# Information Obtainable from Analyses of Scientific Bibliographies

R. T. BOTTLE

WHAT MORE can be obtained from bibliographies than access to the primary literature for which they serve as keys? Because they are ready-made collections of documents about a particular facet of science or technology, they are increasingly being used as source material by historians and sociologists of science. Unquestionably they record the (recent) past; can they be used to predict the future? If the answer is even "possibly," then it would be a very valuable exercise to analyze them.

Analyses of bibliographies would provide an alternative to asking the opinion of experts about the likely course of future events, a sophisticated version of which is the so-called Delphi technique. The experts are, of course, familiar with all aspects of a subject, including its literature, and may well have assessed intuitively trends in the literature which will form part of their judgment of that subject's future.

Literature trends can be quantitatively observed without expert knowledge of a specific subject and possibly provide an industrial company, for example, with predictive material without the need to divulge its interest in the subject to an outside expert. These two approaches exhibit distinct analogies with the so-called fundamental and chartist approach to investment analysis—both of which are considered notoriously unreliable! The major difference between the approaches is the time scale involved; while investment analysis charts can point quite quickly to a change in the underlying factors involved, the publication of research is the culmination of a process where the operative decisions were perhaps made several years previously—indeed a decision to abandon further work on the project may have been made long before even a draft manuscript of an eventual publication had been

R. T. Bottle is with the Department of Librarianship, University of Strathclyde, Glasgow, Scotland.

written. In the case of a scientific research program there is usually sufficient momentum established quite early to continue the publication rate for quite some time after it has been decided to terminate it.

Trends observed in bibliographies right now, therefore, will be the reflection of events which occurred some years earlier. They can often be successfully extrapolated a few years into the future; it may even be postulated that, because of the momentum which a research program acquires, they will continue until such time as a constraint affects them (an obvious analogy with Newton's laws of motion). An example of a constraint is a change in policy by a major publisher, and this will be independent of the research scene. As publishers normally plan four or five years ahead, this period represents approximately the maximum time for which extrapolation procedures are valid.

While the literature abounds with qualitative prophecies of doom[1] that science will become buried under its own literature and that communication between research workers will break down and so forth, it is difficult to find many quantitative predictions made some time ago which can now be tested for validity. In an earlier publication[2] I was rash enough to suggest that it *might* be possible to predict the future size of *Chemical Abstracts* from the equation $\log C = 4.415 + 0.0464n$ where $C$ is the number of columns and $n$ is the number of years since 1960. This prediction, which owes much to an earlier study by Strong and Benfey,[3] was followed quite well up to 1962, but then the observed growth changed to a lower rate and was subject to much wider fluctuations than previously (see figure 1). In 1957 Crane intimated by means of a graph that the number of papers abstracted by *Chemical Abstracts* would rise to about 125,000 by 1965.[4] The number actually published in 1965 was nearly 170,000.[5] Perhaps errors of about one-third are acceptable or the best we can expect from such bibliometric predictions projected beyond two or three years. (If Crane could not predict the future of *Chemical Abstracts* any better than this, what hope is there for lesser mortals?)

As a MEDLARS bibliography on kinetocardiography dating back to the 1964 tapes had been provided for a colleague, the number of references by year of publication were counted and noted in table 1. The leader of the research group working on this project was asked if he felt interest was increasing, decreasing or remaining much the same. As his reply was that interest was increasing, it is probable that failure to observe this trend in the number of publications is a result of the Nixon administration's cutbacks in research funding in 1969-70. There are

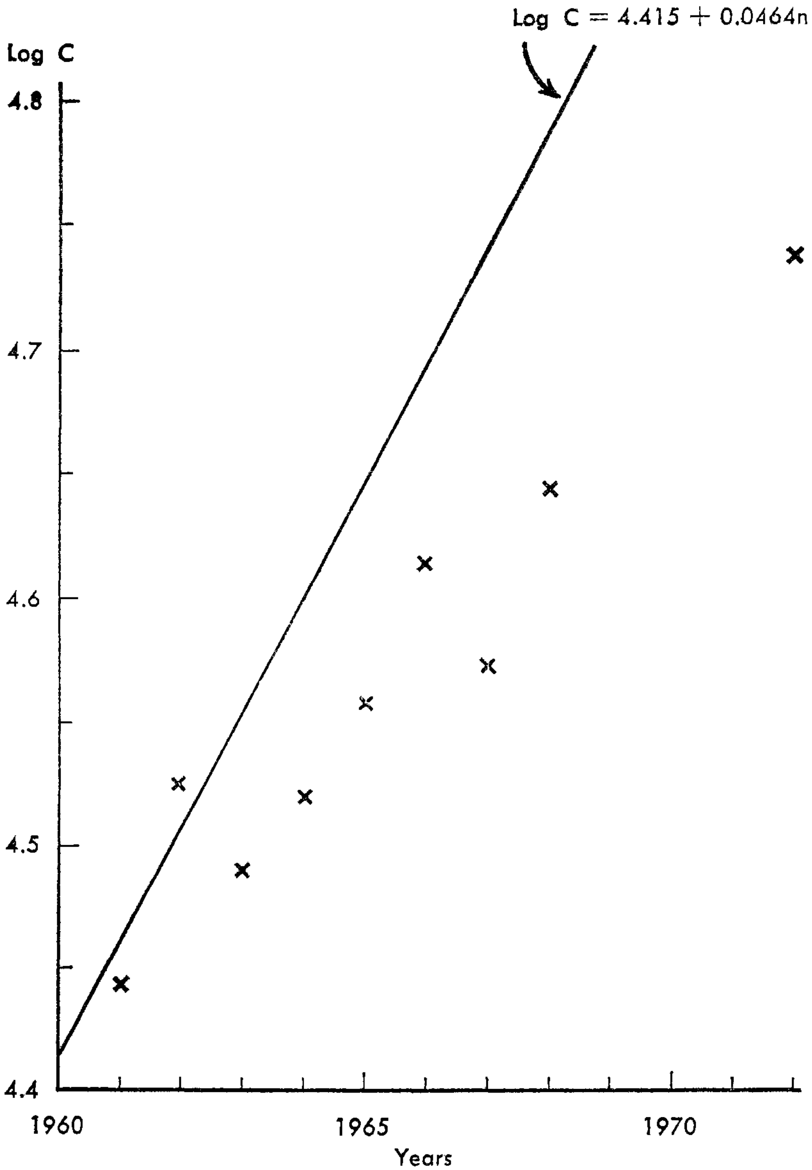


FIG. 1. Predicted and Actual Numbers of Columns in *Chemical Abstracts* vs. Year

then considerable limitations to using this type of study to predict future trends, but there is, however, much other interesting information which can be gained by studying scientific bibliographies, much of which can be of direct practical use to information scientists and their clientele.

A popular myth, especially among scientists, is that de Solla Price's 1963 book *Little Science, Big Science* was the start of interest in studying characteristics of scientific literature. It was undoubtedly a most influential and catalytic book and, in the last few years, there has been

TABLE 1

RESULTS OF MEDLARS SEARCH ON KINETOCARDIOGRAPHY

| Year | No. of References |
|---|---|
| 1972 | 50 |
| 1971 | 66 |
| 1970 | 70 |
| 1969 | 62 |
| 1968 | 72 |
| 1967 | 65 |
| 1966 | 39 |
| 1965 | 39 |
| 1964 | 20 |

Tapes were searched from 1964 to end of 1972. Because of time lags the data for 1972 (and possibly 1971) are incomplete.

a considerable upsurge in bibliometric studies. It has, however, long been recognized that scientific literature had a structure and properties. Even in 1882 H. C. Bolton proclaimed, "Chemical literature is characterized by two opposing forces, a tendency to dispersal and an effort to collect the widely scattered publications."[5] More recent, perhaps, is the realization that a knowledge of these properties and structure is important for an understanding of information transfer processes in science.

The characteristics of the producers of information and the forms in which it is disseminated are especially important. Bibliographies (and this term is used in this paper to include abstracting and indexing services) can provide the raw data for analyses which seek to quantify these points. The address lists at the back of *Current Contents* are very useful for ascertaining the types of institution in which publishing scientists work and which are the most prolific.[6] A sample of these scientists can then be traced in the author indexes of other bibliographies and their productivity in terms of papers per year can be determined. In most fields academics produce the bulk of the literature, but in computer science Pritchard states that 49.5 percent of authors were employed by firms.[7] A variant of this technique has been used in an attempt to assess the quality of British university chemistry departments using the corporate index to the *Science Citation Index* to identify the productivity of staff members and then determining the number of citations (other than self citations) received by the average staff member.[8]

TABLE 2

LANGUAGE OF PUBLICATION BY SUBJECT

| Subject | English | Russian | German | French | Japanese | Czech |
|---|---|---|---|---|---|---|
| Chemistry[9] (*Chemical Abstracts*, 1966) | 54.9 | 21.0 | 7.1 | 5.2 | 3.1 | |
| Physics[10] (*Physics Abstracts*, 1964) | 67.7 | 18.0 | 6.4 | 6.1 | 0.2 | |
| Computer Science[7] | 66.5 | 14.4 | 7.25 | 3.175 | 3.45 | 1.22 |
| Biology[11] (*Biological Abstracts*, 1967) | 68 | 14 | 7 | 5 | 1 | 1 |
| Medicine[11] (*Index Medicus*, 1967) | 56 | 3 | 9 | 11 | 3 | 2 |
| Music & Musicology[12] (*RILM*, 1967) | 46 | 2.5 | 12.5 | 9.5 | 2 | 4 |

The above procedures can work quite well in areas where the journal paper accounts for the bulk of the literature. Thus the first stage in analyzing the literature is to determine the proportion which is in the form of papers, reports, patents, theses, etc., and obviously no service which covers only papers can provide this sort of data. For this a bibliography, which has comprehensive coverage of all types of literature, is required. Similarly, a comprehensive bibliography is necessary to study distribution by country of origin or language of publication. Language of publication (especially when matched against the language capabilities of the user) is particularly important to the manager of an information service when planning translation facilities. Details obtained from several bibliometric studies are shown in table 2. Language distributions can change with time; Schwartz and Powers[13] have charted the decline of German and the rise of English for the primary literature of biochemistry, while in 1970 Webb[14] predicted that, if current trends continued, virtually all biochemical research would be published in English by 1974.

Patents and theses have long been a relatively underused literature form. A detailed investigation of patent literature is currently well advanced at the Polytechnic of North London which will hopefully shed light on patent information/literature relationships. Russian writers have discussed the use of patents statistics in forecasting.[15] A Russian study has shown how "genealogies" of individual engineering ideas can be traced through the reference sections of United States and West German patents.[16] The relationship between theses and the quite considerable quantity of information which is republished from them has recently been discussed.[6]

Bibliographies, if they are reasonably comprehensive, serve to monitor the most readily quantifiable output of science and technology—but, as J. Martyn points out earlier in this issue, a document explosion is not necessarily the same as an information explosion. Nevertheless, bibliographies serve as "catalogs of science and of scientists" and as such can serve as a convenient form of raw data for sociologists of science, information scientists and others. The following examples are not intended as a complete catalog of such studies but merely to illustrate their diversity.

A number of sociological studies on the literature of mathematics, social sciences, etc., are reviewed in Crane's recent book on invisible colleges.[17] Although it contains little of direct practical use to the information scientist, it suggests that the reason for exponential growth of the total number of publications and of new authors publishing for the first time in a new area is due to the interaction of members of a research area with other scientists. If such interaction is absent (as, for example, in English literature), a linear growth rate for these parameters would be observed.

About three-fourths of scientific papers have two or more authors and name ordering patterns in such papers, and their implications, have received considerable attention from sociologists. Zuckerman[18] reports that eminent scientists who become Nobel laureates tend not to insist on being first author to the same extent as do less eminent research team leaders, even five or more years prior to their prize. This concern (and occasional acrimony) over the "social symbolism" of the order of precedence of authors would be much reduced if far more journal editors laid down rules that authors' names should be in a strict alphabetical sequence. (With a name near the beginning of the alphabet, I can perhaps afford to champion this "democratic" method better than most!)

Creativity as a function of the scientist's age has been much studied. Lehman[19] presented evidence that chemists reached maximum creativity in their mid-thirties. From a bibliometric standpoint it is, of course, easy to measure a scientist's productivity in terms of his publications, but is is a very different matter to evaluate their "worth" or "creativeness" bibliometrically. The criterion of average number of citations per year by other research workers propounded by Matheson[8] is not too easy to obtain; it does not recognize the scientist who, like Flory, has produced an idea ahead of his time, and it overates the one whose work is frequently cited by critics questioning its validity. Although Matheson excluded self-citations, he noted that they ranged from 24 to 46 percent (mean = 31 percent) for the sample of British university chemists investigated. Citation counting as a measure of research achievement has, however, been strongly attacked.[20] It has, however, also been used as a measure of the acceptance of new scientific ideas[21] and as evidence for the operation of the so-called Matthew effect whereby a discovery by an already eminent scientist adds disproportionately more to his prestige than it would to a scientist who has not yet made his mark.

Several areas of science have recently been subjected to bibliometric studies. A good review charting the growth of physics literature and information services was produced by a group at Aslib.[10] Pritchard has characterized the literature of computer science,[7] Meadows and O'Connor have studied the journal literature of astronomy and astrophysics[22] and Simon has produced several studies of biological fields.[23] A different type of study, but nonetheless of immediate value to the librarian, is Mann's compilation of 2,000 journals coded according to the number of food papers produced per year.[24] Thornberry[25] showed that from 1900 to 1956 there was a steady (linear) increase in the proportion of papers on phytovirology relative to those on plant pathology. No tail-off was observed before the study ended.

Well-indexed bibliographies provide a good basis for assessing the information content of document titles and hence the reliability of existing titles indexes, the potential usefulness of a projected one[26] or the relative frequency of occurrence of synonyms in a subject area.[27] Although titles indexes such as *Internationale Bibliographie der Zeit-*

*schriftenliteratur, Applied Science & Technology Index,* etc., have been in existence for a long time, they do not seem to have been as favorably regarded or as much used by scientists as by librarians; it was not until the 1950s that a combination of factors—such as intolerable delays in the major abstracting services, the introduction of a title index produced by scientists for scientists (the Chemical Society's *Current Chemical Papers*) and possibly Luhn's advocacy of automated titles indexes—induced scientists to regard title indexes seriously. One consequence of this was that they began to pay more attention to the titles which they gave their papers. By the simple technique of counting the keywords per title in different years Tocatlian[28] recently quantified this interesting trend to more informative titles.

The condensation of information ongoing from a primary publication to the secondary literature leads to transmission losses in the several information transfer channels. These losses are either total, due to noninclusion in a specific secondary service, or partial, due to omissions of items of information from a given document by the abstractor or by the indexer. Some methods of estimating transmission losses from studies on bibliographies are discussed elsewhere.[6]

Citation analysis has long been used by librarians with greater or lesser effectiveness for various purposes connected with the management of journal collections. It is an especially useful aid in journal selection policies. A poor example is an analysis of citations in a few American library journals proposed as a basis for journal selection in a library school library.[29] Much more comprehensive data on which to make such a selection were obtained from an analysis of six abstracting services on information science. Although 990 relevant journals were noted, three-fourths of the total information came from just 100 journals.[30] Citation analysis in science has been made easier by the appearance of the *Science Citation Index* and its machine-readable tapes. For example, Martyn and Gilchrist[31] have shown that nearly one-fourth of the citations to British journals were to *Nature* and *The Lancet.* A method of selecting the most prolific journals for a particular topic and for assessing the productivity of journals of marginal interest (provided that they have been indexed by the *Science Citation Index*) has been proposed.[32] These are aspects of the Bradford scattering effect (discussed earlier in this issue) which Douglas[33] and others have suggested is a time-dependent phenomenon and that variations in rankings with time must be taken into account when forecasting a library's future coverage of a field.

Once an abstracting service has got its bibliographic data into machine-readable form, useful statistics can be obtained very readily. (An even greater volume of useless statistics can be churned out just as quickly.) Some of these statistics are helpful in the management of the service,[34] e.g., a ranked list of journals showing their productivity is useful in identifying the core journals of the field. Very detailed summaries of statistical data from *Nuclear Science Abstracts* have been published;[35] one of the many interesting facts produced is that there is a disproportionately high contribution in the area of theoretical physics from the U.S.S.R. BIOSIS often produces analyses of the literature it processes; its annual listings of the percentage of biological serials derived from specific countries is normally to be found in the prefatory pages of the December issues of *Biological Abstracts*. In a study of 1971 literature on nonhuman primates,[36] it was found that of 3,205 papers from 742 primary research publications, more than 45 percent dealt with members of the family Cercopithecidae, but, surprisingly, one-third of the papers made no reference to the specific animal under study. A scattergram by subject concepts and test organisms revealed that the most frequently studied research topic in which nonhuman primates are used is physiology of the nervous system. This has no connection with a previously unpublished study on primates.[37]

Unlike the Russian and French abstracting services, English-language ones are fragmented with respect to the whole field of science and technology. Because of this, considerable overlap occurs. Three of the major services produced in the United States, BioSciences Information Service of Biological Abstracts, Chemical Abstracts Service and *Engineering Index* have recently undertaken a study of the overlap among their respective services. The total number of different journals monitored by the three services in 1970 was 14,592, of which 10,511 were monitored by only one service and only 140 were covered by all three. The biggest overlap in terms of abstracts was between *Engineering Index* and *Chemical Abstracts*—45.3 percent. The overlap of *Chemical Abstracts* with BIOSIS was 28.6 percent.[38] These studies, which are continuing as a "preliminary step in planning for the future" may well have far-reaching effects on users of secondary literature in these areas and on the librarians who provide for their needs.

Errors in bibliographies can occur, and, if not checked with the original, may be perpetuated by subsequent copiers. (Readers familiar with the *Journal of Chemical Education's* "Textbook Errors" feature will have seen numerous examples where an error of fact was contained in at least two textbooks.) Dobell[39] gives an account of how a title in Czech became substituted for the author's name in *Centralblatt für Bacteriologie und Parasitenkunde* (1887, 1, 537) and for fifty years books and papers on dysentery referred to a paper in an obscure Czech journal by O. Uplavici (Czech for "on dysentery") instead of by the true author, Jaroslav Hlava.[40] There is an apocryphal story of the employee of *Chemisches Zentralblatt,* who, being under notice, managed to insert a nonsense abstract which he attributed to S. C. H. Windler.

In the better bibliographies errors are remarkably few and far between. While I have made no systematic study of this aspect, I did observe only one error in a random sample of 183 subject index entries from the 1967 *Chemical Abstracts.* On the other hand, using the *Science Citation Index* one gets the impression of a rather higher proportion of errors. It is to be expected that as abstracting and indexing services become increasingly produced by automatic means, especially where economic pressures dictate the minimizing of intellectual effort at the input stage, such errors will cause a little, but increasing, frustration for the literature searcher. For example, in a citation index it is clearly impracticable to check every citation in every document processed—one must assume that the journal editor has spotted any errors and inconsistencies; he or she in turn will doubtless assume that authors (or referees) have got the bibliographic details of the citation, such as citer authors' names and initials etc., correct. The *Science Citation Index,* did, however, enter correctly all fifteen citations to Greenstein's joke paper on "armpitin"[41] despite such give away citations as Yolk, A. and White, B., On slicing a hard boiled egg, *Popular Mechanics,* 39:251, 1948 or Goose, M., A child's guide to erotica, Golden Days, Garden City, 1963, etc. The "paper" was also indexed in *Index Medicus* (1966) under "Contraceptives" but I could not find it in either *Chemical Abstracts* or *Biological Abstracts.* A charitable explanation is that perhaps indexers have a sense of humour which abstractors lack.

## *References*

1. An early warning was by Joseph Henry in his Annual Report as Secretary of the Smithsonian Institute in 1851 where he prophesied that scientific knowledge "will begin to totter under its own weight" unless "well digested indexes of subjects not merely referring to volumes or books, but to memoirs, papers and parts of scientific transactions" were produced.

2. Bottle, R. T., ed. *Use of the Chemical Literature.* London, Butterworths, 1962, p. 66.

3. Strong, Laurence E., and Benfey, O. Theodor. "Is Chemical Information Growing Exponentially?" *Journal of Chemical Education*, 37:29-30, Jan. 1960.

4. Crane, E. J. "Chemical Abstracts, Yesterday, Today and Tomorrow," *Proceedings of the Chemical Society*, 1957, pp. 334-40.

5. Bolton, H. Carrington. An address to the American Association for the Advancement of Science, 1882, quoted by Van Patten, Nathan. *Journal of Chemical Education*, 27:431, Aug. 1950.

6. Bottle, R. T. "Scientists, Information Transfer and Literature Characteristics," *Information Scientist*. (In press)

7. Pritchard, Alan. *A Guide to Computer Literature*. 2d ed. London, Bingley, 1972, p. 13.

8. Matheson, A. J. "Centres of Chemical Excellence?" *Chemistry in Britain*, 8:207-10, May 1972.

9. *CAS Today: 60th Anniversary Edition*. Columbus, Ohio, Chemical Abstracts Service, 1967.

10. Anthony, L. J., *et al.* "The Growth of the Literature of Physics," *Reports on Progress in Physics*, 32:709-67, 1969.

11. Wood, D. N. "Foreign Serials and Translations." *In* R. T. Bottle, and H. V. Wyatt, eds. *The Use of Biological Literature*. 2d ed. London, Butterworths, 1971, p. 35.

12. Bottle, R. T., and Chase, W. W. "Some Characteristics of the Literature on Music and Musicology," *Special Libraries*, 63:469-76, Nov. 1972.

13. Schwartz, Ellen S. E., and Powers, Wendell H. "Survey of the Quantity and Distribution of Biochemical Literature," *Journal of Chemical Documentation*, 3:37-42, Jan. 1963.

14. Webb, E. C. "Communication in Biochemistry," *Nature*, 225:132-35, Jan. 10, 1970.

15. Polyakov, G. K. "Patent Statistics: Specifics and Properties," *Voprosy Izobretatel'stva*, 1970:55-60, (1-2) (*Information Science Abstracts*, 6:240, 71-2103, Aug. 1971); and Dauranov, R. B. "Quantitative Studies on the Basis of Patent Materials," *Voprosy Izobretatel'stva*, 1970:49-54, (1-2) (*Information Science Abstracts*, 6:240, 71-2101, Aug. 1971).

16. Korrennoi, A. A. "Quantitative Analysis of Information Relations Between Patents," *Voprosy Izobretatel'stva*, 1969:34-39, (1-2) (*Information Science Abstracts*, 6:178, 71-1601, June 1971).

17. Crane, D. *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago, University of Chicago Press, 1972.

18. Zuckerman, Harriet A. "Patterns of Name Ordering Among Authors of Scientific Papers: A Study of Social Symbolism and its Ambiguity," *American Journal of Sociology*, 74:276-91, Nov. 1968.

19. Lehman, Harvey C. "The Chemist's Most Creative Years," *Science*, 127:1213-22, May 23, 1958.

20. May, Kenneth O. "Abuses of Citation Indexing," *Science*, 156:890-92, May 1967; and Comfort, Alex. "Pop Charts for Science," *Nature*, 227:1069, Sept. 5, 1970.

21. Cole, Stephen. "Professional Standing and the Reception of Scientific Discoveries," *American Journal of Sociology*, 76:286-306, Sept. 1970.

22. Meadows, A. J., and O'Connor, J. G. *A Survey in Depth of a Selected Information Field (Astronomy)* (Final Report, OSTI Report 5092). 1971.

23. Three examples are: Simon, H. R. "Analyses of Bibliographies on Biocontrol," *Journal of Documentation,* 26:337-39, Dec. 1970; ———. "Historische Entwicklung und aktueller Stand der entomologischen Literatur," *Zeitschrift für Pflanzenkrankenheiten und Pflanzenschutz,* 79:413-29, 1972; and ———. "Zur Analyse biologischer Bibliographien," *Heidelberger Jahrbüch,* 15:111-33, 1971.

24. Mann, E. J. *Evaluation of the World Food Literature.* Farnham Royal, Bucks., Commonwealth Agricultural Bureaux, 1967.

25. Thornberry, H. H. "Publications Relative to Vital Diseases in Plants from 1900-1956," *Plant Disease Reporter,* 43:371-73, 1959.

26. Bottle, Robert T., and Seely, Catherine R. "Information Transfer Limitations of Titles of Chemical Documents," *Journal of Chemical Documentation,* 10:256-59, Nov. 1970; and Bottle, R. T. "An Experimental Examination of the Effectiveness of Possible Titles Indexes for Design Literature." Paper No. 9 in *Proceedings of the International Symposium on Information Systems for Designers.* University of Southampton, 1971.

27. Bottle, R. T. "Synonyms, Related Terms and Thesaurus Controlled Indexing." Paper presented at Informatics I, Durham, April 11-13, 1973. (In press)

28. Tocatlian, Jacques Jean. "Are Titles of Chemical Papers Becoming More Informative?" *Journal of the American Society for Information Science,* 21:345-50, Sept. 1970.

29. Reference omitted to avoid embarrassing its author.

30. Novakova, Hana. "The Analysis of World Abstracting Services in Informatics," *Bibliograficky Časopis,* 3:118-36, 1969 (*Information Science Abstracts,* 6:67, 71-550, April 1971).

31. Martyn, John, and Gilchrist, Alan. *An Evaluation of British Scientific Journals* (Occasional Publication, no. 1). London, Aslib, 1968.

32. Bottle, R. T. *The Use of Chemical Literature.* 2d ed. London, Butterworths, 1969, pp. 71-72.

33. Douglas, I. A. "Science Libraries," *Nature,* 239:477, Oct. 20, 1972.

34. Pflueger, Margaret L. "Management Information Obtained from Bibliographic Data," *Proceedings of the American Society for Information Science,* 8:355-61, 1971.

35. Vaden, W. M. An Analysis of World-Wide Contributions to *Nuclear Science Abstracts Vol. 22 (1968)* (TID 25004 and subsequent reports).

36. Kennedy, H. E. "Nonhuman Primate Research Literature—A Subset Analyzed," *Biological Abstracts,* 54(6):xxiv, Sept. 15, 1972.

37. Cantuar, M. "Nonhuman Primates: Attitudes to the Pill," *The Tablet,* 10:1010, 1969. (Inclusion of this reference in a citation index is inadvisable as it could be regarded as an overspill from Greenstein's paper.)

38. Kennedy, H. E. "Progress Report: BIOSIS/CAS/Ei Overlap Study," *Biological Abstracts,* 53(9):xxiv, May 1, 1972.

39. Dobell, C. "Dr. O. Uplavici (1887-1938)," *Parasitology,* 30:239-41, 1938.

40. Hlava, J. "O Úplavici. Předlěžne Sdělení" *Časopis Lékařův Českých,* 26(5), 1887.

41. Greenstein, J. S. "Studies on a New, Peerless Contraceptive Agent: A Preliminary Final Report," *Canadian Medical Association Journal,* 93:1351-55, 1965.