The UNISIST Draft on Indexing Principles. Text and Comments

(*Editor's Note*: The Unesco/UNISIST document entitled "Indexing Principles" (SC. 75/WS/58, Paris, Sept. 1975) is reprinted below with the permission of the Division of Scientific and Technological Documentation and Information. We gratefully acknowledge this permission. It should be mentioned that this document is only a first draft. Since these "Principles" are still being discussed we consider it timely to ask for comments and to invite our readers to pay attention to these activities.)

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 independant of any particular information system. As such, they consitute 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 inquiry.

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.

In 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 inquiry) the inquiry 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 inquiries into concepts represented as descriptors or combinations of descriptors, classification symbols, etc. This close relationship between the indexing of documents and the treatment of inquiries is shown as a diagram in Fig. 1.

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 inquiry-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 subseqent retrieval by computer or other menas.

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 the 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 in 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 inquiries.

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 inquiries.

Tools used most frequently in indexing all 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 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 inquiry.

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 accomodate 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.

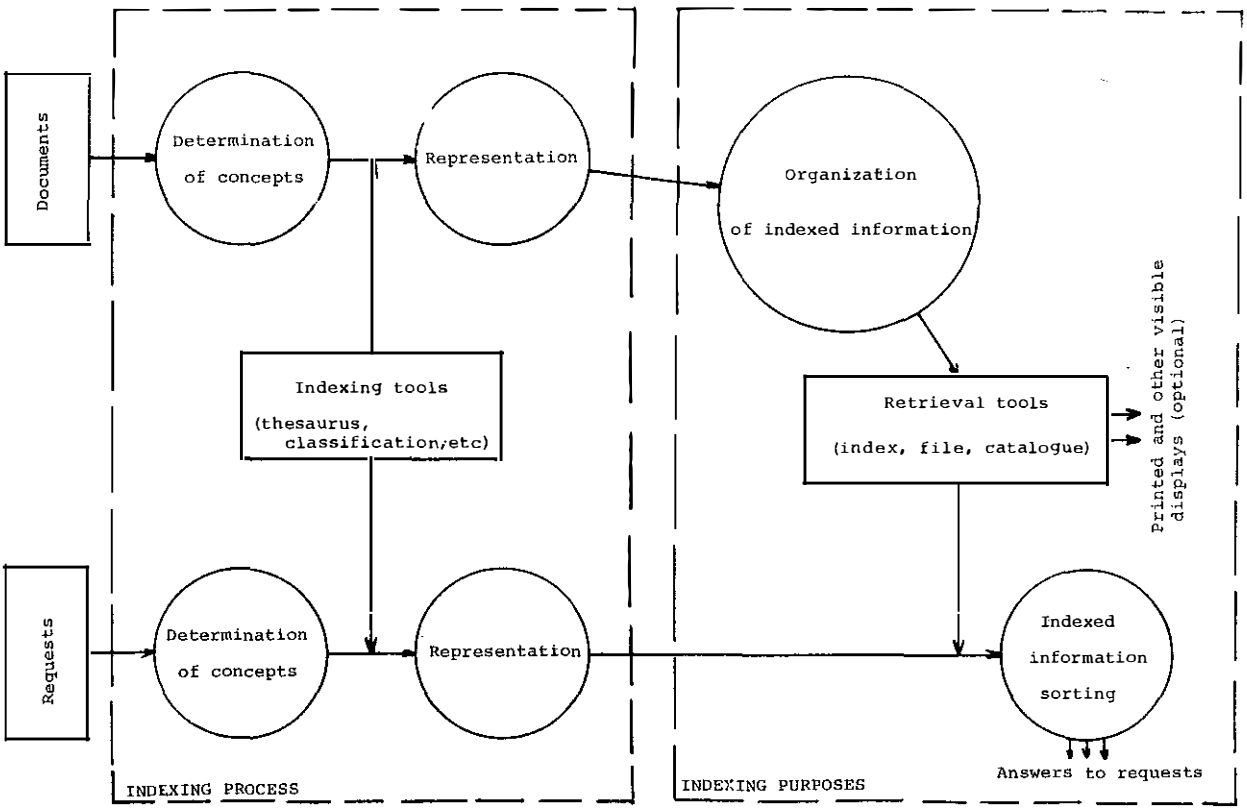

Fig. 1: Relationship between the indexing of documents and the treatment of inquiries

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.

(1) Wellisch, H.: A flow chart for indexing with a thesaurus. In: J. Amer. Soc. Inform. Sci. 24 (1973) p. 185–194.

Comments by A. G. Brown, London

The aim of this paper as stated in its first sentence "the establishment of valid and consistent principles to be followed when representing the subject of a document" is certainly a worthwhile objective. There is a place for a guide-line clearly stating its principles of subject analysis without reference to any particular language or system.

Although the present paper certainly goes some way to providing such, I do not think it goes far enough. The basic weakness seems to be that those responsible for drafting the document have not defined their target population with enough precision.

If they wish to state principles for the benefit of persons with little or no experience in indexing, then they need to define their terms more clearly. At a very basic level, for instance, I think it would be wise to say that indexing here means *subject* indexing. More importantly, that the term is in reality restricted to *co-ordinate* subject indexing. There is no mention of pre-coordinate indexing – although classification schemes are brought in and these will often be used in the context of pre-coordinate indexing.

More specifically, the concepts of exhaustivity and specificity are introduced but are not defined clearly enough for persons inexperienced in the field. There is, for example, no reference to degrees of exhaustivity from summarization upwards (or downwards?). Again the role of indexing languages, their nature and relationship to subject analysis merits closer attention – e.g. they are grouped into "combinatorial" and "categorial" types but we are not told what these terms *mean.*

In short, the aim of the document is most worthwhile. It does cover many relevant concepts. It does not, however, define these precisely enough or set them in a broad enough context – and this *could* be achieved without making the document unwieldy and thus losing the merit of conciseness. Consequently, I feel that it is in danger of falling between two stools – it does not provide sufficient guidance for the beginner yet it is at a level which renders it largely redundant for the more experienced indexer.

Comments by R. Fugmann, Frankfurt

The UNISIST Indexing Principles aim at providing suggestions that are intended to be valid for the indexing of all types of documents. Indexing includes any activity in which essential subject concepts in a document are transformed into a usable and accessible form (para 4), which also includes the allocation of a classification notation. The guidelines are also to be valid for the different technical aids such as printed indices, catalogues, and computer data bases of all types (para 2). This also includes those techniques in which indexing leads to an enrichment of the indexing language vocabulary, which includes also the various techniques of thesaurus formation and thesaurus development. This involves a further field of information science and if guidelines are to be laid down for this in 10 typed pages, drastic simplification becomes necessary. The question arises as to how far such a method of presentation might lead to oversimplification of the indexing problem.

A large part of the problem has been excluded from the guidelines by referring to the UNISIST Guidelines for the Establishment of Monolingual Thesauri. The Indexing Principles are in effect largely based on the ideas as laid down in these guidelines as well as, if only indirectly, in the UNISIST Guidelines for Multilingual Thesauri. It becomes obvious that the main weaknesses of the Indexing Principles are not really inherent, but rather caused by the thesaurus guidelines on which the Indexing Principles are based.

Consistency vs. Predictability

The meaning of any indexing is, for instance, as unclear in the Thesaurus Guidelines as in the Indexing Principles, if the aim is solely to present the concepts in a „usable and accessible form" (para 4). To what extent does the natural language text not constitute such a form and to what extent is translation into an indexing language really necessary? Had this inaccuracy not led to faulty practical instructions, it could be conceded that it could not be the task of a brief instruction to deal with these thoughts. Further, however, (para 5) the quality of indexing is primarily judged according to the *consistency* of the work. If consistency is understood to be the fact that the *one and same document* is given the same set of indexing terms, if indexed by different indexers (or by the same indexer at different times), so is this only an illusory aim of indexing, of interest, if at all, in laboratory tests only. In practice, it is only of secondary interest. For retrieval, for which any indexing is, after all, only a tool, it is of primary importance that it is possible to reconstruct or *predict* how the *one and same concept* would have been expressed in the various stored documents. All these variations occurring in the store have to be known in *advance* on formulation of the search parameters (at least as alternatives) and have to be taken into consideration. If this was only a question of consistency, it could be easily achieved. Automatic indexing would, for instance, certainly achieve the same indexing for the same document with the simplest variants resulting in high consistency. The aim of an indexing language could much more appropriately be described with the achievement of *predictability* and this has already been variously suggested.

The recommendation to the indexer to aim at selecting the most appropriate descriptors from the existing indexing language vocabulary (para 4) gains in importance from this point of view. If this is not done, predictability is reduced at the time of retrieval. It is no longer predictable with certainty with which different,

less fitting, descriptors the concept in question could have been indexed in the store. The careful inclusion of those less appropriate descriptors leads inevitably to a distortion of question formulation and thus necessarily to faulty retrieval.

It would have been the task of the Thesaurus Guidelines, mentioned in this connection, to explain the scope and relational structure a thesaurus should provide (and maintain) in order to be reliable under everyday working conditions and to facilitate the finding of the optimum descriptors contained in the thesaurus. If every important technical term is, for instance, included in the thesaurus without any conceptual analysis, this would, without doubt, lead, in the course of time, to a vocabulary which can no longer be used in the recommended sense. The Guidelines for multilingual thesauri even expressly recommend the abandonment of any conceptual analysis.

The experienced indexer and thesaurus compiler knows that the search for the most appropriate *descriptors* in a vocabulary can rapidly develop into the same major problem which is already well known in searching for relevant *documents*, even if at a different level. However, these indexing principles are primarily intended for the less experienced who may have the impression that this search process for the most appropriate descriptors in the thesaurus does not cause any significant problems. The particular danger is not pointed out which is inherent in the continuous inclusion of new descriptors for the high-grade compound concepts.

Syntax

A further noticeable defect, also caused by the defects of the guidelines for mono- and multi-lingual thesauri, lies in the fact that for an indexing language, only one of its two constituents is made use of, namely the vocabulary. The possibilities and the usefulness of an indexing language *syntax* are only mentioned peripherally and that only through two relatively ineffective tools – roles and links. Considerably more useful and more easily manageable syntactic techniques, such as the bringing together of connected concepts in separate sub-units of a document following defined rules (a technique widely used in punched cards and particularly in computerised systems), have not been considered at all.

When the obvious suggestion is made to make full use of the available indexing tools (para 4) this would also imply the possible use of an indexing language syntax available in mechanised systems. This would also affect favourably the clearness and manageability of an indexing language vocabulary. It could then be relieved of many high-grade pre-coordinated terms without adversely affecting the fidelity of representation that is possible in a particular indexing language. If no mention is made of the possible use of a syntactic tool, and of its favourable effects on scope and structure of a vocabulary, the less experienced would not be aware of the possibilities for making his work easier (and at the same time, improve the quality of his indexing).

Desirable additions to the Indexing Principles

In so far as the deeper problems of indexing can be touched upon at all in such brief guidelines, relatively little can be added. If, for instance, optimum specificity in indexing is recommended, this should be qualified for the sake of clarity: there is no sense in indexing with optimum specificity *every aspect* in each document falling into the scope of an indexing system. Concepts from borderline subjects should only be indexed in a general form or not at all. For instance, a chemical publication or patent often contains specialised terms from physics, engineering, biology, medicine, or statistics. To understand these terms and to represent them adequately would be time-consuming and require much concentration of the chemist and indexing costs would be correspondingly high. This effort would not be worthwhile as hardly any physicist, medical man or mathematician would bother to interrogate a chemical literature data base. One must, therefore, forego highly specific indexing of such terms from borderline subjects in the interest of economy of the indexing job as a whole.

To enumerate the positive aspects of "Indexing Principles" would take at least as much space and the critic has to desist, partly also because these are suggestions which have been current for cataloguing in librarianship for decades and are, therefore, logically included in indexing principles, covering such a wide area. It must, nevertheless, be appreciated that the indexing principles are of relatively general nature and purposely avoid recommendations which would only be valid for specialised subjects and applications.

Conclusion

Many inter-relationships which are significant for the compilation and practical application of an indexing system could not be mentioned in such a short guideline. This, for instance, applies to the effects of store size, frequency of interrogation, available technical tools, type of interrogation concept (individual or general concepts), tolerability of loss and/or irrelevance and many other factors of importance in devising and applying an indexing system. The awareness of these interrelationships, which can only be taught by a textbook, is, however, of considerable importance to a kind of indexing, on which exacting demands are made over a period of time. If the existence of these inter-relationships is not mentioned in the indexing guide, many an indexing project would get on the wrong track from which it could only be retrieved by great efforts and delays. Corresponding explanations should at least be indicated in the Guidelines. A thorough revision of the Guidelines for compilation of thesauri, on which the indexing guidelines are based, appears to be particularly urgent. The unfavourable effects of these deficient guidelines are already widely felt.

Comments by E. Svenonius, London, Ontario

The UNISIST Indexing Principles (Draft 1975) are designed to be "equally valid for manual or mechanized systems (or mixtures of each), whether at the indexing or inquiry-answering stage". The attempt to establish a set of indexing principles that is universally valid is commendable as it is difficult. The attempt must be made

because agreement about what is involved in indexing is a precondition of any cooperative ventures which hope to share the results of indexing. The difficulty comes to the extent that different indexing systems are constructed on the basis of different principles and good indexing performance is equated with following the rules of whatever special system is being used.

The authors of the UNISIST Indexing Principles (Draft 1975) are to be congratulated in succeeding as far as they do in abstracting common principles from particular indexing systems. Probably there would be little disagreement about the definition of indexing: "describing and identifying a document in terms of its subject content," or the description of the process of indexing as a two-stage operation: (1) "establishing the concepts expressed in a document, i.e. the subject"; and (2) "translating these concepts into the components of the indexing language". However, when the indexing operations are specified in more detail, there is a risk that some universality is sacrificed. For instance, the operation of establishing the subject is analyzed into three suboperations, the first of which is "understanding the overall content of the decument, the purpose of the author, etc.". In the case of machine or machine-aided indexing it is only in some euphemistic sense that it can be said a computer understands concepts; when the understanding operation is specified still further to enjoin the indexer to read the full text of the article, do not depend on titles and abstracts alone, here it seems the principles are becoming even more particular, lacking application in some manual and most machine indexing contexts. At this point the question occurs to me, might it not be useful of the UNISIST principles to exclude automatic indexing systems?

It is an interesting exercise to look at the UNISIST Principles from an American point of view, through the eyes of Cutter and his successors. A principle those early American theorists might have felt lacking in the UNISIST Draft is one expressing the overall purpose of an index. Cutter first introduced such a principle as one of the objects of his subject catalog. Later Cutter's formulation came to be known as the Unity principle. According to this principle the purpose of subject indexing is to bring together all documents dealing with the same subject. To an extent this Unity or Purpose principle is covered by the UNISIST injunction to Consistency. But another injunction is implicit here as well, namely that there should be provision for linking synonymous and otherwise semantically related terms. In other words, there should be some measure of vocabulary control. – Of course it can be questioned whether a Principle, like the Unity Principle, has universal applicability.

Another principle which has coloured the development of subject indexing in the United States is a very fundamental one which makes the User the focus of indexing practice. In the UNISIST principles there is a statement that the selection of concepts depends on the purpose for which the indexing data will be used. Obliquely this makes reference to the user. But one might go further to say that also at the second stage of indexing, when the concepts selected for indexing are tooled into a particular indexing language, purpose should be considered. More particularly, every attempt should be made to select terminology likely to be used by the class of reader for whom the indexing is done.

There are two places where I would like to take issue with the Principles – and here my views are personal rather than nationalistic. In the first place I would object to the paragraph which warns against arbitrarily limiting the depth of indexing. The word "arbitrarily" no doubt is used for protection which makes my objection only a quibble. I would like to make it, however. I do not think we know yet what constitutes optimal indexing depth. Recent research suggests that optimal indexing depth may be a function not only of the number of concepts in a document to be indexed but also a function of the purpose of indexing (is good precision or good recall a system specification?). It may also be a function of collection parameters such as the number of documents in the system and the size and distribution of the indexing vocabulary. Finally it may be a function of user query behavior.

The second place where I would like to take issue is with the statement which says that the quality of indexing depends on two factors: the qualifications of the indexer and the quality of the indexing tools. I would like to suggest a third factor affecting indexing quality: the indexing language. It is important to distinguish between an indexing tool, such as a thesaurus or a classification scheme, and an indexing language, with its rules of syntax, semantics and pragmatics. Today we are seeing the mushrooming of a variety of different indexing languages, each with its own semantics, syntax and pragmatics. The rules for constructing a PRECIS string are different from those for putting together a Library of Congress Subject Heading. Independently of the thesaurus used, it is possible to ask which indexing produces the better quality indexing.