



Marlene Burger and Ina Fourie

Abstract

In this brief chapter we want to make you aware of the importance of formulating a policy for the verbal organisation of information. The focus is on indexing and abstracting.

Introduction

To ensure effective retrieval of information, it is essential to formulate a policy for the information organisation and retrieval system (IOAR). If you plan and operate an IOAR system, the onus is on you to determine the policy. On the other hand, if you make contributions to an IOAR system (e.g. a national database), you will have to understand and apply the policy laid down by others. A policy is essential; otherwise there will be chaos in the organisation and retrieval of information. The policy must be based on the purpose of the IOAR system and on approved standards.

An indexing policy lays down guidelines for how indexers should study, analyse and describe the content of entities by allocating indexing terms. An abstracting policy gives guidelines on how abstractors should study, analyse and describe the content of entities as a shortened, accurate synopsis of the original entity. The following is an example of what to consider when a policy for an IOAR system is to be formulated (note the place of indexing and abstracting):

Coverage	What type of entity will be included (e.g. books, journal articles, photos, images, maps). What subjects will be covered?
Extent	What is the expected number of entities to be covered in the IOAR system? (This will influence the choice of vocabulary.)
Users	Who will use the IOAR system (e.g. lay-people, specialists, scientists, academics, schoolchildren)?
Available technology	Will the IOAR be a manual system or a computer system, or will it be run via the Internet? Which hardware (e.g. computer equipment) and which software (e.g. database program) is available for the design of the system? To what degree can technology be used to support the organisation of information (e.g. computer-aided indexing)?
Forms of information organisation	Which form of bibliographic description will be used (e.g. AACR2, Dublin Core)? Which form(s) of verbal subject description will be used (e.g. indexing, abstracting, subject headings)? For each of these, a specific policy is necessary.

Which indexing language will be used?	Will natural language or a controlled vocabulary be used? What will be used in the case of controlled vocabulary (e.g. a thesaurus, a list of subject headings)?
How will the descriptions (indexing and abstracting) be done?	Will the indexing and abstracting be done manually or will an automatic method be used?

The above forms the background against which an indexing or abstracting policy should be formulated. Such policies must take the purpose of the IOAR system into consideration, since they are going to operate as part of the system.

Indexing Policy

An indexing policy is important to the formulation of guidelines on how indexers should study, analyse and describe entities. Soergel (1985:329) contends that knowledge of the indexing policy is important to the assigning of indexing terms and for retrieving information: ‘to use knowledge of indexing criteria for writing query formulations that exploit all the possibilities offered by the indexing and also to compensate, to the extent possible, for any weaknesses in indexing’. An indexing policy pays attention to the following:

- the type of information that is important and should be highlighted (e.g. in a database about nuclear power, any reference to techniques that uses radio-isotopes is very important, even if there is only limited information about it)
- the type of information that should not be covered (e.g. historical information or information that is generally known)
- the specificity of indexing (i.e. assigning specific or general indexing terms)
- the exhaustivity of indexing (i.e. the depth of indexing)
- the point of departure for indexing: user needs (thus the intellectual content of entities) (see chapter 3 for more information)

An indexing policy is important to setting up a standard of indexing and to ensuring that indexing terms are assigned consistently. The following standards can be used as a basis for a policy:

BS ISO 999: 1996. *Information and documentation: guidelines for the content, organization and presentation of indexes*. London: British Standards Institution.

BS 6529: 1964. *Recommendations for examining documents, determining their subjects and selecting indexing terms*. 1984. London: British Standards Institution.

The specificity of indexing terms and the exhaustivity of indexing are two important issues for an indexing policy when it comes to users’ needs. In the following two sections these concepts are discussed in more detail.

Specific versus General Terms

Specificity is the generic level on which the concepts assigned to the entity are expressed. It refers to the degree to which the indexing terms assigned correspond with the precise meaning of the concepts described by the entity. If the indexing terms correspond with the concepts described, one speaks of *specific indexing*. If more general terms are used to describe the concepts, they are referred to as *general indexing terms*.

If very specific indexing terms are used, users are in a position to look for information under specific terms. Users do not then have to look through other information concerned with more general terms. If they require information about African elephants, they do not need to go through all the entities containing information about elephants in order to select specifically those containing information about African elephants.

If natural language is used for indexing, the indexing terms can be as specific as the terms found in the entity. If a controlled vocabulary (e.g. a thesaurus) is used, the specificity of the indexing terms depends on the specificity of the thesaurus vocabulary. The following figure is an example of two thesaurus entries. The one contains very specific indexing terms, while the other contains general indexing terms. The terms indicated here thus determine the terms which will be used for indexing and for retrieving information.

Thesaurus A with general entries	Thesaurus B with specific entries
LIBRARIES RT Archives Community centres Museums	LIBRARIES NT Academic libraries Public libraries Special libraries RT Archives Community centres Museums

Exhaustivity of Indexing

Taylor (2004:250) puts it like this: 'Exhaustivity is the number of concepts that will be considered in the conceptual framework of the system. The number of concepts any given agency's cataloguers/indexers will assign is often guided by local policy.' Lancaster (1998) describes it as the extent to which the breadth of subject matter discussed in a particular document is covered in a representation of that document. It will correspond roughly to the number of index terms assigned or some other measure of access points provided. Cleveland and Cleveland (1990:111) refer to exhaustivity as the depth of indexing.

The exhaustivity of indexing is thus related to the extent to which indexing terms are assigned to the features covered in an entity. The issue is whether indexing terms are only assigned to a few main concepts covered in the entity (low level of exhaustivity), or whether indexing terms are also assigned to features which the entity deals with in some depth (e.g. a few paragraphs) (high level of exhaustivity).

The degree of exhaustivity of indexing is not necessarily related to the number of indexing terms assigned. For example, a lengthy document dealing with the 'advantages and disadvantages of the use of the *Dewey decimal classification* in special libraries' may

only be assigned two indexing terms: *Dewey decimal classification* and *Special libraries*. The indexing may still be exhaustive. Another document may, for example, deal with the ‘advantages and disadvantages of the use of various classification systems in special libraries’, and for this several more indexing terms will have to be assigned to index the document exhaustively. For example Dewey decimal classification, Library of Congress classification, Universal decimal classification and Special libraries.

The level of exhaustivity at which indexing terms are assigned is determined by the indexing policy. While the level of exhaustivity is not related to the characteristics of the indexing language, it may be related to the type of entity indexed (internal reports, for example, will be indexed more exhaustively), as well as the subject of the entity (as indicated in the example of classification systems).

The exhaustivity of indexing also influences the retrieval of entities:

- A high level of exhaustivity leads to a high level of retrieval. All possible entities containing information (even if it is only a little) are retrieved.
- A low level of exhaustivity leads to a high level of precision. All entities retrieved should contain a reasonable amount of information about the subject.

Exhaustive indexing is particularly useful for users who require an overview of all the entities on a subject, since several entities will also be retrieved that may only contain a few paragraphs on the subject. It is less useful for users seeking specific facts or information.

The following (adapted from Soergel 1985:328) is an example of instructions to indexers (contained in the indexing policy) about intended exhaustivity:

Low exhaustivity (few descriptors)

- Use an indexing term only for entities that are highly relevant for a search about an index term.
- Only index the key subjects of the entity.
- Index, for example, only the first ingredient of food.

High exhaustivity (more descriptors)

- Use an indexing term for entities that are only marginally important.
- Index less important (minor) subjects covered by the entity.
- Index, for example, all the ingredients of food.

Abstracting Policy

An abstracting policy gives guidelines to abstractors on how they should study, analyse and describe the content of entities, so as to give an accurate, abbreviated version of the original entity. Such a policy is necessary to maintain the standard and to promote consistency among abstractors. It should take the following into consideration:

- unique nature of the user group
- purpose of the abstracts
- nature of the original entities
- characteristics of abstracts

- guidelines laid down by standards of abstracting
- purpose of the IOAR system

An abstracting policy should meet the following requirements:

- An introduction to the abstracting service, database or other IOAR systems should be included, as well as the purpose of the abstracts and how they will be used. (It is important to remember that abstracts can serve as search points when retrieving information.)
- Criteria to use in selecting entities for abstracting.
- Guidelines on how to analyse an entity.
- Guidelines about the type of abstract (e.g. indicative, informative, critical). They can also determine that a combination of indicative and informative abstracts should be written for reports, for example.
- Guidelines about the type of information that is regarded as important (e.g. purpose of a study, the steps followed, the methods used and the findings).
- Guidelines about information that may not be included (e.g. references to specific table numbers or information not covered in the entity).
- Instructions about the physical presentation and layout of the abstract. These can include instructions about the style and length, that the abstract should be written in one paragraph, or that it can be presented as a structured abstract under different headings.
- The language that the abstract should use and the handling of abstracts in foreign languages.
- Guidelines for abstracting procedures (i.e. how to prepare the abstract).
- Guidelines for special approaches to be followed for different types of entity.
- Guidelines about the form of the bibliographic description and the information that must be included.
- Lists of standard abbreviations that may be used.
- Examples of abstracts for different topics.
- Guidelines for the editing of abstracts.
- Reference(s) to preferred standards to be used, such as the American National Standards Institute's (*American national standard for writing abstracts.*) 1979. New York: ANSI Z39.14:1979).

In order to understand the context of information organisation policies, you must also consult chapters in this book dealing specific issues, such as those in Part Two.

Conclusion

Indexing and abstracting policies are essential to give guidelines to indexers and abstractors. They improve the quality of the work and promote consistency. Keep in mind though that policies are seldom drawn up by individuals – it is mostly a team effort. Policies should also be revised or adapted regularly to incorporate new technologies, changes in the IOAR overall policy, and so on.

BIBLIOGRAPHY

- Cleveland, DB & Cleveland, AD. 1990. *Introduction to indexing and abstracting*. 2nd ed. Englewood, Colo: Libraries Unlimited.
- Cleveland, DB & Cleveland, AD. 2001. *Introduction to indexing and abstracting*. 3rd ed. Englewood, Colo: Libraries Unlimited.
- Fidel, R. 1994. User-centered indexing. *Journal of the American Society for Information Science*, 45(8):572-575.
- Lancaster, FW. 1998. *Indexing and abstracting in theory and practice*. 2nd ed. London: Library Association Publishing.
- Lancaster, FW. 2003. *Indexing and abstracting in theory and practice*. 3rd ed. Champaign: University of Illinois, Graduate School of Library and Information Science.
- Rowley, J & Farrow, J. 2000. *Organizing knowledge: an introduction to managing access to information*. 3rd ed. Aldershot, UK: Ashgate.
- Soergel, D. 1985. *Organizing information: principles of data base and information retrieval systems*. San Diego: Academic Press.
- Taylor, AG. 2004. *The organization of information*. 2nd ed. Westport, Conn: Libraries Unlimited.

ing Klaus and Mariëtte Burger

Part Two

GENERAL APPLICATIONS

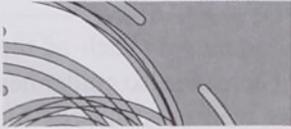
Introduction

Information retrieval systems (IRS) such as *ChemLine*, *Thesis*, *PsychInfo*, *Intelligence*, *Indexing and Abstracting Services*, *PsycInfo* and *Book Indices* are viewed to help scholars and researchers to find the desired material in the resources according to Bergman (1992: 224) an IRS is composed of three major processes: knowledge retrieval, access, sorting, and the transmission, available to users, and finding information of potential value to an active user list of references. To this can be added *Thesis* (1992: 7) mentioned that an IRS should not only find what is already known for records that are required (relevant) and without records that are not required (not relevant). He also has an IRS as a device interpreted between the end user of an information resources and the collection itself. Indexes prepared to describe record titles and their content include items from the information collection. To assure that an IRS meets its purpose, records are created (also called *surrogates*), describing the essential attributes of information. In these records the descriptions of other entities such as people, machines, objects, vehicles and physically described, as well as according to their intellectual content (what they are about, or abstractness). To describe the intellectual content one can use the following:

- persons where names and other attributes are used
- verbal subject descriptions where words are used

For the latter, a verbal description which can be selected from a list of the keywords that can be found in natural language or it can be represented by a controlled vocabulary such as a thesaurus of terms, and subject headings are also verbal representations of the subject content of studies.

When indexing particular entities, such words will be represented as a catalogue record (physical description) of the material.



Ina Fourie and Marlene Burger

Abstract

The intricacies of verbal subject description are discussed. These include a selection of types of verbal subject descriptions, the difference between precoordinate and postcoordinate indexing, decisions to take in verbal subject description and the problems associated with the use of vocabulary. Reference is made to the various verbal methods discussed and analysed in this book.

Introduction

Information retrieval systems (IRS) such as databases, library catalogues, bibliographies, indexing and abstracting journals, portals and book indexes are created to help people who need information to find the desired solution to the problem. According to Ingwersen (1992:228) an IRS 'is constituted by interactive processes between its *system objects*, *system setting*, and the environment, capable of searching and finding *information* of potential value to an actual searcher of information'. To this can be added Harter's (1986:245) argument that an IRS should save time and effort. It should retrieve the records that are required (relevant) and withhold records that are not required (not relevant). He describes an IRS as 'a device interposed between an *end-user* of an information collection and the collection itself ... [whose purpose] is to capture wanted items and filter out unwanted items from the information collection'. To ensure that an IRS fulfils its purpose, records are created (also called surrogates, document representations or entity representations). In these records the documents or other entities (such as people, museum objects, websites) are physically described, as well as according to their intellectual content (what they are about, or aboutness). To describe the intellectual content one can use the following:

- notations where numbers and other symbols are used
- verbal subject descriptions where words are used

For the latter, vocabulary is used which can be selected from an indexing language that can be based on natural language or it can be prescribed by a controlled vocabulary such as a thesaurus. Abstracts and subject headings are also verbal representations of the subject content of entities.

When indexing periodical articles, each article will be represented by a bibliographic record (physical description). For example:

Fourie, I. 2001. The Use of CAI for distance teaching in the formulation of search strategies. *Library Trends*, 50(1):110-127. (Reprinted from *Mousaion*).

The physical description is prepared according to guidelines and standards such as the *Chicago manual of style* or the *Anglo-American cataloguing rules*. There are also guidelines for the description of the intellectual content (e.g. guidelines, as well as standards for indexing, abstracting and the allocation of subject headings) (Cleveland & Cleveland 2001; Lancaster 2003).

The quality of the physical as well as intellectual descriptions affects the efficiency and effectiveness of the IRS, and therefore the preparation of descriptions of the intellectual content of entities is often considered to be one of the most difficult and complex tasks. Blair (1990:vii-viii) puts it as follows: 'The central task of Information Retrieval research is to understand how documents should be represented for effective retrieval. This is primarily a problem of language and meaning. Any theory of document representation, and therefore Information Retrieval, must be based on clear theory of language and meaning.' According to Hjørland and Nielsen (2001:250): '[A]ccess points determine in a rather firm way the objective possibilities that are provided for the talented user (or for any formalized, algorithmic, or automatic procedure). Therefore, it is essential in information science (IS) to develop knowledge about what kinds of subject data exist as well as the strengths, weaknesses, and relative contributions of each kind.'

The description of the intellectual content by means of words is called *verbal subject description*. Verbal subject description is preceded by a process known as conceptualisation, discussed in chapter 3. Verbal subject description can be based on human, intellectual input or automated or semi-automated methods. More detail on automatic indexing and abstracting can be found in the relevant chapters. As words form the basis of this process, one is faced with many linguistic problems such as meaning, synonyms, bias, phrases, compound terms and lay versus scholarly use.

Notations versus Verbal Subject Descriptions

Notations such as the *Dewey decimal classification* (DDC) numbers and verbal subject descriptions, like abstracts, descriptors, subject headings, identifiers, title indexing and PRECIS strings, have both advantages and disadvantages. However, it is important to understand that neither classification notations nor verbal methods of subject analysis can encompass all aspects of subjects in the majority of entities. Even if as many notations, subject headings, indexing terms or descriptors are used as required, these would still represent only the principal subjects of the entity in question. (Macroanalysis is used when an entity is analysed and described in its entirety, or when only the main subject matter is singled out; microanalysis is used when parts of an entity are analysed (e.g. a chapter in a book or articles in a journal) and the subject matter is examined in more detail or in greater depth, i.e. depth analysis). The following examples of an entity on 'computer-assisted instruction in distance teaching in search skills' illustrate the difference.

Examples

The following is an example of a notation:

DDC	374.26 (use of computers in adult education; electronic distance teaching)
-----	--

Other classification systems such as the *Library of Congress classification* (LCC) and the *Universal decimal classification* (UDC) will have different notations to express the same subject.

Examples of verbal subject description include the following:

Indicative abstract	Discusses the use of computer-assisted instruction (CAI) in the formulation of search strategies, and its potential for adult, distance students. The program is used with students from the University of South Africa (Unisa), a distance teaching university. Two versions of the program (1992, 1998) are analysed according to the design teams' experiences and student feedback (collected through questionnaires). The actual design is discussed according to instructional design principles. Selected student feedback is summarised.
Informative abstract	Discusses the use of computer-assisted instruction (CAI) in the formulation of search strategies. The program is used with students from the University of South Africa (Unisa), a distance teaching university. The importance distance teaching holds for life-long learning and programs for adult learners, as well as the potential of CAI for distance teaching is considered. The program was designed by the Department of Information Science. The first version was introduced in 1992, and the second in 1998. They were based on literature reports on the use of CAI in Library and Information Science, as well as principles of instructional design. The latter included determining the need for the CAI, a situation analysis, analysis of the learning content required for the formulation of search strategies, and an analysis of the learners and the availability of technology and support. Objectives and outcomes were set, and methods selected for performance assessment. These include a pretest on two levels of difficulty: Refresher and Challenge. Aspects considered for the actual design, development, implementation and distribution of the program are briefly covered. As part of the formative assessment, feedback was collected on students' experiences, which were mostly very positive. They pointed out sections requiring more examples or further explanation, and the need for online guidelines as part of the tutorial.
Indexing terms, keywords (natural language)	Adult students Computer-assisted instruction Distance students Distance teaching Formative evaluation Instructional design Search strategy formulation Summative evaluation

<p>Descriptors (selected from a thesaurus, e.g. <i>Thesaurus of ERIC descriptors</i>)</p>	<p>ADULT STUDENTS COMPUTER-ASSISTED INSTRUCTION DISTANCE EDUCATION FORMATIVE EVALUATION INFORMATION SEEKING INSTRUCTIONAL DESIGN SEARCH STRATEGIES SUMMATIVE EVALUATION</p>
<p>Identifiers (these are terms that are important enough to select as indexing terms, but do not appear in the controlled vocabulary such as the <i>Thesaurus of ERIC descriptors</i>); they are used in systems based on controlled vocabulary.</p>	<p>Performance assessment South Africa University of South Africa Etc.</p>
<p>Subject headings (there are different lists of subject headings available e.g. <i>Sears list of subject headings</i>, <i>Library of Congress subject headings [LCSH]</i>, <i>MESH</i>)</p>	<p>Adult education – South Africa Computer-assisted instruction Distance education – South Africa Information science – Study and teaching Etc.</p>

The following can also be considered as examples of verbal subject descriptions: broad categories, for example as used by some search engines and directories, titles, PRECIS strings and notes.

Brief Comparison

The methods of subject expression illustrated above are compared briefly so as to indicate relationships and differences. Please note that all of these flow from the initial *conceptual analysis* phase necessary for effective subject description.

Classification Notations

Instead of words, subject matter is expressed by numbers – notations are not language-bound, that is the meaning of a notation is the same in all languages. The function of the classification notation and the terms resulting from verbal subject analysis is different, even though the words describing the notation are the same as those of the subject heading, and so on. For example, the subject heading verbalises specifically a major subject in an entity, while the notation indicates where in the arrangement of its system that subject is to be placed to determine its relationship (usually hierarchical) to other subjects within the discipline. A disadvantage of using notations only is that the user does not necessarily have

knowledge of the classification system and therefore cannot interpret the notations. An index to the notations has then to be compiled as a 'key' to these.

Abstracting

Abstracts give short, concise, accurate descriptions of the intellectual contents of entities so as to promote awareness of entities and information retrieval. Abstracts are often approached through an index, especially if the abstracts have a classified arrangement. They are not meant to replace entities, but serve as surrogates to aid decision making on relevance for a certain information need. Abstracting requires certain language skills and familiarity with subject content. Disadvantages are that they may be affected by errors, policy, omissions and abstractor bias, or that they may be so poorly written that they are difficult to use. It may also be that not all users are equally proficient at using abstracts and other bibliographical tools, so that the user imposes self-limitation on the tools (Cleveland & Cleveland 2001:60).

Indexing

Indexing entails the assigning of indexing terms (also known as descriptors) by using the natural language of the entity or the subject discipline, or by using words from a controlled vocabulary such as a thesaurus. The terms can be combined by Boolean operators or in other ways during the retrieval process to locate the desired information. Citation indexing refers to the situation where the author quoted in an entity (e.g. bibliography, list of cited sources) is used as the point of departure to indicate relationships with other entities. The rationale is that the entities quoted must be on the same subject(s), and that they must be important to the specific subject(s). Book indexes are normally based on the text of a particular entity (i.e. the indexer must use the words appearing in the text); book indexers must therefore have sufficient subject knowledge to add terms that may be useful to the reader. Furthermore, such indexes use natural language, which has its own problems. Terminological control is essential to successful retrieval. Natural language, because of the variety in subject terminology, may cause retrieval problems. To overcome this, some kind of control is necessary: either at the indexing input stage or at the search stage. Book indexes are usually prepared by one person; periodical indexes are compiled by a number of people over a number of years, with shifts in subject emphasis and indexing objectives. Each issue of a periodical may deal with unrelated topics by several different authors, written in different styles, aimed at different users. The periodical index must bring order out of these divergences.

Thesauri

These are controlled vocabularies. They are arranged alphabetically or hierarchically (i.e. formally structured) and can be used for a computerised database or a printed index and are available in printed or electronic format (online thesauri). Thesauri indicate a variety

of indexing terms occurring to the indexer/user. Some of these terms are used as preferred terms in the thesaurus, while others are non-preferred or entry terms. Classification systems are generally hierarchical with secondary alphabetical indexes, while thesauri are generally alphabetical with a hierarchical structure reinforced by cross-references. The relationships between terms in a thesaurus are more specific. In thesauri, descriptors are often dependent on these terms and are intended to be combined with other terms (to express more specific concepts), whereas in a classification system or a subject heading list, the terms can stand alone. In other words, we have on the one hand pre-coordinated systems (e.g. classification notations and subject headings) and on the other hand a post-coordinated language (e.g. descriptors and combinations). In a sense, a thesaurus is a specific language, with words, relationships of words and its own grammatical rules for usage. It is a controlled, collected subset of natural language. A thesaurus is not a dictionary – a dictionary records the standard usage of words and terms, while a thesaurus is the result of the understanding of proper usage. A disadvantage is that not all terms an indexer may need are represented in a specific thesaurus.

Subject Headings

Assigning subject headings by using a standard list is the most common practice. Such lists appear in printed as well as electronic format and are used as standards for the compilation of subject headings. A subject heading list is an alphabetical list of words or phrases that can be used to describe the subject matter of entities. Some lists are general, covering most topics (such as in a general classification system), while others were developed for specific disciplines or subjects. In practice, subject heading lists may be expanded and/or adapted to meet the specific needs of an information agency. Subject headings are used as entry points (access points) in bibliographic databases, catalogues, indexing journals, bibliographies, the Internet and the like. Sometimes subject headings are used in conjunction with indexing terms. It is therefore necessary to understand the difference between thesauri, indexes and subject headings (see the examples above). As with other methods of verbalisation, a disadvantage is that the indexer selects the subject headings based on the text and his or her knowledge of the subject, but these may not concur with the perception of the searcher. *See* and *see also* references are used to guide the user to the 'correct' subject heading(s).

Precoordinate and Postcoordinate Indexing

To understand verbal subject description it is important to distinguish between precoordinate and postcoordinate indexing. According to Harter (1986:247) a precoordinate indexing system is based on a controlled vocabulary in which specific, complex concepts involving a combination of two or more concepts are used as an indexing term in the vocabulary. Coordination takes place at the time of indexing (i.e. at the time of creating the entity representation). 'Database producers' for example combine the concepts of 'databases' and 'producers', or 'computer-assisted instruction' combines 'computer-assisted' and 'instruction'. Subject headings and classification systems are examples of precoordinate indexing systems. In the previous examples used above, DDC, *Eric thesaurus* and

Sears list of subject headings were used for the article on computer-assisted instruction. The relationship between the concepts is thus indicated at the time when the entity representation is created. The relationship is determined by the person creating the entity representation. When retrieving information the same order should be used (e.g. 374.26 to retrieve information about computer-assisted instruction in adult education, or Computer-assisted instruction). Only one order and combination of terms/concepts is possible – in classification this is known as citation order.

Postcoordinate indexing systems can be based on natural language or controlled vocabulary terms/concepts. Coordination of terms/concepts takes place at the time of retrieval, that is the individual concepts such as ‘database producers’ and ‘bibliographic databases’ and ‘South Africa’ can be combined at the time when information is needed. This can be done by using Boolean operators: database producers AND bibliographic databases AND South Africa. Pre-coordinated terms such as ‘database producers’ can also be used in a postcoordinate system. The relationship between terms/concepts is not indicated at the time the record is created, but rather at the time of information retrieval. Any number of combinations can be used during information retrieval. For example: **database producers, database producers AND full-text databases, or database producers AND full-text databases AND South Africa**. It is also important to note that the relationship is indicated by the information seeker. Postcoordinate indexing was made possible by the early uniterm concept of Mortimar Taube, and especially the development of electronic databases and Boolean logic (Chu 2003). The following is but a brief description of the Boolean operators. There are many publications from which more information can be obtained – the book by Taylor (see bibliography) is a good starting point.

Boolean Operators

There are three Boolean operators: AND, OR, NOT. These operators are used to link search terms. Each of the operators has a specific effect on the search results, and one should therefore make sure they are well understood, because using the wrong operator can make a great difference to search results.

‘AND’ Operator

If two or more search terms are linked with the AND operator, it means that all the terms must appear in every record that is retrieved. For example: **Libraries AND South Africa**. Records in which only one of the search terms occurs, for example, will not be retrieved. Any number of search terms can be linked with the AND operator. For example: **Libraries AND South Africa AND CD-ROMs**. The more terms linked in this way however, the fewer records will be retrieved since the search becomes very specific – it may even happen that no records are retrieved.

‘OR’ Operator

If two or more concepts are linked with the OR operator, it means that records will be retrieved if any one, or even more, of the search terms occurs in the record. The more search terms linked with the OR operator, the more records can be retrieved. For example:

Libraries OR South Africa. Any number of terms can be linked with the OR operator, but note that as search terms are added, more records will be retrieved each time.

'NOT' Operator

If two search terms are linked with the NOT operator, the search terms that are preceded by the NOT operator will be excluded. Thus no records in which the search term occurs will be retrieved. This operator should be used very carefully since it can cause useful information to be lost if the search term that has to be excluded occurs together with substantial information. For example: **Libraries NOT South Africa** (no records on South Africa will be retrieved).

Truncation symbols such as ? or * are used to indicate that only the word stem is to be sought. All words beginning with the word stem are then retrieved, for example **manag?** retrieves **management, manager, managers, managing,** and so on. Some information systems also permit wild-card truncation, for example **col?r** retrieves **color, colour.** If more than one type of Boolean operator is used to connect search terms, *parentheses* are necessary to indicate how and in what sequence terms should be combined. If parentheses are not used, the default sequence is used. It is often the following: NOT, AND, OR. For example **dogs OR cats AND breeding** will be used in retrieval as **dogs OR (cats AND breeding)** and **(dogs OR cats) AND breeding** will be used in retrieval as **(dogs OR cats) AND breeding.** *Proximity operators* are used to indicate the position and sequence of words. There are various symbols for proximity operators, and information systems differ in the way in which they indicate these operators. Proximity operators are often used in indexes and some search engines, but less often in library catalogues. There are usually two proximity operators, for example an operator that specifies that words must occur in precisely the same order but that, up to a maximum of the number of words that are specified, may appear between the words – so for **academic (2w) library,** records on academic libraries, academic medical libraries and academic law libraries will be retrieved; the second operator specifies that the sequence of the words is not important, and may, up to a maximum of the number of words that are specified, appear between the words – so for **database (2n) design,** records on database design, design of databases, design of bibliographic databases will be retrieved.

Subject Headings

The various methods of indexing, thesaurus construction, abstracting, and so on, are the main focus of this book, but the assigning or allocation of subject headings (also a form of verbal subject analysis) is not covered in a separate chapter. Subject headings are used widely to access and retrieve information from bibliographic resources; therefore a brief section on subject headings is included in this chapter. Always keep in mind that when using a list of subject headings, the *principles* of the list must first be thoroughly studied.

Standard subject heading lists do not only contain subject headings, but also provide guidelines and examples for allocating subject headings. A subject heading can be described as a word, composite term or group of words (phrases) that represent a subject under which all entities dealing with the same subject can be arranged together in, for example,

a bibliography, library catalogue, or index. The physical appearance of subject headings differs from that of thesaurus descriptors and indexing terms because they sometimes contain subdivisions, qualifiers, dashes, parentheses and so on. Examples of subject headings:

Accidents – Prevention

Belief and doubt

Bestsellers (Books)

Bridge (Game)

Cataloguing – Rock drawings, paintings and engravings

English as a second language

Indexing

International Standard Book Numbers

Pretoria (Gauteng) – Pictorial works

Sewing – Dictionaries

Shakespeare, William, 1654-1616 – Quotations

South Africa – Population – Statistics

World War, 1939-1945

Purpose of Subject Headings

The primary purpose of subject headings is to identify entities on a given subject, but they also have other purposes. For example:

- To identify entities that deal with related subjects. For example **Children's literature** will also be linked to **Children's libraries** and **Libraries and schools**.
- To bring together entities dealing with the same subjects regardless of whether different terminologies are used to describe subjects. **Library technicians** are also used for entities that deal with library assistants, library clerks and paraprofessionals.
- To show the relationships between subjects or disciplines. **Botany** is indicated as related to **Biology**, **Natural history**, **Plants** and **Science**.
- To give access to any subject at any level of analysis (from general to specific). For example **Academic libraries** and **Church libraries** are considered to be specific terms, while **Libraries** are considered to be a more general term, and most general terms are **Books**, **Documentation** and **Books and reading**.
- To give access via a vocabulary that a large group of users have in common, regardless of whether they are specialist or lay users. This purpose is achieved by cross-references from terms not used to terms used (preferred terms), for example 'Flora *see* **Botany**'.

The principles underlying the allocation of subject headings are similar to other verbal methods of subject description: namely the user as focus, the collocation of entities on the same subject, the use of the most representative terms, and specificity.

There are two main types of subject heading. One is *true subject headings*, which indicate the subject matter of entities (topical subject headings, e.g. **Ball games**); name subject headings, e.g. **Shakespeare, William, 1564-1616**; title subject headings, e.g. **Bible**.

N.T.). The other is *form subject headings*, which do not indicate the subject matter of entities, but rather the form in which they are presented (publication formats, e.g. **Atlases, Bibliographies**; literary genres, e.g. **Poetry, Fiction**).

General Guidelines

The following points of departure need to be considered:

- Number of subject headings. Often one subject heading is sufficient, but sometimes it is necessary to allocate more than one subject heading to properly represent the subject of an entity. If the subject matter is not properly reflected by the subject heading(s), valuable information can be lost. In practice there is often a limit on the number of subject headings per entity. With online systems there is no need to impose any limitations.
- General versus specific subject headings. Do not allocate both a general and a specific subject heading if one subject heading encompasses or includes the specific one. You would not, for example, allocate both **Libraries** and **Church libraries** to the same entity.
- Multiple concepts. More than one subject heading is usually necessary for entities dealing with a number of subjects that cannot be represented by a single subject heading. For example, for an entity dealing with Zulu and Afrikaans literature, one would assign **Zulu literature** and **Afrikaans literature** since they are two different types of literature. For an entity dealing with Greek and Latin literature a subject heading inclusive of both, namely **Classical literature**, may be enough. When an entity deals with four or more concepts which are all part of a broader concept, only one subject heading (the broader concept) is needed. For example, for an entity dealing with the history of Argentina, Brazil, Chile and Ecuador, the subject heading will be **South America – History**.
- Multiple elements. An entity with multiple elements is one that covers a central subject from different perspectives and a multifaceted entity covers different elements like form, place and time. If it is difficult to find subject headings expressing these elements and facets, it is often necessary to construct a new subject heading (according to the prescriptions of the list being used). Always consider users' needs in such instances – these needs will decide whether all the elements identified in the concepts should be represented.

The general guidelines to be followed are useful when one formulates a policy for allocating subject headings. A policy is necessary to ensure that different indexers and subject cataloguers allocate subject headings consistently. When using a specific subject headings list, always check how the general guidelines provided should be interpreted and applied. Note the following:

- Only subject headings that reflect general usage or technical terminology should be used.
- Allocate as many subject headings as necessary to describe the subject matter of an entity (taking into account restrictions and system limitations).
- The plural form of nouns is used in countable objects or entities.
- Subject headings are compiled in the language of the bibliographic database or catalogue.

- Different symbols (punctuation) are used in a subject headings list for clarity: they should be used consistently.
- Three types of references are possible: *see* references, *see also* references, and general references.
- Qualifiers and subdivisions are sometimes used.
- There are different types of subject headings.
- Words, phrases and composite terms can be qualified.
- Note special guidelines for personal, place and geographical names.
- Make sure which subject headings may be added by the indexer or cataloguer, for example proper names (e.g. persons, places, battles) and general names (e.g. animals, fruit, diseases).

Necessary Decisions

As indicated earlier, the quality of the verbal descriptions influences the quality of the information retrieval. The following points need to be considered:

- Type of verbal subject description (e.g. abstract, indexing terms, subject headings). Remember this description offers the words that can be searched during information retrieval, and also allows one to decide on the relevance of an entity.
- Length of the description. This will influence the number of words that are available for information retrieval. Longer descriptions such as informative abstracts can facilitate information retrieval. They are, however, more time-consuming to prepare. Lancaster (2003) offers an extensive discussion on the impact of length of description. Hjørland and Nielsen (2001:284-285) also support the idea that a combination of different types of verbal descriptions will contribute to more effective and efficient information retrieval, and will help information seekers to decide on the relevance of entities.
- Specificity. How specific will the words be? In other words, will the words in the representation be as specific as the words used in the text? (This mostly applies to indexing terms, descriptors and identifiers.)
- Exhaustivity. How exhaustive is the coverage of the intellectual content? Exhaustivity is sometimes reflected in the number of indexing terms assigned: the use of more indexing terms is often associated with more exhaustive indexing, although this is not necessarily always the case. Exhaustivity can also be reflected in the length of an abstract and the amount of detail provided (e.g. the difference between indicative and informative abstracts).
- Vocabulary used (e.g. natural language or controlled vocabulary). If controlled vocabulary is used a suitable thesaurus should be selected or constructed according to user needs.

Language Issues

Language is always a key component in information retrieval and especially in verbal subject description. There are, however, many problems with the use of language. Note that language is used to

- represent the content of documents or other entities
- represent the information problems of the users (information seekers)
- instruct the computer to carry out the search and retrieval functions

Languages are Classified by

- vocabulary: strings of continuous characters that form words (e.g. student, student performance, AIS303)
- syntax: rules for combining words to form valid sentences or phrases (philosophy of education; education of philosophy; of education philosophy [not valid])
- logical structure (or taxonomy): describes the nature of the relationships between elements
- domain: can be either a real world or a theoretical world (e.g. discipline such as History or Computer Science, or a work environment like a group of engineers)

Natural language and controlled vocabulary are both indexing languages that are used in verbal subject description. Natural language refers to the words of the author that can be found in the title and full text of an entity, the words used every day in the subject field, or words that an indexer can select from his/her own vocabulary on the subject that is discussed in the entity. Controlled vocabulary refers to terms that are selected from a list of possible indexing terms that has been compiled beforehand (e.g. a thesaurus or a list of subject headings). In the case of indexing terms, we refer to terms selected from a thesaurus as descriptors (e.g. selected from *Thesaurus for ERIC descriptors*).

Both natural language and controlled vocabulary have advantages and disadvantages. The following section is based on the work of Aitchison, Gilchrist and Bawden (2000), Chu (2003), Harter (1986), Lancaster (2003) and Soergel (1985).

Natural Language

The advantages of natural language pertain particularly to the organisation of information (representations of entities) and the disadvantages to retrieval (search terms and search queries).

The following are examples of the advantages of natural language:

- Very specific indexing terms can be assigned. The terms can be assigned as specifically as they are used by the author. If the author of an entity refers, for example, to Persian cats, it can be assigned as an indexing term; a more general term such as 'cats' need not be used.
- Natural language represents reality. The terms used in the everyday world (reality) can also be used as indexing terms. (This is in comparison to controlled vocabulary that is often described as an artificial language.)
- New terminology can easily be added to the vocabulary because it can be chosen from the content of an entity or can even be assigned. It is not necessary to wait until terms have been added to lists of controlled vocabulary.

- It is easier to adapt natural language according to changes taking place with regard to the terminology of the language.
- Natural language makes it possible to keep the differences in meanings of terms (e.g. synonyms and near-synonym terms such as 'diet' and 'nutrition' or 'inflation' and 'price increases'). This is in contrast to controlled vocabulary where a maximum of one or two terms can be used.
- The translation process (thus the process of subject description) requires less time and effort and less utilisation of personnel. It is not necessary to first consult a list of controlled vocabulary or, if a list is not available, to find alternatives. Concepts are translated directly by using the words or phrases of the author, or the words or phrases used by the indexer. It requires less mental effort because it is not necessary to think of alternative words if a specific term does not appear in a list of controlled vocabulary.
- The use of natural language requires no maintenance, as is the case with a list of controlled vocabulary.
- Using natural language to retrieve information makes it very easy, because people can use the terms that they thought of themselves.
- It is not necessary for people who wish to retrieve information to receive special training on how to use a list of controlled vocabulary such as a thesaurus, subject heading list or a classification system.

Most of the advantages mentioned above are concerned with the organisation (thus indexing) of entities. The disadvantages pertaining to the use of natural language are concerned with the retrieval of information:

- The use of natural language requires more work than when controlled vocabulary is used. Although a more specific search for information can take place (remember any term can be used; it does not have to be in a controlled list), those who need the information must think of all the possible terms themselves, instead of using the indexing terms which were assigned during organisation. These may be referred to as search terms. A number of disadvantages include
 - variations in spelling
 - plurals (singular and plural can be used for search terms)
 - different word forms (nouns, verbs or adjectives can be used)
 - synonyms or near-synonym terms that may help to retrieve relevant information
 - broader terms (or concepts), for example the use of 'cats' if a search under 'Persian cats' (a more specific or narrower term) does not deliver adequate information
- To cover all these possibilities, users often have to use dictionaries, reference works, thesauri, classification systems and other lists of controlled vocabulary to ensure that the information required is retrieved.
- There are terms that are difficult to express using indexing terms in natural language, because they are actually descriptions rather than terms. Consider, for example, how you would search for information in a database about cancer (CANCERNET) if you were interested in survival for more than five years. In such a case it would be far easier if an appropriate term in a list of controlled vocabulary could be used.

- In multi-language databases the use of natural language can give rise to problems.
- In some instances, and particularly in the case of full-text databases, the use of natural language can result in too much information being retrieved. (Full-text databases contain the full text of an entity, e.g. journal articles, newspapers and encyclopaedias).
- Information about the same subject is often found scattered in an alphabetical list of terms, because different terms are used. For example 'indexing', 'subject indexing' and 'subject description'.
- If a subject is mentioned implicitly rather than explicitly in a text, it will be more difficult to ascertain using derived indexing (natural language).

Controlled Vocabulary

Controlled vocabulary has advantages and disadvantages. The advantages and disadvantages are distributed between information organisation and retrieval.

Advantages include the following:

- One of the primary objectives of vocabulary control is to ensure that subjects are described consistently by both the indexer and the information seeker. If controlled vocabulary is used, subject description tends to be more consistent since the indexer is guided to acceptable terms. If, for example, a document deals with the rise in food prices, the thesaurus (or other form of vocabulary control) may indicate that the term to be used for subject description should be 'inflation' rather than 'price increases'.
- Vocabulary control also plays an important role in the control of synonyms and homographs, and is thus able to limit the scattering effect of related information.
- Controlled vocabulary deals with the problems associated with spelling variations, word forms, and so on that are experienced with natural language (refer to the previous section for more information). The user in particular needs to consult additional reference works to identify relevant terms that can save a lot of time.
- Controlled vocabulary can also indicate relationships between terms, especially generic relationships, such as 'Burmese cats' that fall under a sub-category of 'cats'. In this way an exhaustive search (i.e. a search where as many relevant and sometimes partially relevant entities are retrieved) is made possible, because the user is led to the terms in a logical manner.
- Since controlled vocabulary can reduce the number of terms which can be used for indexing, it is easier to choose an appropriate term, especially when retrieving information.

The following are disadvantages of using controlled vocabulary:

- Controlled vocabulary can never be as up to date as natural language. It is far easier to include new terminology in titles and abstracts than in lists of controlled vocabulary.
- It takes a great deal of time and money to maintain controlled vocabulary.
- Although a lot of time is saved in the retrieval process, it requires more time to prepare the inputs for the ISR, in other words to translate the concepts.

- Controlled vocabulary often does not keep up with changes in terminology and therefore with reality – that is the terms used in real life.
- It requires special training to use controlled vocabulary. The relationship between broader and narrower terms, for example, is not always apparent to novices.

Quality of Verbal Subject Description

The quality of verbal subject descriptions depends *inter alia* on

- accuracy of the intellectual content's description
- consistency of the subject description
- exhaustivity of the subject description
- specificity of the subject description

It is therefore advisable to work according to a policy for verbal subject description, such as an indexing or abstracting policy (see chapter 5).

Conclusion

Different types of verbal subject description or a combination of these can be used to describe the intellectual content of an entity. Verbal subject description stands in contrast to notational descriptions. When deciding on a type of verbal subject description (e.g. indexing terms, titles or abstracts) one should consider strengths and weaknesses, and their potential impact on information retrieval, as well as the arguments about natural language and controlled vocabulary.

BIBLIOGRAPHY

- Aitchison, J, Gilchrist, A & Bawden, D. 2000. *Thesaurus construction and use: a practical manual*. 4th ed. London: Aslib.
- Blair, D. 1990. *Language and representation in information retrieval*. Amsterdam: Elsevier Science Publishers.
- Chu, H. 2003. *Information representation and retrieval in the digital age*. Medford, NJ: Information Today.
- Cleveland, DB & Cleveland, AD. 2001. *Introduction to indexing and abstracting*. 3rd ed. Englewood, Colo: Libraries Unlimited.
- Harter, SP. 1986. *Online information retrieval: concepts, principles, and techniques*. San Diego: Academic Press.
- Hjørland, B & Nielsen, LK. 2001. Subject access points in electronic retrieval. *Annual Review of Information Science and Technology*, 35:249-298.
- Ingwersen, P. 1992. *Information retrieval interaction*. London: Taylor Graham.
- Lancaster, FW. 2003. *Indexing and abstracting in theory and practice*. 3rd ed. London: Facet Publishing.
- Soergel, D. 1985. *Organizing information: principles of data base and information retrieval systems*. San Diego: Academic Press.
- Taylor, AG. 2004. *The organization of information*. 2nd ed. Westport, Conn: Libraries Unlimited.



Ina Fourie and Marlene Burger

Abstract

The intricacies of book indexing are discussed. These include the clarification of concepts, steps in indexing, the choice of index entries, cross-references, locators, lead-in entries and layout. Indexable matter and what to exclude, working according to an indexing policy and the editing of the index are also considered. Please note that the technical presentation of a book index is discussed in chapter 15.

Introduction

Book indexes are essential to providing micro-level access to the content of an entity such as a book, report, encyclopaedia, yearbook or a hypertext document (i.e. access to the finer detail and pieces of information covered in a document). The importance of book indexing should be obvious from the following incident on which Cleveland and Cleveland (1990:142) report: ‘Several years ago a British politician introduced a bill that would have denied copyright and imposed a fine on a publisher who issued a book without an index, but the bill failed.’

According to Lipetz (1989)

An index can be made very short or very long, very simple in its design or very complex; but one thing an index cannot be is totally exhaustive. As most of us realize and accept – sometimes only intuitively – any text or collection can be indexed to any degree or length one might imagine, and one would still, always, be able to find more entries or aspects that could logically be added to the index. Of course, one tends not to incorporate entries in an index unless they seem “useful”. But deciding what to include, how much to include, how to include it, are problems in indexing that always cry out for rational, exact answers that generally elude us. Still, we know that if there are any answers to be found, they are somehow connected to the question of usefulness, or probable usefulness, of the index that we might produce.

Definition of Concepts

Book indexing is the process whereby an index to the content of an entity is created. It is often seen as an art form because it requires creative as well as writing skills. The book index that is created in the process is a systematic guide to the content of an entity. It can

even be created for fiction. It serves the same function as an IR system: it organises the content of a specific entity in such a way that users can retrieve information from the entity when needed.

Against this background of the *function of an index* (to provide an efficient means of tracing information), the indexer should

- identify and locate relevant information
- discriminate between information on a subject and passing mention of a subject
- exclude passing mention of subjects that offers nothing significant to the potential user
- analyse concepts treated in the text so as to produce a series of headings
- ensure that terms used in the index are appropriate to users of the index, so that they will
 - quickly establish presence/absence of information on specific subjects in an unfamiliar work
 - quickly retrieve information on a remembered item in a known/partially known work
 - quickly identify appropriate entities in a collection
- indicate relationships between concepts
- group together information scattered by the arrangement of the entity or collection
- synthesise headings and subheadings into entries
- direct users seeking information under terms not chosen for index headings to headings used by making cross-references
- arrange entries into a systematic and helpful order

The indexer provides access to the content of the entity by

- conceptualisation (finding out what the document is about, and identifying important points)
- selecting concepts that are worth listing in the index (these are the entries in the index)
- translating the index entries into vocabulary (words) that will probably be used by the readers of the entity, and that will satisfy their need to find specific items of information

When dealing with book indexes, one should also understand a number of other key concepts. An entity (or document), for example, is the medium where a message is recorded. It is applicable to print as well as non-print media such as microfiche, pictures and audio, as well as other human-made objects. The purpose of an entity is to convey information.

An entry (or index entry) in a book index consists of a main heading and all other information. An entry can comprise a combination of single characters (e.g. B12 in a book about vitamins), a word (e.g. oranges), a compound word (e.g. student academic performance) or a phrase (e.g. for legal advisors). It can also include the names of organisations and institutions (e.g. University of South Africa), publications (e.g. *South African Journal for Higher Education*) or people (e.g. Nelson Mandela).

An entry can also include one or more subheadings, cross-references and locators. The layout and use of capital letters or lower case vary considerably. One has to decide on a format and standardise accordingly. The following is an example of an entry:

Book indexing
 policy 40
 purpose 39-40
 steps 40-43 *see also* Web indexing

Main heading: book indexing
 Subheadings: policy, purpose, steps
 Cross-reference: *see also* Web indexing
 Locators: 40, 39-40, etc.

Sometimes the main heading is also referred to as an access point, entry, indexing entry, heading or subject heading. The subheading is sometimes also referred to as the sub-entry or modification. The subheading is subordinate to the main heading and it includes a specific aspect of the main heading. A subheading may comprise a combination of single characters, words, compound terms, phrases, et cetera. A locator may also be called a reference locator. A locator may consist of page numbers or paragraph or line numbers. It indicates where information about a main heading or subheading may be found.

A book index also consists of lead-in entries. They reflect the concepts that readers could use when searching for information. Lead-in entries are also called non-preferred entries. *See* references are used to guide the user from the lead-in entries to the entries preferred by the index. For example:

Back-of-book indexes *see* Book indexes
 ANC *see* African National Congress

See also references are used to guide the readers to related terms which may also contain useful information. For example:

Indexers *see also* beginners; contracts; freelance work; in-house indexes

Reasons for Book Indexes

Book indexes are created for the following reasons. These also show the importance of spending sufficient time and effort when creating indexes:

- To help readers find specific facets of information about a subject as quickly as possible.
- To give access to the content of an entity. For example a textbook is of little use if you have to read through the entire book to find a specific item of information.

- To indicate subjects or items that might be important to the reader; although there might be exceptions, information mentioned in passing is normally not included in the book index.
- To give access to the terminology of the entity as well as the terminology that the user might prefer. Cross-references are normally used to guide readers from the terms or concepts that they have in mind to the terms and concepts that are used as indexing terms.
- To group terms for related matters together; this is done by using subheadings.
- To create entries for concepts covered in the entity and to combine them in such a way that users find it easy to access the information.
- To give an accurate reflection of where the information can be found (e.g. by providing the correct page numbers).
- To present a systematic series of entries so that it is easy to find specific information.

Components of a Book Index

Apart from a list of alphabetically arranged, systematic entries, a good book index should also have an introductory note, which explains the components of the index. Users must know what is covered in the index and what is excluded. Any special characteristics of the index that may affect its use should be explained. Generally only one or two features are explained, taking up only a few lines. For large multi-sequence indexes with complicated locators, abbreviations and typographic indicators, several paragraphs of explanation and guidance may be necessary. One or more of the following features may be explained:

- coverage (parts of the text covered by the index and which not)
- order of entries, if not obvious (e.g. 'broken order', 'institute' for 'institution', 'Mac' and 'Saint' forms)
- make-up of complex locators (e.g. date, volume, part, page)
- use of typestyles such as italic and bold to indicate illustrations, or locators for major references
- special use of capitals
- use of abbreviations in headings (e.g. ASAIB, LIASA) or for locators (e.g. fig for figure, n for notes)
- treatment of multi-element names (e.g. names with prefixes)
- existence of variant forms and spellings of names (e.g. Phillips and Philips; De La Mare and Delamare)
- omission or transposition of definite and indefinite articles (e.g. the, a, an)
- cross-referencing (its purpose)

The following are examples of introductory notes:

- Entries are in word-by-word order, which takes account of spaces and hyphens between words, therefore 'fishing-rods' will precede 'fishmongers'. Locators followed by *n* indicate note numbers, for example 65*n*3.
- The index covers the Introduction, Chapters 1 to 19, and the Chronology. Titles of publications are italicised, for example *Discovering South Africa*. Titles beginning with 'A', 'An' and 'The' are entered under the next word, for example *Sweet Essence of Home*, The. Personal names are entered according to their national or cultural customs; most Western names are placed under surname. Customs vary for multiple-element names, so if one of these is not where you expect to find it, look for it under another element.
- NOTE: The index is alphabetized in word-by-word order, for example data protection is filed before databases. Numbers in headings are filed as though spelled out, for example W3C is filed as *WthreeC*. Prepositions in subheadings are ignored in filing. Page numbers are given in full form, e.g. 124-125. Page numbers followed by *bib* refer to bibliographical references. Where more than one page number is listed against a heading, page numbers in bold indicate major treatment of a subject. The following abbreviations are used in headings in the index: ASI for American Society of Indexers; AusSI for Australian Society of Indexers; SI for Society of Indexers. [This introductory note appears at the beginning of the index to Booth, PF. 2001. *Indexing: the manual of good practice*. München: Saur.]

The other components have been already been introduced and will be mentioned throughout the remainder of this chapter. They are: entries (main entries, lead-in entries, preferred terms), sub-entries, cross-references and locators.

How to Index a Book

Indexing a book is a difficult process, which can, however, be mastered with practice and a critical outlook. Booth and Piggot (1988:7) explain the problems encountered in book indexing as follows: 'Secondly, and entailing much more difficult, but vital decisions, it is the indexer's duty to summarize in a succinct entry, the gist of a page, paragraph, or other section devoted to a single idea, and also to make entries for all significant topics mentioned in that section.'

The indexer should always think of the readers – the people who will use the index. Indexers usually do not know the readers and thus get no feedback about the usability of their indexes. The indexer often has to provide for the needs of a variety of readers, which makes it even more difficult. Encyclopaedias, for example, are often used by laypeople, children and experts; they must therefore contain scientific terms as well as terms for the layperson.

Information One Needs before Starting on the Index

When preparing a book index one needs the following information from the publisher or editor:

- length of the index (i.e. how many pages)
- which house style is preferred (if any)
- whether uniformity with other publications in the same series is necessary

- what supporting material is available to the indexer (e.g. the table of contents, appendices, illustrations and changes to the page drafts of the entity, all or any of which may influence the index)

Apart from the above, which refers mainly to the physical presentation of an index, one must decide on the following matters before starting the indexing process (some made in collaboration with the publisher, others made by the indexer):

- Choice of headings (i.e. which concepts to use); terminology (synonyms, offering a choice of one; linked terms like antonyms (e.g. 'good' and 'evil'); homographs with qualifier added (e.g. 'races (ethnology)' and 'races (sport)').
- Form of headings/subheadings, for example singular and plural, orthography, phrases, prepositions.
- Scope and use of headings/subheadings: sometimes scope notes may be necessary, however rare in book indexes; indentation (one or more indents).
- Proper names, which include personal names, corporate bodies and geographical names.
- Titles of entities, for example books, periodicals, music, art, films, literature.
- Locators: which to indicate and how (e.g. page numbers, volumes, columns, files, footnotes).
- Cross-references, that is *see* and *see also* references.
- Arrangement: word-by-word or letter-by-letter; headings beginning with the same term, for example

```

milk
  cows'
  goats'
milk allergies
Milk (report)
Milk Research Board

```

(The above example includes a term with subheadings, a term with a qualifier, a compound term, and the name of a corporate body. The inclusion and positioning of cross-references should also be considered.)

- Presentation of the index: printed or electronic; layout (e.g. spacing, indentation, run-on text, columns, typography).

Try to get hold of as much supporting material as possible.

Using Standards and Aids

When compiling a book index, it is important to take note of the guidelines offered in standards on indexing. They include

- British Standards Institution. 1996. *Information and documentation – guidelines for the content, organization and presentation of indexes*. BS ISO 999:1996. London: British Standards Institution. (This standard offers guidelines – applicable to all kinds of print,

non-print and electronic materials – for the content, organisation and presentation of indexes, with illustrative examples. Basic principles and practice, on which indexes can be built, are its main concern. It does not deal with detailed indexing procedures or with the automated generation of indexes.)

- *Recommendations for the preparation of indexes to books, periodicals and other documents*. 1988. London: British Standards Institution. (BS 3700:1988).
- *Recommendations for examining documents, determining their subjects and selecting indexing terms*. 1984. London: British Standards Institution. (BS 6529:1964).

More information on standards can be obtained from the website of the International Organization for Standardization at <http://www.iso.ch/iso/en/ISOOnline.frontpage/>

Steps in Book Indexing

Book indexing consists of the following steps:

- conceptualisation, which means analysing an entity to determine what it is about
- selection of concepts or items worth an index entry
- translation of concepts into index entries (e.g. main headings, subheadings)
- indication of the location (e.g. page numbers) where the information appears
- creation of lead-in entries (*see* references) that guide readers from familiar terms to those used in the index
- creation of cross-references (*see also* references)

What to Index

Certain parts of an entity contain ‘indexable matter’ – information that might be worth indexing. The author, book editor, publisher or indexer must decide on the sections of an entity that will be indexed. Must the entire entity be indexed, or only the major sections, or only certain sections? When making such decisions, the potential readers should also be considered.

Sections of an entity which are usually indexed include

- the text (this can include anything from the main sections to only selected sections)
- Introduction
- glossaries
- notes (footnotes, endnotes)
- addenda
- illustrations
- appendices

Feters (1994:14) is of the opinion that word lists do not need to be indexed. References (such as bibliographies, reading lists, further reading and lists of works consulted) are only

sometimes indexed. The following are normally excluded from the indexing process (but not always):

- title page
- name of the person to whom the entity is dedicated
- table of contents
- synopsis of chapters

The type of entity will also determine what is considered to be 'indexable matter'. For example an academic text will be indexed in detail, whereas an index for an annual publication or a journal might only contain the main articles instead of the news items and items of passing interest.

Studying the Physical Entity

One needs to study the physical entity to determine its contents. The subject content must thus be determined. This will be determined by the decision on what is to be considered 'indexable matter'.

Initially, one should make a preliminary study of the entity and skim through the introduction to get a general feel of the text, the potential readers and their information needs. Cleveland and Cleveland (1990:129) advise indexers to write a summary of 500 words of the entity; if an indexer cannot write such a summary, he or she will probably have problems indexing the entity.

The most important matters or components that one has to cover in the initial examination are

- title
- contents page
- introduction
- concluding paragraph and summary
- list of illustrations
- word list
- headings
- words and phrases highlighted in the text

After the provisional examination of an entity, read it a second time in more depth. One can begin to make notes of the main headings and note the major and the less important terms and concepts. It is very important to note new and unique information. (This is difficult if one is not really familiar with the topic.) Each page and each sentence must be considered during the indexing process. Terms which might be particularly important should be underlined. Not every word in a sentence is used for indexing, however. In the sentence 'Children are like rays of sunshine', only the term 'children' is important as an index term. Terms in bold and italics must always be considered. Otherwise one need only take note of particular words if mentioned in a fair amount of detail. Hence the key question: What is a fair amount of detail?

When deciding whether it is worth making an index entry or not, keep the readers' needs in mind and the amount of information in relation to the subject as a whole.

There are no specific guidelines for the number of index entries which may be selected for an index. Sometimes the length of an index is determined by the publisher – this might mean that the indexer can only compile a brief index.

Keep an open mind when studying an entity to determine the concepts for indexing. Indexers must index the content of an entity and not their own ideas about what the entity should cover.

Translating the Content and Creating Entries

Decide on main headings, subheadings and cross-references for all entries, always keeping the readers' needs in mind (e.g. the questions users might ask, the subject they want information about and the words they might use).

Many sources can assist the indexer in the choice and formulation of main headings and subheadings:

- standards
- dictionaries
- word lists
- encyclopaedias (general and subject-oriented)
- guides
- classification schedules
- thesauri in the subject discipline and related disciplines
- house style manuals
- guides to the use of English
- other experienced indexers

The website of the ASI offers an extensive list of links to such sources: <http://www.asindexing.org>

Style for Interpreting Entries

Some publishers have a house style. If not, the indexer can decide on his or her own style. Standards such as the following can give valuable guidelines:

- *Recommendations for the preparation of indexes to books, periodicals and other documents*. 1988. London: British Standards Institution (BS 3700:1988).
- British Standards Institution. 1996. *Information and documentation – guidelines for the content, organization and presentation of indexes*. BS ISO 999:1996. London: British Standards Institution.
- *Chicago manual of style*. 1993. 14th ed. Chicago: University of Chicago Press.

The goal is to formulate the entries as clearly and understandably as possible. The direct format is therefore generally chosen when more than one word is used, for example 'academic libraries' rather than 'libraries, academic'. Exceptions are made for individual

names (e.g. Mandela, Nelson, rather than Nelson Mandela). The opposite is recommended if it helps to group related information together. For example:

- alphabetising, letter-for-letter (instead of letter-for-letter alphabetising)
- alphabetising, word-for-word (instead of word-for-word alphabetising)

Indexes usually consist of two or more columns per page and are printed in a font smaller than the rest of the entity. One column per page takes up too much space.

Generally two basic styles are used to construct a book index: the indent style and the run-in style (or run-on style). The indent style gives a clearer construction, while the run-in style is used primarily to save space. The following is an example:

Indent style	Run-in style
dogs breed from 180-185 breeding clubs 242-248 obedience training 275-281	dogs: breed from 180-185; breeding clubs 242-48; obedience training 275-281

Special typography can also be used, such as italics to indicate the titles of publications or entries for illustrations. Bold can be used to indicate page numbers with key information about a subject. For example:

<i>ASAIB newsletter</i> 225 children's books 8, 14, 118-125 , 138, 177 papyrus 14, 82, 91, 95-99 , 107
--

Formulating Main Headings and Subheadings

There are no precise guidelines for formulating headings or subheadings. Individual indexers will probably have different entries for the same entity. However, there are guidelines that will help with the indexing task when it comes to the following points for example:

- personal names
- geographical names
- names of organisations
- abbreviations and acronyms
- international characters
- capital and small letters

The following tips might also help with the formulation of main headings and subheadings:

- Verify unknown terms.
- Verify inconsistent use of terms by the author.

- Use subheadings if there are more than five pages of references that must be included with the main heading. For example:

Kenya 34-38, 48-52, 60, 66, 70-71, 77-85
 Kenya
 British period 34-38, 60
 Early history 66
 economy 48-52, 60, 70-71
 Kenyatta administration 77-85

- An incomplete entry, such as 'Disraeli', must be verified and reformulated so that it reads 'Disraeli, Benjamin, 1st Earl of Beaconsfield'.
- Keep a balance between key information and less important information.
- Do not index negative information (e.g. do not make an entry for Nelson Mandela if the text indicates that Mandela's presidency is not covered in the study).

Not all entries can (or should) be exactly as they appear in the entity. Sometimes indexers use reference works to adapt index entries (e.g. 'Disraeli' can become 'Disraeli, Benjamin, 1st Earl of Beaconsfield').

Indicating Relationships between Entries

Entries in a book index are arranged in such a way that the relationship between main headings and subheadings is very clear:

learn to index
 bibliographies, 24-26
 correspondence courses, 45
 self-training, 45-47

Recording Entries

Write the entries on a card or use a computer program such as Macrex or CINDEK to generate the entries. Make separate entries, irrespective of the method used. This simplifies the alphabetical sorting of the main headings and subheadings.

Locators

The position where information about a particular heading or subheading can be found, usually given as page numbers, must be correctly indicated. The page numbers are sometimes separated from the heading or subheading by a comma. When a heading or subheading has more than one page number, the page numbers must be separated by a comma. A hyphen can be used to indicate a range of page numbers (e.g. 50-58, 102-110). For example:

indexing
 exhaustivity 43-45
 specificity 24, 34

Making Cross-references

There are two different types of cross-reference:

- *see* references
- *see also* references

See references guide the user from the term not used to the term preferred by the book index. For example:

American National Standards Institute *see* ANSI
 distance teaching *see* distance education

See also references guide users from an index entry to another related entry. There may be more than one *see also* reference for an entry; these are then arranged alphabetically and separated by commas. *See also* references can follow a main heading (i.e. after the page numbers if there are any), or they can be listed as a group of subheadings. For example:

libraries *see also* academic libraries, public libraries
 indexing
 as art 56
 history of 56-67
see also methods of indexing

General *see also* references are used to ensure that users consult the general and specific entries. For example:

churches *see also* names of specific churches
 universities *see also* names of specific universities

When making a cross-reference ensure that the entries referred to really exist. Avoid entries that run in circles. For example:

libraries *see* academic libraries
 academic libraries *see* libraries

Arranging Entries

The entries can be arranged alphabetically and there are two options: word-by-word or letter-by-letter. For example:

Letter-by-letter	Word-by-word
New Amsterdam Newark Newberry New Brunswick New England Newfoundland	New Amsterdam New Brunswick New England Newark Newberry Newfoundland

How to Work according to a Policy

Although there are no fixed rules for book indexing, one needs to clarify a number of issues before starting on the index. Consistency is essential. The indexer can work according to his or her own policy or according to the policy of the publisher, if specified. A policy lays down the rules and guidelines, which must be followed.

It is particularly important to be consistent about the following:

- level of detail provided
- level of vocabulary
- form of terms
- style of entries
- layout
- presentation
- spelling
- punctuation and spacing
- use of capital letters
- word order if there is more than one word
- order of headings and subheadings
- use and form of cross-references
- word forms for preferred terms
- forms of locators (e.g. references to page numbers)

There are many standards, textbooks and other sources that one can consult for information on these matters. See also chapter 15.

Specificity and Exhaustivity of Entries

Specificity of indexing refers to the degree to which an index entry matches the exact meaning of the subject concept (Cleveland & Cleveland 2001:259) or the vocabulary used in the entity. Exhaustivity refers to the degree to which the subject content of an entity is reflected in the entries. It also indicates how much information about a subject can be retrieved and how easily.

The more specific indexing terms are usually assigned (e.g. 'distance teaching' rather than 'teaching'). If a book deals with the purchasing policy of public libraries and academic libraries, the heading will be 'public libraries' and 'academic libraries'; these are preferred to the more general term 'libraries'. A cross-reference is made between 'libraries' and the

more specific terms, namely 'public libraries' and 'academic libraries'. Sometimes the more general term 'libraries' is also assigned if it will benefit the user. If a book deals briefly with a series of related terms, the more general term will be suggested (e.g. 'fruit' to replace 'oranges' and 'peaches').

See also chapter 4 where these two concepts are discussed in more detail.

Synonyms, Antonyms and Homographs

The policy must include guidelines to the handling of synonyms, antonyms and homographs.

Synonyms are terms with more or less the same meaning, such as trespasser/criminal or inflation/price increases. Antonyms are terms that have the opposite meaning, such as pass/fail. Such terms must be approached in a special way regarding headings and subheadings:

- One term is chosen as the preferred term for synonyms. The locator is indicated here. References are made from the non-preferred term. For example:

inflation 75-80, 86 price increases <i>see</i> inflation

Synonyms and near-synonym words can also be linked (e.g. libraries and librarians):

- For antonyms one of the preferred terms is chosen. A reference is made from the non-preferred term to the preferred term. Alternatively, an entry can be made for both terms. For example:

employed and unemployed 56, 60-65 unemployed and employed 56, 60-65
--

- For homographs the meaning of a term is indicated in brackets. For example:

Mercury (planet) mercury (metal)

Form of Headings and Subheadings

The form of headings and subheadings must be standardised; usually nouns are chosen. Pronouns start with capital letters, but other headings use small letters. Descriptive phrases can use subheadings. A heading usually starts with a noun, but sometimes an adjectival noun phrase is more specific and gives a more realistic reflection of the term preferred by the user.

Two related terms can also be combined. For example:

colleges and universities libraries and librarians

The terms will not be linked if the author differentiates between the terms, or if he or she provides a large amount of information about one of the terms, but not the other. Here a main heading must be made for each concept. The two entries must be linked with a *see also* reference.

Singular or Plural

The plural form is preferred for countable nouns (i.e. 'how many?'), such as libraries, dogs or books. Nouns for uncountable nouns (involving mass; i.e. 'how much?') use the singular form, such as gold and water. The singular form is also used for abstract concepts, processes, attributes, quantities and characteristics.

Cross-references

There are no definitive guidelines for making cross-references apart from the following:

- differences in spelling (e.g. alphabetizing/alphabetising)
- words and names in foreign languages
- archaic words
- synonyms and antonyms
- genus to subclass (e.g. fruit/citrus fruit/oranges)
- from subclass to genus (e.g. oranges/citrus fruit/fruit)
- name variations
- abbreviations and acronyms (e.g. ANC/African National Congress)

Editing the Text

The index must be thoroughly edited after all the entries have been made. Check for the following:

- consistency of entries
- duplication of entries
- proper and correct cross-referencing
- spelling (this includes consistency of spelling)
- correct page numbering (or other forms of locators)
- correct alphabetising
- adherence to the standard procedure for filing according to the ALA filing rules

It is very useful to have a second person help check the technical points. See chapter 15 for more detail.

Preparing the Final Index

The final index can be prepared after the draft copy has been proofread and checked. Do not start this process before all the errors have been corrected. The index should be checked

and rechecked thoroughly. Unfortunately, publishers do not always give indexers enough time to check their work.

And finally: always keep a backup copy of the index!

Indexing Software

Traditionally indexers indexed books manually on cards. Although cards may still be used, it is much easier to use an indexing program. Nowadays various single-purpose programs are available for indexing, such as CINDEK, Macrex and SKY Index. These programs carry out certain mechanical tasks, leaving the indexer to provide the intellectual input. Embedded indexing and automatic indexing are discussed in chapters 9 and 10 in detail. The following functions are supported by indexing programs:

- automatically displaying a group of entries that contain a common term or phrase
- checking cross-references
- sorting and arranging terms (including variations in the sorting sequence and arrangement styles)
- generating the printed layout chosen by the indexer
- checking indented and run-on entries
- allowing the indexer to determine spaces for subheadings and sub-subheadings

Word or text processing programs can also be used. Even if these have an indexing module, they are not as effective as single-purpose programs. Other possibilities include embedded indexing (Microsoft Word, Adobe PageMaker and markup languages such as SGML, HTML and XML) and tagging. Input is eased by the provision of a range of sophisticated facilities for

- repeating, rotating, merging, sorting, and styling, of the individual entries
- reviewing and amending the index during compilation
- checking the effectiveness of cross-references
- displaying and producing the final index in the required format (e.g. hard copy, disc, camera-ready copy, marked-up files for electronic transfer)

For more information about these programs, you may like to visit the following websites:

Idxtools <http://www.sid.com.ac.uk/ndv207/idxtools.htm>

CINDEK <http://www.indexres.com>

Macrex <http://www.macrex.com>

SKY Index <http://www.sky-software.com>

Master Indexing <http://www.mindexer@interconnect.com.au>

Although computer programs can help simplify the book indexing process and can improve productivity, human intellectual effort is still needed (Wellisch 1995:247). This statement is

still valid today. Not even the most sophisticated automatic indexing system has been able to produce an acceptable book index.

Conclusion

Book indexing is an important source of freelance work and is gaining in popularity. Good indexers with knowledge of book indexing are in demand with publishers. If one is planning to make a career of book indexing, various courses offered by indexing societies should be considered (for more information, see chapter 31 on training opportunities). The best way to keep up to date with developments in the field of book indexing is to regularly visit the various websites of book indexing societies. A number of associations focus on book indexing, for example

- The American Society of Indexers (ASI): www.asindexing.org
- The Indexing and Abstracting Society of Canada (IASC): www.indexingsociety.ca
- The Australian and New Zealand Society of Indexers (ANZSI): www.aussi.org
- Society of Indexers (SI) (UK): www.indexers.org.uk
- Association for Southern African Indexers and Bibliographers (ASAIB): www.asaib.org.za

BIBLIOGRAPHY

- Bell, HK. 1992. Indexing biographies and other stories for human lives. Society of Indexers Occasional Papers on Indexing, no. 1). London: Society of Indexers.
- Blakewell, KGB. 1987. Reference books for indexers. *The Indexer*, 15(3):131-140.
- Booth, PF & Piggot, M. 1988. *Choice and form of entries*. (Society of Indexers: Training in indexing – Unit 2). London: Society of Indexers.
- Borko, H & Bernier, CL. 1978. *Indexing concepts and methods*. New York: Academic Press.
- Chicago manual of style*. 1993. 14th ed. Chicago: University of Chicago Press.
- Cleveland, DB & Cleveland, AD. 1990. *Introduction to indexing and abstracting*. 2nd ed. Englewood, Colo: Libraries Unlimited.
- Cleveland, DB & Cleveland, AD. 2001. *Introduction to indexing and abstracting*. 3rd ed. Englewood, Colo: Libraries Unlimited.
- Fetters, K. 1994. *Handbook of indexing techniques: a guide to beginning indexers*. Rochester, NY: Fetters InfoManagement Co.
- International Organisation for Standardisation (ISO). 1996. *Information and documentation – guidelines for the content, organization and presentation of indexes*.
- Knight, GN. 1979. *Indexing, the art of: a guide to the indexing of books and periodicals*. London: Allen & Unwin.
- Lancaster, FW. 2003. *Indexing and abstracting in theory and practice*. 3rd ed. London: Facet Publishing.
- Lipetz, B-A. 1989. The usefulness of indexes, in *Indexing: the state of our knowledge and the state of our ignorance: proceedings of the 20th Annual Meeting of the American Society of Indexers*, edited by BH Weinberg. Medford, NJ: Learned Information:111-120.
- Mulvany, NC. 1994. *Indexing books*. Chicago: University of Chicago Press.

- Recommendations for the preparation of indexes to books, periodicals and other documents.* 1988. London: British Standards Institution. (BS 3700:1988).
- Thatcher, VS. 1995. *Indexes writing, editing, production.* Lanham, Md: Scarecrow Press.
- Wellisch, HH. 1995. *Indexing from A to Z.* 2nd ed. New York: HW Wilson.
- Wellisch, HH. 2000. *Glossary of terminology in abstracting, classification, indexing and thesaurus construction.* Medford, NJ: Information Today.



Ethleen Lastovica

Abstract

The indexing of electronic text is discussed. Indexers use a variety of software that has embedded indexing modules. Specific reference is made to the indexing of the *Business blue book of South Africa*.

Introduction

Embedded indexing is a paperless process that has been defined as the inserting, or embedding, of index entries into an electronic document (a computer file) with the software that was used to create that document. The process as described in this chapter is done manually by an indexer, unlike automatic indexing where keywords are automatically assigned to the text by software according to predetermined guidelines. Once subject entries have been embedded in the text, an index of these terms can be compiled to indicate on what page the terms occur (Mulvany 1999).

By considering the indexing of the *Business blue book of South Africa*, the preparation, processes and publication of an index for a conventional book using the techniques of embedded indexing are discussed.

Hiking on the Otter Trail

Embedded indexing can be compared to a hike on the Otter Trail. The guidebook for the five-day walking trail describes the terrain as one of undulating hills with steep cliffs and a magnificent coastline intersected by rivers entering the sea. This creates an image of happy hikers ambling through a pastoral scene with spectacular views. One needs to consider, though, that the cliffs need to be scaled in order to reach the best vantage point for the magnificent views, and that one might have to swim flooded rivers to reach the other bank.

There is an increasing amount being written about embedded indexing and its pros and cons, but until one is confronted with the prospect of undertaking this form of paperless indexing, the words of advice, and warnings about its pitfalls, are not fully appreciated.

Definition / Software

Embedded indexing is done by indexers who select index terms for inserting, that is embedding, into the text of a publication when it is still in electronic form. The software that was used to prepare the electronic document is used. The index can be compiled

once the whole document or publication has the embedded terms in place. Various word-processing and desktop publishing programs, such as Microsoft Word, WordPerfect, Ventura, FrameMaker, PageMaker and QuarkXPress have embedded indexing modules. Conventional book indexes are often created with dedicated indexing software – for example Macrex, CINDEK or SKY Index – but all three of these programs now include an embedded indexing facility.

Embedded indexing software is improving, but is far from perfect. It is important to be aware of the latest developments in the field by not only consulting indexing journals such as *The Indexer*, but also checking for new resources on the Web. Nancy Mulvany (1994, 1999), Jan Wright (1999, 2000), Peg Mauer (2000) and Seth Maislin (1999) are indexer/authors professionally involved in embedded indexing. While Mulvany covers embedded indexing in her 1994 publication, *Indexing books*, the reader must note that the book is 10 years old. More recently she published an article in *The Indexer* (Mulvany 1999) in which she compared the capabilities of four embedded indexing programs with Macrex, which is considered by many as the leading dedicated indexing program. The report duplicated tests done ten years earlier. It is startling to see how differently the five programs alphabetise a list of sample entries. Her conclusion was that, despite improvements, embedded indexing programs are still crude tools.

The type of person who should become involved in embedded indexing is someone who is not intimidated by computers and software, is methodical, and an expert file manager. The indexer undertaking the job should be able to stay calm as the complexity level of the task rises.

Business Blue Book of South Africa

The reality of embedded indexing will be put into perspective by describing the indexing of the annual *Business blue book of South Africa* that is published by National Publishing in Cape Town (2004). It is available as a book or on CD-ROM. The 2004 edition has 916 pages and a 151-page index consisting of approximately 29 000 entries. In the *Business blue book*, relevant facts and legislation needed to conduct business in South Africa are brought together in one volume. New legislation may be incorporated only days before going to press.

Sarah Maddox, a technical writer and indexer, and the former ASAIB representative in the Western Cape, started the embedded indexing project for the *Business blue book* some years ago. She realised the importance of sound documentation as a means of establishing standards and procedural guidelines for the team who work on the book, particularly the indexer. As successive indexers have undertaken the challenge of indexing the *Business blue book*, the procedural guide has proved invaluable, particularly as the editorial team is generally unable to provide help regarding the technicalities of indexing.

The publication is set up using the desktop publishing package QuarkXPress. The indexer uses an Apple Macintosh loaded with this software and Quark Index XTension to create the embedded index. The indexing takes about two months. It is essential that anyone undertaking an indexing assignment like this is totally familiar with the hardware and software used.

Preparations for Embedded Indexing

As with any indexing contract, speed of indexing is crucial. Computers with high capacity are essential because the files for an entire book created with a desktop package are usually large. A modem providing a link to the Internet and e-mail is essential. When quoting for the job one must not forget to include costs for online time because large files take a long time to download and return.

Discussions with the editor and production director must establish a sound understanding of what is required, and the administrative procedures for ensuring the safe electronic transfer of files between editor and indexer.

Conceptual Approach

If indexing an annual, it is advisable to understand the conceptual approach that previous indexers have had, and to become familiar with the terminology they used by studying earlier indexes to the work.

Each chapter of the *Business blue book* is a separate file. It is vital to work only on the current version of a file, which means the editor and indexer cannot work simultaneously on the same file. A progress chart, or spreadsheet, must be meticulously maintained to keep track of who has the current editing rights to a file. The spreadsheet should list chapter title and filename, and include columns for date received, date completed, index tested, date returned and comments.

Procedures

When the editing of a section or chapter of the *Business blue book* is complete and the pages of that section have been typeset, the QuarkXPress file is e-mailed to the indexer for indexing. The chapters may be sent in random order. Besides saving the incoming files on the computer on which the book will be indexed, it is advisable to keep secure originals of these files apart from that computer. 3.5 inch double-sided high-density diskettes (stiffies), CD-ROMs, a zip drive or a flash drive could be used for storing backups. Files should be saved frequently while working on them, perhaps every 10 minutes. At least once a day a backup must be made on one's preferred storage medium and kept apart from the computer in case of system failure. This may seem over-cautious, but should the working file become corrupted it would be time-consuming to reconstruct the index. Embedded indexing generally takes two to three times longer to do than conventional book indexing does with specialised indexing software.

Besides receiving files electronically, page proofs of the chapters are delivered to the indexer. Initially, some might find it easier to mark up the proofs with indexing terms that one intends embedding in the electronic file, but with a little practice it saves time to work directly on the file. The electronic text from earlier editions of a publication retains the embedded indexing terms that were originally inserted into it. The difficulty with working from proofs is that the indexer cannot see what has previously been indexed because, in print, embedded indexing terms are hidden text. Only new or updated text, which the editor

indicates by underlining, needs to be indexed. This is what publishers may consider as one of the advantages of embedded indexing.

Embedded indexing is cumbersome. Until the entire volume has been completed, only separate indexes for each file or chapter can be generated – the amalgamated index for the whole work is only created as the last process before going to press. It is therefore not easy to check, or even remember, where else a topic occurred and how it was indexed. This leads to inconsistency in the index. If indexing an annual, or a new, edition of a work, it may be possible to solve this problem to some extent by consulting the index of the previous volume. If that index was set in very small font, it helps to have an enlarged photocopy of it to make the text more legible. Indexers born into the computer age who have a large computer screen might find it just as easy to have the electronic file of the earlier index open while indexing the new edition.

Indexing Decisions

Many hours can be wasted considering how to deal with previous indexing decisions that one disagrees with. It is not good indexing practice to change the existing indexing terms in the current file or chapter if all other occurrences in the publication are not changed as well. Writers on embedded indexing generally consider changes like this to be impractical because the process would be time-consuming and confusing, particularly as one might not have the entire publication on file at any one time. The best approach is to follow the indexing decisions made for earlier editions. (Embedded indexing is about compromise.) When too many inconsistencies or inaccuracies build up in an index to a long-running publication, a stage will be reached when the document must be completely re-indexed.

Procedure for Indexing an Electronic Document

- Receive file from the editor.
- Update the index by embedding additional indexing terms into the text, or create a new index if the document has not been indexed previously.
- Create a test index, and proofread it.
- Go back to make changes to the embedded indexing terms if necessary.
- Return file with embedded index to the editor.

The embedded indexing terms may be invisible when the computer file is opened, but then the option to show commands should be selected. If this is a revised version of an earlier work, the editor indicates, by using a different colour, or font, what text has been added to the document. It is these segments that must be indexed. The layout defaults in the indexing software being used must be set up, including punctuation, indents and sort options.

Quark Index XTension

Using Quark Index XTension the indexer highlights the word, phrase, paragraph, section or chapter to which an index entry must refer. A pop-up menu allows the indexer to check the

index entry and to confirm or amend it, and define the level, that is a main entry (first level) or a subentry (second level). Page numbers are not assigned at this stage, only locators – the location of the index entry on a page. The tagged entry moves with the text to which it is assigned. This allows reformatting of pages after they have been indexed. Only when all the files are combined as one document, and pagination added at the end of the editing process, is the final index, complete with page numbers, created.

It is all very well to have editorial marks indicating what is new, but a problem arises because it is impossible to show what has been deleted. For this reason the indexer needs to scrutinise the text and list of entries assigned to it. If the whole section to which an indexing term referred was removed, there is no problem because the entry disappears from the already compiled list of entries. If, however, only the beginning or end of a section has changed or been removed, the start or close tags of the entry may be missing. The tagging of the entry will then have to be corrected.

When embedded indexing of a file is complete, Quark Index XTension provides an option to build an index for that file in layout style with temporary page numbers assigned to the entries. This provides the opportunity to check for inconsistencies and spelling errors which must be corrected in the text. The indexed file is then returned to the editor who is quite likely to make further changes to content or layout of the chapter. When all files are finalised, the editor generates the index for the whole book. Locators alongside the index entry are automatically replaced by page numbers, and so the index is ready for checking. Almost without exception, writers on embedded indexing consider this stage to be the most time-consuming of the embedded indexing process. For the *Business blue book* this entails checking for inconsistencies and misspellings across the 78 chapters that were separately indexed. Any errors noted must be corrected within the chapter concerned. The complete index must be regenerated if corrections to embedded indexing terms are made. This is the final stage of the indexer's commitment to the printed version of the publication.

CD-ROM

The *Business blue book* is also published on CD-ROM, and so the indexing process starts again, but this time a microcomputer and a range of Adobe software to create hyperlinks and bookmarks are used. As with the printed version, a progress chart for recording the receipt and return of files is important and regular backups of work in progress are essential. A detailed procedural guide for indexing the CD-ROM must be written to enable someone new to the job to proceed with confidence. In addition the indexer must write technical notes to explain to the readers how to use the CD-ROM.

All files that were created for the printed book are converted to PDF and then reformatted to fit the shape of a computer screen. Only an outline of the steps taken to hyperlink and bookmark the CD-ROM version of the publication follows.

Steps to Hyperlinking an Electronic Document

- PDF files (readable with Acrobat Reader) are supplied by the publisher.
- Files are processed through a template using Acrobat Exchange.

- Text is hyperlinked and bookmarked with Adobe Exchange. This will enable a reader to find a reference in the text by clicking on an entry in the bookmark menu.
- Attention is paid to special hyperlinks for certain files (e.g. e-mail links and forms).
- A simple index to subsections of each Act or chapter is prepared.
- When complete, the simple index is sent to the publisher for conversion to house style and PDF format.
- The index in PDF format is returned to the indexer for hyperlinking and bookmarking.
- Next, Acrobat Catalog, one of the programs supplied with Acrobat Exchange, is used to create an index that enables searches for words or phrases in the *Business blue book* files on the CD-ROM.
- All the hyperlinked PDF files and those created when building the search index are zipped and returned to the publisher for writing to the CD-ROM.

Task complete!

BIBLIOGRAPHY

- Business blue book of South Africa – BBB*. 2004. Edited by L Mitchell. Cape Town: National Publishing. 65th ed.
- Maislin, S. 1999. Ripping out the pages, in *Beyond book indexing*, edited by D Brenner and M Rowland. Medford, NJ: Information Today:43-57.
- Mauer, P. 2000. Embedded indexing: pros and cons for the indexer. *The Indexer*, 22(1):27-28.
- Moncrief, L. 1999. Indexing computer-related documents, in *Beyond book indexing*, edited by D Brenner and M Rowland. Medford, NJ: Information Today:13-24.
- Mulvany, NC. 1994. *Indexing books*. Chicago, IL: University of Chicago Press.
- Mulvany, NC. 1999. Software tools for indexing. *The Indexer*, 21(4):160-163.
- Wright, JC. 1999. Indexing after the millennium 1: getting the tools right. *The Indexer*, 21(1):19.
- Wright, JC. 2000. The world of embedded indexing, in *Beyond book indexing*, edited by D Brenner and M Rowland. Medford, NJ: Information Today:3-12.