

The Evolution of Citation Indexing— From Computer Printout to the *Web of Science*

Jacqueline Trolley and Jill O'Neill

Abstract

Citation indexing was developed in the late 1950s as a new way to monitor, organize, and retrieve the literature. The *Science Citation Index* was one of the first large-scale, machine-generated indexing systems. Over the course of forty years it has become an essential tool for the scientific community. In particular, the *SCI* provided a new dimension in indexing, permitting the researcher to trace the literature both retroactively and retrospectively. Thus the *SCI* complemented traditional bibliographic databases which are designed to assist the researcher with current awareness, to aid in retrieving relevant material from an ever-larger body of literature, and to help separate the more relevant from the mass relevant publications.

Citation indexing was conceived in the early 1950s as a way to monitor, organize, and retrieve published scientific and scholarly literature. While citation indexing had been implicit in legal citators such as *Shepard's Citations*, the concept had not yet been applied to the literature of any field of scientific research. *Shepard's Citations* came into existence in 1873 to provide the legal profession with a tool for tracking subsequent decisions based on cases decided by federal and state courts (Adair, 1955). The *Science Citation Index (SCI)*, launched by the Institute for Scientific Information (ISI) in the early 1960s, was one of the first applications of computers in the production of large-scale, machine-generated indexes.

SCI's development was intimately related to the earlier implementation of *Current Contents (CC)*. In the abstract sense the *SCI* could have been created *de novo*. In practical terms it was the availability of the collection of current journal issues that made it possible to proceed with large-scale experimentation. *CC* (launched as *Contents in Advance* in 1953) was designed to help scientists become aware of what was being published in core journals central to their own investigations as well

as in peripheral journals. It started as a customized service to drug firms in 1957, but by 1958 it was published as *Current Contents/Chemical, Pharmaco-medical and Life Sciences* (Garfield, 1993). It later was expanded into the physical, social, clinical, engineering, agricultural, and arts and humanities editions. Over the course of the late 1950s *CC* demonstrated a multidisciplinary approach to the scientific community.

The Beginnings

The Johns Hopkins Medical Indexing Project was sponsored by the Army Medical Library, later the Armed Forces Medical Library, which ultimately became the National Library of Medicine. Located at the Johns Hopkins University Welch Medical Library in Baltimore, it was established in 1948 to investigate the role of automation in the organization and retrieval of medical literature. In addition to studying machine methods of compiling indexes, the project investigated the human process of selecting subject headings, descriptors, or other indexing terms. The goal was to reduce the human element and thereby increase the speed of cataloging current biomedical articles and including the index entries into the published indexes.

Prompted by a suggestion from Chauncey D. Leake, then chairman of the advisory committee to the Welch project, Eugene Garfield, a member of the project team, investigated the nature and linguistic character of review articles and how they dealt with the literature reviewed. Garfield recognized that they indirectly "indexed" each of the many papers cited. Each sentence in the review, which identified an original published source for a notable idea or concept, was an indexing statement (Garfield, 1993). By capturing these references and organizing them into an inverted list, the researcher could get a view of

the approach taken by another scientist to support an idea or methodology based on the sources consulted and cited. Thus, the addresses of the papers—bibliographic citations—could be assigned by a professional indexer.

In designing the scope of the putative citation index and its cost-effectiveness, Garfield was aware of Bradford's "law of scattering" (Bradford, 1953). Bradford had observed that in any given field of investigation a relatively small group of journals represented the core of the field. However, Garfield discovered that the essential core of journal literature for all fields of scientific research is found in a basic group of five hundred to a thousand journals. Different sets of journals from this basic core will have a greater relevance to one topic and lesser relevance to others. Garfield used the analogy of a comet, "the nucleus representing the core journals of a literature and the debris and gas molecules of the tail of the comet representing additional journals that sometimes publish material relevant to the subject" to describe this (Garfield, 1979). Garfield also observed that the tail of the literature of one discipline consists, in large part, of the core of the literature of another discipline; this is now referred to as Garfield's "law of concentration" (Garfield, 1979). Thus as a complement to Bradford's law, Garfield applied his law of concentration and found that by monitoring the core journal literature in several scientific fields one could optimize the cost-effectiveness of the database. Core journals included those that produced not only the most articles but also the most influential papers as measured by the frequency of citation. However, the Welch Project was terminated in June 1953 before these ideas could be tested.

Early Years at ISI

After the Welch Project, Garfield attended Columbia Library School and began a career as a documentation consultant. He formed DocuMation, Inc., in 1955, which initially published *Management's Documentation Preview*. It ultimately became *Current Contents Management* in 1956, when DocuMation became Eugene Garfield Associates. This firm became the Institute for Scientific Information in 1960.

Eugene Garfield Associates conducted two pilot projects to test the viability of citation indexing. The first involved the creation of a database based on the references cited in five thousand chemical patents held by Merck and other companies. Garfield's collaborator at Merck, Marge Courain, had been a fellow graduate student at Columbia. The references cited in this database were mainly to prior patents, the documentation sources used by patent examiners to support a decision to grant or deny

a patent claim. Garfield and Courain compared the retrieval connections that their experimental patent citation index permitted with those obtained using the Patent Office's classification system. They found that citation indexing retrieved relevant patents that were missed by the Patent Office's current classification system (Garfield, 1979). Eugene Garfield Associates also worked on indexing chemical compounds for the Patent Office under contract with the Pharmaceutical Manufacturers Association.

Several years later, the second, much larger project was launched. In 1960 a grant was obtained from the National Institutes of Health (NIH) and the National Science Foundation (NSF) to build and test a citation index to the published genetics literature. Three test databases were to be created to cover the literature spanning one year, five years, and fourteen years. Each database would include material from a varying number of source publications. The five- and fourteen-year indexes would test the reliability of a narrow, traditional, discipline-oriented *Genetics Citation Index*. The one-year index would test a broader multidisciplinary genetics index, including the emerging field of molecular biology. This index would cover a broadly based set of source publications, since genetics had been invigorated by Watson and Crick's discovery of the DNA double helix in 1953. The emergent field of molecular genetics involved subjects as diverse as crystallography, biochemistry, genetics, and physics. Indeed, some of the early relevant papers in molecular biology were published in the *Review of Modern Physics*. The one-year database drew not only on journals in the field of traditional genetics research but also on a large hardcore interdisciplinary pool of journals ancillary to genetics and molecular biology.

The project employed the automated IBM punched-card system, but workers were still required to key and standardize the varied citation formats. However, the project demonstrated the overall cost-effectiveness of machine-based citation indexing in comparison with traditional human subject indexing. While recognizing the value of natural language title indexing, adopted early on in *CC*, the prime basic objective of the project was to produce the *Genetics Citation Index* proper. The *Genetics Citation Index* would permit the user to determine whether and where any article or book was cited.

During the life of the *Genetics Citation Index* Project the NIH changed its policy of providing grants to all types of organizations. The new policy required that contracts be negotiated with for-profit organizations. Consequently, the NSF was given the task of administering the contract, which also later included a study of coverage by traditional abstracting services. That study

demonstrated the many gaps in article coverage of many journals, especially those with multidisciplinary scope. In particular the letters and other editorial items were often missed. So the *SCI* later adopted a full coverage policy.

At the project's completion the government sponsors chose not to subsidize the development of an ongoing citation index database. Garfield made the financially risky decision to move ahead with the private publication of the already prepared multidisciplinary index. The first edition of the *SCI*, covering 1961 source literature, required six volumes. It was made available for purchase in 1963. The *Permuterm Index* was added to *SCI* as an outgrowth of experience with producing weekly subject indexes for *CC* (Garfield, 1957). Permuterm indexing was designed in 1964 by Garfield and his research collaborator, Irving Sher. It involved pairings of words from article titles in such a way as to give users with limited information a way into the multidisciplinary coverage of *SCI*, even if they did not have a particular paper in mind. It eliminated some of the difficulties associated with keyword-in-context (KWIC) indexes (or rotated indexes), which were popular at the time. Permuterm indexing required that all titles be in English, necessitating translation.

In 1964 the *SCI* was launched on a current quarterly basis with an annual cumulation. In 1970 a five-year cumulation covering 1965 to 1969 was produced. Eventually cumulated citation indexes for 1945 to 1954 and 1955 to 1964 were created. Few people, even at ISI, believed that the costs of these indexes could be recovered, but Garfield believed the leading research libraries of the world would eventually buy these indexes for historical and sociological research. He felt they were essential to the future value of *SCI* as a tool for contemporary history of science and technology. His prediction proved correct.

The *Social Sciences Citation Index (SSCI)* was launched in 1965 and its source literature now goes back to 1956. The *Arts and Humanities Citation Index (A&HCI)* was started in 1975. Since 1980 the *SCI*, *SSCI*, and *A&HCI* have been offered in CD-ROM format. Also in 1975 ISI's new *Journal Citation Reports (JCR)* was included as the last volume of the *SCI*. The *JCR* would eventually become a separate service. *JCR*'s current impact factors and other citation data have a great influence on journal and research evaluation worldwide (Garfield, 1976).

Standard measures of relevance made popular by the Cranfield group led by Cyril Cleverdon could not be applied to the evaluation of citation indexing because by using cited reference searching a researcher was, in fact, able to retrieve papers that at first glance might not

seem relevant to his or her study. Yet these references often proved crucial to research, and users of the *SCI* soon recognized this advantage. The *SCI*, *SSCI*, and *A&HCI* are considered today to be among the most reliable resources for tracing the development of scientific or scholarly ideas beginning with the primordial papers or books on any given topic.

In 1997 ISI launched a Web-based and completely integrated continuation of the *SCI*, *SSCI*, and *A&HCI*. Known as the *Web of Science*, it bridges the cultures of the arts and sciences, providing integrated coverage of all the academic disciplines via the Internet or intranets. Citation networks are an inherently hypertext approach to navigation of the literature: Users can instantaneously search the bibliographic literature independent of time. Bibliographic coupling, called related records, provides an additional method of clustering documents.

It is significant to the history of citation indexing that Garfield began his career as a chemist. In the same period that *CC* was growing from one to seven editions across the academic and industrial spectrum, Garfield also pursued his dream of a unique chemical information service. Thus, in 1960 he launched the *Index Chemicus*. It is now approaching its fortieth anniversary and has culminated in the development of an integrated chemical compound and reaction database fully linked to the citation index.

This short review of the history of ISI and its work in developing citation indexes has omitted numerous details. Many of these have been reported on in a thoughtful investigation by Paul Wouters in his remarkable doctoral dissertation (Wouters, 1999).

References

- Adair, W. C. (1955). Citation indexes for science. *American Documentation* 6, 31.
- Bradford S. C. (1953). *Documentation* (2nd ed.). Washington, DC: Public Affairs Press.
- Cleverdon, C. (1970). Evaluation tests of information retrieval systems. *Journal of Documentation* 26(1), 55–67.
- Garfield, E. (1957). Breaking the subject index barrier—a citation index for chemical patents. *Journal of the Patent Office Society*, 34, 583–589.
- Garfield E. (1976). The Permuterm Subject Index: An autobiographical review. *Journal of the American Society for Information Science*, 27, 288–291.
- Garfield E. (1979). *Citation indexing: Its theory and application in science, technology and humanities*. New York: John Wiley & Sons, Inc.
- Garfield E. (1993). *Essays of an information scientist: Of Nobel class, women in science, citation classics and other essays* (Vol. 15). Philadelphia: ISI Press.
- Wouters, P. (1999). *The citation culture*. Doctoral dissertation, Science and Technology Dynamics, University of Amsterdam, The Netherlands.