig. 4) is deliberately designed also to aid in bibliometric analysis activities. It covers most of the fields that are inverted as phrases or word strings. When in a database, say the SCI and other ISI files in which the corporate source or other data fields are inverted word by word, the Rank function does not apply. In fact, presently only 25% of the Dialog files have their CS = field inverted as phrases or strings. The major portion of these files belongs to the business domains. None of the files are S&T files. This fact suggests that distribution analyses on institutions may only be carried out online by hand or by downloading and reloading the items in different database software.

On the other hand, Rank can present results directly in percent figures and sort data alphabetically. Also, it can provide the analyst with the total number of documents in the file that are associated with the term or name analyzed, i.e., the document frequency. This facility may indicate the weights of the sorted terms in the file which may be compared to the actual frequencies of the same terms in the analyzed data set. For instance, in data set (A) (Fig. 4) the U.S. proportion of the world research on ''asteroids 1992–94'' published in journals is 48.2% (159/330 items). At the same time, the number of occurrences constitutes 159 out of 4,717,249 items, i.e., that the U.S. research indexed on asteroids 1992–1994 covers 0.00003% of the U.S. total research published in journals through time. By knowing the file size, the contributions by U.S.A. and other countries to ''world research in general'' can easily be calculated. This weight parameter applies in principle to all fields that are rankable, e.g., and perhaps more significantly, to authors (AU=), journal names (JN=), or publication years (PY=).

Since we know that data set (A) has a bias towards the Inspec file which, like the other Dialog domain-related files, does not contain a geo-location field (GL=), we should expect the low number of items (330) that contains this field in the ranked data set. As can be observed from column two, only file 434 (SCI Search) has actually been processed. The sample size is only approximately 40% of the population of 783 items. We may indeed do something about this bias and improve the sample size. A fast way is to switch to the RDR data set (B) already generated, in which the proportion of SCI items is much larger (468 items). This improves the coverage to approximately 60% of the original set. A shortened list is shown in Figure 5.

By enlarging the sample size and giving priority to the SCI file via RDR, for instance, U.S.A. as well as Sweden extend their coverages: U.S.A. with 1.5% points to 49.8, and Sweden from 2.1 to 3.4%. Note the interpretative problem introduced by the political reality that ''U.S.S.R.'' has ceased to exist and that ''Russia'' emerges during the period analyzed.

?rank s11 gl detail cont (detailed ranking of data set (B))
RANK: S11/1-783 Field: GL= File(s): 2,434
(Rank fields found in 468 records -- 44 unique terms)

| RANK No. | File | No.Items in File | No.Items Ranked | %Items Ranked | Term |
|---|---|---|---|---|---|
| 1 | 434 | 4717240 | 233 | 49.8% | USA |
| 2 | 434 | 648289 | 55 | 11.8% | FRANCE |
| 3 | 434 | 144477 | 53 | 11.3% | GERMANY |
| 4 | 434 | 911994 | 38 | 08.1% | ENGLAND |
| 5 | 434 | 311511 | 35 | 07.5% | ITALY |
| 6 | 434 | 783524 | 23 | 04.9% | JAPAN |
| 7 | 434 | 533873 | 22 | 04.7% | CANADA |
| 8 | 434 | 253349 | 17 | 03.6% | AUSTRALIA |
| 9 | 434 | 193060 | 16 | 03.4% | SWEDEN |
| 10 | 434 | 218041 | 11 | 02.4% | NETHERLANDS |
| 11 | 434 | 86792 | 10 | 02.1% | CZECHOSLOVAKIA |
| 12 | 434 | 551022 | 10 | 02.1% | UNION OF SOVIET SOCIALIST REPUBLICS |
| 13 | 434 | 117245 | 9 | 01.9% | BELGIUM |
| ... | ... | ... | ... | ... | ... |
| 16 | 434 | 44112 | 8 | 01.7% | RUSSIA |
| ... | ... | ... | ... | ... | ... |
| 22 | 434 | 97937 | 5 | 01.1% | DENMARK |

FIG. 5. Short list of countries producing the items in Data Set (B), ranked according to frequency of occurrence in the set (set 11, Fig. 3). File order: SCI prior to Inspec; GL-field found in 468 of 783 items; 44 unique terms = 44 unique countries.

In order to circumscribe the total population of 783 items in data set (B), one slightly cumbersome method is available: To isolate the unique Inspec items from data set (B) by entering the command (''ss s11 from 2,'' Fig. 3); then to select the individual countries one by one in the CS = field and to produce a list of sets which may be ranked according to their number of hits; finally to add the list to the ranked output (Fig. 5). The resulting total distribution will, however, still be biased since the SCI file contains all the participating countries, whilst Inspec only contains the country associated with the first author. Unfortunately, this pattern is common to most of the domain-related Dialog databases. It is easy, however, to calculate statistically the number of country names omitted in such a file.

It should be noted that a distribution over time by means of ranking the publication year (PY=) is not affected by the nature of the data set, neither is, in principle, a distribution by authors or journal names. However, the latter analyses will display lists in which name form problems may occur, due to various ways of indexing the authors and journal names in the clustered files (see Fig. 8). For files made available both by Dialog, ESA-IRS, and STN, it is possible to carry out a separate frequency analysis via the Zoom or Smartselect facilities provided by these hosts. This issue concerns the Dialog files in which, for instance, the corporate source field (CS=) is inverted word by word, and Rank thus does not apply. The problem then is to isolate the proper set of identical items online from the file provided by the alternative host. The Rank facility does not form part of the CD-ROM versions of the various databases, such as the ISI files, nor is it included in the National Science Indicators database On Disk (NSIOD).

### *3.1. Processing the Overlap of Duplicates*

The overlapping set(s) of duplicates between the two files, i.e., 138 items in the example shown Figure 3 and providing a sample of 18% of the completely tuned 783 items, in principle represents an interesting set for any subsequent publication analysis. As for the data sets (A) and (B), two overlapping sets actually exist containing exactly identical items, though represented differently by the two files in accordance with the current file order. By isolation of the sets, it becomes possible to make adequate use of these differences in analyses. For instance, one may compare the few broad SCI subject categories, in this case eight categories (Fig. 6), with the corresponding more comprehensive Inspec categories. Perhaps more interesting, one can show the distribution of topical aspects contained in the overlap set by means of the 109 very specific descriptors assigned by Inspec (Fig. 7). The isolation of the two overlap sets is done from the sets 10 and 11 (Fig. 3) by the commands: ''ss s10 from 2''; ''ss s11 from 2''; and subtracting the latter selection from the former. This command sequence results in Overlap Set (A) characterized by the Inspec file structure (set 12, Fig. 7). The Overlap Set (B) is obtained by a similar command sequence which results in the isolation of SCI-structured records (set 13, Fig. 6).

This comparative analysis of topical aspects (Fig. 7)

RANK: S13/1-138 Field: SC= File(s): 2,434
(Rank fields found in 138 records -- **8 unique terms**)

| RANK No. | File | No.Items in File | No.Items Ranked | %Items Ranked | Term |
|---|---|---|---|---|---|
| 1 | 434 | 127043 | 105 | 76.6% | ASTRONOMY & ASTROPHYSICS |
| 2 | 434 | 155587 | 24 | 17.5% | GEOSCIENCES |
| 3 | 434 | 415651 | 13 | 09.5% | MULTIDISCIPLINARY SCIENCES |
| 4 | 434 | 1110424 | 2 | 01.5% | PHYSICS |
| 5 | 434 | 31501 | 1 | 00.7% | AEROSPACE ENGINEERING & TECHNOLOGY |
| 6 | 434 | 33181 | 1 | 00.7% | PHYSICS, FLUIDS & PLASMAS |
| 7 | 434 | 74079 | 1 | 00.7% | PHYSICS, NUCLEAR |
| 8 | 434 | 54754 | 1 | 00.7% | PHYSICS, PARTICLES & FIELDS |

FIG. 6. Distribution of eight subject categories in the Overlap Set (B) (set 13 = 138 items) based on SCI categories.

and subject categories (Fig. 6) associated with identical records demonstrates the advantage of having both overlap sets at hand. Clearly, the information provided by Overlap Set (A) (Fig. 7) is much more substantial and useful for scientometric purposes than the list produced by Overlap Set (B). As mentioned above, the overlap is believed to constitute the most relevant and core international research literature on the topic in question (Pao, 1993). The two alternative sets can be switched at will and may thus provide adequate means for a variety of topical trend analyses and citation studies of the central publications. For instance, the SCI-based Overlap Set (B) provides direct access to citation data which make journal impact or bibliographic coupling calculations feasible online. We intend to pursue these issues in a forthcoming publication. In order to observe the statistical validity of the analysis outcome, the results can be compared to similar analyses of the data sets (A) or (B)'s Inspec or SCI-dominated items. Both data sets cover independently a powerful 60% of the original set.

## 4. The Database Specific Issues

From the discussions above, it becomes apparent that the creation of a data set as well as the proper processing and analysis of that set is subject to the following database-dependent issues:

- Availability of adequate data fields in the files, e.g., the GL= or DT = fields;
- availability of sufficient data in existing fields, e.g., only the first corporate source is present in co-authored documents in a multitude of databases, except for the ISI files;

RANK: S12/1-138 Field: /DE File(s): 2,434
(Rank fields found in 138 records -- **109 unique terms**)

| RANK No. | File | No.Items in File | No.Items Ranked | %Items Ranked | Term |
|---|---|---|---|---|---|
| 1 | 2 | 3735 | 124 | 90.5% | ASTEROIDS |
| 2 | 2 | 38865 | 42 | 30.7% | VISIBLE ASTRONOMICAL OBSERVATIONS |
| 3 | 2 | 1433 | 31 | 22.6% | ASTRONOMICAL PHOTOMETRY |
| 4 | 2 | 6493 | 27 | 19.7% | CELESTIAL MECHANICS |
| 5 | 2 | 4603 | 21 | 15.3% | ROTATING BODIES |
| 6 | 2 | 15584 | 19 | 13.9% | ASTRONOMICAL SPECTRA |
| 7 | 2 | 4366 | 19 | 13.9% | METEORITES |
| 8 | 2 | 784 | 16 | 11.7% | PLANETARY SURFACES |
| 9 | 2 | 3819 | 10 | 07.3% | ALBEDO |
| 10 | 2 | 7134 | 10 | 07.3% | ASTRONOMICAL TECHNIQUES |
| 11 | 2 | 8123 | 10 | 07.3% | COMETS |
| 12 | 2 | 7181 | 10 | 07.3% | INFRARED ASTRONOMICAL OBSERVATIONS |
| 13 | 2 | 722 | 9 | 06.6% | METEOROIDS |
| 14 | 2 | 11640 | 7 | 05.1% | CHAOS |
| 15 | 2 | 5103 | 7 | 05.1% | COSMIC DUST |
| ... | ... | ... | ... | ... | ... |

FIG. 7. Distribution of 109 topical aspects in the Overlap Set (A) (set 12 = 138 items) based on Inspec descriptors.

- inversion method of existing fields, i.e., word by word (CS=) or phrase (JN=);
- field code consistency among files provided by the same online host, e.g., JN= always equals journal name while GN= and GL= may signify ''geo name/location'' in different files (ABI-Inform and the ISI files, respectively);
- name forms applied in the author, institutional, country, and journal name fields;
- indexing consistency or rather lack of consistency as to indexing policy across files;
- sudden appearance of novel data in files, e.g., full-text fields are added.

Prior to isolating a proper data set (and performing the subsequent analyses), the analyst should be aware of these issues and check the nature of availability and consistency, in order to repair conceivable damage to the outcome if possible.

Not surprisingly, analysts are tempted to limit their data set collection to the ISI files only. In the pure science domains this limitation may seem adequate, but only in connection with subsequent proportional analyses of, for instance, national contributions over time in selected domains. Some disciplines may, hence, not be analyzable, since they do not conform to the few broad ISI domain categories, as can be observed in Olsen, Foss Hansen, Luukkonen, Persson, & Sivertsen (1994). For instance, for chemistry, Norretranders and Haaland (1990, p. 86) demonstrate a similarity of the proportional distribution of national (Danish) R&D output 1980–1989 between SCI and Chemical Abstracts. For this kind of broad analysis, either of the databases can be applied. But a major problem is that the ISI files cover only the published research output in the international journals and some conferences. In many domains, including the social sciences, characterized by more complex publication patterns than the simple use of journal articles in English, the analyses may be somewhat ''unfair.'' The use of national languages plays an important role in certain disciplines, but the proportion of this channel of communication is biased in the ISI files. An additional problem is the scarcity of subject indexing in these files compared to the domain-dependent databases. Hence, it is not foolhardy to apply the domain-related databases as supplementary sources of data, thus providing overlapping sets of documents or duplicates as demonstrated above.

The lack of fields is important for the selection process as well as for certain analysis types. For instance, the document type field, commonly DT=, is vital when the interest is journal articles only, or concerns the isolation of all other types than articles from a data set (see Fig. 2). The consequence of a non-existing GL= field for processing data has been demonstrated above in relation to the Inspec file (Fig. 4). Similarly, the lack of data in available fields has been shown, e.g., in relation to the lack of corporate names for secondary authors. Selection as well as analysis are biased or made impossible.

Field codes or tags are surprisingly not always kept consistent by the same online vendor. Associated with inversion is the question of which data that actually is contained in a field. For instance, in Inspec and other domain databases, editors of conference proceedings and books may be inverted together with authors. If one blindly truncates a name from the AU = field in Inspec in order to cover both initials and full first names, the set will also incorporate all the items authored by someone else, but edited by the selected person. For instance, by selecting ''SS au=ingwersen, p?'' one obtains 24 publications. This figure is out of tune with actual facts and, if used directly in author-impact calculations, would lead to fatal errors. A check of the name in the alphabetic list of indexed names prior to selecting the name reveals that actually the person has only authored 10 items, but edited 14 contributions produced originally by other authors.

The inversion methods mainly have consequences for the processing of data, as demonstrated when applying the Rank command. A serious consequence is that the counting of the institutions involved in a set is made impossible, e.g., when using the ISI files. The names must be known beforehand and selected individually.

The lack of name form consistency for personal, journal, and institutional names—also within one and same database—poses severe problems. This issue is particular critical when isolating and analyzing R&D performance by individual institutions as well as when measuring the productivity and impact of individuals in a selected area. So-called Top-20 lists of researchers, institutions, or journals cannot be pursued directly; the lists have to be checked and edited in order to be valid. The issue is related to the way name data is structured in the fields of the files. For instance, Hansen, P.—Hansen P—Hansen, Peter. In certain files, both initials and full first names may be found for the same individual, according to what was originally written on the title page. Similarly, and for the same reasons, corporate names may display a variety of forms for each file:

Nat. Phys. Lab., Teddington, UK
Nat. Phys. Labs., Teddington, UK
National Phys. Lab., Teddington, UK
NPL, Teddington, UK

The issue of isolating a particular country or even an institution may be increasingly complex when country names vary within the same database over time, e.g., UK, United Kingdom, or Great Britain. For example, Ingwersen and Wormell (1989) observed in their analysis of chemical research publications applying the Chemical Abstracts file that prior to PY = 1979 it was only possible to isolate Denmark's publications by formulating a search profile consisting of more than 40 different entry points, including the names of the actual Danish R&D institutions. The country name had simply been randomly added

?rank s10 jn cont (ranking of data set (A))
RANK: S10/1-783 Field: JN= File(s): 2,434
(Rank fields found in 783 records -- 238 unique terms)

| RANK No. | File | No.Items Ranked | Term |
|---|---|---|---|
| **1** | **X** | **110** | **ICARUS** |
| 2 | 2 | 101 | INT. ASTRON. UNION CIRC. (USA) |
| 3 | 2 | 101 | INTERNATIONAL ASTRONOMICAL UNION CIRCULAR |
| **4** | **2** | **76** | **ICARUS (USA)** |
| 5 | X | 58 | ASTRONOMY AND ASTROPHYSICS |
| 6 | 2 | 30 | ASTRON. ASTROPHYS. (GERMANY) |
| 7 | X | 28 | ASTRONOMICAL JOURNAL |
| 8 | X | 26 | GEOCHIMICA ET COSMOCHIMICA ACTA |
| 9 | 2 | 26 | MINOR PLANET CIRC./MINOR PLANETS COMETS (USA) |
| 10 | 2 | 26 | MINOR PLANET CIRCULARS/MINOR PLANETS AND COMET |
| 11 | X | 22 | METEORITICS |
| 12 | X | 21 | PLANETARY AND SPACE SCIENCE |
| **13** | **X** | **21** | **SCIENCE** |
| 14 | 2 | 20 | ASTRON. J. (USA) |
| 15 | 2 | 19 | ASTRON. VESTN. (RUSSIA) |
| 16 | 2 | 19 | ASTRONOMICHESKII VESTNIK |
| 17 | 2 | 19 | SOL. SYST. RES. (USA) |
| 18 | 2 | 19 | SOLAR SYSTEM RESEARCH |
| 19 | 2 | 18 | CELEST. MECH. DYN. ASTRON. (NETHERLANDS) |
| 20 | 2 | 18 | CELESTIAL MECHANICS AND DYNAMICAL ASTRONOMY |
| 21 | 434 | 18 | IAU SYMPOSIA |
| 22 | X | 18 | NATURE |
| 23 | X | 17 | ASTRONOMY & ASTROPHYSICS SUPPLEMENT SERIES |
| 24 | X | 13 | ASTROPHYSICAL JOURNAL |
| **25** | **2** | **12** | **SCIENCE (USA)** |

FIG. 8. Frequency analysis of journal names (JN=) of data set 10, Figure 3. File order: Inspec prior to SCI; no. of Items: 783; no. of different journal names: 238; X: Items from both SCI (file 434) and Inspec (file 2).

prior to this date. Obviously, by applying several files in an online cluster, such problems increase further.

For journal names, the problems commonly occur associated with the modes of abbreviation. The ISI files use predominantly the full name, while other databases may apply various abbreviated forms. Figure 8 demonstrates the merger of different journal name forms for identical journals from the Inspec and SCI files. The starting point for the rank analysis is data set (A) (Fig. 3), i.e., that Inspec is the leading file holding all the duplicates.

The journal titles in full on the ranked list (e.g., rank no. 7: ''Astronomical Journal'') derive *both* from SCI and Inspec. The abbreviated title (rank no. 14) derives from Inspec. The latter can thus be deleted from the list. Suspicion ought to arise associated with the one word titles, like ''Icarus'' (rank no. 1) or ''Science'' (rank no. 13), since we may observe similar titles, but with a country added, e.g., ''Icarus (USA)'' (rank no. 4). A close inspection shows that the 76 items for ''Icarus (USA)'' are contained in the 110 items for ''Icarus.'' The reason is that one-word titles in Inspec are seen as abbreviations and, in addition, get the publishing country added. The list must hence be edited according to the practice of abbreviation applied to the involved files. Had we used data set (B) as the set to be analyzed, the resulting list would look identical after a similar editing process. The difference would be that on top of the non-edited list, we would primarily find the name forms in full, deriving from the leading SCI file. The unique journal identification code, CODEN, should be applied when available.

Note that rank numbers 2–3 signify a single journal uniquely analyzed in Inspec and not covered by SCI. Like for the journals ranked as 9–10 and 17–18, the range of articles published by these journals would have been omitted from the analysis if the data set only had consisted of SCI items.

Indexing (in)consistency is a common issue in retrieval of information and consequently in bibliometric analyses. It cannot be avoided. The problem is not only that various files in the same domain apply different vocabularies but also that indexing changes over time. In addition, the specificity of indexing the same items may vary from file to file. In creating a data set, these problems may be approached by designing more topically adequate

search profiles which include synonyms and closely related terms to the area of interest. However, if subject or domain analysis of a data set is the goal, the ranked list of indexing terms must be post-edited in order to include semantic control. Analysts should be aware that only very recently (1991), SCI introduced an uncontrolled descriptor field in which few indexing terms are generated by the author—not an indexer—as well as an automatically generated Keyword-Plus field.

A final issue consists of the recent introduction of full-text fields in the existing databases. Therefore when searching the basic index, as done in the data sets (A/B), in order to obtain a comprehensive set, one must be aware of this full-text possibility in the files involved. ABI-Inform on management and organizational topics has recently been amended with full-text. Results of counting analyses may consequently be heavily distorted due to this fact. In such files, the basic isolation of a data set should hence exclude, or at least isolate, the full-text field (/TX).

## 5. Concluding Resume and Discussion

Comprehensive and structured data set isolation followed by publication counting is a mandatory process prior to citation analyses. In relation to publication counts, the nature and structures of the databases involved limit the parameters available. Fundamentally, the following parameters are commonly available and used, both in relation to data set isolation and publication counting:

| Specific parameters: | General parameters: |
|---|---|
| Disciplines/domains/topics | Time constraints |
| Subject categories | Document types |
| Comprehensiveness/sampling | Languages |
| World/region/country(ies) | |
| Institution(s)/corporation(s) | |
| Journal(s) | |
| Author(s) | |
| Database dependent criteria— e.g., research treatment (Inspec) | |

In relation to both set generation and publication counting, the right-hand parameters pose only minor problems that generally are easily solveable, e.g., the document type issue raised in connection to Figure 2. The major problems occur in connection with the left-hand side parameters. The largest problem related to set generation (and counting) is the issue of corporate source omission for secondary authors in nearly all files except the ISI databases. Only the affiliation of the first author may be isolated and calculated. However, also the ISI files demonstrate severe problems concerned with set isolation and the follow-up analyses. Foremost is the question of comprehensiveness. Although the 17 ISI subject categories may be very suitable for certain types of overall analysis to be made, the limited indexing available in these files makes them, as stand-alone data sources, not suited for specific topical set isolation or in-depth subject and domain analysis. They should be supplemented by the core domain-related databases, as demonstrated (Figs. 7–8). Another facet of comprehensiveness adheres to the level of subject exhaustivity in the isolated data set. The question is: Should the data set consist of documents mainly dealing with the search topic, or should it encompass documents with the topic as a minor or very small aspect as well?

For any subsequent analysis a complete tuning of the initial crude data set is necessary. The tuning process implies the removal of unwanted editorials, letters to editors, and similar insignificant items, as well as the isolation of the variety of document types involved in the file(s). This streamlining operation has been demonstrated in Figure 2.

When working with a cluster of files, duplicate removal as well as reversed duplicate removal (RDR) should take place. Although very easily carried out, the analyst should be careful to check the result of the removal algorithm in each case. One advantage in applying a file cluster is the possibility of isolating the items overlapping the involved files, i.e., the duplicate documents common to the files. This ''overlap set'' may constitute statistically valid and identical samples which may be analyzed according to all the different parameters available in the merged files. On the other hand, the documents uniquely covered by the domain-dependent databases (Fig. 8) often represent the R&D production in the form of articles from specialized journals, conference papers, and reports. This substantial proportion of publications is indeed valid and carries the potential of being cited in the ISI files, although the publications themselves do not form part of the ISI files as source items.

By being aware of the idiosyncrasies of the files, as well as the retrieval possibilities and processing software features involved in the online data set creation, it is believed that the analyst is capable of producing more valid analyses and being more critical towards externally produced analyses.

## References


Christensen, F. H., & Ingwersen, P. (1995). Fundamental methodological issues of data set creation online for the analysis of research publications. In M. Koenig & A. Bookstein (Eds.), *Proceedings of the Fifth Biennial Conference of the International Society for Scientometrics and Informetrics*, June 7–10, 1995, Rosary College, IL (pp. 103–112). Medford, NJ: Learned Information.

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication.* London: Taylor Graham.

Dialog Information Service. (1993, February). Get results with the Dialog Rank command. *Dialog Chronolog,* pp. 27–33.

Dou, H., Hassanaly, P., La Tela, A., & Quoniam, L. (1990). Advanced interfaces to analyse automatically online database set of answers. *Information Services & Use, 10,* 135–145.

Garfield, E. (1979). *Citation indexing: Its theory and application in science, technology and humanities.* New York: Wiley.

Huber, C. F. (1995, April/May). SciSearch on STN. Unique features for sophisticated searching. *Database* pp. 52–62.

Ingwersen, P. (1984). A cognitive view of three selected online search facilities. *Online Review, 8(5),* 465–492.

Ingwersen, P., & Wormell, I. (1989). Databases as an analytical tool in research management: A case study. In B. Cronin & N. Tudor-Silovic (Eds.), *The knowledge industries* (pp. 205–216). London: Aslib.

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

Miller, C. (1990, July). Detecting duplicates: A searcher's dream come true. *Online,* pp. 27–34.

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

Norretranders, T., & Haaland, T. (1990). *Dansk Dynamit: Dansk Forsknings Internationale Status Vurderet ud fra Bibliometriske Indikatorer* [Danish dynamite: The international status of Danish research assessed by means of bibliometric indicators]. København: Forskningspolitisk Råd [Forskningspolitik, 8].

Olsen, T. B., Foss Hansen, H., Luukkonen, T., Persson, O., & Sivertsen, G. (1994). *Nordisk Forskning i Internasjonal Sammenhæng: En Bibliometrisk Beskrivelse av Publisering og Siteringer i Naturvitenskapelig og Medisinsk Forskning* [Nordic research viewed internationally: A bibliometric description of publishing and citations in scientific and medical research]. København: Nordisk Ministerråd [TemaNord 1994: 618].

Pao, M. (1993). Term and citation searching: A field study. *Information Processing & Management, 29(1),* 95–112.

Persson, O. (1986). Online bibliometrics. A research tool for everyman. *Scientometrics, 8,* 69–75.

Persson, O. (1988). Measuring scientific output by online techniques. In A. F. J. van Raan (Ed.), *Handbook of quantitative studies of science and technology.* pp. 229–252. Amsterdam: North-Holland.

Sandison, S. (1989). Thinking about citation analysis. *Journal of Documentation, 45,* 59–64.

Schubert, A., Glänzel, W., & Braun, T. (1989). World flash on basic research. Scientometric datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields 1981–1985. *Scientometrics, 16(1–6),* 3–478.

Seglen, P. O. (1989). Evaluering av forskningskvalitet ved hjælp af siteringsanalyse og andre bibliometriske metoder [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.

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

Wissmann, C. (1993). Techniques of data retrieval for scientometric research in the ISI citation indexes. *Journal of Information Science, 19,* 363–376.