COMPUTER-AIDED PRODUCTION OF BOOK INDEXES

THEODORE C. HINES and JESSICA L. HARRIS

Introduction

The Columbia University School of Library Service offers a course in indexing which is general in scope but includes book indexing both as an important area in its own right and as a type of indexing which is familiar to students and can be used to illustrate indexing principles.

The School also offers courses in information systems, with significant emphasis on computer applications.

For some years now, the authors of this paper have been using the computer as a tool for exploring information handling. This research and development work has been aimed at distinguishing unit operations in information handling, somewhat analogous to the unit operations of chemistry.

We have an advantage in that all of our programming is done by the authors of this paper; that is, the design and execution is done by experienced indexers and bibliographers rather than by systems analysts and programmers. Our theoretical approach cannot be considered adequately tested without production of actual, usable information tools of various kinds: journal indexes; thesauri; concordances; KWIC, KWOC, and other title-derivative indexes; and book indexes.

The research and development programme has now reached the point at which we can make computer programmes for the production of all of these types of tools available for use by students who have no knowledge of computers and computing, and who use the programmes as tools in learning about indexing and information systems. At present, this is done on an individual basis for those students with a special interest in these applications. We are in the process, however, of making use of the programmes a required part of course work.

Goals of Book Index Programme

In the case of the system for computer-assisted production of book indexes, we had two major goals in mind. In the first place we believed that, if we could define and codify a significant proportion of the work involved in producing book indexes so clearly and unambiguously that we could programme it for the computer without loss of index quality, this definition and codification would be an aid to further research into the indexing process. In the second place, we also believed that we could develop a system which would be immediately economic for production of indexes which are designed to be cumulated (such as the subject index to Chemical abstracts) or which are continuously revised for new editions or printings (such as those of many encyclopaedias).

Originally, we did not expect to develop, as we believe we have ended by developing, a system which would be economic and feasible for the production of book indexes generally, even when revision or cumulation is not expected. Of course, the terms economic and feasible must be taken with a grain of salt. The system will not compete with those free-lance indexers who, all too often, are forced to accept less than the going hourly rate for clerical workers for their time. Nor, for that matter, do most free-lance indexers as yet have access to keypunches or terminals, or to computer time on a rental basis when and as needed.

But, in saying that the system is economic, we are doing so against a background of studies in the clerical and professional labour costs of book indexing. We are also taking...
into account the costs of the keypunch or terminal time used and the going rates for rental of computer time on a commercial basis in New York City.

Equipment Requirements

The assumptions that would underlie a decision to adopt this system for book index production are only that input equipment and computer time be available at the normal rates for such equipment today. While free-lance indexers would not be in this position unless they maintained a group of indexers working in-house, most publishers, at least in the U.S., and to a large extent in Britain, should currently be able to meet these requirements.

This estimate does not take into account possible savings to be derived from using the system's computer output to drive graphic arts composing equipment for the production of the final index. We are assuming only the production of a manuscript index for the printer, and our costs are compared with those of producing an index on cards, properly subordinated and consolidated, but not styled for the printer. We know that the output of our programmes could be used for graphic arts output devices, but have not used them thus and have no computer programme for this purpose. We have every reason to believe that this would be economic, too, but have no hard data on this phase of the operation.

Indexing Costs

It should also be noted that we have not included in our costs capital investment in programming. This investment is considerable, but we have done the work bit-by-bit over a long period of time, our programming is modular so that much of the cost would have to be allocated over other programmes for information handling, and we did not keep track of costs for programming.

Detailed costs are not of particular importance here, but benchmark figures are. The actual out-of-pocket cost, not counting overheads, for indexing the McGraw-Hill Modern men of science, Vol. I, in 1967, were very nearly $.25 per entry, with an entry defined as a page reference or an inclusive page reference (e.g. 63-92). Most of the cost was in labour, and we allocated the very modest rates of $2.00 per hour for clerical time and $3.50 per hour for professional time. These rates would of course have changed by now. The labour involved was about 50 per cent clerical and 50 per cent professional and supervisory. While the procedures followed were not the most economic, the book was somewhat over-indexed, which should produce a lower cost per entry, and index design costs were not counted. This study of manual procedures was carried out by Alan Greengrass on an index designed by us. If the figure of $.25 per entry sounds high, we suggest costing actual index production on your own before communicating with us on it. We are willing to admit that more careful management might have lowered the costs by as much as 25 per cent, but not that most free-lancers are much more careful managers. Personnel costs in the U.K. might be somewhat lower. The computer-assisted indexing system discussed below is somewhat cheaper, and results in a manuscript index rather than the card index for which the cost figures were kept.

The Computer-Based Index System

How does the computer-based system work? Basically, the indexer marks or underlines the text and writes in the margins to indicate the entries he wishes made. There are no limits on the length of entries, or on their nature, but the styling of the entries must be precise so that they will file correctly by our computer filing procedures. Non-alphabetical classed groupings of entries are possible, but would require either hand filing of entries at extra cost or development of a numbering system for the classed groups so that filing could be on the basis of the numbers. We have recently found the latter method to work quite satisfactorily in a thesaurus for which we wanted classified sub-arrangement under several of the alphabetically-arranged main terms. The computer filing problem, some published opinion to the contrary notwithstanding, is a complex
As any professional indexer who consults the sub-arrangement of entries in the Index medicus may see for himself. We have discussed computer filing at length elsewhere, and shall not recapitulate that discussion here.

**Keying of Entries**

The keyboard operator keys the entries from the marked manuscript just as the typist would type cards. The programmes do not set any limitation on the kind of input equipment used. We have available to us 2260 consoles, a 2741 typewriter-terminal and 029 keypunches. We also had, for about a year, a Dura Mach 10 paper-tape typewriter. Aside from the latter, all our input equipment has been IBM-manufactured. We use, by preference, the 029 keypunch for input for indexes, since this gives us our entries on convenient, eye-readable cards.

We also prefer to key our input in all capitals. There is no programme limitation on using upper and lower case input, but we believe that we save considerable time and effort by using an automatic capitalization programme segment which gives us upper and lower case output by rule. The rule in this case is that all words not on a stop-list (such words as a, an, the, of, etc.) are always capitalized unless a special symbol is punched before them, and vice versa, with the additional exception that words on the stoplist are capitalized in any case if they are the first word of a subject, though not if they are the first word of a modifier or subdivision. A symbol may be similarly keyed before all acronyms which should appear in all-capitals, such as NASA, or a go list can be used to index these intended to cumulate. We are in the process of developing a routine which will be the reverse of the above: that is, all initial letters not the first word of a subject, and not preceded by a special symbol, will be lower-cased.

We have an alternative means of dealing with such capitalization exceptions as acronyms, which often seems preferable. This is simply to treat them as if they were keying errors and to punch a correction for them after proof-reading so that they are handled with all other corrections and in the same manner. This is likely to be preferable because there are so few instances of this kind, and all of the occurrences of a single entry word (NATO, for instance) can be corrected by punching a single correction card.

The input configuration is quite flexible. Any convenient convention can be used to separate subjects from modifiers or subheads. For an index limited to a single modifier for each entry (e.g., *Chemical abstracts*) or to one level of subordination, subjects may be separated from the modifier by a colon and two spaces, e.g., TUNGSTEN: MEASUREMENT OF DUCTILITY.

The programme permits (and we may use) up to three levels of subject subdivision, and there would be no particular problem in providing for still more levels. Where more than one level of subdivision is required, we generally use the space, slash (solidus), space as a delimiter between subject and modifier, and between modifiers. Any symbol or sequence of symbols not occurring in the entries themselves could be used. This one was adopted because it is easy, logical, and unambiguous, and does not offend the eye in proof-reading input or checking cards.

It is also necessary to distinguish references (page numbers, for example), cross-references, and scope or other notes. Again, any convenient, non-ambiguous characters or sequences of characters can be used. We have used three spaces before references, e.g.,

TUNGSTEN 27

and four spaces before cross-references, scope notes, definitions, or other matter inserted in the index, e.g.,

WOLFRAM SEE TUNGSTEN

WOLFRAMITE SEE ALSO TUNGSTEN

Entries are keyed in the order of the text references; that is, the order in which they are made by the indexer. This has certain advantages, since (except in the case of inclusive page references such as 27-32) the page reference may be keyed only once for each page, and computer-posted to each entry for that page. A programme segment...
would then check the continuity of page numbers to be sure that the keyboard operator has not forgotten to change the page reference. We have not followed this procedure in doing indexes, since we have no regular keyboard operator but depend upon students without keyboarding experience to punch their own data. We have used it, however, in providing both line numbers and speaker designations in a concordance to Goethe's Faust.

We do find that the system allows us to do preliminary indexing of manuscript or galleys, providing page references at a later point. This does not save money and effort, but does permit considerable improvement of quality and saving of time over indexing after page proofs are in and publisher pressure is to have the index out yesterday.

The indexer must style his entries carefully so that they will file correctly, and keying must be accurate and careful.

After keying, entries are filed by the special sort routine, and the filed entries run through a final programme segment. This programme consolidates page references, e.g.,

TUNGSTEN 27-32
TUNGSTEN 44

become

TUNGSTEN 27-32, 44

It also subordinates modifiers under the appropriate subjects:

TUNGSTEN 27-32
TUNGSTEN 44
TUNGSTEN/MINING 211
TUNGSTEN/REFINING 38

become

Tungsten 27-32, 44
mining 211
refining 38

Page references may be right-justified, with or without leading dots, or be left ragged right. The indexer may specify whether he would like modifiers or subheadings for subjects with no page reference consolidated with the subject or not: that is, he may indicate whether he would like

ORES/MINING 311
ORES/REFINEMENT 17, 291, 312-3

to appear as

Ores, mining 17, 291, 312-3

or as

Ores
mining 311
refinement 17, 291, 312-3

The indexer has a wide range of similar options, including numbering (and heading) as pages or leaves, and capitalization or non-capitalization of articles, conjunctions, and prepositions at the beginning of modifiers.

He may specify the number of columns to the page, page and column width, margins, indents, and so on by changing simple English-language definitions at the beginning of the programme.

Corrections are made either by removing offending cards in input and substituting correct cards or by keying enough of the error to identify it uniquely between corrections, followed by whatever is to replace it, e.g., a mispunching of 'Amurican Society of Indexers' could be corrected by

‘Amurican’ = ‘American’

A similar technique can add or delete entries.

Since the programme will accept upper and lower case, special symbols, and so on in input without affecting filing or organization of the material, any indications of italics or accents may be recorded. Unless requested, they do not appear in the manuscript index, but would be available for graphic arts output. Provision exists in the programme for inserting font indications (boldface for subjects and italics for references, for example) by rule, so that much styling normally done by the indexer or editor could be done in this fashion.

Even for upper-and-lower case computer output, overprinting can be used for some accents, for boldface, and for underlining, so that the computer printout is equal in quality to good typing.

There are no limits on the length of any elements, no ‘fixed fields’ in input, and no
labelling of input (except that implicit in the styling conventions given).

Subject index entries may be added to abstracts when keyed. In this case, the abstract number is posted to the main entry and entries for the subjects created for input to the index system, along with author and title entries.

While other book indexing programmes have been written, those we know of are special purpose programmes which cannot handle complex entries, varying subordinations, and different output formats. An example is given in Reference 3.

Several indexes have been handled in this way, including the index to our own book on computer filing. Others are underway, as indicated by the sample (Figure 1) of Gail Persky's index to the American Library Association's Preconference on Subject Analysis, where the example is unedited and the references are to the manuscript papers. The sort routines have been tested on very large bodies of data, including a number of book catalogues and bibliographies. We have plans and preliminary runs at present, for use of the procedures on a consolidated index to the Library of Congress Classification, an index of more than 150,000 entries, and are confident of carrying it through if funding for input, editing, and production becomes available.

PAGE 9

Format

Multiple Necessary in Classification

 Formats

 New for Thesauri

 Frequency of Terms

 Establishing Inclusion in Thesauri

 Fuellhart, P.

 Comparison of Information Science Lexicons

 Garfield, Eugene

 Citation Indexing Technique

 Growing Thesauri

 Definition of

 Guide Cards

 Eliminate Typing of Subject Headings

 for Easy Changes in Subject Catalogue

 Preprinted

 Harris

 Comparison of Lc and Wilson Headings

 Harris, Jessica L.

 Modern Subject Heading Theory

 Study of Type of Adjectival Phrase Headings

 Haykin

 Subject Headings: a Practical Guide

 Haystaq System

 Early Demonstration of

 Hazelton, Robert

 Theory of Three-Dimensional Classification

 Headings

 Not All Analyzable Into Subject and Aspect

 Those Without Syndetic Attachment

 Herner, Saul

 Measuring the Effectiveness of Modern Indexing Media

 FORMAT

 IV, 4

 X, 15

 X, 5

 IX, 4

 II, 10

 X, 16

 VI, 15

 VI, 14

 VI, 21

 V, 19

 V, 28

 V, 2

 VIII, 10

 II, 7

 V, 22

 VI, 7

 XI, 53
Hierarchies
Subject Analysis and User Both Based on Terminology I, 19
Hierarchy
of Library Reference Tools to Match User Specialty Hierarchy I, 16
Rigid in Traditional Classification III, 14
Homography
Problem in Retrieval Systems VIII, 25
Homonyms, Problem in Retrieval Systems VIII, 26
Human Inaccuracies
Evaluative Studies Show Effect of XI, 35
Index Effectiveness
How to Measure XI, 4
What to Measure XI, 4
Index Entry
Concern with XI, 2
Index Language Variables
Aid in Success of Co-ordinate Indexes VII, 7
Definition of VII, 7
Index Medicus
Pre-Co-ordinated Manually-Searched Index VII, 16

Figure 1. Sample page of index to Preconference Proceedings.

The programmes use procedures, or building blocks, Meccano-like sections, which are used very widely in all our other programming—the unit operations discussed earlier. Of course, they deal only with the mechanical and clerical operations of book indexing. The identification and expression of indexable matter remain the province of the human indexer. But much of the handling of the entries, of repetitive typing, and of the styling of entries often done by the indexer or editor has been shifted to the broader back and faster fingers of the computer. At 1,000 entries per minute filing time and at about 20 pages per minute formatting time, a computer can give you quite a lot, including trial runs, before you go broke, even at $5.00 per minute, which is an outside upper limit. All of the figures in this report are loaded, if at all, against computer use.

We are happy with our production system. More important, however, is our feeling that pushing as much as we can regularize on to the computer leaves us freer to investigate what the indexing process itself actually is, and to work on our continuing series of studies in automatic indexing of various kinds. Working with the computer, with its requirement for rigour of definition, has taught us much indeed about filing and about entry form. Some of this we knew intuitively well enough to produce what we still think are good indexes—but intuitive knowledge only gets research started, and exact knowledge is required to advance it.

We do not think any system we or anyone else may devise can replace any professional indexers. We are convinced, however, that such work as we are doing may help indexers to become more professional, better indexers, and may free them from drudgery. Or at least, as a friend of ours has remarked in another context, while computers may not free us from drudgery, we will at least get more drudging done per drudge.

REFERENCES
2 Since this article was submitted for publication, this routine has been written and is used in the index to a forthcoming book by one of the authors (Harris).