History of Citation Indexes for Chemistry: A Brief Review

EUGENE GARFIELD
Institute for Scientific Information®, Philadelphia, Pennsylvania 19104

Received January 11, 1985

Also see Introduction on p.41, Essays, Vol:9

The *Science Citation Index*® (*SCI*®) was the first comprehensive citation index for chemistry. But its use in chemistry was not obvious, even though the *SCI* covers every important journal of chemistry. However, citation-based searching bypasses dependence on chemical nomenclature. Finding applications of synthetic methods and physical-chemical equations is simplified. These are fields where use of traditional indexing is difficult. An extension of citation indexing, co-citation clustering, is now also used for automatic hierarchical classification and mapping of literature. The value of citation indexes to the historian of chemistry will continue to increase as *SCI* coverage is extended back to include the pre-1955 literature.

This year marks the 25th anniversary of the *Journal of Chemical Information and Computer Sciences*. Appropriately, the journal's name has changed since it was founded in 1961 as the *Journal of Chemical Documentation*. When I contributed a paper to the first issue,[1,2] the field seemed to be moving quite slowly. In retrospect, it was difficult to imagine how rapidly problems would be solved that then seemed insolvable.

Over the past 25 years the scope of the journal has broadened as the field of chemical documentation has advanced. This explains why the journal's title was changed in 1975 to reflect the growth of the information and computer sciences. Of the many advances made in chemical information retrieval over the past 25 years, I have been asked to review the history of citation indexes for chemistry.

It is sometimes difficult to discuss problems that are specific to chemical information because they cannot really be separated from the broader problems of scientific infor-

Reprinted with permission from *J. Chem. Info. Comput. Sci.*, 1985, *25*, 170-174. Copyright 1985 American Chemical Society.

mation. Modern chemistry is unavoidably multidisciplinary. *Chemical Abstracts* (*CA*), which calls itself a chemical information service, is used in medicine, engineering, and other disciplines, as well as in chemistry. It can be argued that the first multidisciplinary *Science Citation Index*® (*SCI*®), which covered the literature of 1961, was the first comprehensive citation index for chemical information.

The history of chemical documentation has always been schizoid. We know that chemical information is inherently multidisciplinary in its application, but it also has its peculiar problems. I discussed this extensively, earlier.[3] But there are certain types of information that are specific to chemistry, such as the structure and composition of molecules. These retrieval problems generally are not solved by standard indexing methods. This in part explains the existence of *Current Abstracts of Chemistry and Index Chemicus*® (*CAC&IC*®) and its progenitor *Index Chemicus*® (*IC*®).

Indexing documents to facilitate their retrieval has always been the primary objective of citation indexes. During the formative years of the *SCI*, it was stated repeatedly that retrieval by citation indexing would overcome some of the inherent limitations of then-existing methods of retrieving chemical information. I had worked briefly in physical chemistry, so I was particularly conscious of retrieval problems that were not solved easily by searching *CA*. For example, how do you find all of the papers that mentioned one of Hammett's equations in *Physical Organic Chemistry*?[4]

This question was quite reasonable in the early days of physical organic chemistry. Nevertheless, it was very difficult to do such a search without scanning every article. The problem epitomizes the information needs of workers in a newly developing field. Once the volume of literature on any particular subject gets large enough, traditional indexing systems may be helpful. But in the formative period, when ideas are not expressed precisely, even an informal nomenclature to which one can attach an idea may not exist. Citations symbolize the early expression of those formative ideas. Each new use of a paper is important to the inventor of that idea.[5] Of course, not all ideas we express in papers are supported by formal citations. But when they are, it is easier to find out who has written about them through citation indexes.[6]

My first experiment with indexing chemistry was with patents. I reported on this early experiment in the *Journal of the Patent Office Society*.[7] It is most unfortunate that proposals for a *Patent Citation Index* were ignored.

A significant boost to citation indexing came in the late 1950s from the *Genetics Citation Index*[8] project. Many scientists were already aware of the broad biochemical significance of the emerging field of molecular genetics. Molecular biology was a new field, and some of the most important papers were reported in the *Reviews of Modern Physics*. Our advisory committee members included Joshua Lederberg, Sol Spiegelman, Gordon Allen, and L. Cavalli-Sforza. They decided that only a complete, multidisciplinary input could guarantee that everything important in general and molecular genetics would be picked up.

The result was the 1961 *SCI*. It covered "only" 613 journals, but these included most of the important Western chemistry journals. When the *SCI* for 1955-1964 was created recently,[9] only about 50 chemistry journals needed to be added to its coverage. Many of these were Soviet and Japanese journals. Only six were significant, pure chemistry titles, such as the *Bulletin of the Chemical Society of Japan* and *Zhurnal Obshchei Khimii*. In confirmation of this important expression of my Law of Concentration,[10] when we started *IC*, most of the new compounds and reactions were found by scanning only about 100 journals. (I commented on this in an essay in *Current Contents*® (*CC*®) in 1969.)[11]

While the work on the putative *SCI* was taking place between 1958 and 1961, *IC* also was being planned. It was initiated in 1960 as a current-awareness service for chemists and in particular as an up-to-date molecular-formula index. We were aware of the peculiar problem of chemical substructure analysis and retrieval. But we did not think that citation indexes would help solve such problems directly. We always thought that the *SCI* would "merely" complement the use of molecular-formula indexes or other indexes organized by line notation or nomenclature. It was much later that we began to realize the implications of citation indexing even for such structural retrieval problems. And more recently, we have demonstrated the hierarchical capabilities of co-citation clustering.

We correctly thought that chemists would be needed to do the kind of indexing done in *CAC&IC* to this day. We also believed that the first step in a literature search would be to use its progenitor, *IC*. You would first find the most relevant compound and the paper in which it was reported. The second step would be to look up its applications in the *SCI*. We thought that chemical classification was beyond the power of citation indexing. That is simply because we never attempted to define the term properly.

Another common myth about chemical information retrieval was that only humans could translate chemical names or nomenclature into structural diagrams or molecular formulas. Since then, it has been demonstrated that computers can use the linguistic properties of nomenclature to generate structural diagrams. We can now use the connectivity tables implied by a name to generate graphics for visual displays in *IC Online*. This is now available from the French database vendor, Telesystèmes, which uses Questel and DARC software systems to handle bibliographic and structural data, respectively.[12]

On numerous occasions during the evolution of the *SCI*, we stressed to synthetic chemists that the *SCI* could help them find chemical information. Using the primordial or key reference for a reaction, you can find subsequent uses for it reported in the literature. For example, a chemist can use the primordial reference for the now classical Eschenmoser hydrolysis[13] to find the papers that have used the method. These papers may list additional uses or modifications of the process. Most chemists will agree that if you are going to use a particular reaction, it is helpful to know about other people's experiences with it.

So it is not surprising that the *SCI* has been widely used in chemistry, but its application has been limited because chemists do not adequately identify with it. Among biochemists, the *SCI* has been more popular because of its close connection to the life sciences and medicine where the *SCI* has been adopted universally. For these reasons we have created a separate *Chemistry Citation Index*™ (*CCI*™). This file will be made available online and, if appropriate, also in printed form.

The next stage in the evolution of citation indexing was the development of co-citation analysis as a classification system. The basic assumption of citation indexing is that a unique, highly specific group of papers is formed when they cite one particular article or book. From this assumption it ought to be simple to perceive that the citation of two papers provides an even more unique grouping. Two papers are frequently cited together, or co-cited, because the citing authors have established an important connection between those two papers. By using pairs of papers, citing authors themselves help categorize the current literature into more specific research topics.

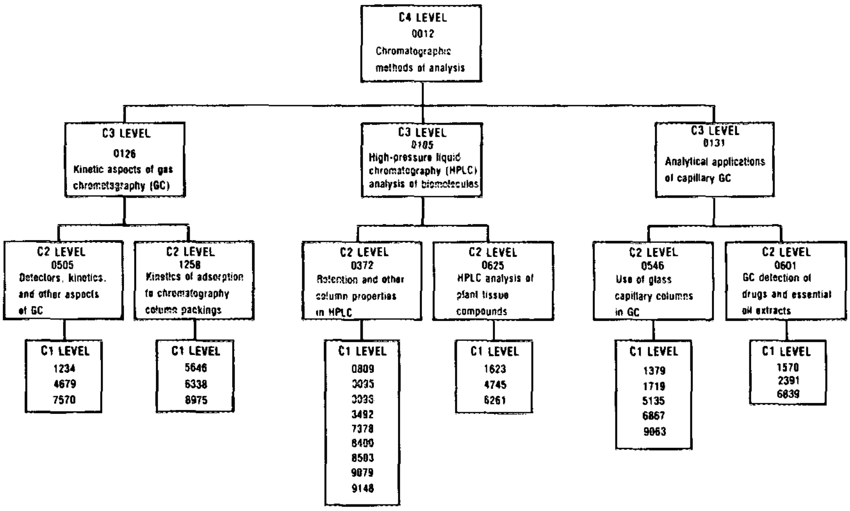
By procedures described elsewhere, we use single-link clustering to form clusters of co-cited papers. Together with the papers that cite them, research fronts are identified. The fronts are named by examining the frequency-ranked list of key words and phrases used in the titles of the citing papers.[14] Once co-citation analysis became a well-understood mechanism for classifying literature, we realized that we had added a new dimension to information retrieval. We had created a system for identifying the emerging scientific research fronts.

At one point, we were going to issue an *Atlas of Science*,[15] which simply provided the bibliographic information for several thousand research fronts, as well as a series of co-citation maps. This bibliography of the core papers of science was to have been a supplement to the *SCI*. I decided to hold back on publishing this primordial atlas. It would have been a new and useful kind of "thesaurus" of current research topics. But I felt we needed to make another quantum leap. We needed to make the transition from bibliographism to encyclopedism. We would do this by including "minireviews" in the atlas. Each minireview would require 750 words or less. It would include all of the key ideas associated with the core papers. It would include a discussion of the relevant current literature for each research front covered in the atlas. The current citing literature would be ranked by relevance, that is, by the number of core papers it cited.

Once we created the first minireviews for the atlas, I knew we had the makings of a large-scale encyclopedia service. After we completed a few minireviews in synthetic organic chemistry, I also realized that we might have a system for emulating Friedrich Beilstein's *Handbook of Organic Chemistry*.[16] By this I mean that co-citation clustering had permitted us to develop increasingly hierarchical classes of chemical information based on the two quantitative criteria used in the clustering procedure—citation threshold and co-citation strength.

Another important step in putting together the *Atlas of Science* was the identification and naming of 10,000 or more new research fronts each year. It took much time and effort to learn how to do this efficiently. The first use of these research fronts was made in the online systems called *ISI/BIOMED*®,[17] *ISI/GeoSciTech*™,[18] and *ISI/CompuMath*®.[19] These ISI online and print products cover the literature of biological sciences,

Figure 1: Hierarchies in clustering. The vast literature of chromatography is illustrated in this hierarchic listing of topics and research fronts. Subtended under the general classification of chromatographic methods of analysis (C4) will be found a series of lower, more specific levels (C3). These include high-pressure liquid chromatography (HPLC) analysis of biomolecules, kinetic aspects of gas chromatography (GC), and analytical applications of capillary GC. Within each of these broader categories are further subdivisions (C2), such as retention and other column properties in HPLC and analysis of plant tissue compounds by HPLC, to name two. Within each C2-level cluster will be found the C1 research fronts; for example, under C2 retention and other column properties in HPLC are the C1 research fronts 0809, 3095, 3096, 3492,7378, 8400, 8503, 9079, and 9148. Full names for the C1 research fronts are listed below the figure.


C4 LEVEL
0012
Chromatographic methods of analysis
C3 LEVEL
0126
Kinetic aspects of gas chromatography (GC)
C3 LEVEL
0105
High-pressure liquid chromatography (HPLC) analysis of biomolecules
C3 LEVEL
0131
Analytical applications of capillary GC
C2 LEVEL
0505
Detectors, kinetics, and other aspects of GC
C2 LEVEL
1258
Kinetics of adsorption to chromatography column packings
C2 LEVEL
0372
Retention and other column properties in HPLC
C2 LEVEL
0625
HPLC analysis of plant tissue compounds
C2 LEVEL
0546
Use of glass capillary columns in GC
C2 LEVEL
0601
GC detection of drugs and essential oil extracts
C1 LEVEL
1234
4679
7570
C1 LEVEL
5646
6338
8975
C1 LEVEL
0809
3095
3096
3492
7378
8400
8503
9079
9148
C1 LEVEL
1623
4745
6261
C1 LEVEL
1379
1719
5135
6867
9063
C1 LEVEL
1570
2391
6839

1234	Studying copolymerization by inverse GC
4679	GC methods to determine equilibrium phenomena
7570	New detectors in GC and HPLC
5646	Kinetics of adsorption and dispersion in packed beds
6338	Adsorption and mass-transfer parameters from chromatography columns and porous catalysts
8975	Ion-exchange and liquid-column chromatography of yeast amino-acid autolysates
0809	Reversed-phase or ion-pair HPLC analysis of nucleotides, bile acids, and aromatic compounds
3095	Use of carbon as packing material for HPLC columns
3096	Use of silica in HPLC
3492	Practical aspects of HPLC
7378	Separation, retention, and selectivity in reversed-phase HPLC
8400	Reversed-phase HPLC determination of proteins and other biological compounds
8503	Theory of retention and other aspects of reversed-phase HPLC
9079	Reversed-phase HPLC with chemically bonded phases
9148	Absorption theory and retention effects in HPLC
1623	Reversed-phase HPLC determination of metal ions and ionic compounds
4745	Ion-exchange and reverse-phase HPLC analysis of organic acids in plant tissues
6261	Sugar-beet metabolism and chemical composition
1379	Fused silica and glass capillary columns in GC
1719	GC/mass spectrometry detection of a trichothecene toxin in foods
5135	Capillary GC analysis by on-column injection
6867	GC analysis by glass capillary columns
9063	Metabolic analysis of trichothecenes and other toxins
1570	Detection of toxic and fatal drug concentrations by capillary GC
2391	GC characterization of essential oil extracts used in brewing
6839	Retention indexes in GC

geosciences, and mathematical and computer sciences, respectively. For each file we created a separate thesaurus of research fronts, by creating separate subdivisions of the *SCI*. While each of these files contains biochemistry, geochemistry, and computational chemistry information, the literature of organic chemistry was not covered extensively. We subsequently created a separate more detailed *Biochemistry Citation Index*™ and then more recently the *CCI*.

Since that time, we have learned not only how to identify research fronts through clustering but also how to create hierarchical clusters that more closely mimic what chemists do when they classify compounds and reactions. For example, we start the clustering process by identifying all highly cited papers and the documents that cite them. We then identify those highly cited papers that are co-cited. A citation threshold is set, and eventually, thousands of research fronts are identified. Each front is associated with two or more core papers. We then cluster these research fronts in much the same way to form about one-sixth as many subspecialty research areas. These areas then are clustered again to identify about 500 subdisciplines of science. By creating these "clusters of clusters" we are moving up the hierarchical scale from the more specific to the more generic.[14] Figure 1 gives examples of each level of clustering.

These research fronts will be updated continuously. Research fronts change rapidly in many cases. Each year there are new and/or emerging fronts. The stability of established research fronts depends on the rate of change of the core papers. As the core literature changes, we have to decide whether or not there is a need to change the name of the front. We also have to determine if it has split into several fronts, merged with other fronts, or been eliminated.

As stated earlier, we decided to create minireviews for each research front we identified. It remains to be seen how these reviews compare with traditional review articles. In 1981 we published a prototype of the comprehensive encyclopedic *ISI Atlas of Science*® envisioned so long ago. The *ISI Atlas of Science: Biochemistry and Molecular Biology, 1978/80* consists of 102 chapters that cover distinct subspecialties or research fronts. Each chapter includes a minireview, a cluster map, a bibliography of the core papers for 1978, and a bibliography of the current citing papers.[15] In early 1985, we published a second prototype *ISI Atlas of Science: Biotechnology and Molecular Genetics*. It includes the core literature for 1981 and the citing literature for 1981-1984.[20]

An encyclopedia of any kind is an ambitious undertaking, but clearly, a continuously current and updated encyclopedia, online and in print, is the only way to deal with the dynamic requirements of modern chemistry. Our initial plan is to create at least 5,000 minireviews each year for as many research fronts in all branches of science. A large percentage of these will be in chemistry. We also contemplate a series of volumes such as the *Atlas of Chromatography, Atlas of Organic Chemistry*, and so on. As each series of volumes is prepared, we will gradually fill in the gaps for the previous years by using the database we have developed covering more than 30 years of literature.

The implications of these files for the study of the history of chemistry are already being felt. I think chemists will readily appreciate the use of co-citation analysis for writing the history of chemistry.

By completing the *SCI* for the 20th century, our eventual goal at ISI, we will help historians resolve many controversies. I am particularly eager to finish the *SCI* for 1945-1954 so that we can determine who cited the 1944 classic paper[21] by Avery, MacLeod, and McCarty, the key historical progenitor to the Watson-Crick paper. It will also be interesting to observe how quickly Watson and Crick's 1953 paper[22] on the structure of DNA was cited. We now know that in 1955 it was cited over 28 times. We recently published the *SCI* for 1955-1964 and have already started work on the postwar period.

Another way in which citation indexes have helped in historiographic research is through our *Citation Classics*® series.[23] To date, we have published over 2,000 commentaries by authors of classic papers. Included in these are hundreds of papers from all types of chemistry journals. This will be expanded significantly over the next several years. The commentaries we receive from the authors of these classics provide a special kind of personal autobiography, of which there is too little in the literature. Recently, we published a commentary by Linus C. Pauling on his 1939 book *The Nature of the Chemical Bond*.[24]

We hope that *Citation Classics* will help young chemists realize that there is more to science than what they learn from the most visible scientists. Science is built on the contributions of thousands of creative individuals, as Ortega y Gasset suggested in *The Revolt of the Masses*[25]—not merely an elite group of highly visible or highly cited persons.

REFERENCES AND NOTES

(1) Garfield, E. "Information Theory and Other Quantitative Factors in Code Design for Document Card Systems". *J. Chem. Doc.* **1961,** *1,* 70-75.

(2) Garfield, E. "Information Theory and All That Jazz: a Lost Reference List Leads to a Pragmatic Assignment for Students". In *Essays of an Information Scientist*; ISI Press: Philadelphia, 1980; Vol. 3, pp 271-285.

(3) Garfield, E. "Where Is Chemical Information Science Going?" *J. Chem. Inf. Comput. Sci.* **1978,** *18,* 1-4.

(4) Hammett, L.P. "Physical Organic Chemistry"; McGraw-Hill: New York, 1940; p 404.

(5) Garfield, E. "Citation Indexes for Science". *Science (Washington, D.C.)* **1955,** *122,* 108-111.

(6) Garfield, E. *Citation Indexing—Its Theory and Application in Science, Technology, and Humanities,* 2nd ed.; ISI Press: Philadelphia, 1983; p 274.

(7) Garfield, E. "Breaking the Subject Index Barrier; a Citation Index for Chemical Patents". *J. Pat. Off. Soc.* **1957,** *39,* 583-595.

(8) Garfield, E.; Sher, I.H., Eds. *Genetics Citation Index*; Institute for Scientific Information: Philadelphia, 1963; p 864.

(9) Garfield, E. "The 1955-1964 *Science Citation Index* Cumulation—a Major New Bibliographic Tool for Historians of Science and All Others Who Need Precise Information Retrieval for the Age of Space and Molecular Biology". In *Essays of an Information Scientist*; ISI Press: Philadelphia, 1984; Vol. 6, pp 27-37.

(10) Garfield, E. "The Mystery of the Transposed Journal Lists—Wherein Bradford's Law of Scattering Is Generalized According to Garfield's Law of Concentration". In *Essays of an Information Scientist*; ISI Press: Philadelphia, 1977; Vol. 1, pp 222-223.

(11) Garfield, E. "Introducing *Current Abstracts of Chemistry and Index Chemicus*". In *Essays of an Information Scientist*; ISI Press: Philadelphia, 1977; Vol. 1, pp 63-64.

(12) Garfield, E. "*Index Chemicus* Goes Online with Graphic Access to Three Million New Organic Compounds". *Curr. Contents* **1984,** *No. 26,* 3-10. In *Essays of an Information Scientist: the Awards of Science and Other Essays*; ISI Press: Philadelphia, 1985; Vol. 7, pp 194-201.

(13) Elsinger, F.; Schreiber, J.; Eschenmoser, A. "Notiz über die Selektivität der Spaltung von Carbonsäure-Methylestern mit Lithiumjodid". ("Note on the Selectivity of the Cleavage of Carboxylic Acid-Methyl Esters Using Lithium Iodide") *Helv. Chim. Acta* **1960,** *43,* 113-118.

(14) Garfield, E. "ABCs of Cluster Mapping. Parts 1 & 2. Most Active Fields in the Life and Physical Sciences in 1978". In *Essays of an Information Scientist*; ISI Press: Philadelphia, 1981; Vol. 4, pp 634-649.

(15) Garfield, E. "Introducing the *ISI Atlas of Science: Biochemistry and Molecular Biology, 1978/80*". In *Essays of an Information Scientist*; ISI Press: Philadelphia, 1981; Vol. 5, pp 279-287.

(16) Beilstein, F.K. "Handbuch der Organischen Chemie" ("Handbook of Organic Chemistry"); Springer-Verlag: New York, 1918-; multivolume.

(17) Garfield, E. "ISI's On-line System Makes Searching So Easy Even a Scientist Can Do It: Introducing *METADEX* Automatic Indexing and *ISI/BIOMED SEARCH*". In *Essays of an Information Scientist*; ISI Press: Philadelphia, 1981; Vol. 5, pp 11-14.

(18) Garfield, E. "Introducing *ISI/GeoSciTech* and the *GeoSciTech Citation Index*—the 50 Most-Active Research Fronts in 1981 in the Earth Sciences Illustrate the Unique Retrieval Capabilities of Our New On-line and Print Services". In *Essays of an Information Scientist*; ISI Press: Philadelphia, 1981; Vol. 5, pp 607-614.

(19) Garfield, E. "*ISI/CompuMath*, Multidisciplinary Coverage of Applied and Pure Mathematics, Statistics, and Computer Science, in Print and/or Online—Take Your Pick!". In *Essays of an Information Scientist*; ISI Press: Philadelphia, 1981; Vol. 5, pp 437-442.

(20) Garfield, E. "Introducing the *ISI Atlas of Science: Biotechnology and Molecular Genetics, 1981/82* and Bibliographic Update for 1983/84". *Curr. Contents* **1984,** *No. 41,* 3-15. In *Essays of an Information Scientist; the Awards of Science and Other Essays*; ISI Press: Philadelphia, 1985; Vol. 7, pp 313-325.

(21) Avery, O.T.; MacLeod, C.; McCarty, M. "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types". *J. Exp. Med.* **1944,** *79,* 137-158.

(22) Watson, J.D.; Crick, F.H.C. "A Structure for Deoxyribose Nucleic Acid". *Nature (London)* **1953,** *171,* 737-738.

(23) Garfield, E. "*Citation Classics*—Four Years of the Human Side of Science". In *Essays of an Information Scientist*; ISI Press: Philadelphia, 1983; Vol. 5, pp 123-134.

(24) Pauling, L. Commentary on "The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry". *Curr. Contents/Phys., Chem. Earth Sci.* **1985,** *25* (4), 16.

(25) Ortega y Gasset, J. "The Revolt of the Masses"; Norton: New York, 1957; pp 110-111.