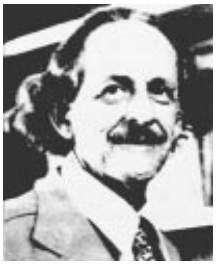
Essays of an Information Scientist: Creativity, Delayed Recognition, and other Essays, Vol:12, p.166, 1989 Current Contents, #25, p.3-14, June 19,1989



Current Comments®
EUGENE GARFIELD
INSTITUTE FOR SCIENTIFIC INFORMATION®
3501 MARKET ST., PHILADELPHIA, PA 19104

Manfred Kochen: In Memory of an Information Scientist Pioneer qua World Brain-ist

Number 25 Go to Reprint June 19, 1989

For the sake of historical accuracy, it is important to point out that even before we published our recent essay on citation behavior,[1] we had planned to reprint the following paper by Manfred Kochen.[2] Then, while preparing an introduction to appear with Kochen's paper in *Current Contents®*, we learned that he had died suddenly in Colorado. We have been deprived of a dear friend and of the chance of conveying to this dedicated scholar this further evidence of our esteem for his work. Perhaps his family and friends will derive some small solace from this posthumous recognition of his contribution to science.

Biographical Data

Fred Kochen was born in Vienna, Austria, on July 4, 1928. After narrowly escaping the threat of Nazism by boarding one of the very last boats to leave Lisbon, Portugal, Fred and his family eventually made it to New York City. In his adolescent years he attended Brooklyn Technical High School, where he was awarded a Mechanical Course diploma in 1947. He met his future wife, Paula, in New York as well: they were married in 1954.

Fred's education included a BS (1950) in physics from the Massachusetts Institute of Technology, Cambridge, followed by MA (1951) and PhD (1955) degrees in mathematics from Columbia University, New York. He was an analyst with the John von Neumann computer project at the Institute for Advanced Study, Princeton, New Jersey (1953-1955), and continued his postgraduate work at Harvard University, Cambridge (1955-1956), where he worked on mathematical models in the behavioral sciences. From 1956 to 1964, he was a research mathematician at the IBM Thomas J. Watson Research Laboratory, Yorktown Heights, New York.

From 1965 onwards, he was a research professor at the University of Michigan, Ann Arbor, and from 1970 to 1989 he was also a professor of information science at the University of Michigan Medical School. While he was a professor at the Medical School, he was (1981-1989) an adjunct professor of computer and information systems, as well, in the School of Business Administration.

Even before his collaborative work with von Neumann at the Institute for Advanced Study, Fred had a long history as a consultant. Between 1948 and 1955, he held part-time consultancy positions in spectroscopy, aeroelasticity, and photogrammetry. In the 1960s he was a consultant to RAND, RCA Laboratories, the United Aircraft Corporation, and EURATOM (Ispra, Italy); in the 1970s, to the Advanced Research Projects Agency, the National Institute of Education, the Science Center of Berlin (Federal Republic of Germany), and the National Science Foundation's (NSF) Division of Science and Technology. Just last year he was a consultant to the US Library of Congress, Washington, DC.

Fred Kochen authored and edited over 250 publications (including eight books), primarily in information science, artificial intelligence, and the behavioral sciences. Table 1 presents a selected chronological list of his works. He was also a member of several

professional and honorary societies: the American Association for the Advancement of Science (AAAS), the American Mathematical Society, the American Physical Society, the American Society for Information Science (ASIS), and Sigma Xi.

The direct connection between Fred and ISI® began in 1972 when I first came across his paper on a proposed world information retrieval system[3] (which he termed "WISE," for "world information synthesis and encyclopaedia"). At that time, I was writing "The World Brain as seen by an information entrepreneur,"[4] for the AAAS meeting on Reorganizing Information Resources to Improve Decision-Making, which was held in San Francisco in February 1974. H.G. Wells's concept of the World Brain[5] had been an inspiration to me early on, and I was delighted it was being recognized as significant by others. Fred and I corresponded from 1973, but our contacts may have been even earlier. For example, Fred was closely associated with Merrill Flood at Columbia. In the early spring of 1954, I had discussed the possibility of taking my doctorate with Professor Flood, and, a few years later, Fred had many contacts with my close friend the linguist Casimir Borkowski.

Kochen the Man: Some Colleagues Remember

When Fred Kochen was nominated to be president of ASIS back in 1986, he interpreted his discipline in these words:

> I am an information scientist. I interpret it very broadly. For me, it includes the study of how brain becomes mind and of the evolution of social organs with mind-like properties, such as scientific communities; how to design and use computer information systems in business; and new roles for information professionals as referential consultants, catalytic brokers, and chief information officers.[6]

That definition of the information scientist is worthy of continued discussion in what may never be a settled debate. But defining oneself as an information scientist only reveals so much about a person, especially one so many-sided as Fred. Feeling that my own remarks could not adequately describe Fred's varied contributions, I asked several colleagues to share some of their thoughts about him.

Henry Small, director of corporate research, ISI, recalls his association with Fred, especially his involvement in clustering techniques and science mapping. (In fact, Fred helped organize an NSF-ISI sponsored conference on clustering held at ISI in September 1986.)

> Fred was very interested in getting a handle on all human knowledge. Some of his work, for example, on discovery and creativity, was quite original. From my view he was more a theorist rather than an empiricist. I remember him telling me that he had anticipated co-citation and some other techniques used here at ISI. On a personal level, he was quiet, and to many he may have appeared introverted. But he was like Derek [J. de Solla Price] in that he was inspirational and had the uncommon knack of bringing people together for a common goal.[7]

Robert Lindsay, research scientist, Mental Health Research Institute, University of Michigan, had this to say about Fred and his connection with the World Brain concept:

> I knew Fred for 30 years, having first met him about 1958. I wasn't intimately involved with Fred's work on the World Brain notion. However, I think that it was a sort of guiding principle of his, rather than a specific project. A variety of things that he worked on were directed towards achieving information systems that would make information readily available to people who needed it. He worked on problems of computer-based matching systems and information retrieval, and the "small world" problem. He approached the World Brain notion indirectly through all his work. He had an enormous range of interests, and breadth of knowledge, and brought great energy and humanity to his work and to his personal relationships.[8]

As a teacher, Fred was one of the best. According to Wynne Chin, one of his students:

Table 1: Selected list of works by Manfred Kochen arranged in chronological order.

Kochen M. Extending the human record. *Bull. Amer. Soc. Inform. Sci.* 14(6):25-6, August-September 1988.
Smit P H & Kochen M. Information impediments to innovation of on-line database vendors. *Inform. Process. Manage.* 24:229-42, 1988.
Kochen M. Advanced information technology and small manufacturers. *Bull. Amer. Soc. Inform. Sci.* 14(4):24-7, April 1988.
Kochen M & Hastings H M, eds. *Advances in cognitive science: steps toward convergence.* Boulder, CO: Westview Press, 1988. 283 p.
Kochen M. How well do we acknowledge intellectual debts? *J. Doc.* 43:54-64, 1987.
Kochen M, Lee C & Westland C. The adaptive man-machine non-arithmetic information processing system revisited. (Williams M E & Hogan T H, eds.) *Proceedings of the National Online Meeting*, 5-7 May 1987, New York. Medford, NJ: Learned Information, 1987. p. 253-9.
Kochen M. Ethics and information science. *J. Amer. Soc. Inform. Sci.* 38:206-10, 1987.
Kochen M, Cohen L & Wulff Y. Information systems and clinical research by residents in internal medicine. *Methods Inform. Med.* 24:85-90, 1985.
Kochen M. Information science research. The search for the nature of information. *J. Amer. Soc. Inform. Sci.* 35:194-9, 1984.
Kochen M. Towards a paradigm for information science: the influence of Derek de Solla Price. *J. Amer. Soc. Inform. Sci.* 35:147-8, 1984.
Kochen M. Information and society. (Williams M E, ed.) *Annual review of information science and technology.* White Plains, NY: Knowledge Industry, 1983. p. 277-304.
Kochen M, Crickman R & Blaivas A. Distribution of scientific experts as recognized by peer consensus. *Scientometrics* 4:45-56, 1982.
Kochen M & Blaivas A. A model for the growth of mathematical specialities. *Scientometrics* 3:265-73, 1981.
Kochen M & Deutsch K W. *Decentralization: sketches toward a rational theory.* Cambridge, MA: Oelgeschlager, 1980. 384 p.
Kochen M. Dynamics of scholarly communication. (Benenfeld A R & Kazlauskas E J, eds.) *Communicating information: proceedings of the 43rd ASIS Annual Meeting*, 5-10 October 1980, Anaheim, CA. White Plains, NY: Knowledge Industry, 1980. p. 233-5.
Kochen M. Can the behavioral sciences contribute to the foundations of information science? (Martin S K, ed.) *Information politics: proceedings of the 39th ASIS Annual Meeting*, 4-9 October 1976, San Francisco, CA. Washington, DC: American Society for Information Science, 1976. p. 81.
Donohue J C & Kochen M, eds. *Information for the community.* Chicago, IL: American Library Association, 1976. 282 p.
Kochen M. *Principles of information retrieval.* New York: Wiley, 1974. 203 p.
Kochen M & Tagliacozzo R. Matching authors and readers of scientific papers. *Inform. Storage Retriev.* 10:197-210, 1974.
Kochen M. *Integrative mechanisms in literature growth.* Westport, CT: Greenwood Press, 1974. 275 p.
Kochen M & Badre A N. Questions and shifts of representation in problem solving. *Amer. J. Psychol.* 87:369-83, 1974.
Kochen M. WISE: a world information synthesis and encyclopaedia. *J. Doc.* 28:322-43, 1972.
Kochen M. Directory design for networks of information and referral centers. *Libr. Quart.* 42:59-83, 1972.
Kochen M. Automatic question-answering of English-like questions about simple diagrams. *J. Assn. Comput. Mach.* 16:26-48, 1969.
Kochen M. Stability in the growth of knowledge. *Amer. Doc.* 20:186-97, 1969.
Kochen M & Deutsch K W. Toward a rational theory of decentralization: some implications of a mathematical approach. *Amer. Polit. Sci. Rev.* 63:734-49, 1969.
Kochen M & Tagliacozzo R. A study of cross referencing. *J. Doc.* 24:173-91, 1968.
Kochen M. *The growth of knowledge: readings on organization and retrieval of information.* New York: Wiley, 1967. 394 p.

I have worked with quite a few professors in my years of graduate studies...[but Professor Kochen] was my first true mentor. With extreme patience and clarity, he would ask penetrating questions about my research that would reflect his intellectual understanding.... In many instances, it would [only] be months later that certain questions he continually raised finally made sense to me.... [H]e had developed an expertise in guiding a student's path towards "intellectual enlightenment"[9]....

Flood, formerly University of Michigan professor of mathematical biology and previous to that professor of industrial engineering at Columbia, fondly remembers Fred both in and out of academe.

For several years, while Fred was manager of a department at the IBM Thomas J. Watson Research Laboratory, I worked with him and his colleagues as a consultant. He generated ideas that kept us all

excited and very busy, many of them well ahead of his time as we look back on them now. One among these projects that I worked closely on with Fred, and others, was the pioneering investigation of automation possibilities for the Library of Congress. Fred continued an active and productive research and teaching career in library automation and library science until his death—a notable tendency of Fred's to stay with an important basic problem.... When Fred decided to return to academic life, it was my great good fortune to bring him to the Mental Health Research Institute in the Department of Psychiatry at the University of Michigan Medical School as a colleague in the Systems Science Group. In addition to his own wide-ranging research and teaching efforts in several schools of the university at Ann Arbor, and his many external consulting and professional society activities, Fred partnered joint research efforts with Anatol Rapoport and Karl Deutsch and me, giving generously of his time and ideas on every occasion.[10]

Belver Griffith, College of Information Studies, Drexel University, Philadelphia, recalls Fred's command of information:

> My main memory of Fred Kochen was of his optimism, openness, and far-ranging curiosity.... Fred [was] always like a lad who knows that tomorrow is his Bar Mitzvah and that it will be a great one.... There was, however, very little that Fred did not understand and almost nothing that he was willing to dismiss out of hand. A rare combination to find in anyone—and one of such great service to his friends, colleagues, and professions.... [We] are grateful for...his enduring vision and for the memory of a mind both deft and graceful.[9]

I don't think I can add much to these succinct and moving statements. I can say that I will not only miss him, but that his pioneering work will be reflected in these essays regularly.

As for Fred Kochen's paper, on reflection and in the aftermath of the International Congress on Peer Review and Ethical Behavior in Science held in May in Chicago, I am persuaded that the issue of citation behavior is a fundamental one. Fred Kochen realized that electronic access in publishing can be useful as well as a curse. Every one of us, veterans and students alike, needs to be reminded how important it is to reflect on the question of how well we acknowledge our intellectual debts.

* * * * *

My thanks to C.J. Fiscus and Peter Pesavento for their help in the preparation of this essay.

REFERENCES

1. **Garfield E.** Citation behavior—an aid or a hindrance to information retrieval? *Current Contents* (18):3-8, 1 May 1989.
2. **Kochen M.** How well do we acknowledge intellectual debts? *J. Doc.* 43:54-64, 1987.
3. ----------, WISE: a world information synthesis and encyclopaedia. *J. Doc.* 28:322-43, 1972.
4. **Garfield E.** The World Brain as seen by an information entrepreneur. Presented at the American Association for the Advancement of Science Symposium on "Reorganizing Information Resources to Improve Decision-Making." February 1974. San Francisco, CA. (Reprinted in: Kochen M, ed. *Information for action.* New York: Academic Press, 1975. p. 155-60; and *Essays of an information scientist.* Philadelphia: ISI Press, 1977. Vol. 2. p. 638-45.)
5. **Wells H G.** *World Brain.* Garden City, NY: Doubleday, 1938. 130 p.
6. **Kochen M.** Candidate's mission statement for ASIS president. 1986. p. 3; 8. (Ballot.)
7. **Small H.** Personal communication. 3 May 1989.
8. **Lindsay R.** Personal communication. 8 May 1989.
9. Manfred Kochen 1928-1989. Remembrances of a scholar and a gentle man. *J. Amer. Soc. Inform. Sci.* 40:223-5, July 1989.
10. **Flood M.** Personal communication. 12 May 1989.

Back to Introduction

Reprinted with permission from *J. Doc.* 43:54-64, 1987.

HOW WELL DO WE ACKNOWLEDGE INTELLECTUAL DEBTS?

MANFRED KOCHEN

Mental Health Research Institute, University of Michigan
Ann Arbor, Michigan, 48109, USA

Authors of scientific articles often read a paper that fails to cite their prior work when they feel it should have. A survey of university faculty shows the extent to which such opinions abound. If justified, they reflect non-use of bibliographic search methods, their inadequacy or non-scholarly use of the result. Principles for the design of a new kind of automated or semi-automated document retrieval system are formulated. They are analysed and shown likely to improve the scholarly quality of scientific work as represented by the bibliographies in manuscripts reporting that work.

INTRODUCTION

WHEN THE EDITOR OF A SCHOLARLY JOURNAL receives a manuscript, he selects qualified referees. He asks for their expert opinions about its suitability for publication. Sometimes he selects the referees from the set of authors cited by the author in the manuscript. The referees evaluate the paper. They judge the quality of the list of references at the end of the manuscript along with its other aspects. We will in this paper also call this list of references the bibliography. During the short time a referee spends in scanning that list, he may think of major omissions. He may question the inclusion of some of the references. But he rarely offers a detailed critique of this bibliography.

A paper that conforms to the norms of scholarly perfection would explicitly cite every past publication to which *it* owes an intellectual debt. Not knowing what he should acknowledge his intellectual debt to is no excuse for omission, any more than ignorance of the law can excuse its violation. Acknowledgement of intellectual debt is not the only function of the paper's bibliography.[1] It indicates the author's actual source of ideas, which may not be the true origin of the idea. It directs the reader to further information. It meets others' expectations about the content of a scholarly paper. There are many other reasons for citing.

In what follows, we ask several questions about these bibliographies.

1. How close or far from such ideal bibliographies are the ones published in journal articles today? We ask that only for one of the functions of a bibliography, however.

2. How important is this aspect of scholarship, and to whom?

3. How should we distribute among authors, referees, editors, readers and sponsors responsibilities for producing good bibliographies or for improving them?

4. Could an automatic or semi-automated reference retrieval system be expected to improve significantly the scholarly quality of bibliographies, by building on what the author provided with his manuscript or perhaps by performing a *de novo* search, given the manuscript? If so, how? A major improvement in one of the functions of bibliography could reduce the effectiveness of another. For example, the author may recommend a work that he has not read; that improves the value of the bibliography in directing the reader to sources, but not its honesty in reporting the sources actually used by the author.

5. If the answer to 4 is Yes, how much effort is it worth expending on it, and how should these costs be distributed?

The primary contribution of this paper lies in the formulation of new principles underlying the analysis to question 4, particularly in the recommendation to develop expert systems. The issue is seen not only as improving the bibliographies at the end of manuscripts submitted to a scientific journal per se, but in improving the scholarship of the entire scholarly research process, in which publication is a final stage. Thus, it raises deeper questions about the role of document retrieval in scholarship and the research process, and the two-way interaction between the processes of literature searching, screening, comprehension,

evaluation, organisation and utilisation on the one hand,[2] and of adding to knowledge on the other.

These issues are also significant for research and practice in information retrieval (IR). One of the basic problems in IR research is how to measure 'recall' or hit rate. This is defined as the fraction of all relevant documents that are retrieved. Estimating the denominator of this fraction, i.e. the number of relevant documents, has challenged information scientists for several decades, because of the difficulty of defining 'relevance', of performing controlled experiments in large collections and of transferring conclusions to real situations. By asking authors to judge errors of omission in bibliographies, we can roughly estimate the number of relevant papers.

These issues are also of practical import for IR for what they imply about citation indexing and related searches. Citation-based retrieval depends critically on proper acknowledgement of intellectual debts.

ON THE QUALITY OF REFERENCES IN PUBLISHED ARTICLES

If the primary criterion of quality is the extent to which all prior publications to which a given article owes some intellectual debt are acknowledged, then quality is probably quite low. It has been estimated that at most 10 per cent of what is published is a genuine contribution to knowledge.[3] Quite possibly, the quality of bibliographies is similarly low. Quality there is also difficult to measure. To estimate it roughly, we conducted a mail survey to determine from a sample of faculty members in a major research university how frequently they encounter published articles in their specialities that they feel neglect to acknowledge intellectual debts.

The first conceptual difficulty arises in defining 'intellectual debt'. If an author uses a hitherto little-known concept, method, or result, without which he could not substantiate the claim advanced in his article, he owes a substantial intellectual debt to the author of that concept, method, result or issue. Those debts are frequently acknowledged. Thus, authors who use the original concept of a fuzzy set have cited Zadeh's 1965 article that introduced it; chemists using Lowry's method for protein analysis have generally cited his seminal paper reporting it; studies based on the result that people can encode into short-term memory only about seven plus or minus two chunks, such as digits in a telephone number to be remembered long enough to dial it, usually acknowledge the seminal paper by G. A. Miller that presented it. And few who contribute to the issue of the 'Tragedy of the commons' fail to cite G. Hardin's pioneering work.

Priority of a discovery is often difficult to establish. Thus, a court settlement awarded the patent priority for the electronic computer to Atanasoff rather than to Eckert and Mauchly. Does that mean that an author of an article or patent that builds on these early designs should cite Atanasoff rather than Eckert and Mauchly? Suppose he independently discovers the same design they used, without having seen their patents, models or papers. Does he owe either, neither or both some intellectual debt? If he has acknowledged his intellectual debt to Eckert and Mauchly, then, logically, he owes an intellectual debt to Atanasoff because Eckert and Mauchly, by the court decision, owe him that debt, and '——owes an intellectual debt to——' (abbreviated by d) can be viewed as being a transitive relation (a hereditary property: IF (xdy) AND (ydz), THEN (xdz)). But painstaking historical scholarship can often uncover obscure antecedents of widely acknowledged seminal reports of discoveries, and this would imply that reference lists cease to be perfect when such priorities are established because they fail to acknowledge debt to the originator.

The most workable definition is probably one that refers to 'reasonable and proper effort' in the determination of priority. Its meaning depends on the consistent interpretation of precedents as well as on advances in retrieval technology. Modern online bibliographic searching makes it reasonable to expect a far more thorough search and identification of prior literature than was possible without it.

Of course, it is widely understood that classics such as the original 'publication' of the Pythagorean Theorem or widely known concepts, such as the definition of a number, generally do not require citation.

A second conceptual difficulty arises in defining the magnitude of the intellectual debt. How big a debt must be incurred to warrant citation? The author of concepts or results that he reported in an older publication may estimate, on encountering the same items in a recent publication, the magnitude of the debt much more highly than the author of the recent article.

What is most inimical to the spirit of scholarship is the deliberate omission of acknowledgement of an intellectual debt so that the author can fraudulently advertise his claim to priority of discovery. It is close to plagiarism. Such omissions should be weighted heavily in assessing quality.

But can it be ascertained that an omission is deliberate and that the intent of the author is dishonest?

Failure to cite an article that should have been cited can be attributed to any one of the following four failures:

1. no attempt was made to search the literature,
2. the literature was searched, but not well enough; the document retrieval system used was not good enough; the queries posed were not good enough; or not enough effort was expended,
3. relevant documents were available but not read or not used,
4. the item that should have been cited was retrieved and at least looked at but not cited: (a) because of an attention lapse or carelessness; (b) deliberately, because the author did not deem it worthy of citing; (c) because the author did not understand it or its relevance; (d) for the less honourable reasons alluded to above.

The questionnaire used in the survey assumes that respondents are experts in their fields; have published in those fields; keep up with current literature; examine the bibliographies of the articles they read; recognise an idea, result or method they encounter in their readings as similar to ones they encountered previously or that they themselves originated; appreciate distinctions between the form, content, clarity and expression of these ideas as encountered and the corresponding attribution to similar ideas. We also assumed that we could recognise and compensate for or against respondents' biases in seeking recognition due to them. The questionnaire was sent to twenty names selected randomly from listings of faculty members in each of three departments: mathematics, history, psychology. Of the sixty questionnaires sent out, twenty-one usable responses were returned. Of these, only one had been in the field of specialisation less than ten years, and seventeen more than fifteen years. Fifteen of them had authored or co-authored more than twenty publications in that field. To keep current with speciality literature, seventeen relied on personal journal subscriptions (not exclusively); fifteen on reprints/preprints; seventeen on libraries; five on online searching; eight on bookstores; sixteen on conferences. Everyone scans the bibliographic references all or most of the time in the articles they read.

Ten of the twenty-one said that in the most recent article they read, the author failed to cite relevant prior work, (seven said no, and four didn't know). Of these ten, two felt that only five such errors of omissions were made, while four thought five or more references were omitted that should not have been. The remaining four didn't know. Six of the ten felt that these omissions were to authors who are widely recognised in the field. On the average, 30 per cent of the articles they read omitted references to prior work that should have been cited, but the variance is very high. Two of the twenty-one felt that 75-100 per cent of the articles in their fields left out works that should have been cited, and four of the twenty-one felt that less than 10 per cent of the articles were flawed in this way.

On the whole, one respondent said that the literature does not adequately acknowledge intellectual debts; eleven said 'somewhat adequately', eight said 'adequately', one didn't know. Five of the twenty-one rarely encountered works that should have cited their own published works; five said this occurred often, and eleven said 'occasionally'. Some of their comments were: 'in my field (adolescent psychology), only a handful cite adequately. I take it to be a part of life. I don't cite adequately, by the way.' 'I have recently become interested in philosophy and have written in that area. I have been shocked at the extent to which intellectual debts are not acknowledged.' 'Mathematics is fairly careful about this. Survey articles are often cited in lieu of direct (original) sources. Most omissions of this type are made by young people not yet in control of the literature, or old people at odds with one another.' 'There is a good deal of superficial, "protective" citation, in which a work is cited (e.g. "see also") without actually taking accounts of its argument or conclusions.' 'Some researchers unwittingly repeat earlier works—things done some fifty years ago. In general, the three or four such cases have obtained these early (uncited) results by more elegant modern methods (mathematics).' 'I think that the problem has decreased with increased speed and scope of information dissemination and retrieval. The older literature was much worse than modern literature in my perception. In part, people are ignorant. They are harried into publication with little time taken for scholarship. It happens to me. But also people are unbelievably peevish about citing competition. Big, established investigators pretend that the others (and their students and their students' students) do not exist. But it's just another fact of life. The situation is hopeless but not serious.'

Suppose that about 30 per cent of the articles published failed to acknowledge intellectual debts.

Suppose that on the average, five prior works that should have been cited were omitted in this 30 per cent of the articles. Suppose further that the average article has ten relevant prior articles in 70 per cent of the citing articles. The other 30 per cent should have cited fifteen instead of only ten relevant prior works. Then, about 10 per cent of the past literature should have been cited but wasn't. That might be about the magnitude of failure to acknowledge intellectual debts. In other words, most (70 per cent) of the quality of bibliographies is quite good in that they cite most of what they should cite. But a significant minority of publications is quite poor, citing only perhaps two-thirds of what it should.

Given the formidable difficulties in judging the quality of a bibliography, it is natural to ask how important the task is. The considerable effort devoted to document retrieval could be viewed as culminating primarily in improved reference lists. But that is only the most visible end-product of the entire scientific research process. Presumably, a more thorough and successful search for relevant, important and valid concepts, methods, results, and issues that appeared in prior publications results in improved contributions to knowledge. Such searches are deeply integrated into all phases of the research process, and scientific inquiry has been suggested as a model for online searching.[4] Having contributed significantly by using prior works that were selected early in the research process, how important is it to check whether the resulting contribution has already been produced by someone else?

IMPORTANCE OF HIGH-QUALITY REFERENCE LISTS

To the sponsor of research, unplanned duplication of contribution seems like an inefficient allocation of his scarce resources. To those concerned with the utilisation of human resources, e.g. employers, it seems as if opportunities for better use of scarce, highly trained talents, were missed. Editors of journals in which there is high demand for the allocation of scarce pages to manuscripts refuse to publish reports of previously published contributions. It would therefore be prudent to search the literature repeatedly at all stages of the research process that leads to a publication.

The list of references in the manuscript that is submitted to a journal is a composite of all these search results. If it omits stating key intellectual debts, it casts justified doubt on the quality of the process and its products. To be sure, carelessness resulting in a low-quality bibliography may do injustice to a good process and product, and a particularly well-prepared bibliography may deliberately mask and conceal a poor process or product, but we expect these to be exceptions rather than a rule. This is a hypothesis that should be experimentally tested. The price of such carelessness should be rejection by a journal, and good referees will rarely be deceived by a good bibliography into accepting a poor manuscript.

Another reason to attach importance to a high-quality reference list in a publication is that it becomes part of the archival record, not to mention its role in citation indexes.[5] As such, it is accepted as an authoritative source for further bibliographic work. Poor bibliographies contribute to the propagation of errors and these are very hard to detect and correct. (There was a good article in the early days of the information retrieval discipline—perhaps in the early 1960s—that demonstrated how an error in a citation was propagated by uncritical, unchecked copying from one bibliography to another. My inability to recall or retrieve the citation to this article is an example of retrieval failure, even given strong clues.) The integrity and quality of the cumulative archival record depends on the quality of bibliographies added to it.

Recognition of priority is a powerful incentive to researchers who engage in the arduous and frustrating, often thankless, effort to add to knowledge. Making light of errors of omission is therefore a disservice to the motivating forces in scholarship. Put positively, ensuring that contributions will be appropriately recognised at least by public citations of intellectual debts can increase the incentives that attract appropriate people to lives as scholars and keep them productively (and happily) engaged in such pursuits. With the trends towards use of lifetime citation counts[6-10] in awards of tenure, awards of promotion, salary increases, etc., these incentives have a material component.

Given that quality of bibliographies is important, how are and should responsibility and authority for such quality control be distributed? Now, authors bear the primary responsibility. Editors rely mainly on referees who are experts in a speciality, and they are not generally aware of any special responsibility for the quality of the reference lists in manuscripts they review. They judge the manuscript in its entirety. Editors have the authority, including that of rejecting a manuscript because of a poor bibliography. Readers have neither responsibility nor authority directly, but they can refuse to read or subscribe to jour-

nals with papers that they judge to have poor bibliographies and, as authors themselves, they can refuse to cite such papers, write letters to the editor, and decide not to submit their manuscripts to such journals. Sponsors can refuse to sponsor work by authors they judge to publish papers with low-quality bibliographies or urge and help their grantees to improve in this regard.

The present distribution of quality control is not commensurate with the importance of quality bibliographies. An improved distribution would give the referees and editors greater responsibility. In the next section we present some new ideas for helping them carry it out.

PROPOSED REFERENCE SEARCH METHODS TO IMPROVE SCHOLARSHIP

The intent is to help an editor to check the quality of the list of references used in a submitted manuscript in the sense discussed so far. Input consists of all or part of the manuscript. It is assumed that these inputs are available in computer-readable form, such as a diskette.

The simplest procedure is to enter the list of references provided by the author and search online citation indexes for recent articles that cite them singly, in pairs, in triples, etc. This is a version of co-citation analysis[11] used at search time. The editor then asks the referees to judge whether the retrieved references that are not in the author's bibliography are serious errors of omission. Publishers are not likely to cover the costs of such searches unless forced to by competitive pressures or by standards set and enforced by professional societies.

The above procedure is biased toward recency and quite costly as well. A more traditional subject search would remedy this. For journals such as *JACM*, which require the author to submit keywords with his manuscript, any of the search systems based on Boolean combinations of the keywords can be used. Again, referees are provided with the resulting bibliographies, in which those items used by the authors are deleted or marked, and they are asked to judge if the remainder contains major errors of omission. Such keyword-driven generation of references may improve the extent to which others' expectations about the content of a paper are met but at the expense of indicating honestly which sources the author actually used. Both of these aspects should, of course, be taken into account. Again the cost of this might be borne by the publishers if that were required by market forces.

Having a complete manuscript in machine-readable form provides opportunities, however, for more sophisticated reference searching than is possible either by references or keywords with citation indexes or Boolean searching, respectively. To be sure, lexical and logical content analysis of clear text is still not well developed and costly. But a variety of simple statistical and lexical/logical methods of the kind proposed in the 1950s[12-14] can be applied. Most of the methods that have been discussed in the literature in relation to indexing[15] can be applied, with modification, to formulate search queries. They are still limited by the way the bulk of literature to be searched is made available for searching: by index terms, full text of titles and occasionally abstracts and references. Thus, the most we can hope to obtain from the analysis of a manuscript is the key concept, methods, findings or issues that its author claims to add to knowledge or that the author used, for which he owes their authors intellectual debts; as an approximation, such concepts, methods, findings and issues must be expressed in the language suitable for searching an online database.

One idea is to look up each text word-stem in a dictionary or alphabetised authority list of search terms, counting the number of times each entry in the list is consulted. Four micro-thesauri of terms, used to write about concepts, methods, findings and issues respectively, are consulted next. A findings thesaurus, for example, contains verbs such as 'find/found', 'show', special adjectives and nouns not in the search vocabularies. The output of each thesaurus is a grouping of terms that can be combined with one or more search terms to express a concept, method, finding or issue, and a Boolean query is composed of those search terms and used to search various databases. Automatic methods for query formulation have been investigated[16] and are relevant here.

Another idea is to divide the number of times a term in the authority list is consulted by the total number of words in the text and compare that frequency with a stored frequency with which that term occurs in general use as well as in documents that have been judged relevant to that term where that is available. If the frequency of occurrence in the document is much greater than the stored frequency—or close to that of a relevant document—then it is used for a search.

A more timely and ambitious undertaking is the development and use of expert systems. That is the main proposal put forth here. An expert system is needed for each speciality. For example, to build an expert system for 'fuzzy set theory and

information retrieval',[17] a knowledge engineer would capture an expert's (say Abraham Bookstein's) expert knowledge of the literature in that topic. This could be in the form of production rules, such as 'If finding (or concept, or method, or issue) *p* is salient, then consult references *a*,*b*,*c*,...' Here *p* can be viewed as a proposition. To illustrate with simple propositions in a well-known domain of discourse, arithmetic and number theory, consider propositions about concepts, findings, methods, and issues respectively:

1. 'A number is *prime* if and only if its only divisors are 1 and itself.' (The primary concept is italicised.)

2. 'Every number can be expressed as a product of powers of primes, in a unique way,' known as the *fundamental theorem of arithmetic* or as the *prime factorisation theorem*. (The names of the finding are italicised: a keyword might be *prime factorisation*.)

3. 'Whether a given *number is prime* can always be found by the method of the *sieve of Eratosthenes*.'

4. *Goldbach's conjecture:* 'Every even number is the sum of two primes.'

Just because a concept is a needed prerequisite does not mean that the first publication to introduce the concept needs to be cited. The concepts of number or prime, for example, are known to just about anyone, and though some people have contributed to their profound explication (e.g. Peano), no individual is regarded as the discoverer of these concepts. It could, indeed, be argued that every new concept is a social rather than an individual product, and no person has a right to claim a discovery as solely his own. Even though discoveries are often made independently by several persons within a few years, when the logic and maturity of the speciality has sufficiently ripened, the practice of crediting an individual—even capriciously as in a lottery—serves as a powerful incentive to discoverers.[18]

These propositions use 'functions', interpreted here in the sense of a programming language used in artificial intelligence such as Interlisp, which are applied to domains and return unique values: e.g. 'Sum of' applies to pairs of numbers as in the list (PLUS 2 3) and, when executed, returns a number, 5, in this case. Predicates are special kinds of function that return only values T (true) or NIL (usually false); 'is prime' applies to integers, as in (PRIME 97), which returns T in this case.

We have implied that concepts, findings, methods and issues are the kinds of entities that comprise knowledge and are added to it, and are used in the production of knowledge. This list is not intended to be complete. Explanations, discourses, critiques, and histories are a few of the other elements found in scholarly publications. 'Concepts' are intended to include ideas (e.g. the idea of an imaginary number), principles and laws (e.g. the principle of duality, the law of effect), definitions, and other seminal mental constructs (e.g. atoms, quarks, libido). By 'findings' we mean theorems together with their proofs, tested hypotheses together with the evidence and statistical inferences for accepting or rejecting them, reasoned conclusions, justified recommendations, principled policies, facts and their sources, trends, and generally justified, true beliefs about the world or ourselves. 'Methods' generally have names, often associated with a discoverer, as do findings occasionally.[8] There are mathematical, experimental, observational and many other kinds of methods. Often a new mathematical method is presented as a theorem. The term 'issues' is used here to include open questions and conjectures, controversies, choice points, foci of debates. Generally, a publication contributes not one but two or more of these four kinds of knowledge. It generally uses all four. Citations acknowledge debts to these, but they serve other functions as well.[1]

Only the first half of any production rule in the proposed expert system has been illustrated, and in a domain in which expertise on references would be too elementary to be useful and hence far from the research frontier. Thus, no author is expected to acknowledge his intellectual indebtedness to Goldbach, Eratosthenes or the author of the paper reporting the first proof of the fundamental theorem of arithmetic. Nor would an expert on number theory be expected to know these references. However, one can quickly move to the leading edge of number theory by asking about the frequency of primes between 1 and any number *n*, the famous prime number theorem. An expert would know that Hadamard and de la Vallee Poussin first proved in 1896 that the ratio of that frequency to $n/\log n$ gets closer and closer to 1 as *n* gets larger and larger, and that more recently, P. Erdos and A. Selberg found more elementary proofs for this finding.[19] This reference, incidently, acknowledges debt to a source of information for both the facts and the references within my assertion. Online bibliographic search systems could provide the expert with (or check his memory of) the most recent exact references and thus build the expert system.

The expertise to be captured pertains primarily to knowledge of the literature. If that is not

to be superficial, it also requires some expertise in the subject matter. Quite often, the most creative contributors to a speciality do not have expert knowledge of the literature. (They often know only as little of the work of others as they need or in fact used to make their own contributions; a Nobel laureate once said, 'If I need a good book in my field, I write one'.) But literature experts (quite frequently Ph.D. students) have quite a bit of the propositional expertise illustrated above. The building of an expert system should capture the expertise of several such Ph.D. students. Indeed, the system of comprehensive Ph.D. examinations could add to as well as draw upon such a corpus of production rules and the network of associated propositions.

How would such expert systems be used? The first step is to determine for each manuscript the concepts, findings, methods or issues it claims to add *or* that it uses in the production of its contribution in a manner requiring citation. The author could be required to identify those passages in his paper, if any, that are transformable into the appropriate propositions. The referees could be asked to formulate such propositions. A computer program could search the text for the names of methods, findings, issues, concepts or phrases in an authority list likely to identify them, and for functions and predicates likely to comprise the sought-for propositions, and output the wanted *ps*. Any of these three methods should generate the propositions of *ps* needed to enter the production rules of the expert system, which then returns an expert-quality list of references.

What are the principles on which to ground the design of such an expert system, that would make it feasible? The first is that production rules capture the knowledge of experts (e.g. in diagnosis and treatment of bacterial infections[20]) and facilitate automatic inference-making.[21] The combination of such production rules with the knowledge of human experts and the output of computerised literature searches is new, as far as I am aware.

A second principle is that concepts, findings, methods and issues can be represented as propositions using functions and predicates as understood in the languages and systems[22] used in 'artificial intelligence' research (e.g. LISP, PROLOG) *and* that such propositional expressions are useful for formulating queries for online searching.

A third principle is the existence of programming languages, such as KEE,[23] for the development of expert systems that facilitate the writing of systems of production rules, with consultation, inferences and question answering available to the user.

These make the construction of such expert systems well within the state of the art. Their development can be expected to improve upon the quality of bibliographies in manuscripts, because it brings to bear the combined power of the indexed literature through computerised searches with that of human experts on the literature in each speciality, as well as inferences from these data and refinements and improvements resulting from their use. If authors don't use them, editors can evaluate the bibliographies in their manuscripts in comparison with bibliographies generated in this way.

CONCLUSION: INFORMATION RETRIEVAL AND SCHOLARSHIP

The development, maintenance and use of aids to improved scholarship, such as proposed in the previous section, is costly and risky. The perception of benefits is subtle. Some practically-minded leaders, including key decision makers in the publishing industry, often relegate scholarship into low-priority categories with inessential luxuries, the pastimes of esoteric academics, the compulsions of perfectionists. Indeed, the magazine articles and reports that busy executives pay most attention to seldom have bibliographies or lists of references. Pedantic scholars are often caricatured as the antithesis of hard-headed, driving, decisive real-world managers.

Yet, scholarship, like the arts, is a worthy end in itself. It epitomises a humanistic tradition, the preservation of which gives more meaning to the pursuit and attainment of practical endeavours. This must be kept in perspective as a high priority, particularly as needs for survival, security, belongingness and esteem[24] are increasingly being met, permitting societies to attend to the needs for self-actualisation and collective well-being. Scholarship is one of the higher forms of self-expression as well as a manifestation of 'group mind', of co-operative advancement of knowledge that enriches civilisation.

Because subtlety, sophistication and high culturedness is needed to appreciate the importance of scholarship as an end in its own right, that message must also be supported in a way that reaches the more practically minded: scientific and the associated technological advances on which depend our competitive position in world markets, in geopolitical arenas and military theatres, our living standards, etc., rely heavily on a high-qual-

ity record and system of communication. As pointed out above, the quality and integrity of the record—and therefore its use for effective communication and making good use of the work of others—requires good scholarship in the sense discussed. The practical importance of this is enormous.

As knowledge continues to double every decade or two, to become more specialised, and to require more study to attain its leading edges, scholarship becomes correspondingly more difficult. The very technologies that are helping to advance knowledge can, however, also be used to manage it by amplifying the productivity of scholars. The ideas proposed here contribute to our knowledge about how to do this. Enterprising publishers or leaders in other information industries should consider seriously the recommendation to launch the steps recommended here toward expert systems that support the evaluation and improvement of bibliographies in the manuscripts submitted to high-quality scholarly journals.

ACKNOWLEDGEMENTS

Thanks are due to D. West for discussing some of the ideas in this paper, and to the National Science Foundation for partial support under grant IST-8301505.

REFERENCES

1. **Moravcsik M J & Murugesan P.** Some results on the function and quality of citations. *Soc. Stud. Sci.* 5:85-92, 1975.
2. **Walker D E.** The organization and use of information: contributions of information science, computational linguistics and artificial intelligence. *J. Amer. Soc. Inform. Sci.* 32:347-63, 1981.
3. *Coping with the biomedical literature explosion—a qualitative approach*, 22-23 May 1978, Pocantico Hills, NY. New York: Rockefeller Foundation, 1978.
4. **Harter S P.** Scientific inquiry: a model for online searching. *J. Amer. Soc. Inform. Sci.* 35:110-7, 1984.
5. **Garfield E, ed.** *Science Citation Index; Social Science Citation Index; Arts & Humanities Citation Index.* Philadelphia: Institute for Scientific Information.
6. ----------. Is citation analysis a legitimate evaluation tool? *Scientometrics* 1:359-75, 1979.
7. ----------. What's in a name? The eponymic route to immortality. *Essays of an information scientist.* Philadelphia: ISI Press, 1984. Vol. 6. p. 384-95.
8. ----------. How to use citation analysis for faculty evaluations, and when is it relevant? Part 1. *Ibid.* p. 354-62.
9. **Geller N L, De Cani J S & Davies R E.** Lifetime-citation rates to compare scientists' work. *Soc. Sci. Res.* 7:345-65, 1978.
10. ----------. Lifetime-citation rates: a mathematical model to compare scientists' work. *J. Amer. Soc. Inform. Sci.* 32:3-15, 1981.
11. **Small H & Griffith B C.** The structure of scientific literatures. *Sci. Stud.* 4:17-40, 1974.
12. **Luhn H P.** A statistical approach to mechanical encoding and searching of literary information. *IBM J. Res. Develop.* 4:309-17, 1957.
13. **Rogers P J & Tanimoto T T.** A computer program for classifying plants. *Science* 132:1115-8, 1960.
14. **King C W.** *Table look-up procedures in data processing.* Yorktown Heights, NY: IBM Corporation, 1962.
15. **Cleveland D B, Cleveland A D & Wise O B.** Less than full-text indexing using a non-Boolean searching model. *J. Amer. Soc. Inform. Sci.* 35:19-28, 1984.
16. **Salton G, Buckley C & Fox E A.** Automatic query formulations in information retrieval. *J. Amer. Soc. Inform. Sci.* 34:262-80, 1983.
17. **Bookstein A.** Fuzzy requests: an approach to weighted Boolean searches. *J. Amer. Soc. Inform. Sci.* 31:240-7, 1980.
18. **Merton R.** Singletons and multiples in scientific discovery. *Proc. Amer. Phil. Soc.* 105:470-86, 1961.
19. **Kac M & Ulam S M.** *Mathematics and logic.* New York: Praeger, 1968. p. 4.
20. **Shortliffe E H, Buchanan B G & Feigenbaum E A.** Knowledge engineering for medical decision-making: a review of computer-based clinical decision aids. *Proc. IEEE* 67:1207-24, 1979.
21. **Kochen M.** Adaptive mechanisms in digital concept processing. *IEEE Trans. Ind. Appl.* 83:305-14, 1964.
22. **Quaile M A.** *The friendly dandelion primer.* Pittsburgh, PA: University of Pittsburgh Learning Research and Development Center, 1984.
23. *KEE training manual.* Menlo Park, CA: Intellicorp, 1985.
24. **Maslow A H.** *Toward a psychology of being.* Princeton, NJ: Van Nostrand, 1968.