PREPARATION OF INDEXES

A S Raizada
A M Arora
D Elhence
Insdoc, New Delhi-12

Indexing is an essential activity in any information and documentation service. Computerised indexing yields economical, faster and accurate results. Enumerates the areas of further work in computerised indexing. Presents cost analysis for preparing monthly and yearly indexes to Indian Science Abstracts.

INTRODUCTION

Considerable literature exists on indexes and the art of indexing. At least one periodical is completely devoted to indexing. In the field of documentation and information work, indexing forms a major operation. It is an essential and costliest bottleneck in these services. Even traditional cataloguing work in a library may be viewed as an activity resulting in a series of indexes. Consequently, indexing, with the aid of computers drew considerable attention particularly in America and a substantial portion of literature on computerised documentation pertains to it. The terms like KWIC, KWOC, KWAC, etc. are common terms these days not only with librarians and documentalists but also with the community of the users of information service. Therefore, it is only natural that indexing should occupy a high priority in the study programme of Insdoc on computerization of information processing.

INDEX

One of the most difficult and essential tasks is that of defining the activity or the concepts in any field of enquiry. It is difficult because a concept or an activity can be defined in more than one way and each definition is equally true, with the result that no single definition is all embracing. The difficulty lies in providing a comprehensive set of definitions. For the present purpose, an Index may be considered to be an ordered set of items of information extracted from a larger set of items of information, having a different order, with a view to provide direct access to information in the larger set. In other words, an index is an information store pulled out of bigger information store, the arrangement of information items in two stores being different. An Index provides direct access to an item of information without the complete scanning of the information store such as a document, a book, a periodical or a library. Thus an index has three basic components viz. 'information' item', 'arrangement', and 'facility for direct access' which may be called address or location.

INDEXING

In this sense, indexing is a process of extracting and selecting items of information from the information store, ascertaining and fitting the address or location of the item in the store and arranging the selected items along with their addresses in a desired sequence. Hence, the process of indexing can be viewed to consist of the following operations:

1. Scanning of the information store and the selection of information items or index terms;
2. Ascertaining the address of the index term in the information store and fitting it to the index terms;
3. Sequencing all the selected terms in the desired sequence; and
4. Presenting the sequenced index-terms to suit the mode of communication and dissemination.
PREPARATION OF INDEXES

Thus, basically, indexing implies the building up of an equivalence table.

TYPES OF INDEX

Depending on the agency responsible for the preparation of an index, there are three broad types of indexes viz., manual, computerised, and hybrid. Each of these categories may result in indexes such as author, subject, town, place or geographic indexes. From the sequence point of view, indexes may be grouped as alphabetical, number, or special indexes.

TECHNIQUE OF INDEXING

On account of the selection involved in indexing, there arises the technique and the art of indexing. It comprises the following discrete operations.

1. Reading the document which may be a line, a para, a page, a section, or a chapter of a document.

2. Selecting an index term on some basis. It is here, that human ingenuity, skill, art and expertise is called for mainly on account of vague, undefined, ambiguous rules for the selection of indexing terms;

3. Recording the selected terms on suitable media according to available rules of grammar. These rules too are ambiguous thereby raising the decision problems;

4. Fitting the index term with its address and recording on the same. The address may be the line number, paragraph number, section number, page number or chapter number in a document;

5. Ascertaining that all the documents have been completely scanned;

6. Sequencing these recorded items in the desired sequence which may be alphabetical (for names of persons, places or things) or numerical (for numbers) or special (for class numbers etc.); and

7. Editing the sequenced terms to suit the dissemination requirements like appearance, column width etc. This may involve operations like merging etc.

PRACTICE AND PROBLEMS IN INDEXING

The normal practice in indexing is to select a single word indicating the concept as index term. Frequently, two worded terms are also selected, in this case one of the words is a qualifier or adjectival. Now these words are removed from their sequence of normal occurrence in the text and are put in different sequence in an index to suit a different purpose. In this process, the original semantics of the word are lost. The main problem lies in retaining the original semantics of the word, even in the event of isolation of the term and the changed sequence of presentation. At present, this is being done by retaining the entire context as in KWIC, by providing the context by an artificial number like class number etc., or by reconstructing the context with the use of roles and links. It must be pointed out that none of these is entirely satisfactory.

COMPUTERIZATION

Without tracing the historical course of the application of computers to the preparation of indexes, it may be said that computers are being used in the entire routine, enumerated under technique of indexing, as a whole or in isolation to one or more operations. The result of computerisation is that indexes are available at a faster rate, with greater accuracy and consistency, and with less human tears. These indexes are easier to consult though not upto the mark from classical and traditional viewpoint. The shortcomings of computerised indexes are mainly due to the lack of unambiguous rules of the selection of index terms on the part of humans, and non-availability of text reading machines which increases the conversion load many times if the entire text is to be read and scanned. The prohibitive conversion costs make it essential that only the most significant portion of the text be scanned by the computer. This is normally a title or an abstract. It is seldom the entire text.
Insdoc has computerised the preparation of the Author and the key-word Indexes to ISA (= Indian Science Abstracts). The work on computerised indexes started during 1967 with the second volume of ISA. The preparation of author index was the first to be computerised, followed by the preparation of key-word index. The details of these have been published in the Annals of Library Science and Documentation, 1967. The cumulation of annual indexes has now been undertaken and the Annual index for 1967 has already been compiled. The annual cumulations for vol. 1 and 2, 1965 and 1966 are being undertaken. In this compilation and processing of indexes to ISA the mode has been a hybrid one. This is to say, that the selection of index terms is done manually by expert abstractors. The computer is used for alphabetisation, and editing the alphabetised index for the purpose of printing the index. In this manner, judicious selection by human agency is retained. The computer is used in doing away with the labour of alphabetisation, decreasing the preparation time of the indexes and increasing the accuracy by minimising typographical errors. Further, the computerised output saves considerable time at the column, page and board layout stages and practically no further corrections are needed. The computerised preparation of monthly indexes takes less than six days to prepare the final press copy and costs about Rs. 400/- per issue. The annual index costs about Rs. 1500/- and takes less than six hours on IBM 1620 computer.

It has not been possible to eliminate the defects from these indexes. Mostly the defects being the inadvertent inconsistencies at the selection stage. These can be minimised by formulating proper and unambiguous codes for the selection of index terms and a strict adherence to the code by the indexer.

CONCLUSION

At least one area where the computerization can be effectively used is the indexing of secondary periodicals. It is economical, fast and accurate. The experience gained at Insdoc during the past two years demands further work in the following areas:

a) Identification and formulation of the rules governing the selection of the index terms from the view point of the purpose of the index;

b) Syntactical modifications necessary in the index terms in order to minimise the scattering in the index;

c) Formalisation of the above set of rules;

d) Techniques of keeping the semantics of the index-term invariant in the sequence transformations; and

e) Developing ready made programme-package for 'off-line' indirectly connected computer facilities available to library and information centres in India.

ACKNOWLEDGEMENT

The authors are grateful to Shri B.S. Kesavan, Director, Insdoc, for his encouragement in computerising the ISA indexes and providing this opportunity for the exchange of thoughts. They are thankful to their ISA colleagues for their frequent discussions and useful suggestions which had gone a long way in the task of preparing these Indexes with the aid of computers.