SOME KEYWORDS ON INDEXING

Over the past fifteen or so years, I have been acutely aware of the extensive and expanding report and journal literature on indexing projects. Many of these projects have resulted in the proliferation of commercial indexes, most of which are described by their producers as nonconventional and computer-produced.

Yet the introduction of a new conceptual system for indexing the informational content of documents is a rare event in the literature of chemical documentation, information science, and library science. The great majority of the indexing literature and of the systems used in the production of commercial indexes can be categorized under one or a combination of the following four systems:

1. Classified index—e.g., the Yellow Pages of the Telephone Directory
2. Subject index—e.g., the CA subject indexes
3. Uniterm index—e.g., the DDC or EJC index
4. Keyword index—e.g., from the title, abstract, or whole document

From time immemorial until the 1950's, the intellectual or scholarly world was reasonably content and secure with the various classification systems used for the ordering of knowledge and in the various classified indexes and subject indexes used as the key to the informational content of documents. Indeed, much of scholarship, from the Greek to modern civilizations, was intimately concerned with the classification and unification of knowledge. Aristotle, Linnaeus, the French encyclopedists, Darwin, Dewey, Mendelyeev, Beilstein, and Gibbs are but a very few of those who devoted their time and energies to this endeavor in their specific areas of interest.

Indexing undoubtedly arose as a by-product of and implementation to classification systems, probably first as alphabetical keys to names and places, and finally to topics and subjects in documents. Surprisingly, the dictionary is not a particularly old concept, that is, in terms of centuries. But the dictionary or alphabetical arrangement is the primary essence of indexing.

Over the ages, indexing evolved into an art and science; "good" indexes tended to have a well-defined grammar and logic and to be concerned with semantics. Most importantly, most of the "good" indexes had a purpose relative to a community of scholars who constituted the users of a given index.

Then we discovered the "information explosion." Almost concomitant with our awareness of the outpouring of printed pages, a new method of indexing, the uniterm system, came into full blossom.

With the advent of uniterm indexing in the late 1940's, its expanding use in the 1950's, and the introduction of its modifications in recent years, the label "nonconventional" was attached to the system to contrast it to subject indexing. In retrospect, the denigration of the art and science of subject indexing was made complete with the highly connotative label of "conventional." Because uniterm indexes tended to look like and be like telephone directories, except that each uniterm entry had many "telephone" (accession) numbers, they became naturals for processing in computers and for production by computer printouts. Anything that could be processed in and printed by computers was really "nonconventional," particularly in contrast to a "conventional" subject index on 5 × 3 cards.

The major merit of uniterm indexing has been its basic simplicity. A uniterm is a single word or a simple compound word that can be a key to all documents that used the word. Furthermore, the power of the word or uniterm can be enhanced by coordinating it positively or negatively with one or more words or uniterms, and word order is completely flexible. Disciples of the uniterm system were so entranced with the power of uniterms that many recommended a uniletter system (however, I have never seen a uniletter index).

Because the uniterm system was widely accepted and tended to be oriented to the words in the documents being indexed, keyword indexing became a logical successor. To some, the millenium had arrived with computer production of keyword indexes by the permuting of keywords in the titles of documents. Keyword indexing attained the status of sophistication with the computer production of indexes without the need for indexers and finally with the computer production of keyword indexes and keyword abstracts from the words in documents without the need for indexers or abstractors. Automation has been achieved if you accept the hypothesis that the words in a document are suitable subjects for the retrieval of the document, with or without a thesaurus to buttress the hypothesis.

In places of the grammar, logic, and semantics that played such important roles in the art and science of subject indexing, there is now a method for taking words from the title or contents of a document. What matter that the index is ad hoc and pragmatic, if the method is computer oriented. Let us not be concerned with relationships between a document and a potential reader; the method pulls out so many words from the document that the reader should find at least one or two that he can relate to, and if the document is mired in meaningless words, the system will give the searcher other documents that contain the desired word, either tangibly or intangibly. The important point is that the system is great for pulling words out of a document. It is not fair to confuse quantity with quality, method with judgment, or a computer with a brain.

HERMAN SKOLNIK