New Factors in the Evaluation of Scientific Literature Through Citation Indexing[1]

More than one million citations from the scientific literature have been processed by the Citation Index Project at the Institute for Scientific Information. The Project, sponsored by NSF and NIH, will be described briefly, and new methods of using citation data for evaluation of publications will be discussed.

Summaries of statistical data, compiled by computer methods such as the following, will be given.

1. Frequency of citation of one journal by another.
2. Frequency of current citations to the past literature.
3. Frequency of self-citation by journals and authors.
4. Number of source citations per cited paper.
5. Number of references per source paper.
6. Number of papers published per journal.

Information scientists and research workers are encouraged to use this unique reservoir of information for additional statistics applicable to their fields of work as a basis for comparative studies on the efficacy of various indexing techniques.

E. GARFIELD and
I. H. SHER

Institute for Scientific Information
Philadelphia, Pennsylvania

The National Science Foundation, National Institutes of Health, and the Institute for Scientific Information have been sharing the expenses of a large research study of citation indexing. Over twenty people are engaged in the Project. Citations from the scientific literature are entered on punched cards and processed by computer to produce citation indexes, i.e., ordered lists of references which are accompanied by citations for documents in which they have been cited. In contrast to conventional indexes, which take you back in time, a Citation Index brings you forward in time and thereby writes the "subsequent history" of the particular document under investigation (1). In Fig. 1 is given the portion (taken from a 326,000-item sample) of a Citation Index showing citations to the work of Albert Einstein. In Fig. 2, these entries are expanded to include the titles of the articles. Note the diversity of sources which cite Einstein's 1906 article. The source journals are: J CHEM SOC, J POLYM SCI, and J DAIRY SCI. With Einstein's 1908 article in ANN PHYSIK, there is illustrated an error in

the literature. Dr. Marshall Fixman accredited this ANN PHYSIK article to Einstein, and so it appears here in the Citation Index. However, a check of the original article for the purpose of obtaining a full title disclosed that it was really written by Smoluchowski.

It is interesting to point out that 13 works of Einstein, from 1905 through 1938, are found cited in this limited sample taken from 1960 life sciences literature. Einstein's bibliography is reported to include 310 scientific works, including many repetitions by translation.

Over a million citations from a wide variety of journals have been processed thus far in the Citation Index Project, which makes available for the first time a large reservoir of interdisciplinary data that allows for testing the power of and need for citation indexes.

Our study has been oriented toward the field of genetics. However, the true range of this or any other eclectic subject cannot be properly evaluated without starting from a more general subject coverage. Only after supplying broad comprehensive *input* can one study the effects of limiting the input by selecting citing journals (genetics journals versus nongenetics journals), by author

```
EINSTEIN A--------*05*ANN PHYSIK----------   1 7   5 4 9
        KING  AL          BIOC  BIOP  A       6 0     4 2   3 4 4
        ROSS1  C          NATURE              61  1 8 9    8 2 2
        --------------06-ANN PHYSIK----------   1 9   2 8 9
        ELWORTHY   PH    J CHEM SOC            5 9            1951
        VARADAIAH  VV    J POLYM SCI           6 0     4 6   5 2 8
        WHITNAH CH       J DAIRY SCI           5 9     4 2   2 2 7
        --------------06-ANN PHYSIK----------   1 9   3 7 1
        KING  AL         BIOC BIOP  A          6 0     4 2    3 4 4
        --------------07-JAHRB RADIOAKT ELEKT    4     4 1 1
        DEBEAURE,  0     COMPT  REND         6 0 2 5 0  2 1 4 9
        --------------08-ANN PHYSIK----------   2 5   2 0 5
        FIXMAN M         J CHEM PHYS           6 0    3 3  1 3 5 7
        --------------08-Z ELEKTROCHEM-------   1 4   2 3 5
        MAJUMDAR  SK     NATURWISSEN           6 0    4 7    3 9
        --------------10-ANN PHYSIK----------   3 3  1 2 7 5
        BULLOUGH  RK     J POLYM SCI           6 0    4 6   5 1 7
        COUMOU DJ        J COLL SCl            6 0    1 5   4 0 8
        --------------11-ANN PHYSIK----------   3 4   5 9 1
        ELWORTHY PH      J CHEM S O C          5 9           1951
        GIBBONS  RA      BIOCHEM J             5 9   7 3    2 1 7
        VARADAIAH VV     J POLYM SCI           6 0    4 6   5 2 8
        --------------12-ANN  PHYSIK----------  3 8   3 5 5
        WHITROW GJ  .    NATURE               6 0   1 8 8   7 9 0
        --------------12-ANN PHYSIK----------   3 8   4 4 3
        WHITROW  GJ      NATURE               6 0   1 8 8   7 9 0
        --------------24-Z PHYSIK------------   2 7     1
        BAKANOV  SP      DISC  FARAD           6 0          1 3 0
        --------------26-INVEST THEORY BROWN-¤
        BAKANOV  SP      DISC  FARAD           6 0          1 3 0
        --------------37-J FRANKL INST-------  2 2 3   4 3
        DROZ-VIN.  P     COMPT  REND          6 0 2 5 1  2 2 9 7
        --------------38-ANN MATH------------ 39 6     5
        HOANG  PT        COMPT  REND          6 0 2 5 0  1 1 9 5
EINSTEIN  HA-------*42*TR  AM  SOC  CIVIL----- 1 0 7   5 6 1
        KNISELY  MH      ANGIOLOGY            6 0    1 1   5 3 5
EINTHOVEN  W-------*03*PFLUEGERS ARCH-------   9 9    4 7 2
        DISTEL R         COMPT  REND          6 0 2 5 1  1 1 8 2
EIPEITAUER  E------*56*ZEMENT KALK GIP-----    9    5 0 1
        POWELL  DA       NATURE               6 0  1 8 5   3 7 5
EIRICH F---------*36*KOLLOID Z----+/-B-I)--   7 4   2 7 6
        DEINDOER.  F     IND  ENG  CHE        6 0    5 2    5 9
        --------------56-J COLLOID S C - & ------   1 1  7 4 8
        JONES   MH       CAN  J  CHEM         6 0   3 8  2 3 0 3
        KEOSIAN  J       SCIENCE              6 0  1 3 1   4 7 9
        OVERBERG,  C     J AM CHEM S          6 0   8 2   9 2 9
EISAMAN JL--------*51*JAMA---------------- 1 4 6  1 4 1 7
        ALBRECTSEN       ACT  DER-VEN         6 0    4 0   4 7 4
```

FIG. 1. Computer printout of citations to the works of A. Einstein from CI.

EINSTEIN A   05 ANN PHYS IK   17 3 4 9
                        UBER DE VON DER MOLEKULARKINETISCHIN
                        THEORIE DER WARME GEFORDERTE BEWEGUNG VON
                        IM RUHENDEM FLUSSIGKEITEN SUSPENDIERTEN
                        TEILCHEM

    KING A L   BIOCHIM BIOPHYS ACT   42   344   60
                        BACTERIA AND THEIR FLAGELLA

    ROSSI C   NATURE   189 822 61 .
                        DIFFUSION OF SMALL MOLECULES

——————————————— 06 ANN PHYSIK   19   289
                        EINE NEUE BESTIMMUNG DER MOLEKULDIMENSIONEN

    ELWORTHY PH   J CHEM SOC   1951 59
                        THE STRUCTURE OF LECITHIN MICELLES IN
                        BENZENE   SOLUTION

    VARADAIAH VV   J POLYM SCI   46   3 28 60
                        THEORETICAL EVALUATION OF THE FLORY UNIVER-
                        SAL CONSTANT

    WHITNAH CH   3 DAIRY SCI   42 227 59
                        SOME PHYSICAL PROPERTIES OF MILK. VI. THE
                        VOLUMINOSITY OF CASEINATE COMPLEX IN MILK,
                        AND RECONSTITUTED SEDIMENTS

——————————————— 06 ANN PHYS IK   19   371
                        ZUR THEORIE DER BROWNSCHEN BEWEGUNG

    KING AL   BIOCHIM BIOPHYS ACT   42   344 60
                        BACTERIA AND THEIR FLAGELLA

——————————— 07 JAHRB RADIOAKT ELEKTR   4   411
                        U B E R DAS RELATIVITATS PRINZIP UND DIE AUS
                        DEMSELBEN GESOGENEM FOLGERUNGEN

    DE BEAURE O   C R AC SCIPAR   250 2149 60
                        DEVELOPPEMENT DES CONSEQUENCES DE LA
                        THEORIE DE LINERTIE DE D. W. SCIAMA ET D E
                        Do PARK

——————————————— 08 ANN PHYSIK   25   205
                        MOLEKULAR-KINETISCHE THEORIE DE Q OPALESZENZ
                        VON GASEN KRITISCHEN ZUSTANDE, SOWIE
                        EINIGER VERWANDTER ERSCHEINUNGUN

    FIXMAN M   J CHEM PHYS   33 1357 60
                        DENSITY CORRELATIONS, CRITICAL OPALESCENCE'
                        AND THE FREE ENERGY OF NONUNIFORM FLUIDS

——————————————— 08 Z ELEKTROCHEM   14   235
                        ELEMENTARE THEORIE DER BROWNSCHEN BEWEGUNG

    MAJUMDAR SK   NATURWISSENSCHAFTEN   47   39 60
                        MOLECULAR WEIGHT OF MYCOBACILLIN BY
                        DIFFUSION METHOD

——————————————— 10 ANN PHYSIK   33 1275
                        THEORIE DER OPALESZENZ VON HOMOGENEN
                        FLUSSIGKEITEN UND FLUSSKEITSGEMEISCHEN IN
                        O E R NAHE DES KRITISCHEN ZUSTANDES

    BULLOUGH R K   J POLYM S C ?   46   517 60
                        BIREFRINGENCE AND LIGHT SCATTERING OF' HIGH
                        POLYMERS

    COUMOU DJ   J COLL SCI   15   408 60
                        APPARATUS FOR THE MEASUREMENT OF LIGHT
                        SCATTERING IN LIQUIDS. MEASUREMENT OF THE
                        RAYLEIGH FACTOR OF BENZENE AND OF SOME
                        OTHER PURE LIQUIDS

FIG. 2. A. Einstein's citations expanded to include titles.

------------------- 11 ANN PHYSIK   34   591'
                    BERICHTIGUNG ZW MEINER ARBEIT. F I N E   NEUE
                    BESTIMMUNG DER MOLEKULDIMENSIONEN

ELWORTHY PH   J CHEM SOC   1951 5 9
                    T H E STRUCTURE O F LECITHIN MICELLES T N
                    BENZENE   SOLUTION

GIBBONS RA   BIOCHEM J   73   217 5 9
                    THE PHYSICOCHEMICAL PROPERTIES OF TWO
                    MUCOIDS FROM BOVINE CERVICAL MUCIN

VARADAIAH VV   J POLYM SCI   46   528 60
                    THEORETICAL EVALUATION OF THE FLORY UNIVER-
                    SAL CONSTANT

---------------- 12 ANN PHYS IK   38   355
                    LICHTGESCHWIMDIGKEIT U N Q STATIK DES
                    GRAVITATIONSFELDES

WHITROW G J   NATURE   188 790 60
                    GENERAL RELATIVITY AND LORENTZ-INVARI'ANT
                    THEORIES OF GRAVITATIONS

---------- ------- 12 ANN PHYSIK   38   443
                    ZUR THEORIE DES STATISCHEN GRAVITATIONS-
                    FELDES

WHITROW GJ   NATURE   188 790 60
                    GENERAL RELATIVITY AND LORENTZ-INVARIANT
                    THEORIES OF GRAVITATIONS

- - - - - - - - - - - - - 2 4 Z PHYSIK   27   1
                    ZUR THEORIE DER RADIOMETERKRAFTE

BAKANOV SP   DISC F A R SOC   130 60
                    THE MOTION O F A SMALL PARTICLE IN A NON-
                    UNIFORM GAS MIXTURE

- - - - - - - - - - - - - 26 INVEST T H E O R Y BROWN MOV□   26

BAKANOV SP   DISC FAR SOC   130 60
                    THE MOTION OF A SMALL PARTICLE IN A NON-
                    UNIFORM GAS MIXTURE   ·

----------------3 7 J FRANKL INST   233   43
                    O N GRAVITATIONAL WAVES

DROZ-VINCE P   C R AC SCI PAR   251 2297 60
                    ISOTHERMIE APPROCHEE ET QUANTIFICATION POUR
                    LA METRIQUE DE ROSEN

--------------- 38 ANN MATH   39   63
                    THE GRAVITATIONAL EQUASIONS AND THE PROBLEM
                    OF MOTION

HOANG PT · · ○   AC SCI ▣-▣○   2501 1956 0
                    SUR LES EQUATIONS DU MOUVEMENT EN RELATIV-
                    ITE GENERALE

FIG. 2. A. Einstein's citations **expanded to include** titles. (Continued)

(geneticists versus nongeneticists), by type of reference (journal references versus nonjournal references), by subject of article (genetics articles versus nongenetics articles), etc. Similarly, only initial broad and comprehensive input can enable one to later study the possibilities of selective output.

The file of citations is alphabetized by the first reference author and printed as the Author Citation Index, or 'is alphabetized by reference publication to give the Journal Citation Index. These citation indexes can help answer such questions as:

1. Where has an individual article or author been cited?
2. Who else is publishing on a given subject-?
3. What journals are publishing articles on a given subject?
4. What papers belong in the historical introduction to a paper in progress, in a review article, in a chapter of a book, etc.
5. What are all the publications by a given author?
6. What is the distribution of citations throughout the years, throughout the journals, etc?
7. Has a contemplated piece of research already been done ?

Some of these uses are difficult or impossible with conventional indexes. Others are similar to typical reference uses for large-scale indexing services.

In the course of our processing, certain "vital statistics" have been gathered. For example, data taken from journals published in the first third of 1962 and included in *Current Contents* -Life Sciences edition showed an average of 2.1 authors and 5.4 pages per paper. The same journals published an average of 17.3 articles per issue. However, we are more interested in certain "impact" factors such as how often a particular paper, author, or journal is cited compared to corresponding average values in a given Citation Index file.

In a Citation Index compiled from the references in one issue of one journal, there are few cases of more than one source citing the same reference. In this case, therefore, the source/reference S/R [2] ratio is almost equal to 1. As the input grows, more multiple citations are encountered, and the average number of sources per reference (S/R) increases. Thus, in a Citation Index compiled from 60,000 citation, the S/R was about 1.2. 'In a Citation Index of 326,000 items, the S/R ratio was 1.5 instead. In Fig. 3 is shown (for the 326,000-citation sample) the percentage of references cited one time, two times, three times, etc. From these data it can be calculated that 95% were cited one to three times. The range in this compilation extends from the bulk of articles which are cited only once to one article by Oliver H. Lowry which was cited 305 times. It is seen then that most Citation Index searches will yield a manageable number of citing papers.

In this Index, 8% of all citations are first-author

² S = total number of source articles processed multiplied by the average number of references per source article. In other words, S is the number of source lines in the Citation Index.
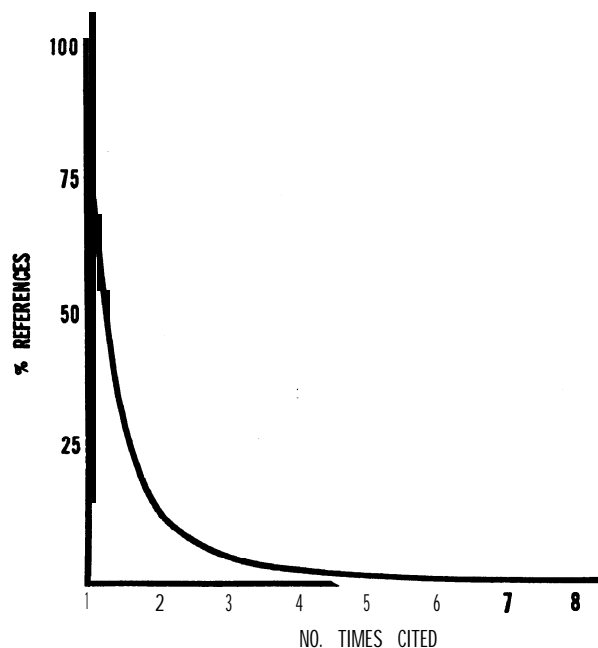

FIG. 3. Per cent references in 326,000-item citation index vs. number of times cited.

self-citations. Similarly, 20% of all sources which cite a journal are found to be self-citations by journal. That is, of the citations to a particular journal, 20% arise from that same journal title — obviously this varies for each journal considered.

Statistics taken from 1.4 million references which appeared in the 1961 literature disclose an average of 13.7 references per article. Of these references, an average of 2.2 are to nonjournal publications. Thus, 84% of these references are *to* journal articles. Of the approximately 102,000 source items or articles processed to produce the file of 1.4 million references, 11.9% were items containing no references at all. This value of 11.9% is, of course, swollen by the inclusion of brief notes, editorials, queries, and letters to the editor which often contain no references. The latter items can be found through the use of a first-author-alphabetized bibliography of the processed source items which, it is planned, will accompany the Citation Index Proper (2). It should be noted that these items are not usually indexed by the conventional services.

The combined use of the author bibliography and the Citation Index aids the user in searching backward and forward in time. The typical search starts by looking up a particular paper in the Index and'finding the citations to that paper. The user can also take note of papers by the same first author either in the Citation Index or in the author bibliography. The likelihood of a paper escaping detection because it neither contains references nor was cited subsequently is quite small and is, in fact, the product of the probabilities of each contingency. .

The 326,000-item sample of 1960 literature revealed

that the first year preceding the citing source publication is the most heavily cited, and that half the references are from the eight-year period preceding the source. Ours, and all other absolute figures reported on this subject, however, must be corrected for the varying numbers of articles actually published in earlier years. One hundred citations to a year in which only 1,000 articles were published is comparable to 10,000 citations to a more recent year in which 100,000 articles were published.

One way of describing the utilization of a journal is to list the titles and distribution of the publication which cite the journal in question. For example, Fig. 4 shows the number of citations from the sample Index to AM J HUM GENET. This example includes a high proportion of self-citations by this journal — 37% rather than the average of 20%.

In order to have a single term which describes the utilization of each journal, we define a "utilization factor" (Fig. 5). The utilization factor is the product of two terms, K and p. The slope K (derived from Cole's (3) approach to reference scattering) is obtained from the plot (Fig. 6) of the logarithm of the cumulative sum of citing journal titles as a fraction of the total titles versus the cumulative proportions of the citing articles (grouped by journal titles). This slope characterizes the distribution of citations among various journal titles. The term p is that *proportion* of the total source journal titles processed which actually cite the journal in question. The

Source Journal Title	No. Times Citing AJHG
AM J HUM GENET	59
ACTA GENET STAT MED	24
GENETICS	7
NATURE	7
EUG QUART	6
CANCER	5
HEREDITY	4
P NAT AC SCI US	4
NATURWISSENSCHAFTEN	4
P SOC EXP BIOL MED	3
J HERED	3
SCIENCE	3
JAP J GENET	2
J GENET HUM	2
AM J CLIN PATH	2
BRIT J HAEM	2
J BIOL CHEM	2
P JAP AC SCI	2
J GENET	1
ARCH PATH	1
BRAIN	1
J DAIRY SCI	1
ANN END	1
ARCH BIOCHEM BIOPHYS	1
ANN INT MED	1
AM J DIS CHILDR	1
ARCH DIS CHILDH	1
NEUROLOGY	1
EXPERIENTIA	1
CANC RES	1
GENET IBER	1
PED CLIN	1
ACTA PATH MICROBIOL SCAND	1
HELV PAED ACTA	1
TOTAL	157

Fig. 4. Total number of times source titles cite the *American Journal of Human Genetics* in the 326,000-item citation index.

JUJ = Journal Utilization by Journal

JUJ = K × p

Where

K = Slope of the best straight line obtained from the plot of the log of the cumulative sum of source titles as a fraction of total titles (ST/ΣST), versus the cumulative tallies of title occurrences (from Slide 6) as a fraction of the total occurrences (S/ΣS).

And

p = Proportion of the total journal titles processed which cite AM J HUM GENET

NOTE: K is analogous to the "Reference Scattering" term described by Cole, PF, J DOC, 18, 58 (1962)

Fig. 5. Utilization factor.

product of K and p gives a value indicative of the utilization of the journal by journals.

Users of the Citation Index are struck by the enormous variety of data that can be extracted from our file. One of the most interesting correlations is the "journal impact factor." In the usual citation count methods, as, e.g., Gross and Gross (4), the importance of a journal is determined by the absolute number of citations to it. The J AM CHEM SOC ranks first on such a list.[3] However, this count is largely a reflection of the fact that more articles are published in this journal than most. This approach is not much more sophisticated than ranking the importance of a journal by the quantity of articles published. The first step in obtaining a more meaningful measure of importance is to divide the number of times a journal is cited by the number of articles that journal has published. This linear relationship is valid at least as long as self-citations are not eliminated, as Westbrook (5) did, nor multiple citations omitted, as by Raisig (6). When this calculation is performed, the J AM CHEM SOC no longer ranks first. In our own citation counts, the PROC NAT ACAD SCI, NATURE, SCIENCE and other journals move towards the top — including some journals which publish many articles and others which do not. However, even these rankings may not hold in each field. A leading chemist may prefer to publish in the J AM CHEM SOC, whereas a leading geneticist may prefer the PROC NAT ACAD SCI or NATURE. Indeed, the number of genetics and biochemical articles in the PROC NAT ACAD SCI has grown considerably of late.

Taking the analysis one step further, librarians and information scientists can organize collections of frequently used papers. It seems utterly foolish to be sending out bound volumes of journals which are being borrowed for a small group of frequently cited articles. The citation data now available makes such a determination possible without intimate knowledge of the subject matter. Thus, we can say with reasonable certainty that any biochemistry librarian would be well advised to have Lowry's

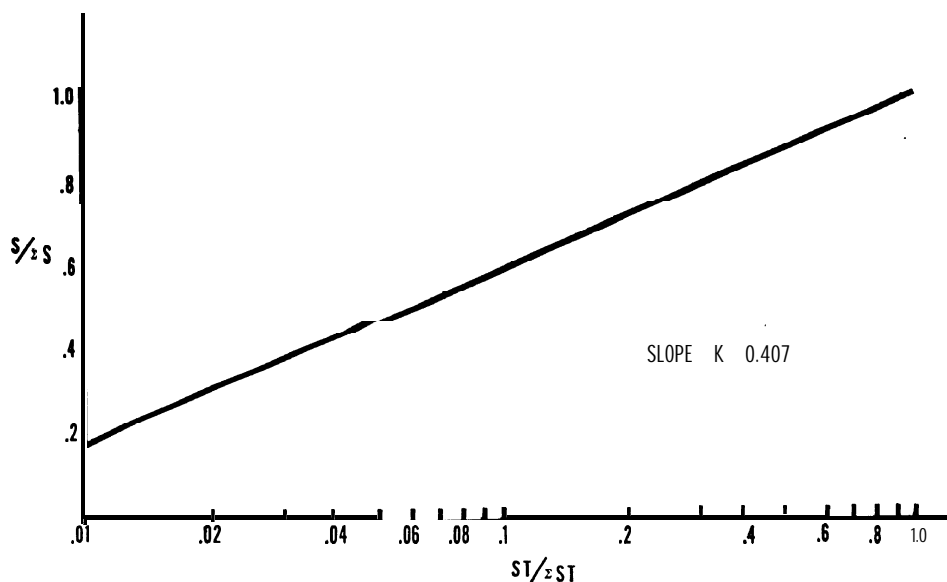[3] As modified by Brown, C. H., Scientific Serials, p. 101 (1956).

FIG. 6. Source-scattering among journals which cite the **American** Journal **of** Human Genetics.

article on **protein** analysis readily available, since it is the most frequently cited paper in the field.

On the other hand, this same information should be used with caution for personnel selection and evaluation (e.g., by the Nobel committees). I might add, however, that many people are interested in this application of citation indexing. Citation indexing will obviously facilitate the evaluation of milestone papers. Indeed, if time were available we could demonstrate how one might use citation networks for extremely interesting historical and sociological studies.

It is hoped that the file of citations for the 1961 literature can be expanded in our continuing research work so that it will one day approach completeness. This is very important, since there has not yet been a definitive study of the number of articles published each year-or the number of significant journals published. If significance can be related to the number of times a journal is cited, then there is a high probability that the number of significant journals has been vastly overestimated. Certainly, a very small number of journals account for a very large percentage of all citations (Fig. 7). In our
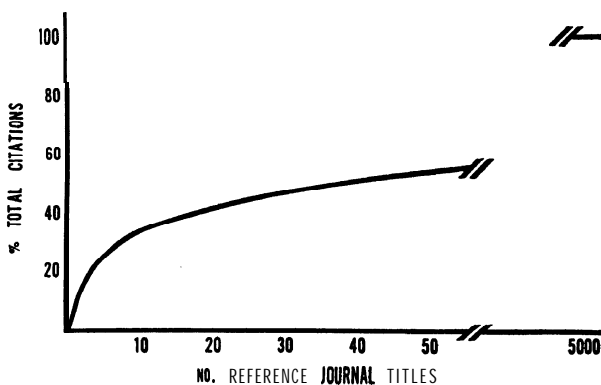
326,000-item sample, for instance, 10 reference journal titles account, for one-third and 36 titles account for one-half of the total citations. These few journals are but a small fraction of the estimated 5,000 reference journal titles cited in this file.

Further, by going back and processing one year in each decade comprehensively, it will also be possible to obtain definitive information on the growth of the literature. If the literature has been doubling every ten years, then one would expect 640,000 articles to have been published in 1960 if 10,000 were published in 1900. The cumulative total of papers published in the last 100 years would then have been about 10,000,000. By using the rule of thumb, in part due to discussions with Professor Derek Price, there should be an average of one citation each year to every article ever published. The use of an average of 15 citations per published paper would seem to confirm this initial estimate.

In conclusion, it should be emphasized that the basic purpose of the project is not to take a statistical inventory of scientific publication. That is, indeed, an important byproduct of the work. The main objective, however, is to develop an information system which is economical and which contributes significantly to the process of information discovery — that is, the correlation of scientific observations not obvious to the searcher. Citation indexes can provide new insights impossible through descriptor-oriented systems.

FIG. 7. Per cent total citations vs. number of reference journal titles.

References

1. GARFIELD, E. 1955. Science, **122** : 108–11.
2. GARFIELD, E. 1958. Unified indexes to science. *Proc. ICSI*, 1: 461–474.
3. OLE, P. F. 1962. *J. Doc.*, 18: 58.
4. GROSS, P. L. K. and E. M. GROSS. 1927. *Science, 66: 385.*
5. WESTBROOK, J. H. 1960. *Science,* **132** : 1229.
6. RAISIG, L. M. 1960. *Science,* 131: 1417.