Index Traps and Pitfalls*

Human Generated/Computer Generated

1. CHARLES L. BERNIER

Subjects vs. Concepts vs. Topics vs. Words

I find that people, including indexers, do not know what a subject index is. Strangely enough, it is an index to subjects. Documents, unless written by schizophrenics, have subjects. Let us say that a person has listened to a debate. I ask him its subject. He instantly says, e.g., 'Free trade vs. tariff protection'. Movies have subjects. In this country, the subject used to be 'Boy meets Girl'; now it's 'Skin meets Skin'. In the Soviet Union the subject used to be 'Boy Meets Tractor; Girl Meets Quota'. Today, it could be 'Breshnev Meets Nixon'. They've just met. Subjects differ from concepts, topics and words. Subjects are what the guy is writing about, working on, investigating, talking about, painting, photographing, teaching, debating, reporting, researching, developing, or inventing. Indexers flood users with useless entries if they index concepts, topics, and words rather than subjects.

Users

Index-users are largely uninformed about indexes. They cannot detect if more than half of the subject entries are missing. Indexing density (number of subject-index entries per abstract) for Chemical abstracts was 2.3 in 1935; in 1965 it was 6. Yet, through all of the thirty years, not one of tens of thousands of chemists who used CA ever complained that entries were missing—more than half of them in the earlier years!

Users do not understand cross-references, introductions, notes, search arrays and strategy, nomenclature responsibility for scattering, principle of maximum specificity, reading all modifying phrases under a heading, documentation of search, and even how to find indexes. Lack of knowledge about indexes prevents informed feedback from users. If most users raised hell every time they had to use a poor index, there would be fewer poor indexes.

Another thing, users blame indexes when the fault is clearly theirs. One critic of CA indexes repeatedly wrote to point out incorrect index-reference numbers. Checking showed that all the references about which he complained were in fact, correct. He had simply failed to find material actually in the abstract or original. Had he taken time to look rather than write, he would have saved making himself look foolish.

Users may feel that they must 'outwit' the indexer, not realizing that indexers cannot invent words and generally use terms in common use.

Since users are rarely connoisseurs of indexes, they do not know what they have the right to expect from indexes. Users do not know that indexes should (1) collect like entries; (2) provide guidance to related terms; (3) enable users to say, 'Yes' or 'No' rather than 'Maybe' after reading an entry, and (4) explain index policies. There are golden opportunities for the ASI in instruction of users.

Lack of numerical specifications

Current feeling is that all indexes are about the same. To paraphrase George Orwell, 'All indexes are equal—only some of them are more equal than others'. Imagine shopping for a car. You ask the salesman about

* Two papers read at the 1973 Annual Meeting of the American Society of Indexers.
the motor. He says he’s not even sure if the car has a motor—it may be pedalled. You ask him how many horsepower and he tells you that there are no horses in the car. You ask him to stop horsing around and tell you on what kind of fuel it runs. He’s not sure whether it runs on furniture polish or lighter fluid. Today, when a user shops for an index, things are even more vague. Almost nobody can tell indexing density percentages of cross-references; control of synonyms; percentage of incorrect references; internal vs. external guidance; kinds and amount of internal guidance; precision of modifying phrases; percentages of missing, incorrect, and unnecessary index entries; whether the publisher has tampered with the index; if the index is copyrighted; and even who compiled it. Lack of numerical specification leads to incredible variations in quality of indexes. Diamonds from Kimberley are quite different from Woolworth’s; girls need to learn what scratches what—not whom. Index users need simple tests, perhaps based upon the load that indexes shift to their shoulders. The ASI has an unparalleled opportunity here.

The question game

For some time, it has been customary to assume that users always approach an information system with questions. When they fail to get answers or receive only partial answers, it has been fashionable to charge users with inability to ask the right question. This is a neat way to duck responsibility for system failure. The user may be so uninformed that he accepts this undeserved responsibility.

For questions of discovery—in which the asker knows not if there are any answers—there may be no questions asked at all. Users may scarcely know that they have a problem, let alone a question. Many simply stumble or wait around until the solution or a datum that leads to a solution comes along. Sometimes when a datum or information arrives, users invent, on the spot, the problem that this datum or information solves. I call this ‘instant relevance’. Two ideas collide to generate hope—hope that is nourished by collision with still other ideas. Often, the entire process is too vague to be called asking questions and getting answers. Creativity generally comes, not by asking questions and getting answers, but by collision of ideas generating hope. All of us have many highly individualized areas of interest. We selectively and almost magnetically collect relevant data. A datum that solves a problem or gives hope causes us to act. This is cogent communication.

Loading indexes with responsibility

In the last quarter century, it has been a popular game (as in Games People Play) to load indexes, indexers and index users with responsibility for failure of better knowledge transfer. The false argument goes something like this: The major problem faced in knowledge transfer today is inability to find what has been written; we lack indexes and especially good indexes. Human indexers do such a poor job that we need computerized indexing. Humans do such a poor job, largely, because indexing is repetitive, mechanical, and boring. It is a task fit only for computers. Index users search so poorly in manual indexes that they need computerized services. The song and dance continues: If we really want to solve this crucial information problem and get the information we really need, we must develop computerized systems.

The truth of the matter is that indexing is not repetitive because indexers index the same thing repeatedly only by mistake—which is rare. Indexing is not mechanical because machines cannot index—as yet. Indexing is not boring. CA’s turnover of indexing chemists was 0.5 per cent per year; then, industrial turnover of technical personnel was 6 per cent; and governmental, 9 per cent. Chemists indexed at CA because they were genuinely interested in and enjoyed indexing the chemical literature.

The truth is that all of us are drowning in relevant reading material turned up by

The Indexer Vol. 8 No. 4 October 1973 225
allegedly inadequate knowledge-transfer systems. Why should we develop vastly superior retrieval systems that will drown us even more rapidly?

The truth is that, to test information systems, human indexers have been created by fiat—instant indexers—people who do not know how to index. Hence, the new system tested appears to be remarkably good by comparison. Experienced human indexers checking other human indexers in chemistry regularly change 20 per cent of the entries. Of this 20 per cent, only 5 per cent is of serious consequence, e.g., missing entries, misunderstandings, inappropriate headings, and scattering. The 15 per cent of change remaining is of a minor nature, e.g., wording of modifying phrases. During editing of subject-index entries fewer than 3 per cent of the entries are transferred from one subject to another. This means that more than 97 per cent are under the correct subject heading before editing, an excellent record for human indexers.

The truth is that information retrieval started out to be precisely that—answering specific questions. When that goal was not achieved, reference retrieval invaded the field of indexing. To justify invasion, human indexing was discredited. This wasn’t difficult, however, belittling was pushed much too far. At present, there is no computer subject-indexing—only associative retrieval, computer search of man-made index entries, and concordance building.

The truth is that a major problem that we face today is cogent communication—doing something about what we read. Those who research computer indexing could do more good, in my opinion, if they worked on cogent communication and got off the kick about computer indexing. It would be nice to have computer indexing—but not necessary.

Invisible loss

Invisible loss of relevant information is an extremely serious problem in computer search. Loss can also occur in the use of manual indexes. I wanted references on sucrose in etiology of ischemic cardiovascular disease. I asked for correlation of the MeSH terms: ‘Sucrose’, ‘Ischemia’, and ‘Cardiovascular diseases’ as Boolean products. Nothing came out through the computer terminal. So I dropped ‘Ischemia’. Still nothing came out. Finally, I dropped ‘Cardiovascular diseases’ and ran a deck search on ‘Sucrose’ alone. Four relevant items printed out in the first ten-item limit. Had I not known this to be an active area of research and had I not persisted, I would have missed pertinent material. Invisible loss can be reduced by the education of users, negotiation of questions, expert search formulation, and, I hope, by association retrieval.

Publisher naiveté

Some publishers view indexes as undifferentiated blobs on the budget, failing to distinguish reeking ‘Musk Oil’ from subtle Rogers & Gallet ‘Sandalwood’, or sonnets from doggerel, or diamonds from paste. For example, concordances of the KWIC and KWOC ilk have replaced indexes of high quality because of cost and romance of computers.

One publisher pulled cross-references out of excellent indexes because he couldn’t figure out how to have the new computer put them in. Thus, he loaded the hapless users with looking in two sources and facing blind leads, and loss of leads. Publishers introduce stereotyped subheadings because they are unable to adapt the computer to the infinite flexibility and subtlety of research and English. The ASI has a marvellous opportunity to show all publishers excellence in indexes.

Teaching indexing

Since accepted, numerical specifications for indexes are lacking, teaching indexing can become an exercise in witchcraft. Choice words are stewed in a brew of links, roles, weights, Boolean operators, etc., lubricated with zealot’s zeal and administered to the unsuspecting student under the label of ‘information retrieval’, a term that he may not
even suspect is phony. He may not know that actually retrieved are references to documents, which documents, if read with understanding, may inform him about a question he wishes he hadn't been forced to ask.

The ASI has a great opportunity to establish teaching standards for selection and paraphrasing subjects, choice of subject heading words and translation of these words into standard subject headings, writing modifications, and use of cross-references. Design of tailor-made and even exotic indexes can be taught. Index use should also be taught; here is a vast wasteland. Ph.D.s who have indexed at CA for a year have remarked, 'When I was out in industry, I wish that I had known all that I know now about indexes'.

**Education in subject area**

Free-lancers will probably burn me for this—I hope in effigy. Subject indexers who know less about a technical subject, e.g., chemistry, make more mistakes and more horrible mistakes in indexing than do those who know more about the subject. This is from experience. Seventy-five per cent of the subject-index entries of a new indexer have been changed during checking. Twenty per cent of an experienced indexer's entries are regularly changed. It is possible to teach a chemist indexing, but nearly impossible to teach an indexer chemistry. Ph.D. chemists make fewer errors and mistakes in indexing abstracts on chemical research and development than do master chemists, and these, in turn, make fewer than do bachelors in chemistry. Those without any degrees make the most serious errors and horrible mistakes, and many more of both. Because we have no numerical specifications, these mistakes of understanding pass undetected and unchallenged.

**Delays**

Delays in indexing can kill indexes and their organizations if computer buffs have convinced the boss that information is available instantly at the touch of a button. Deliberate deterioration of quality of indexes enables computer-search systems to appear sooner than do good manual indexes. By elimination of index editing and elimination of cross-references from the index, publishers and computerniks can create decadent search systems that appear sooner than do manual indexes of high quality. Since users are so generally uninformed, publishers can get by with the decadence which shifts added work on to the uncomplaining user. The ASI has many golden opportunities.

**Generic search**

One department in which computer-search systems shine is generic search—and especially combination-generic search. For example, if we want all documents on 'Side reactions of antibiotics in pulmonary diseases', a computer search in which exploded search (or tree search) is available is doubtless fastest. Perhaps two hours vs. two weeks may be the ratio.

**Word pollution**

All of us are faced by the abominable snow job—words, words, words. We have become erudite gastropods rather than erudite gourmets. There are too many words to become informed; data are dispersed as raisins in rice pudding; conclusions die beneath torrents of words; and explanations are exhumed by the word-weary. We are clever enough to invent new forms of writing and to revive useful old ones, such as aphorisms. While these are not directly the problems of indexers and the ASI, I do plead for the very bright minds in this field to work toward solutions.

---

2. **CAROLYN M. FLANAGAN**

'Next to knowing a thing comes the faculty of knowing where to look for it; and indeed, it is a question as to which form of knowledge should take precedence.'

'The capacity of human memory is closely limited, even in the most highly developed instances, and with the vast and continually increasing stream of knowledge and record
of experience in books, and in the flood of periodical literature, the most highly trained memories have long ago recognized the severity of this limitation.

'To know where to look for information about a desired subject is, then, the aim of practical study of current literature.'

The foregoing statements are not new, they are not mine, they appeared as part of the Preface to the Engineering Index volume III copyrighted in 1901.

My conversation with you today is based upon experience in attempting to create a communication link for engineers between the information they need and the information that appears in the approximately five hundred thousand technical articles published annually.

Unlike the monograph which is self-contained and usually a learned treatise on a small area of learning, or even the abstract journal in a single discipline, the scope of coverage of Engineering Index involves overlapping areas among disciplines which must be mediated and satisfied.

We have learned, and there is an ANSI Standard that says, that 'An index is a systematic guide to items contained in or concepts derived from a collection. These items or derived concepts are represented by entries arranged in a known or stated searchable order, such as alphabetical, chronological, or numerical'. We also know that the essential functions of indexing are to identify specific items entering a collection, to locate items in the collection, bring like items together or indicate relationships between items, and to correlate the language of the user with that of the indexer.

Your own Society is fostering a much needed project in the development of criteria for the evaluation of indexes in which the principal elements are structure or indexing scheme, depth of indexing, accuracy, and consistency.

The methodology for indexing engineering literature by subject with which I am most familiar was originally based on the Library of Congress List of Subject Headings and followed the basic rules for structuring established by LC. However, the headings were found to be too general and as a result both Main Headings and subheadings were amplified in consultation with engineers and subject specialists in order to provide direct access to specific terms, with cross-references suggesting the means to further search. Changes in terms have been made where definite usage has dictated the need and terms have been added because of new technological developments.

We chose a logical pattern in structuring our controlled vocabulary for indexing with the basic premise that this vocabulary reflect the nomenclature derived from the literature of the many specialized areas of engineering and technology. We also built in a constraint in the form of a time frame which did not allow changes within any one calendar year.

We soon discovered that you can control vocabulary but as a secondary service you cannot control the magnitude of primary information which needs to be indexed under a designated subject heading. Typical examples can be cited. The term Satellite appeared in the literature in the early 1950's but was not used as an indexing term until 1955 when three items were indexed under that heading. There were twenty-two items in 1956, and after Sputnik in 1957, we produced two pages of references. This bunching of information is not conducive to easy search and the patience of the user begins to give out.

Environmental Engineering has become the primary interest of the nation. In 1967 we thought of it as Air Conditioning or Bio-engineering or Human Engineering. Now in 1972 it has grown into Air Pollution, Architecture, City Planning, Ecology, Power Generation, Refuse Disposal, Sanitary Engineering, Sewage Treatment, Water Pollution and many process terms which appear in the literature of all engineering disciplines.
The latest Engineering Index has twelve pages of abstracts and thirty-five 'see also' references under Environmental Engineering. The patience of the user is now completely exhausted.

We have been criticized recently because the word Energy has not been included in our Subject Headings. But Energy is the capacity for doing work. Also bear in mind that engineering is the practical application of science and the use of internal, kinetic or potential energy forms the basis for much of applied science and thus for engineering. We have preferred to be more specific, but must now include the term Energy Conservation by popular demand.

Semantic problems will always be with us and the only solution seems to be a type of audit trail that may keep us reasonably flexible.

Additionally, problems come up in handling the foreign literature. The abstractor/indexer knows the language and is capable of translating it into good English but he is often called upon to translate material which is outside his area of specialized knowledge. Consultation between the translator and a subject specialist becomes a necessity, but even then some peculiar index references appear.

Depth of indexing very often is controlled by a dictate from on high because it is one of the areas in which cost savings can be effected. When you are producing a data base that requires five million key strokes per month, any small economies are welcome and the precision of indexing terms is scrutinized very closely and kept within a maximum of five for each article indexed.

And now for consistency. This is one of the greatest problems in creating an index, mainly because, as we have noted before, the capacity of human memory is closely limited. Indexers fall into traps of linguistic habit. They go along merrily using pigeon-holes that are correct and comfortable and then begin to have some doubt when new terminology is introduced, and it is too late or too difficult to modify the index under construction.

To this point, I have been concentrating on the index which appears in published or book form. When the same indexing policy and methodology are used to create a data base for computer search, our problems are magnified and made more apparent. The problem of consistency is highlighted in conducting retrospective searches. Updated, changed, and/or modified vocabulary lists cannot be used to structure strategies for demand searches against back files. There are numerous other machine problems but the machine itself may be our biggest trap.

And now, for the best way to stay out of subject index trap: know your subject; analyse the content of your material; organise your vocabulary; be as specific as possible; do not 'crowd' your subject areas; follow the ground rules established at the administrative level; and above all—attempt to keep the needs of your user or reader in mind.

**BIOGRAPHICAL PROFILES:**

by ANNE J. RICHTER, of the American Society of Indexers

When I was last in London I called at the British Museum to see Mr. Richard Bancroft, then Chairman of the Society of Indexers, who received me most graciously and suggested that I might wish to go down to Putney to meet Mr. G. Norman Knight, the Society's distinguished founder and President. This I was delighted to do and no sooner had I returned to my hotel than Mr. Knight was on the phone inviting me for luncheon a day or two hence. I met him at Scio House, saw the row of bound volumes of The Indexer from its first issue, his extensive reference shelf, and all the working equipment of the scholarly indexer that he is. Also present was Mrs. Elizabeth Wallis, the lively assistant editor of The Indexer.

In the course of our conversation about the mutual aims and achievements, the problems and pleasures, of our two Societies (which seem much the same on both sides of the Atlantic), it was proposed that I prepare brief biographical profiles of the current officers, board members and committee chairmen.