...ternational Standard ISO 5963

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION•МЕЖДУНАРОДНАЯ ОРГАНИЗАЦИЯ ПО СТАНДАРТИЗАЦИИ•ORGANISATION INTERNATIONALE DE NORMALISATION

Documentation — Methods for examining documents, determining their subjects, and selecting indexing terms

Documentation — Méthodes pour l'analyse des documents, la détermination de leur contenu et la sélection des termes d'indexation

First edition — 1985-12-01

UDC 001.815 **Ref. No. ISO 5963-1985 (E)**

Descriptors : documentation, subject indexing.

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work.

Draft International Standards adopted by the technical committees are circulated to the member bodies for approval before their acceptance as International Standards by the ISO Council. They are approved in accordance with ISO procedures requiring at least 75 % approval by the member bodies voting.

International Standard ISO 5963 was prepared by Technical Committee ISO/TC 46, *Documentation*.

Users should note that all International Standards undergo revision from time to time and that any reference made herein to any other International Standard implies its latest edition, unless otherwise stated.

Printed in Switzerland

Contents

Documentation — Methods for examining documents, determining their subjects, and selecting indexing terms

1 Scope and field of application

1.1 This International Standard describes recommended procedures for examining documents, determining their subjects, and selecting appropriate indexing terms. It is restricted to these preliminary stages of indexing, and does not deal with the practices of any particular kind of indexing system, whether pre-coordinated or post-coordinated. It also describes general techniques for document analysis which should apply in all indexing situations. These methods are, however, especially intended for indexing systems in which the subjects of documents are expressed in summary form, and where concepts are recorded in the terms of a controlled indexing language. In this context, a controlled language usually refers to a subset of terms selected from natural language, and regulated, for example, by a thesaurus. These methods would apply, however, to systems in which concepts are represented for retrieval purposes by symbols chosen from the schedules of a classification scheme.

1.2 The techniques described in this International Standard can be employed by any agency in which human indexers analyse the subjects of documents and express these subjects in indexing terms. They do not apply to agencies which employ those automatic indexing techniques in which terms occurring in texts are organized into sets or classes according to criteria which can be established by a computer, for example frequency of occurrence and/or adjacency in the text, although the aims of these systems are the same.

1.3 This International Standard is intended primarily as a guide to indexers during the stages of document analysis and concept identification. It may also be helpful for the analysis of users' enquiries and their translation, for retrieval purposes, into the controlled terms of an indexing language, and it could function as guidance to abstractors during the preparation of abstracts. It should be borne in mind, however, that although these tasks are analogous they are not identical.

1.4 This International Standard is intended to promote standard practice

a) within an agency or network of agencies;

b) between different indexing agencies, especially those which exchange bibliographic records.

2 References

ISO 2788, *Documentation — Guidelines for the establishment and development of monolingual thesauri.*

ISO 5964, *Documentation — Guidelines for the establishment and development of multilingual thesauri.*

3 Definitions

For the purposes of this International Standard, the following definitions apply.

3.1 document: Any item, printed or otherwise, which is amenable to cataloguing or indexing.

NOTE — This definition refers not only to written and printed materials in paper or microform versions (for example books, journals, diagrams, maps), but also to non-print media (for example machine-readable records, films, sound recordings), and three-dimensional objects or realia used as specimens.

3.2 concept: A unit of thought.

The semantic content of a concept can be re-expressed by a combination of other and different concepts, which may vary from one language or culture to another.

3.3 subject: Any concept or combination of concepts representing a theme in a document.

3.4 indexing term: The representation of a concept in the form of either

— a term derived from natural language, preferably a noun or noun phrase, or

— a classification symbol.

NOTE — An indexing term can consist of more than one word. In a controlled indexing language, a term is designated either as a *preferred term* or as a *non-preferred term*.

3.5 preferred term: A term used consistently when indexing to represent a given concept; sometimes known as "descriptor".

3.6 non-preferred term: The synonym or quasi-synonym of a preferred term.

A non-preferred term is not assigned to documents but is provided as an entry point in an index, the user being directed by an instruction (for example USE or SEE) to the preferred term; sometimes known as "non-descriptor".

3.7 index (plural "indexes"): An alphabetical or systematic listing of subjects which refers to the position of each subject in a document or collection of documents.

3.8 indexing: The act of describing or identifying a document in terms of its subject content.

4 Operation and purpose of indexing

4.1 Indexing is not concerned with the description of a document as a physical entity (for example by stating the form, publisher, date, etc.), although these factors may be included in a subject index if this information would enable a user to determine more accurately whether or not a given document is likely to be relevant to his enquiry.

4.2 During indexing, concepts are extracted from documents by a process of intellectual analysis, then transcribed into indexing terms. Both analysis and transcription should be performed with the aid of indexing tools such as thesauri and classification schemes.

4.3 Essentially, indexing consists of the following three stages, although these tend to overlap in practice:

a) examining the document and establishing its subject content;

b) identifying the principal concepts present in the subject;

c) expressing these concepts in the terms of the indexing language.

Each of these stages, together with a section on quality control, is considered below in clauses 5 to 8.

5 Examining the document

5.1 The thoroughness with which a document can be examined depends to a large extent upon its physical form. Two different cases can be distinguished, i.e. printed and non-print documents.

5.2 Printed documents 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 texts. A complete reading is often impracticable, nor is it always necessary, but the indexer should ensure that no useful information has been overlooked. Important parts of the text need to be considered carefully, and particular attention should be paid to the following:

a) the title;

b) the abstract, if provided;

c) the list of contents;

d) the introduction, the opening phrases of chapters and paragraphs, and the conclusion;

e) illustrations, diagrams, tables and their captions;

f) words or groups of words which are underlined or printed in an unusual typeface.

All these elements should be scanned and assessed by the indexer during his study of the document. Indexing from the title alone is not recommended, and an abstract, if available, should not be regarded as a satisfactory substitute for an examination 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 needed by an indexer.

5.3 Non-print documents, such as audio-visual, visual and sound media, including realia, call for different procedures. It is not always possible in practice to examine a record in its entirety (for example by running a film). Indexing is then usually carried out from a title and/or synopsis, though the indexer should be allowed to view or hear a performance of the medium if the written description is inadequate or appears to be inaccurate.

6 Identification of concepts

6.1 After examining the document, the indexer should follow a systematic approach to the identification of those concepts which are essential elements in a description of its subject. Agencies should establish check lists of those factors which are recognized as important in the field covered by the index.

The questions listed below illustrate general factors which such a check-list should establish:

a) Does the document deal with the object affected by the activity?

b) Does the subject contain an active concept (for example an action, an operation, a process, etc.)?

c) Is the object affected by the activity identified?

d) Does the document deal with the agent of this action?

e) Does it refer to particular means for accomplishing the action (for example special instruments, techniques or methods)?

f) Were these factors considered in the context of a particular location or environment?

g) Are any dependent or independent variables identified?

h) Was the subject considered from a special viewpoint not normally associated with that field of study (for example a sociological study of religion)?

These are offered as examples of general factors which are likely to apply in any subject field. Other questions may need to be formulated within a special discipline.

6.2 The indexer does not necessarily need to represent, as indexing terms, 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 terms will be used. Various kinds of purpose can be identified, ranging from the production of printed alphabetical indexes to the mechanized storage of data elements for subsequent retrieval by computer or other means. The identification of concepts may also be affected (as noted above) by the item being indexed. For example, indexing derived from the texts of books, journal articles, etc. is likely to differ from that derived from abstracts or synopses. The two characteristics of an index most likely to be affected by these matters are exhaustivity, and specificity.

6.3 Exhaustivity refers to the number of factors (such as those associated with the questions in 6.1) which are represented by the terms assigned to a document by the indexer.

6.3.1 An indexer who follows 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 an index occur independently in the same document. These should be treated separately, and if necessary by different subject specialists.

6.3.2 The breadth of interest covered by an index should not be interpreted too narrowly. With the growth of information networks it needs to be borne in mind that the indexing data created initially for one group of users (for example scientists or technologists) could usefully be studied by other groups of users (for example economists). With this potential use in mind, it is recommended that indexers of scientific and technical literature, for example, should not overlook other facets of a subject, for example its social or economic aspects.

6.3.3 In selecting concepts, the main criterion should always be the potential value of a concept as an element in the expression of the subject of the document and in its retrieval. In making the choice of concepts, the indexer should 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. Within this context, the indexer should:

a) choose the concepts which would be regarded as most appropriate by a given community of users, bearing in mind the purpose of the index,

b) if necessary, modify both indexing tools and procedures as a result of feedback from enquiries. Such modification should not be taken to a point where the structure or logic of the indexing language is distorted.

6.3.4 No arbitrary limit should be set 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, related to the expected needs of the users of the index. The imposition of an arbitrary limit is likely to lead to some loss of objectivity in indexing, and to the distortion of information which would be of value during retrieval. If it is necessary within a given agency to limit the number of terms, the selection of concepts should be guided by the indexer's judgement concerning the role of each concept in expressing the overall subject of the document.

6.4 Specificity refers to the extent to which a particular concept which occurs in a document is specified exactly in the indexing language. Loss of specificity occurs when a particular concept is represented by a term with more general meaning.

Concepts should be identified as specifically as possible. More general concepts may be preferred in some circumstances, depending upon the following factors:

a) the extent to which the indexer considers that over-specificity might adversely affect the performance of the indexing system. An indexer may decide, for example, that very specific models of equipment may be represented by more general terms such as the name of the maker and perhaps of the family of models, especially when these concepts occur only in the fringe areas of the subject field covered by the index.

b) 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.

7 Selection of indexing terms

7.1 When concepts are being translated into indexing terms, the indexer should observe the following practices (see also the annex):

a) Concepts which are already represented in the indexing language should be translated into their preferred terms.

b) Terms which represent new concepts should be checked for accuracy and acceptability in reference tools such as the following:

— dictionaries and encyclopaedias recognized as authorities in their fields;

— thesauri, especially those constructed in accordance with ISO 2788 or ISO 5964;

— classification schemes.

Subject specialists, especially those with some knowledge of indexing or documentation, may also be consulted.

7.2 The indexer should be familiar with these tools and their working rules and procedures. In particular, he or she should be aware that these tools may impose certain constraints. For example, a prescribed list of subject headings, or the schedules of a classification scheme, may not permit the exact representation of a concept encountered in a document. If concepts are represented by classification symbols, it needs to be understood that these marks usually indicate a wider or a narrower context (i.e. a main class) which may not be entirely appropriate for the document in hand.

7.3 If an indexing language incorporates a thesaurus the number of terms assigned to the document, and the multiplicity of entries, can be reduced without loss, since generic and other *a priori* relationships can be established directly from the thesaurus itself. When using a thesaurus, the most specific term available should be selected to represent a given concept.

7.4 Some indexing systems employ roles, links, weights, etc. The indexer should be familiar with any special rules associated with these mechanisms.

7.5 In practice, the indexer will frequently encounter concepts which are not present in an existing thesaurus or classification scheme. Depending upon the system in use, these concepts may be handled in various ways, for example

a) expressed by terms or descriptors which are admitted into the indexing language immediately;

b) represented temporarily by more general terms, the new concepts being proposed as candidates for later addition.

8 Quality control

8.1 The quality and consistency of indexing depend upon factors such as

a) the qualifications and expertise of the indexer;

b) the quality of the indexing tools.

In an ideal situation, the indexing terms assigned to a document and the level of exhaustivity attained during indexing should be consistently the same regardless of the indexer employed. These factors should, furthermore, remain relatively stable throughout the lifetime of a particular indexing system. It is not always possible to achieve this standard of consistency in practice, but the goal of consistency, and hence predictability, is an important factor in the performance of an indexing system, especially when information is exchanged between different agencies in a network.

8.2 Complete impartiality on the part of the indexer is a necessary factor in achieving indexing consistency. Subjective judgement in the identification of concepts and the choice of indexing terms will inevitably affect the performance of the indexing system. Consistency is more difficult to achieve within a large indexing team, or when indexing is performed by teams of indexers working in different locations, as in a decentralized system. In these situations, a centralized checking stage, with feedback to indexers, is recommended.

8.3 The indexer should have adequate knowledge of 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 language.

Agencies handling documents in foreign languages should have recourse to language specialists.

8.4 Quality of indexing can be achieved more effectively if indexers also have direct contact with users. They could then, for example, determine whether certain terms or descriptors are likely to produce false combinations and so create irrelevant output.

8.5 The quality of indexing also depends upon the hospitality of the indexing language employed. This should freely admit new terms or changes in terminology, and also respond to new needs of its users. A policy of frequent updating is regarded as essential.

8.6 Where possible, indexing quality should be tested by analysing retrieval results, for example by calculating recall and precision ratios.

Annex

Flowchart of the indexing operation using a thesaurus

(This annex does not form part of the standard.)