Back to Contents Page

Chapter Three

The Design and Production of a Citation Index

There is little, if anything, that is simple at all levels of reality. Citation indexing is no exception. At the conceptual level, it provides a simple way around the semantic, intellectual, and economic problems of traditional subject-term indexing (see Chapter One: "A Conceptual View of Citation Indexing"). At the implementation level, however, citation indexing loses much of that simplicity, particularly when it is being implemented on a large scale.

The basic implementation complexities lie in the areas of coverage and production, which are related in a direct way by economics: the more efficient the production process, the more literature one can afford to cover. Though the nature and magnitude of the complexities vary somewhat according to the scope of the index (see Chapter Two: "An Historical View of Citation Indexing"), the best illustration of them, and of their relationship, is the *Science Citation Index*. Because its scope is so much broader than any other citation index to scientific literature, it demonstrates better than any other what is involved in realizing the single most important attribute of the citation-indexing concept—the ability to provide an integrated view of the scientific literature that is unrestricted by disciplinary boundaries.

THE COMPLEXITIES OF COVERAGE

There are three measures of the coverage of a citation index. One is the number and variety of journals from which references are obtained for indexing. Another is the number, variety, and time frame of the references.

In terms of the first measure, *Science Citation Index* (*SCI*) covers several thousand journals from literally every scientific discipline. As for the second measure, all references listed in all original articles, editorials, letters, meeting reports, and notes are indexed without restriction. This means that the cited material listed in *SCI* is not

limited by either journal, publisher, or publication type: everything an author references is listed, regardless of where it was published or whether it took the form of a journal article, book, thesis, letter, or report. Nor are the reference citations limited by time period. If references are made to works by da Vinci or Copernicus, they will be included.

By combining these two measures, the annual coverage of *SCI* can be defined as consisting of approximately 500,000 source articles and some 7 million references from 3000 to 4000 journals and multiauthored books of all scientific disciplines. The references identify over 3 million unique cited items, which consist of both journal and nonjournal material, deal with all the subjects of science, and stretch as far back in time as the authors' work took them.

The third measure of coverage is qualitative. Ideally, a comprehensive citation index to the journal literature of science might be expected to cover all the scientific journals published. For a number of reasons, however, this is impractical—and may even be impossible. One reason is that no one knows how many journals are published, because there is no agreement on what constitutes a journal. Some serials appear only once a year—a frequency that throws considerable doubt on any claims that they are journals. Many so-called scientific journals that appear more frequently publish little, if any, material that is a serious attempt to help solve research problems. And many more journals do not last long enough to earn serious consideration.

Another thing that makes the ideal impractical is economics. After eliminating all the serials that suffer from the shortcomings just described, there probably are something on the order of 10,000 left whose intent, frequency, and endurance qualify them as scientific journals. If each of these journals publishes an average of 100 articles a year, the total universe to be covered by a comprehensive citation index would be 1 million source articles a year. Considering that the average article covered in *SCI* requires the creation, entry, storage, and manipulation of a computer record some 1000 characters long, just the data entry and computer costs would make the economic feasibility of complete coverage rather shaky under most real-world circumstances. Economics, therefore, dictates that even a comprehensive citation index must be selective.

Because the problem of coverage is one of practical economics, the criterion for what is covered is cost effectiveness. The cost-effective objective of an index is to minimize the cost per useful item identified and to maximize the probability of finding any useful item that has been published. One factor in achieving this objective is the efficiency with which the index is produced. Another factor, since it costs as much to index a useless item as a useful one, is the utility of the items covered. A cost-effective index must restrict its coverage, as nearly as possible, to only those items that people are likely to find useful.

This is not as impossible as it sounds. The trick is to identify the journals that publish the highest quality material. Expert practitioners in a field can do this easily enough. The difficult part of the job lies in trying to make the coverage as complete as possible by expanding it beyond the core of journals whose importance to a given field is obvious.

In 1953 S.C. Bradford described the difficulty when he wrote in *Documentation* (1), "Articles of interest to a specialist must occur not only in the periodicals specializing in his subject, but also, from time to time, in other periodicals, which grow in number as the relation of their fields to that of the subject lessens, and the number of articles on his subject in each periodical diminishes." In simpler terms, the pursuit of complete coverage of the literature pertinent to a given field takes one farther and farther afield; and the farther away you go, the more journals must be added to the collection to improve coverage. A physical analogy of the situation described by Bradford would be a comet, with the nucleus representing the core journals of a literature and the debris and gas molecules of the tail representing the additional journals that sometimes publish material relevant to the subject. The tail becomes wider in some proportion to the distance from the nucleus.

Bradford first demonstrated this law in a study of the literature of electrical engineering. Others showed that it held true for other segments of the literature as well. On the strength of Bradford's insight and other subsequent findings, information scientists developed a rule of thumb that said that somewhere between 500 to 1000 different journals are required to obtain 95% of the significant literature published in a given field. In other words, an index attempting to identify 95% of the significant journal literature in a single, given field would have to cover 500 to 1000 different journals.

On the surface, Bradford's law would seem to imply that it is economically impossible to provide anywhere near complete coverage of all the literature of science in a single index. Too many journals—500 to 1000, multiplied by the number of disciplines involved—would have to be covered. But that is not the way it works.

All that Bradford talked about was the number of journals involved in publishing the literature of a single field. He did not say that each group of journals was unique to its field, and he did not say anything about how much the journals in one field might overlap other fields.

It turns out that there is a very significant degree of overlap. The evidence of it takes the form of numerous studies showing that relatively few journals are involved in the publishing of an overwhelming majority of the material important enough to be referenced or abstracted. One study of the *SCI* data base (2) shows that 75% of the references identify fewer than 1000 journals, and that 84% of them are to just 2000 journals (see Figure 3.1). The same study also showed (see Figure 3.2) that 500 journals accounted for 70% of the material indexed in *SCI* in 1969 and that almost half of the 3.85 million references published in *SCI* that year came from ony 250 journals (see Figure 3.3).

The same kind of concentration has been shown in studies of the two major services abstracting the chemical literature. A study of the 1974 edition of *Current Abstracts of Chemistry and Index Chemicus*™† (3) shows that 20 journals accounted for 68% of the new compounds announced, that 40 accounted for 88%, and that only 43 journals accounted for 90% of the compounds. A study of *Chemical Abstracts* (4) showed that only 8% of the journals it covers were responsi-

†A trademark of the Institute for Scientific Information.

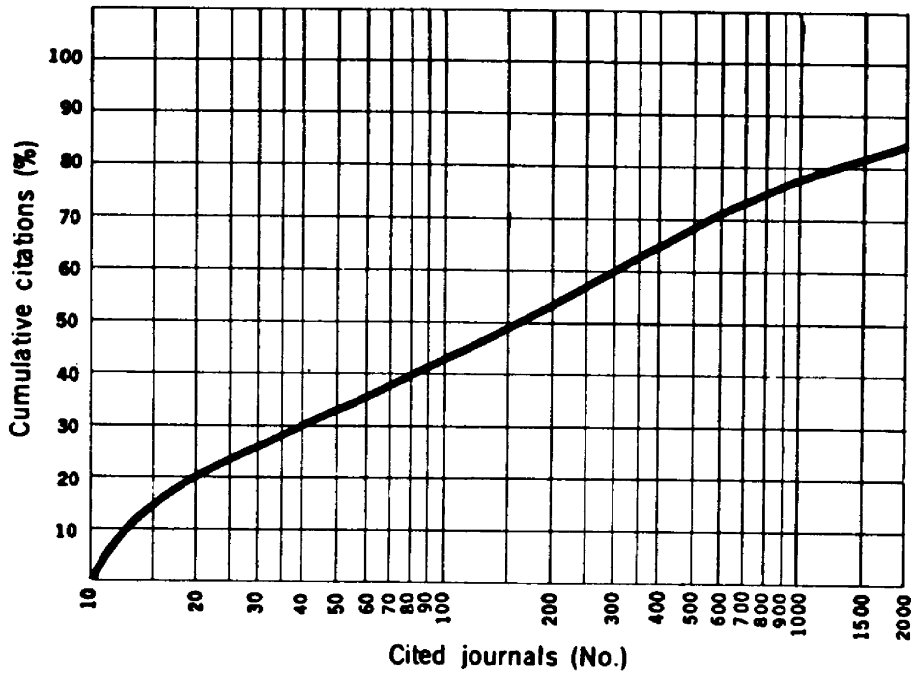


Figure 3.1 Distribution of citations among journals cited in *SCI* shows that 75% of the references identify fewer than 1000 journals and that 84% of them are to just 2000 journals.

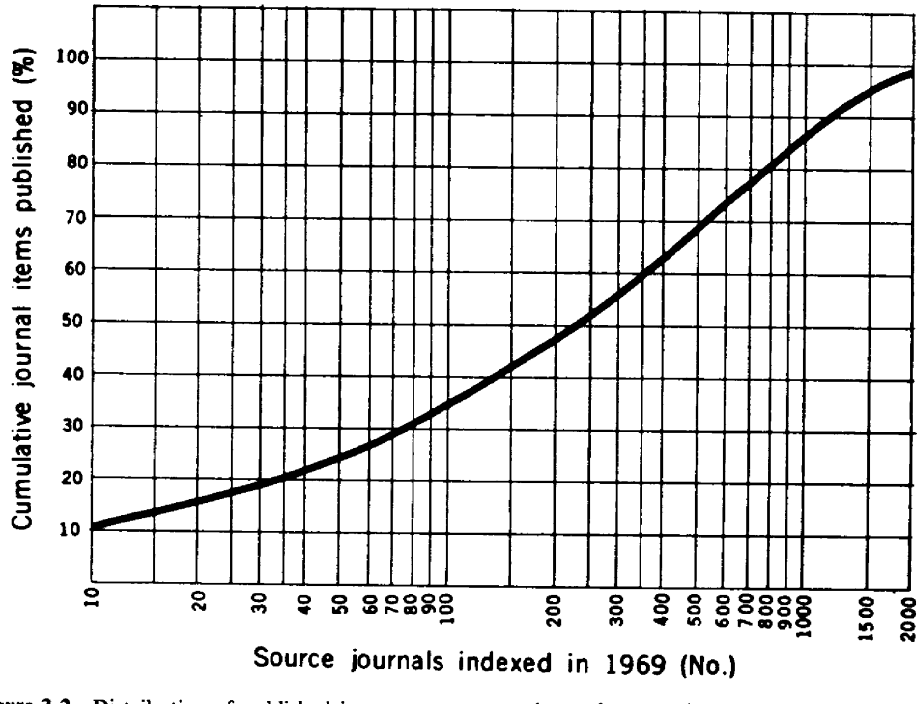


Figure 3.2 Distribution of published items among source journals covered by *SCI* in 1969 shows that 70% of the material indexed was published by only 500 journals.

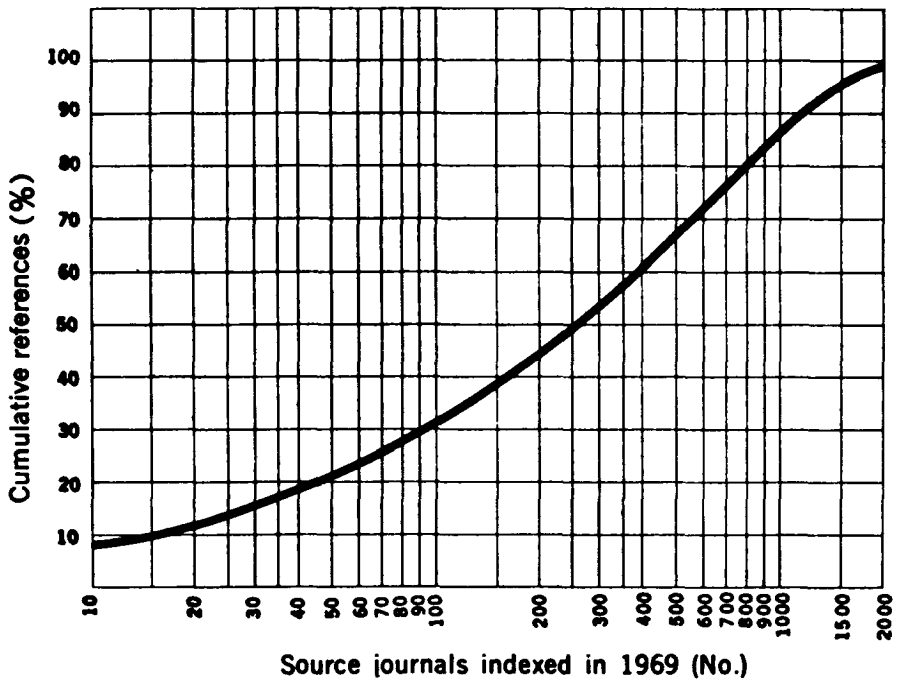


Figure 3.3 Distribution of references among source journals covered by *SCI* in 1969 shows that almost half of the 3.85 million references came from only 250 journals.

ble for the publication of 75% of the items it considered important enough to abstract.

This type of evidence makes it possible to move from Bradford's Law of dispersion to Garfield's law of concentration (5), which states that the tail of the literature of one discipline consists, in a large part, of the cores of the literature of other disciplines. So large is the overlap between disciplines, in fact, that the core literature for all scientific disciplines involves a group of no more than 1000 journals, and may involve as few as 500. In less abstract terms, this means that a good general-science library that covers the core literature of all disciplines need not have any more journals than a good special library that covers all the literature of a single discipline.

Consisting of the total output of some 3000 journals, the coverage of *SCI* is well past the core of all scientific journal literature, and into the area where cost effectiveness becomes critical. For that reason, the judgments that must be made about the quality of the journals being covered, and being considered for coverage, are taken very seriously and are based on the best information available.

Interestingly enough, the best information available consists of the statistics on how frequently journals are cited that can be generated from the *SCI* data base (the collection of computer records from which each edition of *SCI* has been compiled). Since authors refer to previous material to support, illustrate, or elaborate on a particular point, the act of citing is an expression of the importance of the material. The total number of such expressions is about the most objective measure there is of the

material's importance to current research. The number of times all the material in a given journal has been cited is an equally objective and enlightening measure of the quality of the journal as a medium for communicating research results.

Two kinds of citation data are used to measure journals. One is a straight citation count: the total number of times a journal has been cited in a given year. The other is something called "impact factor." When a journal, as opposed to a single article, is being measured, the total number of items published by the journal influences the number of times it is cited; the more it publishes, the greater the number of opportunities it has to be cited. Given a large and small journal of equal quality, the large one will be cited more frequently than the small one. The impact factor discounts this advantage of large journals by showing the average citation rate per published item. This is done by dividing the number of times the journal has been cited by the number of items it has published.

Both of these measures are used in a continuing series of studies to keep *SCI*'s coverage cost effective. They are used to monitor journals already being covered; to spot journals not now covered that merit consideration; and, when cost considerations make it necessary, to decide between competing journals that have equal editorial board support for inclusion. One notable example of the value of these measures in keeping *SCI*'s coverage cost effective took place in 1968. An analysis of journal citation rates showed that the Russian journal *Teploenergetica,* which was not covered by *SCI,* was among the 500 most frequently cited journals in the world. It was added the next year.

New journals sometimes impose particularly difficult coverage decisions. It is desirable to cover the worthwhile ones as quickly as possible. Since it usually takes two to three years for the citation rate of a published item to peak, unless it is in a particularly "hot" field, citation counts are not usually relevant to the evaluation of a new journal. In this situation, the people who worry about *SCI* coverage look at such factors as the reputation of the publisher; the geographic representation of the journal's editorial board; its reliability in meeting scheduled publication dates; and its format and bibliographic standards as reflected in article titles, references, authors' addresses, and abstracts. If it is published in a language other than English, the inclusion of English abstracts or summaries is vital. While the journal may translate article titles into English on its contents pages, it may omit this information in the abstract. New journals that score well on these counts are submitted to the editorial advisory board. Those that are added are then monitored annually by citation analyses reported in the *SCI* volume called *Journal Citation Reports.*® *

The purpose of all these activities is to make sure that no significant new or old journal is omitted from coverage. If a journal is picked up in a few years, we will go back to process earlier volumes for the five-year cumulations.

Once we are certain that the best are included, it is extremely difficult to decide which of the hundreds of remaining journals to add. Fundamentally, the decisions are economic ones, since the continued existence of a research journal is an obvious indicator that it is important to someone. In 1978, ISI made the basic decision that

**Registered trademark of the Institute for Scientific Information.*

any journal suitable for coverage in any *Current Contents* edition should eventually be included in the *SCI* or *SSCI,* since this would ensure uniform processing of over 5000 journals in the system.

The key step to ensuring complete coverage within specialties is to do a field-by-field citation analysis. In this way all significant journals within the field are ranked by impact and citations. It is difficult to imagine an important journal escaping this citation net. But as the coverage of the system is increased to the point where even large numbers of mediocre journals are processed as source journals, then one must evaluate many journals that have purely regional or local value. The failure to include a particular journal may often be interpreted as having political significance, since coverage in *Current Contents* and the *SCI* is often a matter of prestige.

PRODUCTION EFFICIENCY

The other side of the cost-effectiveness coin, production efficiency, is more critical than is generally recognized. The production of citation indexes is more involved than is generally appreciated. Although citation indexing eliminates the expensive intellectual effort associated with traditional subject-term indexing (see Chapter One), producing a citation index of appreciable size is a massive materials-handling and information-processing job.

The job of producing the *Science Citation Index,* as well as the *Social Sciences Citation Index,* begins with "editing" or screening the individual journals that are covered. Every article or editorial item must be examined to determine whether it should be covered. Every item other than minor news notices and advertisements must be marked in some way to simplify the huge job of converting the information into machine language. This so called "pre-edit" process also helps standardize the information that enters the system. Pre-editing involves coding each item as to type, that is, an article or a technical note or editorial. The first and last page of each article must be identified and labeled. In many journals, especially in the social sciences and humanities, extensive marking of cited references is required. Reference formats may differ not only from journal to journal but even from article to article. Pre-editors also identify titles that must be translated into English. In addition, considerable editing must be done of titles, author names, organizational names and addresses, and references.

Titles must be marked and edited to show where they begin and end; eliminate unnecessary words; add pertinent footnote annotations; and standardize punctuation, numerical expressions, and proper names. Scientific notation must be edited to meet rules of standardization and computer processing requirements.

Author names and addresses must be underlined, and each name must be coded to distinguish between primary and secondary authors. Author names must be standardized, too; this includes non-English names, for which the rules of standardization are quite involved. The organizational names in author addresses also must be standardized.

References interspersed throughout the text or split between the text and foot-

notes, as well as footnotes that contain multiple references, can be the toughest part of the editing job. Most often found in social sciences journals, these types of references require extensive editing notation to identify, integrate, and complete them, and may require the help of a professional translator if they involve non-English citations.

Editing time per journal issue varies anywhere from half an hour to three days. Journals dealing with the social sciences are generally the most time consuming because their bibliographic standards tend to be archaic and their references are frequently complex, often citing exotic types of nonjournal material, such as rare documents, legislation, and laws. It is not unusual to find references scattered throughout the text of a social sciences journal, which means the editor must scan the entire article. Footnotes containing multiple references are common; and the format of references, regardless of where they are found, is eclectic enough to make reformatting the rule rather than the exception. The impact of these problems on productivity is great enough to justify a continuing and sizeable effort to educate editors about the reader and the economic advantages to be gained from adopting simpler, more standardized format rules.

The next production step is putting the edited material into the computer. With *SCI* and *SSCI* this job is done by over 100 data entry operators working two shifts, five days a week. These specially trained "indexers" use keyboard-display terminals connected directly to a central magnetic disk memory. The journals move through this process in large batches. Of course, recording formats have been specified in advance and job control numbers have been assigned to each batch. As part of the job control procedure, individual journals are logged into the system, when assigned to a batch, by name, volume, issue, month, year, accession date and number, and a status-and-date statement. After that, the status-and-date statement in the system log is updated every time the journal moves from one operation to another.

Once a journal has been assigned and logged, it goes to a data entry operator, who verifies and updates the log to let the system know what journal is being worked on and where it is. The operator then works through the journal article by article, keying the pertinent information from each into the system in a three-part sequence.

First comes the basic information that identifies the article: its type, title, page numbers, and primary author. The middle part of the data entry sequence involves additional author information: the address of the primary author and the names and addresses of any secondary authors. The last part of the sequence deals with the references cited in the article.

When all the information about all the articles in the journal have been entered into the system, the operator lets the system know the journal is finished by updating its log. Another operator then goes through the entire data entry sequence again, character by character, to verify the work of the first operator.

Periodically, the verified records created for batches of journals are automatically transferred, under the control of someone working at a supervisory terminal, from the magnetic disk to a magnetic tape. As the records are transferred, they also are reformatted for computer processing.

The data entry workload for *SCI* and *SSCI* can be defined by a variety of numbers, all of them large. Over 2000 source articles, involving some 25,000 cited references, are processed each day. With the record length per source article averaging 1000 characters, the total number of characters entered each day exceeds 2 million. And, if the verification operation is included, the total number of keystrokes per day is something on the order of 4 million.

At this point in the production cycle, the computer takes over, and the information goes through the sequence of processing operations shown in Figure 3.4.

The first step in this sequence illustrates the primary difficulty of preparing scientific information for computer processing. Despite the pains taken in the editing operation to identify, clarify, and standardize everything, and the character-by-character verification performed in data entry to assure keying accuracy, the first thing the computer must do to the tapes from data entry is edit them to make sure that all the records are complete and properly formatted. Some 1% are not and must be recycled through editing and data entry a second time.

Besides checking the content and format of the individual records, which are organized by journal, the computer also checks the journals against a year-to-date file of all the journal issues that have already been processed. Duplicates are recycled back through the journal control people to work out the problems. Those that are not duplicates are copied onto the year-to-date file.

The tapes from data entry are edited this way on a daily basis and accumulated into a weekly data base, which is edited again to verify the daily checks of content and format.

The edited weekly data base is then coded to show what journal records go into what index. Working from information in a master journal file, the computer codes each journal for one of three categories: *SCI* only, *SSCI* only, or both *SCI* and *SSCI*. It then looks at all the articles in the *SCI*-only journals and determines, on the basis of their title words and references, which ones qualify for inclusion in *SSCI* in addition to *SCI*. In 1977, this procedure was expanded to deal with the production of the *Arts & Humanities Citation Index*.

The records on the coded weekly data base are then sorted into four data categories: source data (bibliographic descriptions of the published articles from which the references are taken), citation data (the references made in the source articles), corporate data (names and addresses of the organizations with which the authors of the source articles are affiliated), and patent data (bibliographic descriptions of patents that have been cited in source items). All these categories, except for the one of patent data, correspond to major sections in both *SCI* and *SSCI*; the patent data is included only in *SCI*.

The rest of the weekly processing cycle consists of one refining operation and the creation of separate data bases for each of the two indexes. The refining operation involves the "post-editing" of the source-data file to assure the accuracy of the title information. Post-editing is also concerned with maintaining the currency of a title-word index included in *SCI* and *SSCI* to help users who do not have the name of an author with which to start a search. This is done by checking every key word in every

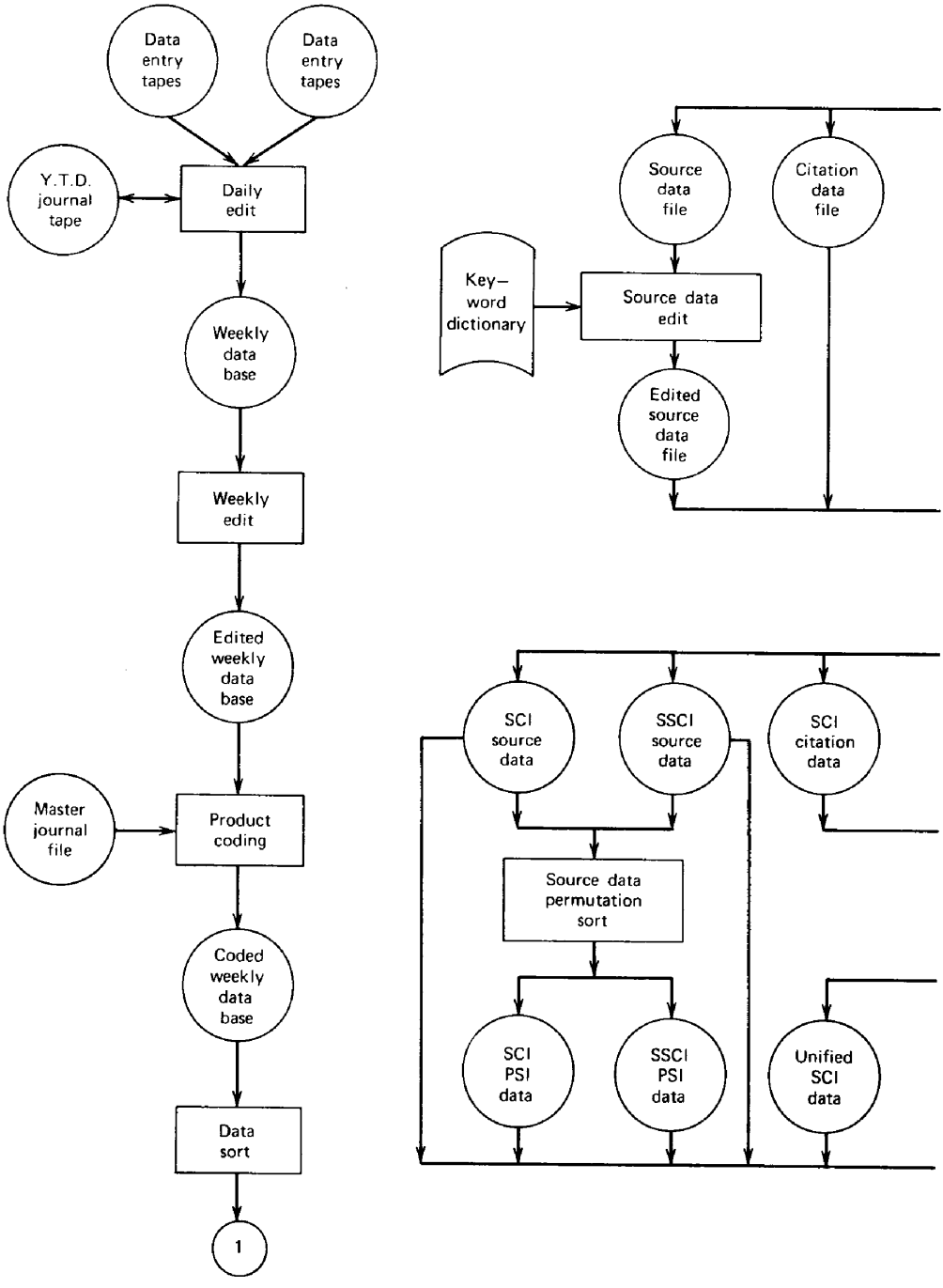
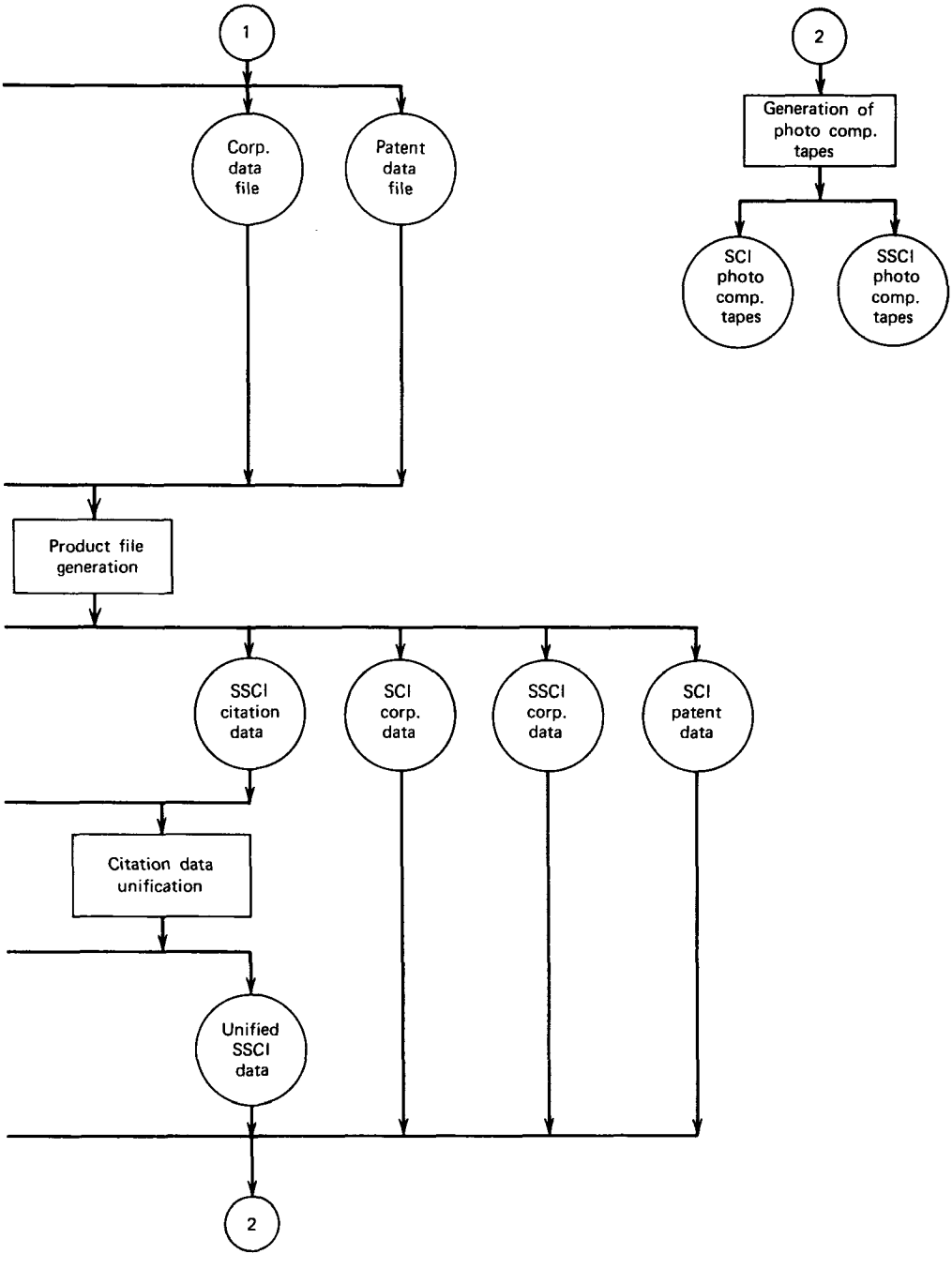


Figure 3.4 Computer processing operations in the production of *Science Citation Index*.

title on the file against a key-word dictionary. Words not found in the dictionary are passed on to editors, who determine whether the words are misspellings of valid words or are words not yet included in the dictionary. Misspelled words are corrected; new ones that are judged to be important are validated and added to the dictionary.



The final operation in the weekly processing cycle is generating separate *SCI* and *SSCI* counterparts of the four files of source, citation, corporate, and patent data. This is done on the basis of the product coding performed earlier.

The rest of the computer processing is done on quarterly, triannual, and five-year cycles. The quarterly cycle is concerned with integrating a three-month cumulation

of weekly, *SCI* data files and preparing them for publication. It starts with two operations: the unification of the citation data file and the production of a file for the *Permuterm® Subject Index,* or *PSI,* which is the only component of *SCI/SSCI* for which no file is created during the weekly processing cycle. ("Permuterm" is a registered trademark of the Institute for Scientific Information constructed from the phrase "permuted terms.")

The unification of the citation-data file is concerned, again, with accuracy—this time, with the accuracy of the authors and journals cited in the references. To maximize this point of accuracy, the records on the citation files are sorted by author name, journal, volume, year, and page numbers, so that multiple references to the same citation are batched together. The computer then looks for inconsistencies in the names of the authors and journals cited by each batch of references. Any that are identified are resolved according to rules built into the computer program. The result of this operation is citation data in which the variations that inevitably creep into the spelling of author and journal names have been minimized, if not eliminated.

The *Permuterm Subject Index* is the title-word index mentioned earlier in connection with the editing of the source-data file. Producing a file from which the *PSI* can be published involves permuting all the key words in the source article titles and sorting the word pairs into alphabetical order.

All the *SCI* data files are then consolidated and reformatted by a routine that organizes the material into pages and specifies formats and type fonts. The tapes produced by this routine are used to drive an automatic photocomposition machine, which turns out reproduction-quality page proofs from which offset negatives and plates can be made for printing.

At the time the last quarterly index is prepared, the material for that quarter is consolidated with what had been published in the preceding three quarters to produce a cumulative annual index. Every fifth year, the material for the annual index goes through an extra processing operation in which it is consolidated with the material published in the four preceding annual editions to produce a five-year cumulative edition. This involves considerable changes in the *PSI,* which is refined and made more specific by looking at the frequency with which subject terms occur.

The same thing is done with the *SSCI* files on a four-month cycle.

There are a number of additional accuracy checks in this final stage of production. The first is a detailed check of the first statistically significant batch of pages produced by the photocomposer. The accuracy of the weekly data bases and the effectiveness of the computer routine that merges them are checked by matching a random sample of articles that should be covered in the initial pages against the page proofs. The effectiveness of other key computer programs also is checked in this initial sample by looking for discrepancies and omissions in names, cross references, formats, special signs, and features new to the index. If everything is all right, the rest of the pages are produced. These too are checked, but for such things as print quality, the number of columns per page, and the sequence of columns—all things that can go wrong in the photocomposition stage of production. Only then is the job

released to the printer; the printer's work, too, is spot checked, but for all the things that can go wrong in the printing process.

Even at computer speeds, the amount of information processing required to produce a ciation index the size of *SCI* is staggering. Nearly 200 computer hours are required to go from the raw material that comes in daily from data entry to the weekly *SCI* files that are ready for quarterly processing. Another 25 hours of computer time is needed every quarter to go from the consolidated weekly files to the photocomposition tapes. In the last quarter, when the material for the entire year is being consolidated, more than 230 hours is needed to produce the photocomposition tapes, and the five-year cumulative edition takes some 2800 more hours of computer time.

SSCI, which is one quarter the size of *SCI,* uses an additional 60 hours or so of computer time to get to each of the first two sets of triannual photocomposition tapes and approximately another 60 hours for the last, annual cumulative set.

ROLE OF TECHNOLOGY

Theoretically, it is possible to produce a citation index without the aid of computers, though one of the advantages of the concept is the very neat match between its production demands and computer capabilities. From a practical viewpoint, however, computer technology (or, more accurately, information-processing technology) is critical to the cost effectiveness of a comprehensive citation index of the *SCI* type. In keeping with the coverage-production relationship mentioned earlier in the chapter, it improves the scope and depth of coverage that is economically practical by reducing the cost per item indexed.

Exploiting the potential of computer technology for this purpose is a matter of continually searching for production efficiencies among the technological advances. Some of the efficiences are built into the lower cost per unit of processing offered by succeeding generations of equipment and can be realized merely by upgrading the equipment periodically. Other, more significant efficiencies call for the ability to innovate from the improved functional base provided by the new equipment. The impact that key-to-disk data entry equipment has had on *SCI* production is a case in point.

For a data entry operation as big as the one involved in *SCI,* key-to-disk systems are more efficient than the older keypunch. Job control procedures are easier to implement; keying is done at electronic, rather than mechanical, speeds; and a lot of punched-card handling is eliminated. In addition, each terminal operator has access to a central disk memory, around which ISI has built a production innovation that increases efficiency far beyond the level made possible by the superior speed of key-to-disk systems.

The innovation, called Keysave (6), consists of using the shared disk memory to store an historical file of reference citations from the *SCI* data base. The increase in efficiency comes from reducing the amount of keying necessary to enter and verify reference citations. Instead of keying the full citation, the operator keys in a 14-

character code abstracted from the full citation. Each of the citations on the historical file has attached to it the same sort of coded identifier. If the code the operator enters matches one in the file, the full citation is brought up on the terminal display, where it is verified visually and entered on the disk with a single keystroke.

Every time an operator matches a reference citation against one in the historical file, the number of keystrokes required to enter the citation is reduced from an average of 70 to 14, and the keystrokes normally required for verification are eliminated completely.

The match rate achieved depends on the number of citations in the historical file, which is limited by the size of the central disk memory available with the system. Initially, the file contained enough citations to produce a match rate of 75% on the references that cited journal material. Some changes in the design of the file increased the utilization efficiency of the central memory enough to push the match rate to 85%. Whether the rate can be raised still higher is uncertain, depending upon available memory capacity and how efficiently it is used. Match rates vary considerably from journal to journal. A journal in molecular biology will have rates in excess of 90%, while a rate less than 20% is common for journals in the social sciences and humanities.

Another example of how technology can be used to reduce costs is provided by a system improvement recently implemented in the editing operation. This involves the coding of organizational names and addresses, which pose accuracy and standardization problems in any index.

The entire editing operation associated with making sure that organizational names and addresses are accurate and consistent has been reduced to looking up and writing down two alphanumeric codes. One identifies the name of the organization; the other, the specific department or other organizational unit. The computer uses these codes to pick up full names and addresses from its own file of organizational data.

The productivity impact of this way of handling organizational names and addresses will not be limited to the editing operation. It also will be felt in data entry, where the use of the codes will significantly reduce the number of keystrokes needed to enter organizational names and addresses.

The role of technology in the production of a citation index goes beyond cost cutting into quality improvement. In some cases, the two can be combined, such as with the system for organizational names and addresses, which will enhance the quality of *SCI* by raising its level of standardization. Such an improvement is not an abstract achievement. More consistency and accuracy in organizational names and addresses make it easier for users of the index to contact the authors of source items for reprints of useful papers or for additional information.

More often than not, however, such quality improvements do not go hand in hand with cost reductions; they must be important enough to justify an increase in production costs. A computer-based system that automatically monitors the arrival of journals and tracks them through the processing mill according to a planned schedule, for example, is more expensive than doing the same thing manually. But it does a better job, which produces the important qualitative benefit of increasing the timeliness and comprehensiveness with which *SCI* covers its defined journal base.

FRUITS OF THE LABOR

The object of all this attention to the niceties of coverage and productivity is a multivolume, five-part index to that portion of the scientific journal literature published each year that is most likely to be useful—no matter what particular discipline or speciality is being researched.

The two key parts are the *Citation Index* and the *Source Index.* The *Citation Index* (see Figure 3.5) connects items published during the year with past items they have cited in references. It is organized alphabetically by cited author, using the last name of the first author. Under each cited author are listed, chronologically, the items that have been cited in references. Under each cited item are listed the sources of the references.

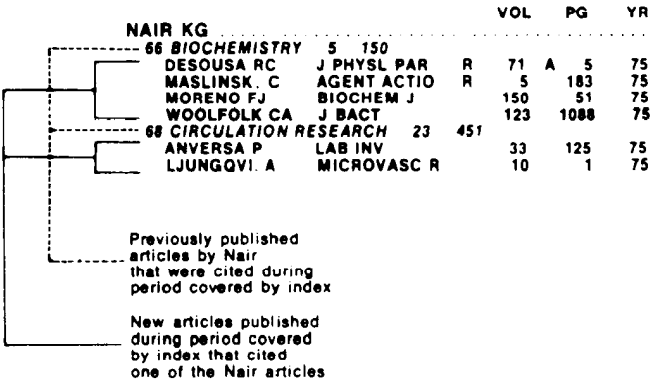


Figure 3.5 Typical entry from the *Citation Index* section of *SCI.*

Even anonymous items that have been referenced are included. Listed in a separate section, they are organized by journal, organization, and title.

Both the cited and source items are described in the same way (with minor exceptions for anonymous items): by the last name and initials of the first author and the name, year, volume, and page number of the publishing journal. The only thing missing that might be pertinent is the title of the item. In the case of a cited item, the title is unnecessary since it is reasonable to assume that the user must know what it is if he is searching for items that cited it. That, of course, is not the case for the source items, and that is the reason for the *Source Index.*

A straightforward author index to the items published during the year, the *Source Index* (see Figure 3.6) also is organized alphabetically by the last name of the first author. For each source item listed, there is a full bibliographic description: full title; last names and initials of all authors; address of the first author; name, year, volume, and page numbers of the publishing journal; language in which the item was published; and number of references made in the item.

In the *Source Index* for *SSCI,* and the *Arts & Humanities Citation Index,* this description is supplemented by one additional piece of bibliographic intelligence: a list of the reference citations that appeared in each item. Again, as in the *Citation Index,* the reference citations do not include the title. But they do provide a type of abstract

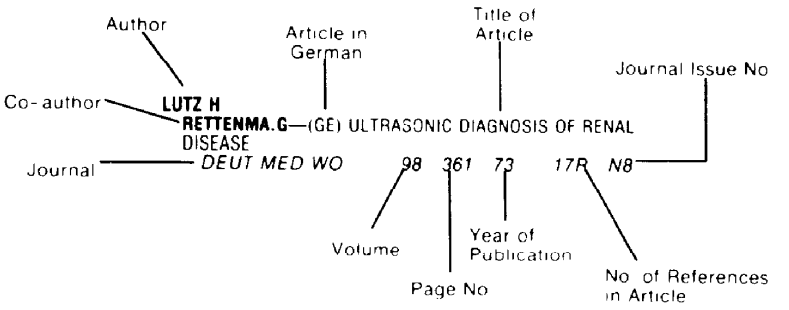


Figure 3.6 Typical entry from the *Source Index* section of *SCI.*

of the item that is useful in making preliminary judgments about its quality, pertinence, and subject orientation; and that is useful, too, in finding other citations with which to continue and refine a search (7).

The inclusion of reference citations in the *Source Index* entries is an expensive feature, which significantly increases both the size of the index and the cost of printing it. But, if the feature's effectiveness, as measured by subscriber utility, turns out to balance the cost, it will be added to the *Source Index* of *SCI* as well.

The other three parts of the *SCI* are the *Patent Index,* the *Permuterm Subject Index,* and the *Corporate Index*—all of which have the same functional relationship to the *Source Index* as does the *Citation Index.*

The *Patent Index* is conceptually the same type of index as the *Citation Index,* except that it deals with patents rather than journal items. Organized by patent numbers that have been referenced, rather than by authors, it provides the same partial description of source articles as the *Citation Index* and must be used in conjunction with the *Source Index* if a more definitive bibliographic description of them is required. Figure 3.7 shows a typical entry in the *Patent Index.* Besides the patent number, the reference citation shows the name of the patent holder, the country that issued the patent, and the year in which it was issued.

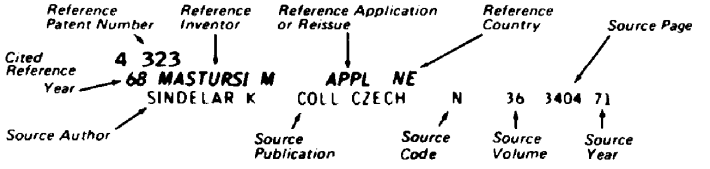


Figure 3.7 Typical entry from the *Patent Index* section of *SCI.*

The *Permuterm Subject Index* (*PSI*) is a title-word index, but one that permits more than the usual amount of search precision by enabling the user to search on a combination of two or more terms. This is done by going one step beyond the usual practice, in simple title-word indexes, of making every significant word in every title an indexing entry. Under each of these entries in the *PSI* are listed all the words with which the entry word has appeared in some title or other (see Figure 3.8). Next to each word in the list is the name of the author of the article whose title contains that particular pair of words. The author's name and the title words permit the searcher to find a complete bibliographic description of the article in the *Source Index.*

The reason for including a subject-word index in a citation index in the first place

is to give people a way of taking advantage of the multidisciplinary coverage of *SCI* even when they do not have enough information about a field to perform a citation search. The *PSI* gives them the option of bypassing the *Citation Index* completely, or they can use it to identify an article that will provide them with a starting point for a citation search. If several annual editions or five-year cumulations of *SCI* are available, an article identified by *PSI* three or four years back can function as the

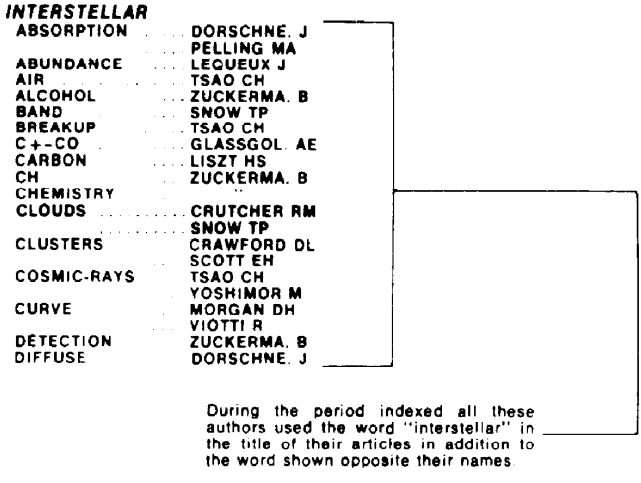


Figure 3.8 Typical entry from the *Permuterm Subject Index* section of *SCI*.

starting point for the citation search; the bibliographic description of it in the *Source Index* will give all the information needed, saving the time and trouble of obtaining useful reference citations from the full text of a paper.

PSI also serves the secondary purpose of easing people into citation searches, which are unfamiliar to many, by letting them start out with the more traditional search technique. In many cases, they begin conducting citation searches in an attempt to improve the productivity and efficiency of their subject searches.

The *Corporate Index* looks at the journal articles published during the year from the viewpoint of the organizations with which the authors are affiliated. Each organizational entry (see Figure 3.9) shows the articles that the staff has had published during the year. They are identified in the same way as entries in the *Citation Index:* by author and by the name, volume, year, and page number of the publishing journal. Again, the author's name permits the searcher to find a complete bibliographic description of the item in the *Source Index.*

After years of arranging the *Corporate Index* alphabetically by name of organization, it was decided in 1978 to provide a geographical arrangement comparable to *ISI's Who is Publishing in Science*®* . While retaining an alphabetic cross-reference system, the geographic arrangement eliminates the ambiguity of similar names located in different countries. The geographic arrangement permits one to obtain a picture of scientific publication by country or city without sacrificing the ability to

**Registered trademark of the Institute for Scientific Information.*

		VOL	PG	YR
MAX PLANCK INST BIOL,TUBINGEN,WEST GERMANY				
BISSWANG H	BIOC BIOP A	321	143	73
BRAUN V	J BACT	114	1264	73
ENGELRAE M	BIOC BIOP R	53	812	73
HENNING U	FOL MICROB	18	268	73
	P NAS US	70	2033	73
SORSA V	NATURE-BIOL	245	34	73
TICHY H	GENETICS	74	S276	73
ZARYBNIC V	VIROLOGY	54	318	73

Figure 3.9 Typical entry from the *Corporate Index* section of *SCI*.

observe the patterns of individual institutions. The cross-reference file is especially useful for multinational or multiregional organizations.

The five *SCI* indexes open up the journal literature to exploration from a variety of viewpoints for a multiplicity of purposes. The combination of search flexibility and comprehensive, multidisciplinary coverage produces a powerful tool for literature research.

REFERENCES

1. **Bradford, S.C.** *Documentation,* 2nd ed. (London: Lockwood, 1953).
2. **Garfield, E.** "Citation Analysis as a Tool in Journal Evaluation." *Science* **178**:471–479, 1972.
3. **Garfield, E., Revesz, G.S., and Batzig, J.H.** "The Synthetic Chemical Literature From 1960–1969." *Nature,* **242**:307–309, 1973.
4. **Wood, J.L.** "The Parameters of Document Acquisition at Chemical Abstracts Service." Paper presented at the American University 8th Annual Institute of Information Storage and Retrieval, Washington, D.C., February 14–17, 1966.
5. **Garfield, E.** "The Mystery of the Transposed Journal Lists—Wherein Bradford's Law of Scattering is Generalized According to Garfield's Law of Concentration." In *Essays of an Information Scientist,* Vol. 1 (Philadelphia: ISI Press, 1977). Pp. 222–223.
6. **Garfield, E.** "Project *Keysave*—ISI's New On-Line System for Keying Citations Corrects Errors." *Current Contents,* No. 7: 5–7, February 14, 1977.
7. **Garfield, E.** "Bibliographies, Citations, and Citation Abstracts." In *Essays of an Information Scientist,* Vol. 2 (Philadelphia: ISI Press, 1977). Pp. 190–191.

Back to Contents Page

To Chapter 4