Presented at the American Chemical Society National Meeting, Philadelphia, PA. April 6-11, 1975

March 20, 1975

"The Entrepreneur as a Doctoral Candidate"

In preparing for this talk I had the difficult problem of selecting a topic that would interest you and me. That's not always easy for someone who has to speak regularly. Lately I've used a ploy that some of you may want to use. I prepare abstracts of six different talks I could give and then I ask the audience which they would prefer to hear. This might sound democratic. But I've also learned from experience that my audiences always select the same talk unless I omit it from the list. You may be interested to know that the topic is "How to Forecast Nobel Prize Winners." Well I'm not going to use that technique today but by an associative process I did begin to think about all I've said and written on the use of citation analysis to identify milestone papers in writing the history of scientific disciplines.

We have a high correlation between subjective ratings of important or milestone papers and the frequency of their subsequent citation. This assertion must be tempered by saying it is, on average, a correct assertion. Thus the citation description model of the macrostructure of science holds up well. It will take much more refined techniques to properly identify all milestone or significant papers. This is what we call the microstructure of science. For a variety of reasons, not all such papers were heavily cited. I am mindful of this because the citation history of my own work would not fall into the general pattern we define as citation distinction. Indeed, it is somewhat ironic that this should be true in two separate areas of information research I've done. During the time Science Citation Index® was a research idea it was barely noticed by the scientific or information community. However, once it had been applied in the form of the published SCI®, most users of SCI that had occasion to cite the SCI did so anonymously.

Hagstrom is a good example. (1) Thus citation indexing, like many other methods, was institutionalized rather quickly. On the other hand, my paper on the use of the SCI data base for the testing citation method of evaluation of journals (2) has produced a better than average citation rate. This is because so few people have access to the JCR as yet.

Thus, in my own case the work I regard most highly is least cited. I know this is true for other individuals. How extensive this phenomenom is we have not yet been able to determine. I wonder whether Watson and Crick would agree that their 1952 paper in Nature[3] represents the pinnacle of their work? I know that Oliver Lowry[4] correctly asserts that his most important papers are not his most cited. But that does not say his most important are not heavily cited. Keep in mind that I am not saying citation analysis cannot detect the significant though infrequently cited paper. Back in 1964 we produced a "computerized" history of DNA[5] which showed some papers that were infrequently cited but were significant in breaking down the genetic code.

Examples of this kind have given me reason to question the assertion by Cole[6] that there is no validity in the Ortega hypothesis[7]. The latter theory asserts that advances in science depend in part on the contributions of mediocre scientists. While we may all stand on the shoulders of giants, they in turn depend upon many average scientists. Whether they depend upon dwarfs is another question.

All this is leading up to a discussion of some work I did which is rarely cited but which gave me fantastic satisfaction. I refer to a paper on mechanical translation of chemical nomenclature[8]. This was the subject of my doctoral dissertation. Since I'm so often asked this question, I'd like to tell you how I happened to take a degree in linguistics rather than library science.

I entered the field of documentation, now information science, from chemistry by joining the Johns Hopkins University Indexing project in 1951. I stayed until its demise in 1953. By the middle of 1954 I had already accumulated a Master's Degree in Library Science and sufficient graduate credit to satisfy the minimum requirements for a Ph.D. But it proved impossible for me to find a faculty member at Columbia University who would enable me to write a dissertation on the use of machine methods in scientific information. The only sympathetic ear was that of Professor Merrell Flood, but in order to take a degree with him, I would have had to take undergraduate training in industrial engineering. In retrospect, I see more clearly how relevant systems work has been in my career.

I also tried to form an interdisciplinary faculty group, but I was also not interested in spending ten years trying to satisfy an inter-faculty group that would supervise my work. By that time my family had already been convinced I was going to be a student forever. I left Columbia disappointed. But in 1954, through my friend and colleague, Casimir Borkowski, I met Professor Zellig Harris at the University of Pennsylvania, Department of Linguistics. His work in structural linguistics was already well known to scholars. But in the field of scientific information he was unknown. In 1956 I wrote a paper on the application of structural linguistics to mechanized indexing(9) and showed it to Harris. Though it was never published, Harris becama sufficiently interested in the field of information retrieval to accept some huge grants from NSF over a ten-year period. Helen Brownson had a lot to do with this. Most of this work is now continued primarily by Naomi Sager at NYU[10]. Some of you may recall transformational and discourse analysis.

I suppose it was prestige that made me seek a Ph.D. I ultimately worked out a doctoral program with Professor Harris which commenced officially in 1958. We had agreed on the amount of course work and my ultimate dissertation topic. By then I was quite preoccupied with problems of chemical indexing. We were encoding all new steroids for the U.S. Patent Office under a contract with the Pharmaceutical Manufacturers Assn.
My interest in chemical documentation had, of course, begun at Johns Hopkins. It is difficult for me to explain why I never went to work for E.J. Crane at Chemical Abstracts. He and I discussed this both in Baltimore and Columbus but it never worked out. Maybe Charles Bernier will better recall those days.

By the time 1960 rolled around, ISI® was publishing *Index Chemicus*®. The original purpose of this service had been *merely* to index compounds by molecular formula. So it was natural for me to want to find a way of calculating molecular formulae in the simplest way possible. Until that time *everyone* assumed that it was necessary to draw a structural diagram in order to calculate a molecular formula. Even my good friend, Ascher Opler[11], who wrote the pioneering paper in 1956 on "New Speed to Structural Searches", assumed this was the case. That is why he first wanted to represent the compound in a topological matrix which later was called a connectivity table.

My linguistic studies convinced me that the "meaning" of chemical nomenclature had to include enough information for calculating molecular formulae straight away. Otherwise how could we do this so quickly in our heads for simple compounds? I told Harris my theory and he accepted it as my doctoral thesis, the first in the new field of chemico-linguistics. Thanks to the recognition by Prof. Allen Day of Penn's Chemistry Dept. that it was a non-trivial problem, the topic was agreed upon in the graduate school. However, I had to agree, not merely to write a dissertation on my theory. I also had to prove it worked. If it did not I would have to choose another topic, no matter how long I spent on the research.

Recognizing that the dictionary work alone might take me several years unless I got help, I proposed that the theory be proven with respect to acyclic compounds.

During the next few years I got into the detailed problems of discourse analysis for my target language -- chemical nomenclature. The details are not essential to this story. When I was ready for actual computer trials I got the help of John O'Connor in programming Univac I which was then in use at Penn. But I could never get on the computer in time so I bought time at the Franklin Institute computer center.

The outcome of all this was "an algorithm for translating chemical nomenclature into molecular formulas"(12). When I submitted it to the department it was exactly ten pages. My substitute adviser was dumbfounded by this. Dissertations in linguistics are written by the pound -- not the page. I spent a whole semester filling it out with interesting theoretical statements and formal analyses of chemical morphology, etc. But in late 1960 I had already made the first successful computer run in calculating a molecular formula directly from a systematic name[(8)]. I had done this manually hundreds of times earlier in the year.

Professor Harris had gone off to Europe that year and turned me over to this other faculty member who was determined not to make a decision. I spent the next four months writing constant revisions of my dissertation. He did not know that I was using a tape typewriter. So whenever he suggested changes I was able to produce new versions literally overnight. Then towards the end he would make changes that would have delayed any student without access to a Xerox machine. One time I reproduced ten copies of 75 pages overnight so he finally capitulated - especially after I called Harris one night in Rome insisting that he read my dissertation so I could graduate in June. The next day a cable came in and the bureaucracy finally relented.

As it turned out ISI was never able to finance the research necessary to complete this work. NSF was not very kindly disposed to us in those days. We also were up to our ears in the Genetics Citation Index project so I had to put chemical nomenclature work on the back burner. We still do not input compound names for CAC -- on the contrary, we now input WLN for each compound and that is what we use to compute the molform. However, the double bond checking routines that we used for so long were included in my algorithm.

About eight years ago I saw the proposal Chemical Abstracts made to NSF regarding chemical nomenclature translation research. Naturally I felt envious that they should get this support when it was clearly an operational development they needed more than ISI. That's what made it applied for them and academic for us.

However, I was very glad someone was doing this and read with mixed feelings the first reports of this research in 1970[13]. A more recent paper in the Nov. 1972 Journal of Chemical Documentation[14] shows that this work is finally coming to fruition and I am glad to congratulate the CA group on their accomplishment.

Returning to the main point of my essay. Here is a topic of research which has multi-million dollar economic significance. There are only a few people in the world interested in it so the number of times this kind of work will be cited is bound to be small. Clearly it is the kind of thing that is less cited than e.g. papers on WLN but there is an important connecting thread. Perhaps historians will decide that the Opler and Norton notion of a connectivity table for chemical compounds has been the most important concept in this field[11]. Most people seem to think that Sussenguth was the first one to use this concept[15]. But clearly none of these chemical information milestones have had any major discernible impact outside the field and that is what the historian seeks and seems to find in large-scale citation analyses. The microstructure of science is very different than its macrostructure.

So much for the history of mechanical translation of nomenclature. Let me digress now to make some observations on the future of chemical and scientific publication. This has been much in the news these days, that is, C&E News! Joel Hildebrand, my freshman chemistry professor, has caused a lot of soul-searching with his re-discovery of the ancient idea of publication by abstract. I've had some correspondence and contact with him in recent years and I know why he is making these proposals. Unlike James Stemmie who in *C&EN* Jan. 13, p. 33-34,[16] seems worried that some important ideas will be lost to posterity if we adopt any changed systems, Hildebrand is trying to tell us that the system is overloaded with useless information -- he is talking about information pollution on a large scale. I have recently[17] asserted that the abuse of the page charge system may be aggravating this pollution problem. And I regret to say *Chemical Abstracts* may be equally guilty. *CA* does this unwittingly in its hopeless aim to be complete. Consider that 25% of the abstracts in *CA* are of Russian material.[19] At ISI we have compiled data that shows this is absurd in relation to the significance of Russian research. They are polluting the waters of science with a lot of mediocre and unrefereed material. Probably another 10% of *CA* falls in this category. No doubt others do it too, but the data shows how clearly the Russians are the worst offenders. Is anyone in the ACS prepared to debate whether or not the *J ACS* is superior to the *Zhurnal Obshchei Khimii*? Or how would you even compare the Abstracts of the ACS meeting to the abstracts of unpublished papers that the Russians are now loading into the *Journal of Physical Chemistry*. Undoubtedly it gives the Russians significant political leverage to assert they account for 25% of *CA*'s coverage. Maybe they will even claim *CA* should pay them a royalty for abstracting without their permission. After all, doesn't *CA* abstracts constitute a substitute for

depository. Each abstract requires the same space and work. But at least someone was willing to pay for that so-called dignified journal. If depository. Each abstract requires the same space and work. But at least someone was willing to pay for that so-called high priced journal. If librarians were as indiscriminate as they are accused of, then why aren't they buying the original Russian journals and abstracts? I'm sure that Earl Coleman would be delighted if libraries bought his translation journals without the slightest evaluation. He knows how hard it is to sell the best that the Russians publish. He would court disaster to publish everything without regard to quality.

It is a rather interesting observation that 10% of *CA*'s budget is about $2 million. If they cut back on Russian material they would find the same $2 million they want the Russians to pay us for pirating *CA*.

At ISI we have very mixed feelings about *CA*. On the one hand we resent their high price because a chemistry dept. is generally apt to say we can't afford *SCI* but we must buy *CA*. If for no other reason, they couldn't get ACS accreditation without it. On the other hand, the higher *CA* becomes the more easily we can convince buyers that *SCI* or *CAC* is good value. However, given my choice I would much rather see *CA* priced lower. So I have a real concern for their cost-effectiveness. In fact, given my druthers, we would provide for *CA* a citation index to the chemical literature that would complement *CA* searches. The combined use of *CA* and *SCI* is happening increasingly. But it would be nice if we could accelerate the use of *SCI* by chemists as was suggested by the Hannay Committee many years ago.[19]

The recent paper by Parry, Linford, and Rich in the Information Scientist(20) shows a clear trend towards such complimentary use of large data bases. This will only increase as the cost of on-line services declines.

I recently did a search of the CA data base using our PSI to identify pertinent search terms and then follow-up the output from CA by checking the items retrieved in the SCI! This is frequently done when people use Medline & SCISEARCH but obviously the inclination to do so is tempered by the vast differences in per hour rates.

As a closing act I thought I would refer to miniprint which has now come into the limelight. I've passed around some samples in miniprint I prepared about fifteen years ago. As the cost of paper goes up, CA and ISI may well have to adopt such methods. Whether users will accept miniprint more readily than microform is hard to determine, but there is a whole new technology opening up there, now that OED has become so successful in this medium. Ralph Shaw and Albert Boni experimented with miniprint long ago. I just rediscovered it when I was thinking about ways to cut down on indexing costs. Maybe it's still not too late for CA to try it. After all, the most successful publishing venture of the past decade has been in the miniprint edition of the Oxford English Dictionary.

References:

1. Hagstrom, W. O. "Inputs, Outputs, and the Prestige of University Science Departments," Sociology of Education 44, 375-397 (1971).

2. Garfield, E. "Citation Analysis as a Tool in Journal Evaluation," Science 178, 471-479 (1972).

3. Watson, J. D. & Crick, F. H. C. "A Structure for Deoxyribose Nucleic Acid," Nature 171, 737 (1953).

4. Lowry, O. Personal Communication to D. J. D. Price. Quoted in Garfield, E. "Citation Frequency as a Measure of Research Activity and Performance," Current Contents No. 5, 5-7 (1973).

5. Garfield, E., Sher, I. H., Torpie, R. J. The Use of Citation Data in Writing the History of Science. (Philadelphia: Institute for Scientific Information, 1964) 86 pp.

6. Cole, J.R. & Cole, S. "Ortega Hypothesis," Science 178, 368-375 (1972).

7. Ortega y Gasset, J. The Revolt of the Masses (New York: Norton, 1932) pp. 84-85.

8. Garfield, E. "Chemico-Linguistics: Computer Translation of Chemical Nomenclature," Nature 192, 192 (1961).

9. Garfield, E. "Proposal for Research in Mechanical Indexing," Unpublished (1956).

10. Sager, N. "Syntactic Formatting of Science Information," AFIPS Conference Proceedings 41, 791-800 (1972).

11. Opler, A. & Norton, T. R. "New Speed to Structural Searches," Chemical and Engineering News 34, 2812-14 (1956).

12. Garfield, E. An Algorithm for Translating Chemical Names to Molecular Formula (Philadelphia: Institute for Scientific Information, 1961) 68 pp.

13. Vander Stouw, G. G., Naznitsky, I., & Rush, J. E. "Procedures for Converting Systematic Names of Organic Compounds into Atom-Bond Connection Tables," Journal of Chemical Documentation 7, 165-169 (1967).

14. Vander Stouw, G. G., Elliott, P. M., and Isenberg, A. C. "Automatic Conversion of Chemical Substance Names to Atom-Bond Connection Tables," Journal of Chemical Documentation 14, 185-193 (1974).

15. Sussenguth, E. H. "Graph Theoretic Algorithm for Matching Chemical Structures," Journal of Chemical Documentation 5, 36-43 (1965).

16. Stemmle, J. T. "Control of Scientific Papers," Letter to the Editor of Chemical & Engineering News 53, 33-34 (1975).

17. Garfield, E. "Page Charges -- For-Profit and Non-Profit Journals -- And Freedom of the Scientific Press," Current Contents® No. 7, 5-7 (1975).

18. Baker, D. "World's Chemical Literature Continues to Expand," Chemical and Engineering News 49, 37-40 (1971).

19. Anonymous, "ACS Report Rates Information System Efficiency," Chemical & Engineering News 47, 45-46 (1969).

20. Parry, A. A., Linford, R. G., & Rich, J. I. "Computer Literature Searches--A Comparison of the Performance of Two Commercial Systems in an Interdisciplinary Subject," Information Scientist 8, 179-187 (1974).