

# The FUTURE of Citation Indexing:

An Interview with Eugene Garfield

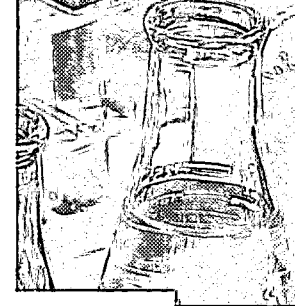
By Péter Jacsó

Dr. Eugene Garfield, the founder and chairman emeritus of the Institute for Scientific Information, spent a few days in Hawaii, giving a guest lecture at the University of Hawaii and an interview to *ONLINE*. This is the print version. A digital version with hot-linked cited references is also available [[www2.hawaii.edu/~jacso/extra](http://www2.hawaii.edu/~jacso/extra)].

**Q** You outlined the concept of citation indexing in your seminal 1955 article in *Science* [1], then several years later implemented the idea by first publishing the *Science Citation Index* and then related products for the social sciences and the arts and humanities. The

whole World Wide Web, which is interwoven through billions of links (the technical equivalent of cited references) among digital documents, is based on the very same concept. Looking up cited documents electronically or at least their abstracts, while reading an article, is far more convenient than consulting printed journals or indexes. Do you sense an increased awareness of the importance of comprehensive citation tracking among scientists for subject searching?

**A** Undoubtedly many researchers use this capability, but that does not mean they are conscious of citation indexing



per se. We don't know the full extent of their awareness of what makes the links possible, but in sites like HighWire and others, it is explicitly stated that you can be linked both to the cited references within an article as well as the citing papers by direct links to the SCI in the ISI Web of Science.

**Q** In a 1965 paper [2] you explored and outlined the limits and potential of automated citation indexing, which has now been partially accomplished in the autonomous citation indexing system by people who were not even born then. ResearchIndex (formerly CiteSeer) and CiteBase are working prototypes for documents available in open access Web sites. These models suggest that it may be possible to eliminate or reduce the most expensive part of citation indexing—the human labor—along with some costs and thereby widely extend the scope of traditional citation indexes. How do you see the future of traditional and autonomous citation indexes?

**A** The Research Index is undoubtedly a significant advance—if it can be expanded to cover the whole range of science and technology. So far, it is dependent upon whatever journals and other materials are available free on the Web. That is a significant limitation that might be overcome if commercial publishers make their full-text data available to those who operate new databases created with the same software. However, there are problems even for material that is free. I need to know more about how CiteBase deals with the lack of standardization in cited references as well as the numerous errors made by authors.

Another aspect of the open access idea for indexes is the question of user education. In the commercial world, this is part of the marketing effort. If Research Index were marketed that way, it would have much greater support. Word of mouth only goes so far, and I don't have to tell you how limit-

ed the library budget is for providing user education.

In the meantime, there is also the question of the legacy literature. Who will provide the cost of processing all the back years of literature and deal with the lack of standardization, which gets worse as you go backward in time?

If all the indexing and abstracting services went out of business tomorrow

ISI in unifying variations. Indeed, they have now become more conscious of author errors because the Web edition makes them stand out. A paper that is cited correctly 500 times occupies only one blue line in the Citation Index, whereas the dozen or more variant citations will each occupy another dozen unlined lines. By the way, in spite of the relatively small percentage of errors and even more vari-

Most users are unaware of how many errors are already corrected.

(displaced by autonomous indexing), libraries would still need to access their existing files for coverage of whatever literature is still not covered by electronic services. Hopefully, the percentage will go down as the years go by, but it is hard to imagine that this would happen in the next 10 years. I hope that it will so that even I might one day benefit from access to historically important materials.

**Q** Some of the cited references are not hotlinked in World of Science (WoS) because they were not covered by ISI. Many of them are not linked because of sloppy citations in the source journals, where the volume, the pagination, the author or journal name are misspelled. The naked eye can easily recognize these variations or errors in references to the same items, but ISI's software doesn't seem to unify these items. You spent a lot of brain juice and money on developing programs to correct and normalize these misspelled citations, and a lot of ink to educate editors and authors about the implications of this sloppiness. Roughly what percentage of the citations have typos that scatter the results and cripple the links in WoS?

**A** Behind the scenes of the SCI production, users are generally not aware of the considerable degree of work done by

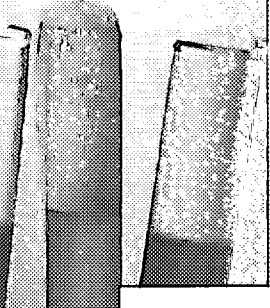
ations, ISI is increasingly devoting greater resources to quality control.

Your question is a highly complex and inherently ambiguous one. Do you mean for example, what percentage of citations prevent the retrieval of the citing papers when one conducts a search? Or do you mean what percentage affect bibliometric studies, which may rely solely on correctly cited references that are included in the ISI source record?

The mere fact that you can identify, by eyeballing, obvious variations in one or more elements of the citation suggests that only a small percentage of papers are not retrieved. However, only more detailed studies could reveal scattering due to variant spellings in Russian or Spanish names. Even in those cases, experienced searchers know that they must check for such variants. But most errors are due to erroneous volume or page numbers.

Henk Moed in his commentary in *Nature* [3] suggested a 7 percent error rate based on using an automatic procedure to match ISI's source data with the cited reference file. These include errors in pagination due to the citation of other than the first page, as is often the case in chemistry. However, some internal studies at ISI may interest you. The most cited paper in the history of science by Oliver H. Lowry *et al.* has been cited about 270,000 times. (This and about three or four other papers have been cited over 100,000 times).





An additional 10,000 citations to the Lowry paper contain variants and errors—about 4 percent. One might ask who would want to retrieve so many hits, but in a combination search with keywords or other cited papers, as in a co-citation search, that is reasonable. How typical is this sample? Other samples need to be studied. There is a large literature on error rates, but each author may define errors differently. Few of these errors prevent retrieval in the Web of Knowledge, as I discuss in my essay [4].

In the arts and humanities, I would expect much lower figures for

would be the same as a National Citation Facility. Cross-Ref uses the DOI to find the full text of a single article. ISI also uses DOI, and together with the Webfeat system it is able to integrate seamlessly across the entire suite of databases at every institution.

One should not trivialize the size of the error correction process. Most users are unaware of how many errors are already corrected. Further, ISI has recently augmented the editing process so that in 2004, even more corrections will be made. However, if readers know of serious errors in WoS, they should not hesitate to inform me, and every

undertook with my Russian colleagues Alexander Pudovkin and Vladimir Istomin to greatly enhance the ease of managing the output of World of Knowledge (WoK) searches. Your readers can access articles on HistCite at my Web site [<http://eugenegarfield.org>] and also dozens of HistCite databases [<http://garfield.library.upenn.edu/histcomp/>]. HistCite provides the user with an easy way to identify the core literature retrieved in a WoK-marked list and arranges the material in a pure chronological order, which WoS does now approximately in reverse-chronological order. HistCite also helps the user edit variations in cited references. Finally, it generates Historiographs (maps) of the key literature to aid visualization for the evaluation of a topic and to review the evolution of the subject.

**Who will provide the cost of processing  
all the back years of literature and deal  
with the lack of standardization,  
which gets worse as you go backward in time?**

journal articles, but in checking cited books, one has to take into account variant citations to book title abbreviations, as well as the citation of individual pages.

**Q** How large is now your “Forever Dictionary,” which includes the authentic source information as verified/corrected by ISI? Do you still harbor the idea of the online National Citation Facility for authors to verify their citations, maybe in cooperation with CrossRef, which has about 10 million digital object identifiers (in addition to the traditional bibliographic data elements) that can be accessed for free (one at a time) by anyone?

**A** The modern equivalent of the “Forever Dictionary” reported in my essay is about 33.5 million, which is even larger than the 32.8 million 1945–2003 source items listed for SCI/SSCI/AHCI.

The expansion of Cross-Ref is highly desirable, but it is doubtful that it

reasonable effort will be made to correct them. This is especially important if the paper in question is highly cited.

**Q** While I see strategic plans and vision statements popping up left and right, it is delightful to talk to someone who has had visions and a strategy to implement them for decades. What new idea do you have now for enhancing citation indexes in the Web era?

**A** In the future, I hope that National Citation Indexes like the Chinese *SCI* or the Brazilian *SCI* would be available for other regions. They should ideally include English translations of vernacular titles and abstracts. I hope that in the future mechanical translation systems will supply non-English users with titles in their own languages. I would also like to see citation indexing of source books and dissertations, but that is a topic for future discussion.

Let me conclude by discussing HistCite, a development I personally

#### REFERENCES

- [1] Garfield, Eugene. “Citation Indexes for Science: A New Dimension in Documentation Through Association of Ideas.” *Science*, 122(3159), p.108-111, July 1955 [[www.garfield.library.upenn.edu/papers/science\\_v122\(3159\)p108y1955.html](http://www.garfield.library.upenn.edu/papers/science_v122(3159)p108y1955.html)].
- [2] Garfield, Eugene. et al. “Can Citation Indexing Be Automated?” *National Bureau of Standards Miscellaneous Publication*, 269, p.189-92, December 1965. No. 114 [[www.garfield.library.upenn.edu/essays/V1p084y1962-73.pdf](http://www.garfield.library.upenn.edu/essays/V1p084y1962-73.pdf)].
- [3] Moed, Henk. F. “The Impact-Factors Debate: The ISI’s Uses and Limits,” *Nature*, 415 (6873), p. 731-732, 14 February, 2002 [[www.rtn.lt/mi/0202/citation.pdf](http://www.rtn.lt/mi/0202/citation.pdf)].
- [4] Garfield, E. “Journal Editors Awaken to the Impact of Citation Errors. How We Control Them at ISI,” *Current Contents* #41, p.3-11, October 8, 1990. Reprinted in *Essays of an Information Scientist: Journalology, KeyWords Plus, and Other Essays*, Volume 13, p.367, 1990 [[www.garfield.library.upenn.edu/essays/v13p367y1990.pdf](http://www.garfield.library.upenn.edu/essays/v13p367y1990.pdf)].