Essays of an Information Scientist, Vol:3, p.42-44, 1977-78 Current Contents, #7, p.5-7, February 14, 1977

Project *Keysave* ™ --ISI's New On-Line System for Keying Citations Corrects Errors!

Number 7, February 14, 1977

At a seminar in 1969, one of ISI®'s vice-presidents, Phil Sopinsky, overheard Derek de Solla Price conjecture that "80% to 90% of the literature cited in a current year had been cited previously." Sopinsky's systems intuition immediately recognized unnecessary redundancy. Each of those repeated citations was unnecessarily recorded in its entirety. Indeed, many citations were being keyed two, three, or more times. This involved an enormous waste of time and energy in creating ISI's data bank.

Under Phil Sopinsky's guidance, Project *Keysave*™ was launched. Citation files for the years 1966, 1967, and 1968 were merged into one large file. This file was matched against the citation file for 1969. The match rate between these two files was found to be 66%! However, to verify Price's original conjecture was one thing. To develop a means for exploiting it was another.

A great deal of work would be saved even if one continued to key the full citation, that is author, journal, volume, page and year. However, this information would now be matched against an on-line master file of citations. Whenever there was a match we could eliminate manual verification. This is a process which involves rekeying the identical information for each citation. *Keysave* eliminated verification on as many as 90% of citations depending upon the journal involved.

However, Phil Sopinsky realized that we could use our knowledge of citation redundancy to reduce the initial keying effort as well. For many years we have known that 14 characters of information is all one needs to identify most articles uniquely. By keying just these 14 characters we could identify each citation on a file containing both the 14 character code and the complete citation. Consequently, whenever a match was obtained, we would eliminate keying the remaining characters of the citation. Only in a small percentage of cases would the entire citation be keyed.

The non-*Keysave* system involved keying as many as 48 characters; 18 for the author's name; 20 for the journal title; and 4, 4, and 2 for the cited

volume, page number, and year. Under the old system, if the identical article was cited by two different authors, the journals might be abbreviated differently. Authors' initials might be omitted as well. Under the *Keysave* system, however, each citation requires only 14 characters; the first four characters of the first author's last name, and the same 4, 4, and 2 for volume, page number, and year. When the computer 'recognizes' this abbreviated citation, it instantaneously informs the data-entry operator, by flashing a special symbol on the terminal screen. When the citation does *not* match the *Keysave* library a bell sounds, indicating to the operator that there is no match. Then the entire citation must be keyed and subsequently verified by another operator.

The advantages of the *Keysave*™ system are threefold: increased productivity and less boredom for the data-entry operator; decreased expense for ISI; and the greater standardization of citations. A fourth benefit is that the new procedure also corrects errors in citations--a significant qualitative improvement of *SCI*® and its related services.

Project *Keysave* was initially proposed for use at ISI in March 1972. Further studies were then conducted to determine feasibility. Two prime topics considered were the number of citations to be stored, and their chronological distribution.

A new test file was accumulated for 1964-69. The 1970 file was matched against this five-year file of approximately 8 million cited articles. A match rate of 72.5% was attained. We knew from *SCI* statistical data that about 25% of all citations are to papers published in the previous two years. So we felt confident that we could use our own source data files to augment the *Keysave* file with accurate data. Eventually, by a combination of source and citation data, we developed a file of about 5 million citations that gave us a 60% match rate. This file became operational in October 1974.

If one considers that we key over 6 million citations in a year, a match rate of 60% means we save most of the keying and complete verification of 3.5 million citations per year! However, the savings are more dramatic when individual journals are studied. For example, in keying 1,918 citations in the *Journal of Virology* during the first quarter of 1972, the match rate was 91%. Of 5,857 citations keyed from the *Journal of Bacteriology*, 89% matched. The same figure obtained for *Virology* on the basis of 2,880 citations. And 88% of the citations in the *Journal of Molecular Biology* matched in a sample of 3,974 citations.

Obviously, we get the best results with journals that have high impact since impact is based on the number of citations to articles published in the two years prior to the year studied. Journals in the social sciences produce

the worse results since they involve citation of books and other non-journal material. Hopefully, we can develop *Keysave* so that it will include book references as well. In that way, for some journals we will have a 100% match rate. If present hardware prevents expansion of the *Keysave*™ disc file we can also augment the capabilities of the system by using an auxiliary tape file to correct errors in citations subsequent to keying. By using such tapes in batch-mode we can correct errors for all but the most obscure items.

For a variety of reasons it is our responsibility to take every reasonable step to assure accuracy in our files. Since there has been so much discussion about the use of citation data for research evaluation,[1] it is essential that we take every reasonable step to improve accuracy. Citation errors creep into the literature for all sorts of reasons.[2] Not the least of these is the failure to eliminate printer's errors in proofreading. This happened to me recently in the citation of my own work[3]--rather embarrassing for a citation expert. But in spite of the known errors in the literature, we should not forget that the vast majority of citations are accurate and even those which contain small errors are still recognizable with a minimum of human or computer effort.

Phil Sopinsky

Phil Sopinsky is ISI's vice-president for computer services. He has been with ISI ten years, most of them spent in the development and management of computer systems and data processing. Understandably, his position gives him a key responsibility for the accuracy and timeliness of ISI products and services.

Phil is a native of Philadelphia and came to ISI after employment with the United States Army Signal Corps, the Curtis Publishing Company, and Food Fair Stores Inc. He is a graduate of Temple University and of the United States Naval Reserve.

Phil lives in Elkins Park, a suburb of Philadelphia, with his wife, four children, and 114 clocks. The last represents a 20-year hobby of collecting and restoration--of which many ISI employees and friends have taken happy advantage.

1. **Roy R.** Comments on citation study of materials science departments. *Journal of Metals* **28**(6):29-30, 1976.
2. **Garfield E.** Errors--theirs, ours and yours. *Current Contents*® No. 25, 19 June 1974, p. 5-6.
3. ----------. Significant journals of science. *Nature* **264**:609-15, 1976.