Essays of an Information Scientist, Vol:7, p.515-522, 1984

GENETICS CITATION INDEX

Experimental Citation Indexes to Genetics
with Special Emphasis on Human Genetics

Prepared by the

Institute for Scientific Information
Philadelphia 3, Pa.

Eugene Garfield, Ph.D., *Director*
Irving H. Sher, Sc.D., *Project Director*

PREFACE

Dr. Garfield's article on citation indexing which appeared in *Science* in 1955 first brought this technique to my attention and was my first introduction to the organization now known as the Institute for Scientific Information. Citation Indexing seemed a clever idea at the time and I wondered whether it would ever come to fruition.

A few years later the suggestion recurred and I was puzzled how to find out whether there had been any follow-up on Garfield's first suggestion. I had no idea how to look up the literature in the documentation field and from past experience with subject indexing in science had little confidence in the utility of a literature search.

This was the very incident that convinced me of the need for the citation index -- it was parallel to many others in my own research activity. How often I have run across some older reports on methods or on some curiosities of bacterial variations and been frustrated in attempts to find later work on the same subject and, especially, critical enlargement on the earlier work.

For many reasons genetics is an especially apt field for the introduction of citation indexing. It is inherently interdisciplinary, cutting across biochemistry, statistics, agriculture, and medicine so that geneticists need insight into a wide range of scientific literature. While there have been many revolutionary developments, many facets of genetics still rely heavily on older work. The principles of *Drosophila* research of 40 years ago are first finding their application in human cytogenetics today. Geneticists have tended to be perceptive about the historical development of their concept and to fulfill their responsibility in furnishing the appropriate citations in their bibliography. Their concern with parent-offspring relationships perhaps makes geneticists more perceptive to the understanding of the structure of scientific activity that is inherent in citational references. It was, therefore, most gratifying that the review panel of the NIH and NSF concurred in supporting this trial in the field of genetics.

Citation indexing is, of course, only one aspect of literature searching. There will be many disappointments in its use--but a negative result within the scope of the index is perhaps more meaningful than with any other technique. Other methods generally place great reliance on subjective classification with which the final user can rarely be entirely familiar. Citation indexing can uncover unexpected correlation of scientific work that no other method could hope to find, and a successful match can often be located with great speed and assurance. The chief limitation is perhaps merely the scope of the indexing effort in the sample--in a given year there may have been no literature on a given reference. A cumulative index to all of science would, of course, be a large undertaking but of course no larger than the problem to which it is addressed. In fact the machine basis of this approach should make it far less costly and more expeditious than any other technique now apparent. Until a complete index is available we may not know the full value of the technique, but the present sample is a noble effort which should give many investigators substantial help in their present retrieval problem and show the way to an ultimate, even more satisfactory, result.

My own contribution to the project has been too limited to inhibit me from commending Dr. Garfield and his associates for organizing and implementing a project which has required an unimaginable attention to detail, technical skill, enthusiasm, and above all, an irrepressible concern for meeting the real need of scientists. To flourish, science has many needs but none are more vital than responsible communication with history, society, and posterity embodied in what we casually call the scientific literature.

Joshua Lederberg
Stanford University

THE GENETICS CITATION INDEX EXPERIMENT

INTRODUCTION

In January 1961, with the partial support of National Institutes of Health grant RG-8050, the Institute for Scientific Information began a research investigation of citation indexing.[1] Since the Genetics Study Section of NIH was particularly interested in new approaches to the documentation of the burgeoning genetics literature, genetics became the focal point of the project. An Advisory Board of geneticists was also selected to guide the project. From the outset it was recognized that a universally acceptable definition of genetics was all but impossible especially since human and biochemical genetics are highly interdisciplinary. Defining the genetics literature was therefore quite difficult. It was decided that a comprehensive and interdisciplinary approach was needed as well as an arbitrary, acceptable definition of the genetics literature to serve as a point of departure for automatic selection.

In May 1961, as a result of contract C-201 with the National Science Foundation, which provided additional partial support, the interdisciplinary and comprehensive approach recommended by our Advisory Board became feasible. A selective approach to the literature, on the other hand, would have required fine judgments as to what is or is not genetics. The comprehensive approach permitted us to begin with an essentially clerical or automatic procedure wherein all references in all articles of each issue of the source journals were processed.

In the comprehensive 1961 Science Citation Index there is, therefore, no question as to what references are or are not covered. However, the reader will always have to distinguish between comprehensiveness of sources and of references. In a citation index, unlike conventional subject or author indexes, one must distinguish source year and reference year. These distinctions are emphasized differently in each of the experimental indexes that follow. The distinction between source and reference diminishes with the passage of time. It becomes apparent that any current source ultimately may become a reference.

The three different sections of this Genetics Citation Index are in fact three different indexes. One reason for including all three is to illustrate the problems in choosing between greater chronological or broader journal coverage in the source material.

The 1961 Genetics Citation Index is based on a single source year and covers 613 source journals. The second index is based on five source years for 38 hard-core genetics journals, while the third index covers more than 14 source years for 3 key genetics journals. The resources of the project did not permit us to control these variables in such a way as to produce three indexes of equal size or information content. Each index stresses certain features. The 14-year citation index presents an historical picture of the field that is not as easily obtained from the other indexes. Nevertheless, an examination of the 1961 Genetics Citation Index reveals that a remarkably comprehensive recapitulation of the genetics literature is provided even by a single year's citation indexing.

In these citation indexes there is no artificial separation of the "old" and "new" literature. Our studies confirm the frequent utilization by scientists of the older literature. Over 50% of the cited references in the 1961 index are more than five years old, in spite of the fact that the literature has grown more voluminous through the years. Certainly the most frequently cited papers are "classical" rather than contemporary.

While the so-called "genetics parcel" is fully explained in the introduction to the 1961 Genetics Citation Index section, let me emphasize in what respects it is selective rather than comprehensive. There is an important distinction between being comprehensively selective (as are the conventional

indexes) and being selectively comprehensive. To create the experimental 1961 Science Citation Index we prepared a punched-card for each of 1.4 million cited references appearing in 102,000 source articles. In that respect it is comprehensive. However, our objective here was to produce a genetics citation index in contrast to a general science citation index. Establishment of several basic criteria was necessary. Economic considerations, primarily, determined that up to 20% of the reservoir of literature processed could be included. The genetics citation index parcel (1961 Genetics Citation Index) is therefore derived from a comprehensive, interdisciplinary reservoir of citations. The extraction of 20% *output* from the massive *input* became an exciting and challenging task. The reader will have to judge the validity of the results. To help you in this evaluation, a complete listing of the original 102,000 source articles, arranged alphabetically by first author, immediately follows the 1961 Genetics Citation Index. Though the genetics parcel was derived from the data in these source articles, not all of the sources processed will be found in the extracted parcel of the 1961 Genetics Citation Index.

The genetics parcel contains 19% of the citations in the total 1961 Science Citation Index file, yet these citations refer to only 6.7% of the reference authors in that index. It would appear that the reference authors in the genetics parcel are cited more frequently than those in the total 1961 index. This may, in part, be due to the extensive bio-medical coverage in the index and, in part, to the mechanics of the selection procedures.

The genetics parcel for 1961, therefore, is selective, but if an author is listed then all the citations to those papers in which he was first author are included--if they were cited in 1961. The limitation to first author holds since, for purely methodological reasons, only the first author in each reference was processed. This in no way affects the value of the citation index for its original purpose--to locate sources which cite specific reference works. However, any author who is always listed as a second co-author could not be selected in this experiment. Further, those authors who were selected might have co-authored additional papers with first authors who were not selected.

The 5-year and 14-year citation indexes, in contrast to the selective 1961 Genetics Citation Index parcel, are comprehensive. All citations in all papers were not only processed, but also included. However, the same rule concerning first author applies. If this is not kept in mind, one may draw erroneous conclusions concerning the "impact" of an author. Naturally, this attribute would also affect the retrieval value of the index if one expects to find all references to an author's works in one place. This problem could have been overcome by doubling the size of the index through repetition of references under all cited authors--first, second, etc. For economic and editorial reasons this solution was not practical in these experiments.

Alternatively, a journal citation index arrangement, in contrast to an author citation index, would have placed more emphasis on the individual articles, where it belongs, rather than on the authors. The journal arrangement of the index was seriously considered but would have required excessive editorial effort in standardizing variant journal title abbreviations used in the literature.

However, the lack of standardized reference journal abbreviations in the literature did not essentially affect our accuracy in dealing with multiple sources citing the same reference. Thus, while various abbreviations for a reference journal may appear throughout the index, each easily recognized, the computer program did standardize most variant presentations of the reference journal and/or author for a given reference article. In fact, this so-called "unification" procedure actually corrected many errors in reference journal titles and in reference authors' names and frequently chose the longer and more meaningful variation.

The reader will generally find this experimental Genetics Citation Index quite easy to use and understand. Those of us who helped to prepare it are well aware of features which may prove annoying such as truncated source author names which may be only eight characters in length. Similarly, source journal abbreviations are limited to only eleven characters. We realize that the size of print is much smaller than the average index and that surely it would have been most desirable, though not economi-

cally feasible, to include the full titles of citing and cited articles. However, in spite of such shortcomings, which can be eliminated in future citation indexes, I am confident that the Genetics Citation Index and the companion 1961 Science Citation Index will justify many times over the cost of preparation--over $400,000.

During the past few years, while these indexes were in preparation, I have had the unique advantage of having at my disposal, not only the files from which this published genetics index is derived, but also earlier experimental files. For example, over 325,000 references from the 1960 literature were processed for various methodological studies.(2) I have rarely been disappointed by a search of these files and on numerous occasions information was uncovered which would otherwise have remained buried. However, the discovery of "buried treasure" in the scientific literature is a function of the user as well as the index. As with a conventional subject index, experience in using a citation index increases one's ability to find information quickly. Unlike the use of a language-oriented index, training in medical, chemical, or other nomenclature is not required to use a citation index effectively. However, it is assumed that the average searcher knows where he wants to begin. At times the user will have to find a "starting point" elsewhere--in a book, encyclopedia, article, or classical index. The citation index can then be used to locate "what has happened since."

Generally one will find a few source papers which cite a particular paper in which one is interested. The citation index is highly specific in that sources retrieved have a direct relationship to the starting point. By various techniques a search may be readily expanded in order to obtain more extensive bibliographies. For instance, one may select relevant articles from the bibliographies of the sources disclosed by the first step of the search and use these erstwhile sources as new entry points into the citation index. This "cycling" procedure may be repeated. You will find that with "cycling" one uncovers relevant material through all the years preceding the citation index. Another technique is to take advantage of the fact that authors will frequently write more than one closely related paper. Therefore, when using the citation index, additional articles by a given author can be checked for relevancy and the sources citing these articles may be picked up. Additional articles by a given author may be found in the list of source articles even if the articles contained no bibliographies.

Starting with a given target reference paper one may find that it is not cited at all. Do not give up! Refer to the target paper and examine the citations contained in it. Locate the most directly related reference citation in the paper. Then, in the citation index, locate new sources which cite this reference. If not, try another relevant reference from the same target paper. Since the average paper has a bibliography of fifteen items it is rare that a pertinent source is not located, provided there has been some work done on the subject.

Naturally a citation index will be most helpful when used in conjunction with a good library. Since the scope of these indexes is quite broad many source articles may be encountered which are not readily available. The Institute for Scientific Information will assist in the procurement of articles in any source publication processed in this experiment.

Since two years have elapsed since the publication of the source articles in the 1961 experiment it should not be surprising if many articles located are known to you. We believe this will confirm the value of the method. Only when citation indexes are issued currently will they compete as current information sources with *Current Contents, Biochemical Titles* or other tools. However, even at this late date, I am confident you will find sources of information that would have been buried or missed in conventional indexes. It is on this basis primarily that citation indexing should be evaluated. Citation indexing is not recommended as a substitute for conventional indexes. Rather, I believe the citation index adds a new dimension to the pursuit of scientific knowledge. The citation index is your road map of the literature. Where it takes you is primarily your decision. What you find depends on you. Only you can measure its relevance. What may be relevant to one man is irrelevant to another. Two otherwise unrelated scientific observations may be correlated in the citation index through a common reference or through a chain of source-reference links. However, you must select the starting point and detect the correlation if it exists.

The detailed history of this project will be covered elsewhere, both in genetics[4] and documentation journals.[5][6] However, the detailed explanatory descriptions of the various experimental indexes contained in this volume will, together with earlier papers,[2,3,7,8,9] provide those interested with a more complete story. In addition, more background information will gladly be provided to anyone who wishes to contact me directly.

Inquiries are cordially invited from journal editors or publishers who would be interested in the use of our files to conduct special studies of citation patterns to and from their journals.

Eugene Garfield, *Director*
Institute for Scientific Information
Philadelphia 3, Pa.

REFERENCES

(1) Garfield, E., "Citation Indexes for Science," *Science* 122, 108-111 (1955).

(2) Garfield, E. and I.H. Sher, "New Factors in the Evaluation of Scientific Literature Through Citation Indexing," *American Documentation* 14, 195-201 (1963).

(3) Garfield, E., "Citation Indexes in Sociological and Historical Research," *American Documentation* (in press), October 1963.

(4) Garfield, E. and I.H. Sher, "Dissemination and Retrieval of Genetics Information Through Interdisciplinary Citation Indexing," *XIth International Congress of Genetics, The Hague*, September 1963.

(5) Garfield, E., I.H. Sher, and C. Voytko, "Citation Index for Genetics: A Programme for Research and Evaluation," in S.R. Ranganathan and A. Neelameghan, *Documentation Periodicals--Coverage, Arrangement, Scatter, Seepage, Compilation*, pp. 181-185, Bangalore, 1963.

(6) Garfield, E., I.H. Sher, and C. Voytko, "Citation Index for Genetics--A Research and Evaluation Program," *Proceedings Second International Congress on Medical Librarianship*, June 1963 (to be published).

(7) Garfield, E., "Citation Indexes--New Paths to Scientific Knowledge," *Chemical Bulletin* 43(4), 11-12, (1956).

(8) Garfield, E., "Breaking the Subject Index Barrier--a Citation Index for Chemical Patents," *Journal of the Patent Office Society* 39, 583-595 (1957).

(9) Garfield, E., "A Unified Index to Science," *Proceedings of the International Conference on Science Information*, November 1958, Vol. 1, pp. 461-474, Washington, 1959.

ADDITIONAL PERTINENT REFERENCES

Adair, W.C., "Citation Indexes for Scientific Literature," *American Documentation* 6, 31-32 (1955).

Seidel, A.H., "Citation System for Patent Office," *Journal of the Patent Office Society* 31, 554 (1949).

Tukey, J.W., "Keeping Research in Contact with Literature: Citation Indices and Beyond," *Journal of Chemical Documentation* 2, 34-37 (1962).

Weinberg, A.M., et al., President's Science Advisory Committee, *Science, Government, and Information (The Responsibilities of the Technical Community and the Government in the Transfer of Information)*, Part 3, "Citation Indexing Should be Useful," p. 35, Government Printing Office, 1963.

SYNOPSIS

WHAT IS A CITATION INDEX?

A citation index is a directory of cited references where each reference is accompanied by a list of source documents which cite it. The most characteristic feature of the citation index is that the user begins a search with a specific known paper and from there is brought forward in time to subsequent papers related to the earlier paper.

HOW IS A CITATION INDEX PREPARED?

The 1961 Science Citation Index was prepared by processing 613 journals published in 1961. For *every* reference appearing in *every* article in the 613 source journals a separate IBM punch-card was prepared containing both the reference data and the source data. The 102,000 source articles yielded 1.4 million reference cards. The punched-cards were converted to magnetic tapes. The tapes were sorted and otherwise processed on IBM 1401, 1410, and 7074 computers. While the source data in the 1961 Science Citation Index is limited to the year 1961, references published in any period of recorded history are included. The 1961 Genetics Citation Index was extracted from the total 1961 Science Citation Index by computer selection.

The source coverage of a citation index can be extended by processing sources from additional years. In the 5- and 14-year Citation Indexes produced as part of this experiment, the range of source years is increased but the number of source journals covered is decreased.

HOW IS THE CITATION INDEX USED?

To locate source documents which have cited a particular paper, it will suffice to know the name of the first author, year, and page of the target reference paper. It is generally not necessary to know the title of the publication in which the paper appeared. In the citation index the cited or reference author is easily located on the left. For each paper by that author there is a dashed line which continues to the column reserved for the year of the reference publication. The journal abbreviation (or non-journal acronym), volume and page are found to the right of the reference year. Indented under each dashed line are data identifying source articles which have cited that specific reference. When a given reference has been cited by several sources they are arranged alphabetically by source authors.

The primary use of the citation index is the location of citations to a specific reference work. A secondary application of the index is its use as a conventional author index to help identify an author's publications. Note, however, that only papers in which he appears as first author can be found under a specific author's name in this experimental index, and then only when the work was cited within the source coverage. We therefore caution readers that frequency data for individual scientists can be low since senior investigators may be the last ones named in a paper.

APPLICATIONS OF THE CITATION INDEX

The fundamental question one can answer quickly through the citation index is, "Where and by whom has this paper been cited in the literature?" The significance of the answer lies in the policy of scientific publication that one cites appropriate references--methodological, ideational, historical, or otherwise. We have found this rule is, in general, followed. Naturally there are violations and abuses, but these are the exceptions and not common practice.

The citation index is invaluable in preparing historical introductions to scientific papers and in preparing critical reviews and books. The citation index finds further use in tracing new applications of theories, methods, instruments, chemicals, etc., and in the location of corrections, errata, amendments, refutations, letters, editorials, discussions, translations, reviews, etc.

The sociological applications of citation indexes for personnel evaluation, faculty promotions, awards, etc., are legitimate to the extent that one judiciously uses the citation index, as a retrieval tool, to facilitate the location of criticisms of a man's work. Qualitative judgments must temper quantitative data. The citation index can also be used to quickly identify scientists currently working in special branches of science either for personnel or communication purposes.

A great deal has been said about the use of citation indexes for ego gratification. Though this topic might best be covered in a purely psychological or sociological study of scientists, it is discussed here because well-meaning individuals have "rejected" citation indexing on the grounds that ego gratification is its only value. We do not believe citation indexing will be supported purely for ego gratification of subscribers. Ego gratification is not the only motivation for a scientist who wishes to determine whether his work has been applied or criticized by others. Certainly, each scientist evaluates the citations to his works differently. One man's work stimulates another. The citation index then facilitates feedback in the communication cycle. Any author may choose to ignore citations to his own work. Nevertheless, he may wish to retrieve publications which cite the works of other scientists in whom he is interested.

HOW TO BEGIN A SEARCH

Scientists and librarians are experienced in the use of conventional subject indexes. To find articles on biochemical studies in which a particular dye indicator was used, one would normally expect to look under a subject heading for the dye. In the citation index the "subject" is the information contained in a starting point--the target reference. Using a paper on the synthesis of the dye, or any other known paper involving the use of the dye, one can trace subsequent or previous papers on its use.

Similarly, one can find current modifications of a work, as, for example, the use of Einstein's formulas for measuring molecular dimensions. His 1906 paper was cited in 1959 in the *Journal of Dairy Science* in a study on the molecular properties of milk.

The citation index is a relatively sophisticated searching tool. Some knowledge of a starting point is assumed. A target reference sometimes must be first identified through the use of a conventional encyclopedia, book, or subject index. The citation index is then used to answer the question, "What has happened since?" We believe this question is fundamental to research activity.

Various search strategies are discussed elsewhere but a few actual searches will quickly demonstrate the simplicity of the citation index system. The conventional cross-referencing structure, *see* and *see also* references, is absent because it is not required. Nomenclatural ambiguities and rules are completely eliminated.

The reader who is interested in more detailed information should read the introduction to each of the several sections in this Genetics Citation Index as well as the papers which have been cited in the bibliography.