Indexing Principles
UNISIST

INDEXING PRINCIPLES

Paris, September 1975

First draft

Photographic reproduction of the original manuscript
1. Object of the present document

2. The operation and purpose of indexing

3. First stage of indexing: establishing the subject
   3.1. Understanding the document
      3.1.1. Written documents
      3.1.2. Non-written documents
   3.2. Identification of concepts
   3.3. Selection of concepts
      3.3.1. Exhaustivity
      3.3.2. Specificity

4. Second stage of indexing: representing concepts by elements of the indexing language

5. Quality control

6. Conclusion

Appendix A: Diagram showing the relationship between indexing and searching
1. Object of the present document.

This paper has as its goal the establishment of valid and consistent principles to be followed when representing the subject of a document. For indexing and retrieval purposes, concepts in documents can be represented either by terms selected from natural language(s), (e.g. keywords), or by symbols (e.g. class marks).

These principles have been conceived so that, as far as possible, they are independent of any particular information system. As such, they constitute a unified set of rules or recommendations which should promote:

- easier cooperation between different information services;
- the development of compatible but more specific rules within the context of a particular information system.

2. The operation and purpose of indexing.

Indexing is regarded as the fact of describing and identifying a document in terms of its subject content. Consequently, the present paper is not concerned with the description of documents as physical entities (e.g. by stating the form, number of pages, etc.), although a statement of these factors by an indexer is necessary if it is considered that this information will enable a user to determine more accurately whether or not a given document would be relevant to his enquiry.

During indexing, concepts are extracted from documents by a process of analysis, then transcribed into the elements of the indexing tools, such as thesauri, classification schemes, etc.

Indexing decisions, concepts are recorded as data elements organised into an easily accessible form for information retrieval. These records can appear in various forms, e.g. back-of-the-book indexes, indexes to catalogues and bibliographies, machine-held files, etc. In using these tools for retrieval (i.e. when identifying a set of documents, or a part of a document, relevant to a given enquiry)
the enquiry itself is treated in a similar fashion - i.e. it is analysed into individual concepts, and these are then translated into the components of the indexing language.

Indexing procedures can be used, on one hand, for organising concepts into tools for information retrieval, and also, by analogy, for analysing and organising enquiries into concepts represented as descriptors or combinations of descriptor-, classification symbols, etc. This close relationship between the indexing of documents and the treatment of enquiries is shown as a diagram in Appendix A.

The general principles set down in this document are equally valid for manual or mechanised systems (or mixtures of each), whether at the indexing or enquiry-answering stage.

Essentially, indexing consists of two stages:
- establishing the concepts expressed in a document, i.e. the subject
- translating these concepts into the components of the indexing language.

3. First stage of indexing: establishing the subject.

Establishing the subject of a document can itself be divided into three stages:

- understanding the overall content of the document, the purpose of the author, etc.

- identifying the concepts which represent this content, purpose, etc.

- selecting the concepts needed for retrieval.

In practice, these three stages tend to overlap.

3.1. Understanding the document.

Full comprehension depends to a certain extent on the form of the document. Two different cases can be distinguished, i.e. written documents and non-written documents.
3.1.1. **Written documents.**

These represent the usual case in libraries and information centres where the stock consists largely of monographs, journals, reports, conference proceedings etc. Ideally, full understanding of these documents depends upon an extensive reading of the text. For economic reasons, however, this is not usually practicable, nor is it always necessary. Nevertheless, the indexer should ensure that no useful information has been overlooked. Important parts of the text need to be considered carefully, particular attention being paid to:

- the title
- the introduction, and the opening phrases of chapters and paragraphs
- illustrations, tables, diagrams and their captions
- the conclusion
- words or groups of words which are underlined or printed in an unusual typeface.

The author's intentions are usually stated in the introductory sections, while the final sections generally state how far these aims were achieved.

All these elements should be scanned by the indexer during his study of the document. Indexing directly from the title is not recommended, and an abstract, if available should not be regarded as a satisfactory substitute for a reading of the text. Titles may be misleading; both titles and abstracts may be inadequate; in many cases, neither is a reliable source of the kind of information required by an indexer.

3.1.2. **Non-written documents**

A different situation is likely to arise in the case of non-written documents, such as audio-visual, visual and sound media. It may not be possible for the indexer to examine these intensively, so that indexing from a title or synopsis then becomes a necessity.
3.2. Identification of concepts

After examining the document, the indexer needs to follow a logical approach in selecting those concepts which best express its subject.

The choice of concepts can be related to a schema of categories recognised as important in the field covered by the document, e.g. phenomena, processes, properties, operations, equipment etc.

"For example, when indexing works on 'Drug therapy', the indexer should check systematically for the presence or the absence of concepts relating to specific diseases, the name and type of drug, route of administration, results obtained and/or side effects, etc. Similarly, documents on the synthesis of chemical compounds should be searched for concepts indicating the manufacturing process, the operating conditions, the products obtained, etc."

3.3. Selection of concepts

The indexer does not necessarily need to retain, as indexing elements, all the concepts identified during the examination of the document. The choice of those concepts which should be selected or rejected depends on the purpose for which the indexing data will be used. Various kinds of purpose can be identified, ranging from the production of printed alphabetical indexes to the mechanised storage of data elements for subsequent retrieval by computer or other means.

The kind of document being indexed may also affect the product. For example, indexing derived directly from the text of books, journal articles, etc. is likely to differ from that derived only from abstracts.

The two characteristics of an index most likely to be affected by these parameters are: - exhaustivity
- specificity
3.3.1. Exhaustivity

An indexer following the procedures outlined above should be able to identify all the concepts in a document which have potential value for the users of an information system. In some cases two or more themes within the field covered by the index occur independently in the same document. These should be treated separately, and if necessary by different subject specialists.

It is important to realise that the breadth of field covered by the index should not be interpreted too narrowly. With the growth of information networks, it may happen that the indexing data created initially for one group of users (e.g. scientists and/or technologists) could usefully be studied by other groups of users (e.g. economists). With this potential use of mind, it is recommended that indexers of, for example, scientific and technical literature, should not overlook other aspects of a subject, e.g. the social and/or economic.

In selecting a concept, the main criterion should always be its potential value as an element in expressing the subject content of the document. In making a choice of concepts, the indexer should constantly bear in mind the questions (as far as these can be known) which may be put to the information system. In effect, this criterion re-states the principal function of indexing. With this in mind, the indexer should:

- choose the concepts which would be regarded as most appropriate by a given community of users.
- if necessary, modify both indexing tools and procedures as a result of feedback from enquiries.

Such modification should, nevertheless, not be taken to a point where indexing is distorted.
There should not be an arbitrary limit to the number of terms or descriptors which can be assigned to a document; this should be determined entirely by the amount of information contained in the document. Any arbitrary limit is likely to lead to some loss of objectivity in the indexing, and to the distortion of information which would be of value during retrieval. If, for economic reasons, the number of terms has to be limited, the selection of concepts should then be guided by the indexer's judgment concerning the relative importance of concepts in expressing the overall subject of the document.

In many cases the indexer needs to include, as part of the indexing data, concepts which are present only by implication, but which serve to set a given concept into an appropriate context.

3.3.2. Specificity

As a rule, concepts should be identified as specifically as possible. More general concepts may be selected in some circumstances, depending upon the purposes of the information system. In particular, the level of specificity may be affected by the weight attached to a concept by the author. If the indexer considers that an idea is not fully developed, or is referred to only casually by the author, indexing at a more general level may be justified.

4. Second stage of indexing: Representing concepts by elements in the indexing language

To ensure that concepts are organised in a usable and accessible form, full use should be made of indexing tools. The same applies when dealing with enquiries.
Tools used most frequently in indexing fall into two broad categories:
- a "combinatorial" type represented by thesauri, subject heading lists, etc.
- a "categorial" type where concepts are represented by indexes or symbols of a classification.

The indexer should be familiar with these tools and their working rules and procedures. In particular, he should be aware that these tools may impose certain constraints upon recommended practices. For example, a prescribed list of headings, or the schedules of a classification scheme, may not permit the exact representation of a concept encountered in a document.

If the indexing tool is a thesaurus (cf IS 2788) the number of terms assigned to a document, and the multiplication of entries can be reduced without loss, since generic and other \textit{a priori} relations can be established directly from the thesaurus itself. When using a thesaurus, select the most specific descriptor available to represent a given concept (1).

Some systems make use of roles, links, weights, etc. The indexer needs to be familiar with any special rules associated with these systems.

If concepts are represented by classification symbols, it needs to be understood that these marks usually indicate a wider concept (i.e. a main class) which may not be entirely appropriate to the document in hand.

These two kinds of indexing tool (i.e. thesauri and classification schemes) can be used together to allow retrieval via one or the other. Either may prove to be more economical or effective, depending on the nature of the enquiry.

In practice, the indexer will frequently encounter concepts which are not present in an existing thesaurus or classification scheme. Depending on the system in use, these concepts may be entered into the system immediately, or the indexer may have to use more generic descriptors, the new concepts being proposed as candidates for a later edition.

5. Quality control

The quality of indexing depends on two factors:
- the qualifications of the indexer
- the quality of the indexing tools.

For a given information system, the indexing data assigned to a given document should be consistently the same regardless of the individual indexer. It should, furthermore, remain relatively stable throughout the life of a particular indexing system. Consistency to this standard is particularly important if information is to be exchanged between agencies in a documentary network.

An important factor in reaching this level of consistency is complete impartiality in the indexer. Almost inevitably, some elements of subjective judgment will affect indexing performance; these should be minimised as far as possible. Consistency is more difficult to obtain with a large indexing team, or with teams of indexers working in different locations (as in a decentralised system). In these situations, a centralised check stage is advisable.

The indexer should preferably be a specialist in the field covered by the documents he is indexing. He should understand the terms encountered in documents as well as the rules and procedures of the specific indexing system.

Quality control would be achieved more effectively if the indexers also have contact with users. They could then, for example, determine whether certain descriptors produce false combinations, and also create noise at the output stage.
Indexing quality is also dependent upon certain properties of the indexing method or procedure. It is essential that an index should be able to accommodate new developments in terminology, and also new needs of users: that is, it must allow frequent updating.

Indexing quality can be tested by analysing retrieval results, e.g. by calculating recall and precision ratios.

6. Conclusion

These recommendations should permit indexing suited to any normal retrieval system. Every system can, however, be refined further to meet specific needs of its users through the development of more specific rules, provided that these are formulated in the light of these general guidelines.
The diagram illustrates the indexing process and purposes. It starts with documents, which are determined to have concepts. These concepts are then represented. The representation is organized into indexed information. Indexing tools, such as a thesaurus, classification, etc., are used in this process. There are also retrieval tools, such as an index, file, catalogue, which can be used (optional). The process and purposes are connected by arrows, showing the flow of information. The ultimate purpose is to provide answers to requests.