Changing technologies: impact on information: the case of string indexing

Timothy C. Craven

Since 1974 I have been working in the area of generating index displays by computer. I would like to use index displays, and specifically string index displays, to demonstrate the impact of new technology, especially of increased computing power, on information work.

In string indexing, a human indexer provides a string of terms—usually words or phrases in ordinary language and usually with additional coding symbols. Generally, each string describes a different item to be indexed. From each string, computer software generates a set of index entries, each entry describing, more or less completely, the item indexed, but under a different access point. When sorted and merged, and possibly enriched with cross-references, the index entries may combine to form an index display of considerable elegance. PRECIS is a well known example of such a string indexing system.

Computer power originally made sophisticated string indexing practicable by taking the burden of generating, sorting, and displaying multiple index entries from single descriptions off the shoulders of human indexers or clerical staff. The continuing increase in this power, I suggest, is now making possible a further step—the customizing of index displays to meet the needs of individual users and uses.

The question of producing different displays from the same input for different large groups of users has already been addressed in the context of translating indexing from one human language to another. The PRECIS/Translingual Project is an important application of this sort of translation to string indexing. But current increases in computer power in terms of storage and processing capacity make much more personal and ephemeral customizing possible.

Index displays have traditionally not been seen as ephemeral, even when produced with the aid of a computer, because of the time and money required to produce them. Thus they have been embodied in durable forms such as print, printouts, microfiches, and microfilm, while online searching has been carried out, typically, by means requiring users to put in a good number of carefully thought out commands in order to interact with the machine system. Even when an index display is available online, as in some library public access catalogues, it generally shows the same, durable structure from one user or use to the other.

An ephemeral index display need not be, and probably will not be, a display of an entire index, but only of the parts of the index useful to a particular user with a particular need at particular time. Since this specific user need will probably not be known in advance, the mechanism for generating the index display from a more primitive form of data must be so quick as to be almost instantaneous; otherwise, many of the advantages of scannable, browsable full-screen displays over other approaches to retrieval will be lost.

Microcomputers may prove extremely useful in providing some of the quick-response capability needed, but the necessary components are quick retrieval and quick manipulation and display of data, rather than any particular variety of hardware.

A number of questions call for answers about this kind of quick-response customizing. First, what options would users really like to have in string index displays? There is certainly evidence—for example, from the Wollongong University Subject Catalogue Study—that different users may have decided preferences for different forms of display. KWIC indexes may be loved or loathed.

The need for further investigation leads to a second question: how should different index display formats be categorized by researchers? As well as the complexity of description and the number of variables involved, there are serious problems of terminology and a lack of theory standing in the way of an adequate categorization.

The complexity of the task facing the theoretical categorizer suggests that a third question may be even harder to answer: how can different index display formats be categorized for users to specify their options? The emphasis on speed makes this question still more thorny. Perhaps some automatic method for index format selection could be developed, based on user and query characteristics?

A fifth question is one step back from the actual display. What form of underlying representation of indexing data should be adopted to permit the effective and efficient production of the desired index displays? Partly this will depend on the kind of index displays desired and how often they are desired. Partly also, however, it may depend on the answer to a sixth question: what form of input should be required?

Text of an address given to the Indexing and Abstracting Society of Canada, Toronto, May 1984.

The Indexer Vol. 14 No. 4 October 1985 255
Initial answers to these questions, and others like them, are likely to be partial, and oriented towards getting something that works at least some of the time. Answers to the sixth question, relating to input, will be of especial interest to indexers. One extreme would be no human intervention at all—whether an artificial-intelligence dream in which the machine system understands all and interprets all, or a quick-and-dirty approach along the lines of KWIC and KWOC. Another extreme would set the human indexer to complex coding of input—with fields and subfields, role and facet operators, substitute phrases, conditions and exceptions—to deal explicitly, if clumsily, with many possible varieties of output. In the former extreme, indexers go into other lines of work; in the latter, they must add new linguistic and logical competencies as well as a thorough knowledge of input coding conventions.

Perhaps the sixth question should really ask what forms of input should be permitted. That is, just as different kinds of display formats might be selected for output, so a variety of different inputs should be handled and the results stored in a common structure. This, like the partial solution of a multiplicity of restricted index-display customizing systems, would permit a gradation of indexing jobs and indexing competencies.

The same increase in computer power which makes possible the customizing of index displays may also put new aids in the hands of the indexer. Automatic checking of terms against a thesaurus or dictionary for standardization and for spelling-error correction can be carried out ever more promptly. Complex structures of term interrelationships can be represented diagrammatically on request. Effects of indexing on index displays in various formats can be monitored continually. Such a toolkit is, of course, two-sided: as the number and sophistication of the tools increases, so may the amount of knowledge required to use them to good purpose.

Some general truths about the effects of changing technologies, especially of increased computing power, on information work can be seen in the points I have raised on string indexing. In theory, new technologies make possible much more personalized activity by information workers and users while at the same time allowing for sharing of data. The standards for the shared data are likely to become quite complex, though initially there may be a number of partial, competing standards and, in any event, many workers and users may never be aware of the full complexity. A large number of levels of competency will be recognized and provided for in serving users, in classifying jobs, and in educating information workers. Unfortunately, hardware is now far ahead of theory and research in this area, with the theoretical and research problems being quite thorny; and the short-term result is likely to be ad-hoc, temporary, and perhaps clumsy solutions to specific practical problems.

Select bibliography
Hunt, R. et al. Wollongong University subject catalogue study: PRECIS, LCSH and KWOC: report of a research project designed to examine the applicability of PRECIS to the subject catalogue of an academic library. Part III. Section 1: methodology and results: interviews and library survey; Section 2: conclusions and recommendations. Wollongong, NSW: University of Wollongong Library: 1977.
Verdier, V.; Austin, D. Research into the translingual potential of PRECIS In: Third European Congress on Information Systems and Networks: overcoming the language barrier; Luxembourg; 1977 May 3-6. Volume 1. Munich, FRG; Verlag Dokumentation; 1977, 322-337.

Timothy Craven has been a member of the faculty of the School of Library and Information Science, University of Western Ontario since 1974.