Deconstructing indexing standards

Drusilla Calvert

Discusses the usefulness, practicality and clarity of standards for indexing with particular reference to the differences between BS3700:1988 and the forthcoming ISO999. Based on a paper given at the Australian Society of Indexers' International Conference 1995.

Standards for something as subjective as indexing exist at the very most as the basis for discussion and as an exemplification of good practice. They are not there for promoting consistency for its own sake. The needs of the user must always have priority — and if the user understands an apparently 'inconsistent' index better, then that is what should be offered.

Most indexers will quite happily follow a publisher's house style, although they might point out that, for instance, normal practice is not to capitalize the first letter of all headings. Without getting into impossible levels of complexity, all possible permutations of indexing cannot be covered by a standard.

What exactly is 'wrong' indexing practice? I think we would all agree that circular and blind cross-references are wrong; that a reference to page 57 when the material referred to is on page 75 is wrong, but what else is wrong? We look at an index and decide from our experience — and prejudice — whether it is a good or bad index. Experts often disagree on the quality of an index . Until recently, there has been little useful research, so we are left with nationally or internationally supported standards (based on the collective opinion of representatives of indexing and related professions) to give us something to go on.

This paper looks at standards with particular reference to the forthcoming International Standard ISO 999 (ISO 1996) and the British Standard BS 3700:1988 (BSI 1988) which has been used as the basis of ISO 999.

Should an indexing standard necessarily become out of date? Is there any reason why good practice in 1900 (or even 1600) shouldn't be good practice in 1996? The media may change, but does the basis on which we make indexing decisions change? How much should we adapt standards to take account of automated sorting, to accommodate non-displayed as well as displayed indexes? Is machine indexing a completely separate activity or can the same principles be applied to it as to human indexing?

All standards are in fact 'out of date' by the time they appear — in that the decisions were made several years before publication. Some recommendations are obviously dated: BS has recommendations about the 'typescript' and a 'keyboard operator working at speed' which are omitted in ISO; BS mentions preparing copy for publishers while ISO broadens this to 'final processing'.

ISO 999 (ISO) differs from BS 3700 (BSI) in various ways, and it is in looking at these differences that we can dissect the specific provisions and consider their relevance and practicality. I discuss below a personal selection of the similarities and differences.

Terminology:

First of all, the two are really very similar. The differences lie initially in changed terminology; for example — there is no reference to 'thesauri' in ISO. References to thesauri in BS 3700 are replaced by references to authority files and controlled indexing languages; 'location references' (BS) are replaced by 'locators', and there are additional mentions of 'collection indexes' (e.g. indexes to collections of abstracts, or library materials) throughout ISO.

The actual definition of an index is different:

BS: 'A systematic arrangement of entries designed to enable users to locate information in a document.'

ISO: 'Alphabetically or otherwise ordered arrangement of entries, different from the order of the document or collection indexed, designed to enable users to locate information in a document or specific documents in a collection.'

ISO also introduces 'qualifier' and 'scope note' and defines these concepts.

Quality:

On the quality of an index the emphasis is changed in ISO:

BS: 'The indexer's knowledge of the subject matter and language of the document indexed will clearly affect the quality of the index, as will the quality of the indexing tools used, e.g. thesauri and other guides'.

ISO: 'The indexer's knowledge of the principles of indexing, gained through study and experience, determines the quality of the index. Other important factors include the indexer's knowledge of the language and subject matter of the document and the quality of indexing tools used, e.g., controlled indexing languages'.

Consistency:

BS: 'consistency is more likely to be achieved if
a) a reliable thesaurus is used; ...'

ISO: 'consistency is more likely to be achieved if
a) indexing policies and authority files are established and followed;
b) reliable indexing resources are used, e.g., dictionaries, controlled vocabularies, consultations with experts; ...'

Again on consistency ISO adds references to the importance of checking and proofreading as well as editing.

Singular or plural terms:

The choice of form of heading as singular or plural is modified for different languages; for instance German and French use the singular for all headings, whereas English language practice is to use plural for countable objects and singular for non-countable concepts.

Incidentally, the International Standard gives no guidance on

non-roman scripts except on conversion of words from other writing systems (e.g. Japanese kana script to Roman script — 7.3.7) and there is only passing reference to languages other than English, French and German.

Personal names:

ISO recommends using 'generally' as full a form as possible. BS and ISO both recommend using the form used in the document — ISO limits this recommendation to 'single document indexes' and adds 'when preparing a name authority file for the indexing of multiple documents, indexers should choose the most recent, or the most commonly used, form of personal name as the heading and add "see" cross-references from other forms.'

This recommendation, which goes back to AACR2 (LA 1976), is obviously sensible, but in practice I find it is often very difficult to know — or to have any way of finding out — details of a new or unfamiliar author, or even his nationality.

Capitalization and inversion of articles:

Both BS and ISO recommend inversion of the article as in:

BS: 'Ideal Husband, An'

ISO: 'hunting of the snark, The'

Incidentally, this last use of capitalization is sanctioned by BS 1749 (Standard on alphabetical order and the filing order of numbers and symbols) on which BS 3700 is meant to be based, but doesn't appear in BS 3700 itself.

Elided numbers

BS makes no specific mention of the treatment of numbers such as 243-244, but, as we shall see later, uses the form of Oxford University Press' Hart's Rules in examples (the above example would be 243-4 in this format).

ISO specifically states that numbers should not be elided, because maximum clarity is achieved by presenting first and last numbers in full. I don't agree with this. Apart from the waste of space involved in including numbers in full, what is most important in an elided number is the first page number, followed in importance by the fact that this extends to other page numbers, followed again by the actual number on which the passage ends.

Both BS and ISO agree that 3*ff* and 3 *et seq* as well as 3-5*passim* are to be avoided. There has been some correspondence in the (British) Society of Indexers' *Newsletter* recently in which several respected indexers have defended *passim*. Perhaps we shall see this resurrected in a future standard, if users remember what it means by then.

Filing rules

Filing, or sorting, is a hornet's nest. All standards seem to disagree with all others. The authors of BS 1749 (the filing standard adopted by BS 3700) were accused of being over-influenced by library cataloguing. In the early days of computer enthusiasm some sorting systems recommended copying, rather than manipulating, ASCII sorting. This faintly ludicrous idea has been abandoned by indexers, but is still a feature of some older database programs, so it needs to be dealt with, as it is in ISO 999.

It is in the section on filing that BS and ISO differ radically. The British Standard is concerned primarily with alphanumeric sorting and has a detailed order for filing characters:

'The basic order of filing characters should be:

a) spaces, dashes, hyphens, diagonal slashes, all of which have equal filing value; followed by

b) ampersands, unless it is considered more useful to file as spelt out in the appropriate language; followed by

c) arabic numerals and Roman numerals; followed by

d) Roman alphabet letters

Other characters

a) the apostrophe should be disregarded and the word in which it occurs treated as one word

b) punctuation marks other than those listed (above) should be disregarded unless given a function by the indexer, as, for example, to separate a surname from a forename or to enclose a qualifier. They then acquire a filing value imposed by the indexer

c) signs and symbols should be treated according to their importance in the text and the likelihood of their being found in an alphabetical sequence. If they are few, they may be arranged as if spelt out in words. If many, a conventional order may be used as demonstrated in 4.3 in BS 1749:1985.

d) modified, additional and combined Roman alphabet letters used in languages other than English should be filed as the nearest equivalents of the English alphabet.'

ISO 999 synthesizes this into general filing rules:

'For either machine or manual sorting, all characters and symbols normally have a filing value. These filing values may derive from an established system such as ISO 646, but such systems, designed for character representation rather than for sorting, generally file all upper case letters before all lower case letters, and intersperse accented letters with non-alphabetical symbols. For sort purposes, therefore, upper and lower case versions of the same letter should be given identical values so that a single alphabetical sequence is obtained. Similarly, when modified characters are used, e.g., Ä, Â, â, à, å, á, ã, they should be given values to enable them to be sorted according to local practice.

Punctuation symbols used to distinguish types of index headings may be given special values in order to sort the headings in the order required.

Symbols and numerals which are to be ignored in filing should be given a null value.

The value given to a blank space will depend on whether letter-by-letter or word-by-word order is required.

Software used for index sorting should be capable of accommodating all the above requirements, if necessary by permitting the user to edit the sort translation table.'

The idea of a sort translation table (*Figure 1*) is that any character can be given any filing value. This means that the sort translation table can be adapted for, for example, Scandinavian indexes as well as giving a specific meaning to any symbol.

The thorny (for British indexers) problem of Mc and Mac in BS is omitted in ISO, which obviously has to address not only the entire English speaking world but also different languages.

Word-by-word sorting is preferred by both standards, although the Chicago manual of style (still the most influential standard for the United States) still prefers a hybrid form of letter-by-letter.

Arrangement of subheadings:

The ordering of subheadings has been the cause of many arguments. Both BS and ISO state:

'within an alphabetical index, subheadings may occasionally be more suitably arranged by a method other than alphabetical'

and later:

'The main alphabetical order of an index may be helpfully modified by numeric, chronological or some other systematic arrangement ...'

The examples of non-alphabetical order in the British standard are not particularly helpful, for example, Henry IV comes before Henry VII in alphabetical order as well as a numeric order.

I think it might have been more helpful to give examples where alphabetical order is quite out of the question. For example, any string player would be completely bewildered at finding the order:

cello, double bass, viola, violin

rather than size/ pitch order:

violin, viola, cello, double bass.

Similarly, (to keep to the musical theme) consternation would be caused by a sequence like this:

duos, nonets, octets, quartets, quintets, septets, sextets, trios

Let alone this sequence:

eight, five, four, nine, one, seven, six, three, two, zero

Similarly, the readers of a biography would be confused by the following:

Jones, Henry
adolescence 4-6
birth 2
childhood 3-4
death 445
marriage 230

Subheadings within a biographical entry in practice often combine two kinds of sorting within a single entry: chronological for biographical information, alphabetical for characteristics or the author's works.

This is by no means an exhaustive list of the range of non-alphabetical orders which we imagine is expected by users.

Computer software has to be flexible enough to manipulate the order which the user will be most likely to look up — and this is always going to depend on the context, so each instance has to be assessed independently.

Presentation of indexes

Both BS and ISO to some extent leave the format to the 'final processor'. It is recommended that, even if a run-on layout is required, the indexer should produce a set out layout, since (presumably) this is so much easier to check. As you will see later, this puts a great responsibility on the publisher or final processor. ISO says that the indexer should notify the publisher of possible errors or inconsistencies in documents, but to what extent should we as indexers be expected to proofread?

There are many problems which we have faced as a result of writing an indexing program (MACREX). Although we strive to give choices for every conceivable situation, we do have to make decisions on all kinds of things which are not covered by standards. How, for instance, do you deal with the fact that a hyphen is the same character as a minus sign and is often used as synonymous with an en-dash? How do you decide on the order of a sequence like this?

xiv, 14, **14**, *14*, 14n

The answer is that in the former case you make a decision based on the fact that minus signs are unlikely to be used in indexes as frequently as hyphens, and in the latter case you take the advice of the person who presented you with the problem and hope no-one will say you're wrong. (This is a real example.) All we can really do is to give as many choices as possible and use as the 'default' what we hope is the most likely form.

Cross-references:

Both BS and ISO deal with the definition, content, nature and positioning of cross-references, both 'see' and 'see also'. Neither mentions 'see under' or 'see also under'. (In the examples below ISO text differing from BS text is shown in parentheses.) In regard to the position of cross-references, both BS and ISO say:

'see also' cross-references should normally follow the location reference(s) (locator(s)) relating to the heading or subheading from which they refer'

ISO adds a reference to multiple headings:

'Where the direction of the cross-reference is to multiple headings, these should be listed in alphabetical order, separated by semicolons e.g. bears 100, 217, 923 see also badgers; koala bears; raccoons'

BS/ISO: 'see also' cross-references should precede location references (both locators and subheadings) in those types of index where they may be overlooked or found only after perusing unwanted references ... (In this case they should be clearly distinguished from the remainder of the entry, e.g. by enclosing them within parentheses.)

Here are some alternatives:

1. "see also" following page numbers as part of main heading

economics 33, 144, 195, 229, 499, 502 *see also* assets; banking; business firms; commerce; transport; wealth
bibliographies 208
mathematical models 160
statistics 155

2. ISO with recommended turnover indent

economics
(*see also* assets; banking; business firms; commerce; transport; wealth) 33, 144, 195, 229, 499, 502
bibliographies 208
mathematical models 160
statistics 155

3. BS usage

economics 33, 144, 195, 229, 499, 502
bibliographies 208
mathematical models 160
statistics 155
see also assets; banking; business firms; commerce; transport; wealth

This leads us on to:

Set out and run-on subheadings

set out layout

Here is the BS example of set out layout.

Aristotle:
debt to Plato 23, 46
literary criticism in 35
74, 89-93,
101-97
on Aeschylus 101-4,
279
on Aristophanes 195
on Euripides 104-26
187 , 265-6
on Homer 103,
190-4, 206
on Sophocles 127-83
275-7, 306,
309-10
Antigone 155
Oedipus Tyrannus
140-9
origins of tragedy
in epic 196
in revelry 197

However, the colon after the main heading is not followed in the index to the standard itself, 'squashed' or 'elided' page numbers follow Oxford University Press' Hart's Rules and turnover lines are indented further than the deepest subheading (again not followed in the index to the Standard)*.

The reason for the highly indented turnover lines is to avoid confusion. The following alternative does not follow this practice: is the extra 'confusion' caused worth setting against the saving of two lines (19 as opposed to 21 using the same line width — 1900 as opposed to 2100 lines)?

Aristotle
debt to Plato 23, 26
literary criticism in
35, 74, 89-93, 101-97
on Aeschylus 101-4,
279
on Aristophanes 195
on Euripides 104-26,
187, 265-6
on Homer 103, 190-4,
206
on Sophocles 127-83,
275-7, 306, 309-10
Antigone 155
Oedipus Tyrannus
140-9
origins of tragedy
in epic 196
in revelry 197

Run-on layout

The following is what is given in BS3700 as the 'run-on' layout (publishers and final processors please note). As is pointed out in the draft American standard (ANSI/ NISO 1996?) this is no such thing: it is a hybrid — the first level of subheading is set out, the second level of subheading (sub-subheadings) is run on and the third level of subheading (sub-sub-subheadings) is run on from the second level, using parentheses to distinguish this level from the previous one.

The differences between standard and index were because BSI editors insisted on following their own house style rather than the standard or the indexer's and committee's wishes — Ed

Aristotle:
debt to Plato 23, 26
literary criticism in 35,
74, 89-93, 101-97; on
Aeschylus 101-4, 279;
on Aristophanes 195; on
Euripides 104-26, 187,
265-6; on Homer 103,
190-4, 206; on
Sophocles 127-83,
275-7, 306, 309-10;
(*Antigone* 155; *Oedipus*
Tyrannus 140-9)
origins of tragedy: in epic
196; in revelry 197

What interests me is what happens in the rest of the index. Say you have another entry:

Plato, Aristotle's debt to 23, 26

Plato, philosophy 224-38

How do you incorporate that in the above layout? Like this:

Plato:
Aristotle's debt to 23, 26
philosophy 224-38

— that is, a standard set-out layout; or like this:

Plato: Aristotle's debt to 23,
26; philosophy 224-38

— run on from main heading, or like this:

Plato:
Aristotle's debt to 23,
26; philosophy 224-38

— run on from first subheading?

And what about:

tragedy, origins, in epic 196
tragedy, origins, in revelry 197

Would you set it out like this:

tragedy:
origins: in epic 196; in revelry 197

— run on from first subheading, or like this:

tragedy: origins (in epic 196; in revelry 197)

— run on from main heading? or like this?:

tragedy:
origins: in epic 196; in revelry 197

The ANSI draft simplifies matters by saying that run-on layout, where required, should be limited to a heading and single level of subheading:

Aristotle: debt to Plato 23,
26; literary criticism in
35, 74, 89-93, 101-97;
origins of tragedy 196

and that a so-called 'hybrid indented/ run-on layout' like this can be used when there is a further level of subheading: