A Computer-generated Index Technique* 

BRENDA HALL 

Indexing is not a new science. The preparation of indexes had reached a high peak of skill in the English world of letters by the mid-eighteenth century. In those more leisurely days an indexer had time to bring to his work much art as well as the necessary craft. Confronted with today's overwhelming volume of print in all fields of knowledge, however, indexers would be in danger of failing their public if forced to depend solely on conventional manual methods. The alternative—a marriage of convenience with the computer—seems to some indexers to be a treasonable alliance. But there are those who have been taking their first timorous steps in the direction of computer-generation of indexes.

It may help to understand progress in computer-generated indexing techniques if the usual process in making an index by the manual method is broken down into its components.

A Text to be indexed
B Trained indexer
C Material from A abstracted by B on to cards, making as many separate entries under different headings as are necessary to cover each concept
D Product of C arranged alphabetically by B
E Product of D edited by B†
F Product of E typed into copy for printer
† The indexer may work on C, D and E more or less simultaneously; it is broken down into separate processes for the convenience of readers who may not know how an index is prepared.

The above processes are gone through in small or greater degree whether the desired end-project is a simple index of names or a highly organized index to a detailed work of scholarship. The latter type of index is sometimes called an articulated subject index.

The main advantage of an articulated subject index is that it provides a well-organized display of indexing phrases which it is easy to scan visually, because the technique of multiple indentation helps the reader to find what he is looking for quickly. Articulated subject indexes are used successfully with both large and small indexes.

*This paper is compiled from material originally published as two separate articles in the Cartographic Journal (Vol. 7 (2), page 70, and Vol. 9 (1), page 8). The material is here reproduced by permission of the Editor of the Cartographic Journal.
The selection of the headings and modifying phrases, the manipulation of language, and word order to give the smoothest reading and the correct emphases, and the arrangement of common material in a logical system of heading, subheading, and sub-subheading, all in alphabetical order (i.e. the processes of A-F above), all these have been undertaken intuitively by the experienced indexer, taking into account the nature of the material and the needs of the reader who will consult the index. It is only recently that it has been discovered that indexes of this kind have a logical basis which can be expressed in mathematical terms and so be simulated by computer program. This logical basis is the point of common ground from which a happy marriage of art and science may spring.

Studies carried out in the Post-Graduate School of Librarianship in the University of Sheffield, under the direction of Dr. Michael Lynch, and with the assistance of grants from the Office for Scientific and Technical Information (Department of Education and Science) have led to the devising of a novel computer technique for the production of articulated subject indexes, the feasibility of which has been demonstrated in a number of trials. Some of the trials were based on samples submitted by the author of this paper, who acts as indexer of the Cartographic Journal, with the encouragement and support of the British Cartographic Society.

The first program used for these trials was written by Janet Armitage in Sheffield whose work formed the essential basis of subsequent studies. A later and more sophisticated version of Miss Armitage's program was developed by J. Howard Petrie, also in Sheffield. This is not the place to discuss (nor is the author of this paper competent to describe) the technical details of the program from the point of view of the computer scientist. Readers of this publication may however find the following account of the processes involved from the indexer's point of view both interesting and, it is hoped, reassuring.

The process may be summarised as follows:

Indexing phrases are derived by the indexer from the text to be indexed (compare stages A-C above). As many phrases are constructed as are necessary to extract all the material for the index. (In what follows, these phrases are referred to as the 'input'.) The input phrases, prepared on this basis, are then punched on to cards or tape ready for computer processing. The computer, having been instructed by the program then selects each heading in turn, produces an entry for each, and arranges them in appropriate alphabetical order. (Compare stages C-E above). The computer is linked to a line printer and the 'output' may then be compared to stage F above, providing the basis of the printer's 'copy'.

The indexer first then, brings his expertise to bear on the abstraction of relevant material from the text in order to derive the indexing phrases which are to be punched on to cards. For example, it may be that the material to be indexed starts as the multiple concept

problems and techniques in land use mapping in Malawi, page 123

The indexer wishes to have entries under Land use, Mapping, land use, and Malawi. By the manual method he would write something like the following three cards:

Land use mapping problems and techniques in, in Malawi, 123
Mapping, land use problems and techniques in, in Malawi, 123
Malawi problems and techniques in land use mapping in, 123

To achieve a comparable result in the computer-generated index one 'input' phrase is written and subsequently punched in exactly the following format

PROBLEMS AND TECHNIQUES IN <LAND USE <MAPPING>> in <MALAWI>, 123
The computer selects and alphabetizes the headings to give the following three entries:

**LAND USE MAPPING**
**PROBLEMS AND TECHNIQUES IN - , IN MALAWI, 123**

**MAPPING, LAND USE**
**PROBLEMS AND TECHNIQUES IN - , IN MALAWI, 123**

**MALAWI**
**PROBLEMS AND TECHNIQUES IN LAND USE MAPPING IN, 123**

The first obvious advantage is that for the effort of writing one phrase three entries are derived and arranged in alphabetical order. The second is that the ‘nesting’ of the parentheses (which it will have been noted are the symbols which indicate to the computer what words are to be selected for headings) is most helpful in subject matter where composite terminology is a characteristic (e.g. air photography, hydrographic charts, topographic mapping).

A further and sophisticated feature of the program is its capacity to select material common to several entries to produce alphabetically arranged and indented sub-headings. For example, in the following input phrases:

**USE OF <BENCHMARKS> IN <DATING> OF OLD MAPS**

**SIGNIFICANCE OF <WATERMARKS> IN <DATING> OF <ORDNANCE SURVEY MAPS>**

<DATING> is common to both. The program generates the following entries:

**BENCHMARKS**
**DATING OF OLD MAPS, USE OF - IN OF ORDNANCE SURVEY MAPS, USE OF - IN**

It will be seen that the common material has been selected by the computer to give a sub-heading, i.e. dating. A similar selection of common material will be brought forward to appear as a sub-heading under WATERMARKS and ORDNANCE SURVEY MAPS.

It may be desired to include the following material in the index:

representation of <rivers>, <roads>, <railways>, <urban areas> and <rural areas> in road maps.

The needs of the user of the index dictate that entries should be found under each of the bracketed terms, but not that all the material will be found repeated in each case. The typical entry should read:

Rivers representation of, in road maps
not
Rivers representation of roads, railways, urban areas, rural areas and, in road maps

The earlier program required the indexer to write separate input phrases for each heading, five in all. By constructing the input phrase in the following way the five desired output phrases are generated from one input phrase:

representation of !<rivers>, <roads>, <railways>, <urban areas> and <rural areas> ! in road maps

It is not, however, possible to use this technique for punctuated phrases or names or for phrases incorporating ‘nesting parentheses’. Separate input phrases have to be written, for example, if it is desired to generate headings from the following:

<Smith, D. G.>, <Jones, E. R> and <Wilson, R. A.> paper on automated cartography production of <hydrographic charts>, <aeronautical charts> and <bathymetric charts>

Even with these limitations, however, this is a useful addition to the program.

Successful use of the program depends on (a) the prolific use of noun phrases separated by connective words like ‘of’, ‘for’, ‘in’, ‘by’, ‘to’ (referred to hereafter as function words), and (b) expertise born of study and practice in manipulating the word and phrase order in the input phrases. This is because the program operates on the basis of six hierarchical rules; if the first rule operates, rules two to six do not come into force, if the first rule does not apply, but the second one does, then rules three to six do not come into force, and so on. These rules are:

1 If the subject heading is first in the input phrase the modification appears in the natural order;
2 if the subject heading is part of a complex noun phrase the remainder of the noun phrase must occur first;

3 if the subject heading is followed by 'of' the preceding component comes first;

4 if the subject heading is preceded by 'of' the preceding component comes first;

5 if none of these apply, the most common component occurring in all phrases with this subject heading is selected to come first (to constitute a common subheading);

6 if the first component cannot be determined by rule 5 then the original order is preserved.

The Cartographic Journal which was the subject of some of the trials of this program is a learned periodical appearing twice yearly in issues of between 60 and 80 pages each, numbered consecutively (so that issue 2 of each volume starts with the page following the last page number of issue 1). The contents include learned articles, reviews, Cartographic Society proceedings and news, and miscellaneous items of interest to cartographers. The scope of the index is thus very diverse.

The earlier program was used to prepare an index to Volume 5 of the Journal. The later program was first used to prepare an index to Volume 6. Work of this comparatively limited scale and undertaken in a short span of time did not produce any very intractable problems for the indexer, other than those which are dealt with below. The program was then used to produce a cumulative index for Volumes 1-8. Problems arising at the input stage and the output stage will be discussed separately, but first it may be convenient to add a word about the actual mechanics of punching cards.

In the trials on the Cartographic Journal index the input phrases were prepared and punched by the indexer herself, who was fortunate enough to be given access to punching equipment at the Cripps Computing Laboratory at the University of Nottingham. The principal point to be made is that it is essential that the exact format has to be observed down to the last comma and space, and in particular, placing of < > symbols. Mistakes cannot be rectified except by punching a new card. This takes time and very accurate checking through, because a mis-type can cause a faulty entry and the computer cannot be programmed to correct a mistake; the most it can do is include in the output 'FAULTY ENTRY' and go on to the next card. Thereafter the only way to get the output correct is to replace the faulty cards with correct ones, and process the whole batch again. This is expensive and for practical purposes really rather impossible; the process can be stopped however, at the end of the feed-in of input, and faulty cards be detected from the print-out of the input material, provided facilities exist for this. What this amounts to is that an exceptionally high standard of accuracy is needed in preparation of the punched cards, which naturally takes quite a lot of time. The keyboards of the punching machines used are designed to provide upper case characters only, and so expert typists sometimes find it happens that mistakes arise out of old habits rather than lack of basic typing expertise. Bureaux specializing in punch card operating do provide this kind of service, but, of course at a cost.

Problems arising at the input and output stages may now be discussed.

A. Input stage.

The production of a cumulative index presents problems different from those which arise in handling a limited project such as a single volume. In this case the problems were aggravated because:—

1 Data were derived of necessity from five separate sources:

   (a) the existing index to Volumes 1-4
2 A great deal of work was involved in editing and streamlining the input material for the following reasons:

(a) the literary forms of the components listed in 1 (a)-(e) above were all different because of the different conditions in force at the times the work was originally done;
(b) it was necessary to establish consistency of form both in content and in punctuation of all words and phrases to be used as headings;
(c) recurrent material had to be traced through all issues and comprehensive entries compiled.

The need for consistency in form and punctuation may be illustrated. A journal compiled as a symposium of the work of different authors and editors may present particular difficulties. For example:

Smith, D. G.; Smith, David G.; Smith, David; Smith, D.; Smith, Dr. David; Smith, Dr. David G.; Smith, Dr.

If these variants were all to occur, the indexer operating manually might have cards written for all seven forms, particularly if they arose at widely spaced intervals of time. But they would be all filed together alphabetically, and in editing before printing, the queries as to identity would arise, investigation would take place, and the final entry would be determined according to an adopted principle, for instance, 'all known initials are given but no titles', so all entries would appear under Smith, D. G.

But in preparation of the computer index theoretically no alphabetical filing is necessary since the program does this part of the work. However, if the material is prepared over a long time, one of two situations may arise. Either the indexer will write eight separate input phrases giving rise to eight headings in the output, possibly some separated by intervening 'Smiths' with other alphabetical demands, or, the indexer, growing suspicious as yet another Smith is encountered, has to trace the earlier entries in order to collate them. Ideally the preparation of the input should not involve this type of search and no alphabetical filing should be necessary. But in practice, handling material of the scale of a cumulative index for the eight volumes of the Journal revealed the need for frequent searching of already-prepared input phrases. Therefore some alphabetical filing of input phrases had to be adopted, albeit reluctantly.

Need for consistency of punctuation may also be briefly illustrated. One paper may write of 'place names' and another of 'place-names'. Input phrases may be derived on widely separated occasions. In a manually prepared index two cards might be written but would be filed together, for example:

Place names, standardisation of Scottish Place-names, transliteration of Arabic

In the production of the edited typescript only one heading will appear, with both phrases listed below it. Given two input phrases prepared one with hyphen and one without, the computer reproduces them separately, treating them as separate concepts. Unless they are filed together during the preparation of the input the discrepancy may not be revealed until it is too late. A modification to the program could be made to remove this type of difficulty, which exists at present.

In the matter of preparation of recurrent material the input phrases may take longer than the preparation of similar data by the manual method. For example, in approxi-
mately every second issue of the Journal there are references to the Cartographic Society Annual Report for the year. A manual index would contain one card:

British Cartographic Society
Annual Reports
1965/6 ......
1966/7 ......
1967/8 ......

To yield the same information underneath a single heading the program requires separate punched cards for each entry:

<British Cartographic Society> Annual Reports
1965/6
<British Cartographic Society> Annual Reports
1966/7
<British Cartographic Society> Annual Reports
1967/8

Preparation of a cumulative index raises considerable problems of terminology for the indexer. The sort of problem encountered is illustrated by material arising from papers on automation in cartography. Scarcely touched on in Volume 1, it is getting under way in Volumes 2 and 3, and is in full flood by Volume 4. It starts to splinter into separate concepts. Automation applied to cartography becomes automated cartography, computer mapping, automated plotting, automated draughting, automated contouring, and so forth. Even the most intuitive indexer cannot always foresee how subjects will develop. Solutions to this problem are at the root and heart of successful cumulative indexing.

The problem was more acute when working with the computer program in preparation of the input phrases, because again it became necessary to search through previously prepared phrases for trends in terminology, and to edit to bring cognate material together. In practice in preparation of the cumulative index for Volumes 1-8 of the Journal it became necessary to devise a system to make easier the search. Two methods were tried.

1 input phrases were numbered in sequence and each phrase occurring as a heading was listed alphabetically with the number shown against it. This meant that a heading occurring frequently might occur in eight or ten phrases scattered through the input, each of which had to be consulted when adding additional material.

2 a system of modified alphabetical filing of the input phrases was adopted. It must be remembered that many phrases contained more than one concept to be used as a heading in the index, for example:

(a) <accuracy> of <sampling methods> in <population mapping>
(b) experimental <sampling methods> in <population mapping>

The practice followed was to file the input phrase under the alphabetically earliest heading it contained, whether or not it was the first in the phrase: (a) would therefore have been filed under Accuracy and (b) under Population mapping. But a further control had to be established because of the composite nature of the phrases. A search for population mapping would not automatically lead the indexer to look under Accuracy. Therefore a strictly alphabetical list was compiled of subject headings as they emerged, together with each other heading found in association with them:

```
A                    B
Accuracy             Sampling methods
Data banks           Automated plotting
Population mapping   Population mapping
```

Each item in list B had its own place in this alphabetical list in column A, with its particular associates under B:

```
A                    B
Population mapping   Accuracy
Data banks           Sampling methods
```

Both these methods described in (1) and (2) had their advantages. The first was tried out successfully in preparation of the index for Volume 6. The second was preferred for the cumulative index because other reasons (see above) already pointed to the need for alphabetical filing of input phrases.
No doubt other controls could have been devised. Experience suggested that if an index generated this way was to be truly integrated and systematic, control of terminology and retrieval of input phrases during preparation of the input was most necessary. This added much to the indexing time and work involved, and was by way of making necessary a data retrieval system for a data retrieval system. This cancelled out the saving of indexing time which the program otherwise allows.

In fairness to the program, however, it must be stated that if the technique were used for a cumulative index right from the beginning, it would be possible to edit the input phrases for each volume as they were prepared, using either the previous volume index or a print-out of the volume file which is in effect a keyword out of context index. Edit programs are available for deleting entries, but these would not be suitable for use on files of such a large scale as are needed for the cumulative index. The expansion of vocabulary could also be accommodated, using this technique, by deleting entries and inserting new input phrases using revised terminology. A modification to the program to bring about both these improvements could be introduced, and would provide a machine-stored authority file. But in either case a great deal of work is involved.

Before consideration of the computer output, it is relevant to refer briefly to problems arising from unstandardised terminology. This is closely linked to the problems of vocabulary control discussed above, and is a matter of such general concern that it is perhaps an impertinence for an indexer to animadvert at all on such a topic. However, the indexer of a periodical like the *Cartographic Journal* has the task of bringing to an integrated whole a précis of the thoughts and concepts of many different scholars who write using terms acceptable to themselves but not necessarily identical with those used by others to describe identical or very cognate matters. A very few examples from the recent work will illustrate the difficulty:

<table>
<thead>
<tr>
<th>Air photography</th>
<th>Aerial photography</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cartographic bibliography</td>
<td>Cartobibliography</td>
</tr>
<tr>
<td>Geologic maps</td>
<td>Geological maps</td>
</tr>
<tr>
<td>Line features</td>
<td>Lineal features</td>
</tr>
<tr>
<td>Topographic maps</td>
<td>Topographical maps</td>
</tr>
</tbody>
</table>

The list can be extended almost indefinitely. The position for the manual indexer is complex, but using the computer program the difficulties are highlighted by the problems of tracing input material and establishing control of vocabulary and punctuation.

B. Output.

It is a generally accepted principle in orthodox indexing, endorsed by the British Standards Institution pamphlet on preparation of indexes\(^{(1)}\), that alphabetisation should be either letter-by-letter or word-by-word, e.g.:

- **Letter-by-letter**
  - Newcastle
  - New Orleans
  - New York

- **Word-by-Word**
  - New Orleans
  - New York
  - Newcastle

Which is adopted depends on the character of the material and the type of index desired. On the whole, word-by-word is suited to the material derived from the *Journal*. But in the computer-generated index there are complications arising from the hierarchical nature of the six selection rules which are intrinsic to the program, and the additional fact that in the operation of these selection rules the function word ('of', 'in', 'for', etc.) is subordinated to the first component of the modifying phrase which follows that function word. The application of these rules affects the output in three ways:

1. given the following input phrases:
   - use of *<marginalia>* in *<dating>* *<ordnance survey maps>*
   - use of *<representation>* of *<railways>* in *<dating>* *<ordnance survey maps>*
   - use of *<price>* in *<dating>* *<ordnance survey maps>*
   - use of *<watermarks>* in *<dating>* *<ordnance survey maps>*
In fact, however, the order which appears in the output is 4, 3, 1, 2. This is because at present the modification sort key covers only the first two subject word groups (and related function words). However, a modification could be introduced to make the alphabetisation word-by-word, and to extend it over all the word groups in each input phrase.

3 A more complex problem arises on the selection of material which is common to several entries when the program can and should ideally bring it forward automatically as a sub-heading, and in many situations this is carried out successfully. But given the following input phrases:

- advantages, disadvantages of <mapping> by <computer>
- principles of <mapping> by <computer>
- calcomp plotter for <mapping> by <computer>
- census data in <mapping> by <computer>

the sub-headings appear as desired under ‘computer’:

Computer
- mapping by, advantages, disadvantages of calcomp plotter for
- census data in choice of scale in
- principles of

But under ‘mapping’, where ‘by computer’ is common to all five phrases the output is:

Mapping
- advantages, disadvantages of, . . . by computer
- by computer
- calcomp plotter for
- census data in
- principles of, . . . by computer

The reason for this is that the rule requiring the program to select common material is the fifth of the six rules. If any of the rules 1-4 have come into operation the remainder are not checked. This gives inconsistent and unsatisfactory results. With further work, however, the sophistication of the algorithm which is basic to the program could be increased to ensure that words such as ‘advantages’, ‘principles’ did not become the first elements of a modifying phrase, and so would be disregarded in the selection of common material.

It may be asked why the input cannot be prepared using multi-case fonts instead of the upper case only which was used in Nottingham and Sheffield. In fact the machinery available for these trials did not make this possible; it would in theory have been possible to punch the input on tape which would have provided an output with built-in sym-
bols indicating to the printer when to use capitals and when to use lower case. But punching equipment of this kind was not available. Even if it had been, a higher level of interpretation would have been required than most printers are able to give at commercially reasonable rates.

Summary

To sum up, the work at Sheffield, and the trials of the program on the index for the Journal must be regarded as academic research of which the index is in a sense a by-product, though it is hoped a very useful one. In preparing indexes for small and homogeneous batches of material which can be produced over a short time only it is simple to use and yields a reasonably satisfactory output. For large-scale indexes, particularly those dealing with very diverse material, derived over a long period of time, experience with the material from the Journal has shown that at present it needs a cumbersome system of controls, and at the preparation of the input stage it certainly did not save any indexing time; rather the reverse, because in practice the time spent on setting up controls, editing, searching, and checking for inconsistencies in content and punctuation was at least as great as that needed to write, file, and edit a manually prepared index. Even the punching of cards took longer than the preparation of a typescript because mistakes took longer to rectify.

Just as a manually produced articulated subject index, using multiple indentation and sub-headings is a much more complex structure, requiring a great deal more indexing expertise, than is an alphabetical list of paper titles, so the program is a very sophisticated concept of how to use automated techniques for data retrieval in depth. Its development is a considerable tour de force, and with the various modifications which the experience with the large samples from the Journal, inter alia, has revealed to be desirable, and some which as indicated could be introduced, most of the remaining problems could be removed. Indexing time would then be much reduced, and vocabulary control would be much simplified. Successful operation will continue to depend on the exercise of indexing expertise to bring the end product closer to the standard set by the linguistically polished and conceptually accurate excellence of the best manually produced indexes. The skills of the indexer should be willingly given in support of further research.

Acknowledgements

As before I am indebted to the Council of the Cartographic Society for its encouragement and support for this experimental work. Dr. Michael Lynch, Janet Armitage, J. Howard Petrie and Michael Snell of the Post-Graduate School of Librarianship at Sheffield have co-operated and encouraged with unfailing courtesy, though they would wish their thanks to the Council to be recorded as well, for allowing the program to be tried on the production of a cumulative index for the Cartographic Journal. It should be recorded here that the grant which supported their programme of research was made by the Office for Scientific and Technical Information. The thanks of the Society are also due to Dr. Eric Foxley of the Cripps Computing Laboratory at Nottingham University for giving access to the punching machinery which was used throughout.

Reference

1 B.S.I. publication Nos. 1749, Alphabetical Arrangement, and 3700 Preparation of Indexes.

The Indexer in 1886

‘Did any of my readers ever busy him- or herself with index-making? There requires one or two pressures to persevere in it: enthusiasm of a high order, which is unflagging, or the necessity of earning daily bread.’ —From Canadian archaeology, an essay by William Kingsford (Montreal, Wm. Drysdale & Co., 1886, p. 14).

Kingsford (1819-98) had a long and distinguished career as an engineer in Canada, the United States and Italy before turning his hand to historical and bibliographical studies.