The Permuterm Subject Index: An Autobiographical Review

Eugene Garfield
Institute for Scientific Information
Philadelphia, PA 19106

The *Permuterm Subject Index* (*PSI*) section of the *Science Citation Index* (*SCI*) was designed more than ten years ago and has been published both quarterly and annually since 1966. There is, however, no 'primordial' citable paper about the *PSI*. It has been described and discussed from different standpoints in a number of papers (*1,2*), but none of them provides the formal description usually accorded a new bibliographic tool. This article is intended to provide such a reference point for future workers in information science.

The *PSI* was designed in 1964 at the Institute for Scientific Information (ISI) by myself and Irving Sher, my principal research collaborator at the time. In the subsequent development of the *PSI*, contributions were also made by others, including Arthur W. Elias, who was then in charge of production operations at ISI. In the early sixties we were too preoccupied with the task of convincing the library and information community of the value of citation indexing even to consider the idea of publishing a word index. But it was a logical development once we added the *Source Index* containing full titles.

The value of the *PSI* as a 'natural language' index is now well recognized and exploited by its users, but this was not the original reason for its development. The *PSI* was developed as one solution to a problem commonly faced by uses of the *Citation Index* section of the *Science Citation Index* (*SCI*). While the typical scientist-user could enter the *Citation Index* with a known author or paper, other users with a limited knowledge of the subject often lacked a starting point for their search. Before publication of the *PSI*, we told users whose unfamiliarity with subject matter left them doubtful about a starting point to consult an encyclopedia or the subject index of a book. If these failed, we told them to use another index, such as *Chemical Abstracts*, *Biological Abstracts*, *Physics Abstracts* or *Index Medicus*. Once the user identified a relevant older paper, it could be used to begin a search in the *Citation Index*. Users of the *SCI*—and librarians in particular needed some tool with which a starting point, or what used to be called a target reference, could be quickly and easily identified.

In those days the information community was pre-occupied with Key-Word-in-Context (KWIC) indexes. The development of the KWIC index, which was subsequently vigorously marketed by IBM, undoubtedly had an enormous impact (*3, 4, 5*). But I was never happy with the KWIC system for a number of reasons.

First, Sher and I felt that the KWIC index was highly uneconomical for a printed index. KWIC's use of space is prodigious, and it can be extremely time-consuming to use in searches involving more than one term.

Another aspect of the KWIC system (as used for example by *Chemical Titles*) that disturbed us was its indiscriminate use of stop-lists to eliminate presumably non-significant title words. In our view, it caused considerable loss of information on many subjects of interest to some users, if not to all. Consider the effect of deleting terms like METHOD and BEHAVIOR. In order to retain much of this information, but still prevent the useless entries generated by "terms" like THE and WHICH, we developed the concept of the semi-stop list to be used in addition to a full-stop list.

The full-stop list for the *PSI*, which contains words that are completely suppressed, was and is quite small. The semi-stop words such as METHOD, BEHAVIOR, CAUSE, REPORT and TECHNIQUE are suppressed as primary terms (main entries), but not as secondary or co-terms (subentries). In addition, certain frequently used two-word phrases, which have been identified through statistical analysis of word frequencies, are kept together and treated as a single term rather than being allowed to permute separately. Such phrases as GUINEA-PIG, NEW-YORK, ESCHERICHIA-COLI and BIRTH-CONTROL appear in the *PSI* as hyphenated terms, thus reducing look-up time in many types of searches. This is done by computer in the *PSI*, while in indexes like *Chemical Titles*, it is done by a manual process called "slash and dash."

Finally, the KWIC format was rejected because a number of studies had demonstrated that users of scientific indexes generally specify two or three terms when they use coordinate indexes. We reasoned that the optimum system would precoordinate any two terms, no matter how far apart in the title.

Over ten years of *PSI* experience has confirmed that "specificity" *per se* does not guarantee efficiency of a word as a search term. If used frequently enough, a seemingly highly specific term like DNA becomes as inefficient as more general terms that are used less frequently. The converse also holds; consider the term CREATIVITY. It is general, but because of the comparatively low frequency with which it occurs in the scientific literature, it is an efficient search term. Therefore, pairing—together with precoordination—becomes essential for high-usage terms, and merely convenient for low-usage terms. Triple coordination—and even higher-level coupling—may also be desirable if two terms occur together with a third frequently enough. But the threshold must be correlated with cost of processing and printing, not only with economies in users' time. The ideal system would handle three or more terms, but this proved too costly. We therefore settled on two terms, although recently precoordination of three terms has been built into the five-year cumulative 1965-1969 *PSI*, and an improved three-term precoordination routine will be achieved in the five-year cumulative, *PSI* for 1970-74, to be published by ISI in 1977.

The choice of name for the *Permuterm Subject Index* was quite deliberate. Ohlman suggested the term *permuted* from *cyclic permutation* used in mathematics (*6*). It was in that sense appropriate to KWIC indexes. *Permuterm*, however, is a complete permutation of all title words to produce all possible pairs, including of course, the inversion of every pair. As I and others have noted before, KWIC indexes are more appropriately called *rotated indexes* (*7, 8*). For example, ISI's *Rotaform Index* section of the *Index Chemicus* is a rotated formula index. The *Chemical Substructure Index* (*CSI*) is also a cyclic or rotated index. Using the Wisewesser Line Notation, the *CSI* rotates the line notation to create a main entry for every substantive constituent in each notation.

For each title in the *PSI* with *n* title words, *n*(*n*-1) word-pairs are created by permutation. After applying the full-stop list and semi-stop list, this usually produces about 40 word-pairs for the typical seven-word title. It is by no means unusual for the *PSI* to contain over 100 word-pairs for titles with 11 or more words.

In the *PSI*, every significant word in the title is permuted [not merely rotated, as in a KWIC index (*7*)] by computer to produce all possible pairs of terms. Every word is potentially both a primary term and a co-term. On the printed page, each permuted word-pair is arranged alphabetically by primary term. All co-terms occurring with a particular primary term are idented as subentries and listed in alphabetical order under the primary term. Dashed lines lead from the co-term to the author, whose name can be used to locate in the *Source Index* section of the *SCI* the complete bibliographic data, including the title for the article.

As part of ISI's quality-control precoordination and spelling-variance unification procedures, every incoming term—that is, every word in every title—is passed against the established *PSI* vocabulary. In this computer comparison, wrong and variant spellings are corrected and coordination tests for accepted word-pairs are applied. Terms which are truly new are selected for human review and added to the vocabulary. Naturally, many author- or ISI-produced errors are identified and corrected in this process.

From the earliest days Sher and I were aware of the enormous potential of the *PSI* vocabulary for scientific lexicography. Besides allowing very specific searches on terms that would never have appeared in thesaurus-controlled indexing systems, the use of actual title-words reflects terminological innovation long before anyone but specialists in the affected field are aware of the changes. Every year nearly two-thirds of the words *added* as primary terms to the *PSI* vocabulary are "new" in the sense that they occurred only once or not at all in titles processed the previous year (*9*). This does not, of course, mean that two-thirds of each year's vocabulary is "new".

The cumulated vocabulary of the *PSI* comprises an author-generated word-index to all the significant articles of science and technology—including letters, technical notes, and proceedings of meetings. It is a pity that the *PSI* vocabulary has not yet been used by lexicographers to identify and define new scientific terms and usages (*10*). A dictionary based on the *PSI*, which could be updated quarterly, would be the first current dictionary of new scientific terms based on primordial sources.

From the outset, we were aware of the shortcomings of title-word indexes: the lack of resolution of obvious (and not-so-obvious) synonyms and the unavoidable fact that morphological variations of the same primary terms, *e.g*, CLASSIFY and CLASSIFICATION, appear separately in the index. Even the plural of a noun may be separated from its singular, *e.g.*, SUGAR and SUGARS. In Ohlman's permutation index to the proceedings of the ICSI 1958 conference, this problem was alleviated somewhat by restricting sorting of the first six characters of each term. However, use of this procedure is impractical for an index as large as the *PSI* (*3*).

Such problems were of minor importance as long as the *PSI* was regarded merely as a supplement to the *Citation Index*. We found that many scientists preferred a title-word index because it enabled them to retrieve a work by a word or phrase remembered from its title, or by subject words they knew to be relevant.

It was inevitable that librarians and others would begin pressuring us to make the *PSI* a search tool in its own right. Our response began with provision of cross-references and eventually led to certain standardizations, especially in the case of spelling variations. Today the so-called source-data edit procedures at ISI are quite systematic and comprehensive (*11*), and the *PSI* does stand on its own as both a current and retrospective subject index.

As early as 1969, I reported at Amsterdam on ISI's efforts to develop automatic procedures for hyphenating word-pairs into phrases, a process we called "precoordination" (*12*) to produce bound terms like BIRTH CONTROL. Such terms would be hyphenated automatically, provided they occurred with sufficient frequency. It was remarkable to discover that punctuation could be ignored if a given word-pair occurred above a certain very low threshold, about two or three times. One would not find too many titles in which the terms BIRTH and CONTROL were separated by a comma, such as "Season of birth, control of disease, and WHO statistics." Linguistic analysts have agonized over the problem of differentiating such items, but it is rarely a real problem.

Besides increasing the specificity and thus the informational value of the *PSI*, the main objective of pre-coordination is to reduce the number of permutations required. This did not prove to be as easy as we had first imagined. We have since found that precoordination is best performed by source-data edits, which requires constant monitoring of term-pair frequencies.

An important objective of permuted index display should be to minimize post-coordination by the user. For example, while BIRTH-CONTROL provides one level of precoordination, the resulting term is of such high frequency that one ought to be able to precoordinate BIRTH-CONTROL at a second level, with terms indicating drugs, devices, methods, etc., so as to narrow the focus of retrieval to less than ten articles for most searches. Obviously, the value of precoordination increases five-fold for a five-year cumulation, in which certain terms might occur dozens or even hundreds of times.

In closing this belated report on *PSI*, we should not overlook the application of the *Permuterm* concept in controlled or manual indexing systems. We first used *Permuterm* in a controlled indexing situation during the production of *Current Contents /Chemical Sciences*. Since then, we have used the method in producing the yearly index of the *Journal of the Electrochemical Society*, and some industrial organizations have used our *Permuterm* programs to generate their own indexes. Further, our on-line searching experience has demonstrated that *PSI* can be (and now *is*) used to facilitate searches of other data bases, such as MEDLINE, precisely because it displays term pairs that one might not think of or cannot find in thesauri such as MeSH. Otherwise, *Permuterm* indexing has had little application outside ISI.

A proper evaluation of *PSI* by the information community has yet to be published. Meanwhile, we can only report that *PSI* has been steadily gaining increasing acceptance among *SCI* subscribers. Most users today know how to optimize their use of the *SCI* with the most appropriate word index available for the time period covered in the search, whether for the period prior to 1965, when *PSI* first became available, or thereafter. Since 80 percent of *SCI* subscribers now also subscribe to *PSI*, it seems reasonable after more than ten year's development, to incorporate *PSI* into the *SCI* system. Thus in the future no user of the *SCI* will lack its complement, the *PSI*.

1. **Weinstock, M.** 1971. "Citation Indexes." *Encyclopedia of Library and Information Science*. 5 Vols. New York: Marcel Dekker, 1971;5:16-40.
2. **Garfield, E.** 1971. "Automation of ISI Services: *Science Citation Index (SCI), Permuterm Subject Index (PSI)*, and *ASCA*." *International Association of Agricultural Librarians and Documentalists, IVth World Congress, Paris, 20-25 April 1970*. Paris: Institut National de la Recherche Agronomique, 1971; p. 107-112.
3. **Citron, J.; Hart, L.; Ohlman, H.** 1959. "A Permutation Index to the Preprints of the International Conference of Scientific Information." Reprint No. SP-44, Revised edition. Santa Monica, CA: System Development Corp. 1959 December 15; 37 pp.
4. "Keyword-in-Context Index for Technical Literature." Report RC 127. New York: IBM Corp., Advanced System Development Division, 1959. Also published in: *American Documentation*. 1959;11:288-295.
5. **Stevens, M.E.** 1965. "Automatic Indexing: a State-of-the-Art Report." National Bureau of Standards Monograph 91. Washington, DC: Government Printing Office, 1965, March 30.
6. **Ohlman, H.** Personal communication, 1975, November.
7. **Garfield, E.** 1972. "Indexing Terminology and Permuted Indexes." *Journal of Documentation*. 1972; 28(4):344-345.
8. **Heumann, K.** et al. 1954. *The Chemical Biological Coordination Center of the National Academy of Sciences*. Washington, DC: National Research Council. 1954: p. 18.
9. **Weinstock, M.; Fenichel, C.; Williams, M.V.V.** 1970. "System Design Implications of the Title Words of Scientific Journal Articles in the *Permuterm Subject Index*." *The Social Impact of Information Retrieval: The Information Bazaar, Seventh Annual National Colloquium on Information Retrieval. May 8-9, 1970*. Philadelphia, PA: The College of Physicians of Philadelphia, 1970;181-200.
10. **Garfield, E.** 1969. "Permuterm Subject Index, the Primordial Dictionary of Science." *Current Contents*. 1969 June 3;22:22.
11. **Fenichel, C.** 1971. "Editing the Permuterm Subject Index." *Proceedings of the American Society for Information Science. 34th Annual Meeting*. Denver, CO. 7-11 November 1971;349-353.
12. **Garfield, E.** 1970. "Citation Indexing, Historio-Bibliography, and the Sociology of Science." *Proceedings of the Third International Congress of Medical Librarianship, Amsterdam, 5-9 May 1969*. Amsterdam, The Netherlands: Excerpta Medica. 1970; p. 187-204.

Note: Permuterm Subject Index (PSI), Science Citation Index (SCI), Citation Index, Index Chemicus, Rotoform Index, Chemical Substructure Index, Current Contents/Chemical Science are registered trademarks of the Institute for Scientific Information.