What is a thesaurus?

Leaving aside the general dictionary definition (which tends to refer to a 'storehouse of knowledge'—true, but not helpful on this occasion), it is most appropriate for our purposes to start by thinking of a thesaurus simply as 'a list of verbal terms, usually in alphabetical order, usually representing simple classes and with indications as to the relationships between the classes'

In other words, it is an A-Z list of words (single words or words in phrases) each selected to be the preferred term or descriptor representing a particular concept, and with the immediate hierarchical (a priori) relationships of each term shown against it. Supplementing these and integrated with them in the same alphabetical sequence are the other verbal terms by which those concepts may be known (such as synonyms, near-synonyms, alternative spellings, and abbreviated forms) and which may be used as first look-ups by searchers for information; each of these non-preferred terms (non-descriptors, lead-in terms) is listed with a reference to the preferred term which should be used for it.

Architecture
   NT Buildings
Buildings
   BT Architecture
   NT Civic buildings
      Industrial buildings
Civic buildings
   UF Civic centres
      Town halls
   BT Buildings
Civic centres USE Civic buildings
Factories USE Industrial buildings
Industrial buildings
   UF Factories
   BT Buildings
Town halls USE Civic buildings

In this example the preferred terms in the list are printed in bold type, the non-preferred in light type; some thesauri distinguish one from the other by the use of capitals for preferred terms, others make no distinction. The hierarchically-related classes are indicated here by BT (broader term) and NT (narrower term); in some thesauri the direction see also is used instead, with no distinction between the broader and narrower classes. Here the non-preferred terms carry the instruction USE, leading to the preferred term; an alternative is to use a see reference. Preferred terms in this list are followed by the non-preferred terms that they are used for (UF). Other (non-hierarchical, but conceptually associated) relationships can be indicated by RT (related term), as in Birds RT Ornithology, or by a see also reference.

The level of specificity in the example is low—the concept of 'town halls' not being distinguished from that of 'civic centres', for example, both being represented by the preferred term 'Civic buildings'. For any thesaurus the compilers will choose a level of specificity suitable for the purpose and so in another case we might find further distinctions made (narrower classes specified).

Civic buildings
   BT Buildings
   NT Civic centres
      Town halls
Civic centres
   BT Civic buildings
Factories
   BT Industrial buildings
Industrial buildings
   BT Buildings
   NT Factories
Town halls
   BT Civic buildings

The thesaurus, as described so far, can be seen as performing the same function as, and being in similar form to, the subject headings lists used for compiling alphabetical subject catalogues in libraries (such as the Library of Congress subject headings used by some university libraries); other compilations known as term lists, term indexes, lists of descriptors, keyword lists, controlled vocabularies and other similar names can all be regarded as thesauri of one kind or another. Many thesauri, however, consist of more than a straight alphabetical listing of terms; they contain also a hierarchical display in which the preferred terms are set out to show the complete subject coverage of the thesaurus in a

The Indexer Vol. 15 No. 3 April 1987
systematically arranged (classified) manner. They consist, in effect, of an alphabetical list of terms and a faceted classification of those terms.

**Architecture**
- Buildings
  - Civic buildings
  - Civic centres
  - Town halls
- Industrial buildings
- Factories

In this example, indentation is used to show the subordination of one class term to another. Some thesauri show subordination by typographical variation (capitals, bold, italic and so on), while others (like EUDISED\(^3\)) provide a graphic display to show connections between terms. By reference to the hierarchical display, a user can easily see which concepts are provided for within a particular subject area and the indexer maintaining the thesaurus can add new chosen terms in their correct (logical) positions. In some cases (such as the BSI *ROOT thesaurus*\(^4\)), this hierarchical display is the major part of the thesaurus, containing the greatest detail about terms, their equivalents and contexts; the alphabetical listing then acts as an index to the hierarchy.

Some thesauri (the BSI *ROOT thesaurus* is again an example) are also intended to act as a means of entering details of documents in a classified catalogue and of shelf location and therefore the hierarchical display is allocated notation (numbers or letters or both) so that each concept has its own *class number*.

It is important to realize that an alphabetical thesaurus, even when not produced with an accompanying hierarchical display, should be firmly based on an underlying classification or set of classifications; only then can the subject field be comprehensively represented, term relationships accurately indicated and new concepts incorporated in their proper places as the thesaurus is developed. (Specialized thesauri cannot be left unrevised for more than a few years, because of new technical developments in their subject fields, changes in the subject areas of interest among their client groups, newly coined terms, changing acceptability of abbreviated forms and so on.) The alphabetical listing is, in effect, a classification scheme with the terms in indicative order (A–Z) rather than systematically set out, and the better the underlying classified structure, the greater the potential of the thesaurus.

A well founded thesaurus includes a comprehensive listing of relevant terms whose meaning is usually indicated by the structure, but with supplementary scope notes (sometimes indicated by SN) giving definitions where needed, contextual notes for homographs (words spelt the same but with different meanings) in order to distinguish them, directions for consistent choice of one term from several with the same (or almost the same) meaning. These features are dealt with further under *Thesaurus use by indexers* below.

The purposes of thesauri in the information retrieval context are set out in British Standard 5723:1979, *Guidelines for the establishment and development of monolingual thesauri*; these relate particularly to the need to use controlled vocabulary in information retrieval systems and to ensure consistent practice between indexers for those systems. This standard gives guidance on the construction of thesauri, but states that it is not concerned with the preparation of back-of-the-book indexes, though the procedures recommended in it may be useful in that context.

The major indexer's standard, BS 3700:1976, *The preparation of indexes to books, periodicals and other publications*, directs in clause 5.1.3 that indexes should be compiled using a logical, balanced and consistent pattern which can be easily recognized by users in the field, and it is in this respect that the recommendations for the construction of thesauri are found useful. For example, guidance is given on selecting the most suitable word-form for an indexing term, the inclusion of abbreviations and acronyms, the treatment of homographs, variant spellings, loan words, slang terms and transliterated terms.

Other publications giving advice and instruction on the making of thesauri are Atchison and Gilchrist *Thesaurus construction?*, Orna *Build yourself a thesaurus?* and Townley and Gee *Thesaurus-making?*; indexers responsible for developing a system to index and retrieve information from a collection of documents will find these useful.

The particular focus of this article is the uses to which existing, published thesauri may be put in the indexing of a single book or periodical; it will be helpful first to give a few examples of the lists currently available.

### Published thesauri

There is a large number of thesauri in publication, each one typically covering a restricted subject area. Some have been compiled for use specifically within particular industries and therefore employ the technical vocabulary of those industries; others are intended for use with certain forms of document, such as audiovisual materials. One which has a wider coverage is the BSI *ROOT thesaurus*\(^4\). Though its major provision is for engineering and technological subjects, it includes some terms for the social sciences and humanities.

Most thesauri are limited to a single discipline, field of knowledge or area of activity (such as education in EUDISED\(^3\), health sciences in DHSS-DATA thesaurus\(^10\); and broadcast news and current affairs in BBC *Data's thesaurus of terms*\(^11\)) and have been developed to provide (often using computerized systems) for the infor.
mation retrieval needs of user groups specializing in one subject or a collection of related subjects. These needs are not always adequately provided for by the large, general schemes such as the Universal Decimal Classification\(^2\), the Dewey Decimal Classification\(^1\), or the Library of Congress subject headings\(^3\). It should be noted, however, that a general system can help to provide terms and structure for systematic and alphabetic displays in specialized thesauri; Aitchison\(^14\) has recently described how the Bliss Bibliographic Classification\(^15\) was used for the development of thesauri such as the DHSS-DATA thesaurus\(^16\).

Lists of thesauri added to the thesaurus reference collection in Aslib's Information Resources Centre are published in the September and March issues of the Current awareness bulletin\(^17\); this collection now contains many specialized listings and is a valuable resource of technical terminology.

**Thesaurus use by indexers**

Thesauri can be used by book and periodical indexers in several ways, grouped here under four headings: subject familiarization, meaning and authenticity checks, vocabulary control, and standard vocabulary for successive indexes.

**Subject familiarization** Although indexers may aim or tend to specialize in a particular subject field, most will from time to time need to index books on subjects with which they are less familiar. Background reading in encyclopaedias and standard texts can provide an appropriate information base from which to operate, but there may still be an uncertainty regarding specialized terminology and the pattern of conceptual relationships. Examination of the hierarchical display in a thesaurus for the subject area concerned is helpful in providing a 'picture' of the subject and its major concerns, with an overview of the constituent classes and their relationships to each other; the relative significance of terms can also be evaluated, giving the indexer some guidance as to what should be indexed.

**Meaning and authenticity checks** Dictionaries and glossaries—both general and specialized—are part of the indexer's reference library, but because of the time-consuming nature of compilation and publication are always slightly out of date as far as neologisms and new meanings are concerned. Thesauri, because of their origin in the information retrieval needs of specialized groups, are often more rapidly published, more comprehensive, more detailed and more frequently updated (by new editions or by supplements) and therefore give the indexer a surer base from which terms can be selected for a book or periodical index.

In the instance of texts in which authors coin their own words and use existing terms idiosyncratically, indexers can refer to thesauri to check on usage and to obtain guidance on the terms which would be acceptable to, and used by, workers in the field. Those workers will be using their normally preferred terms when they consult an index, not the various 'eccentric' versions used by authors; the authors' terms will obviously be the preferred terms for references in their own book's indexes, but the normal terms must be included as lead-in terms with a see reference.

**Vocabulary control** It is difficult for indexers to be totally familiar with all the technical vocabulary of each subject area within which they work. Thesauri can assist in extending their knowledge because they can give, for any technical term in their special fields, its context and meaning, its synonymous and abbreviated forms, and its relationships (hierarchical and associative) with other listed terms.

Thesaurial listings should not, of course, be slavishly followed—it is always the text being indexed which principally determines the identity of the preferred terms included in the index. They do show, however, for their subject fields, the accepted practices for forms of word (such as singular or plural, noun or verbal phrase), spelling variations, abbreviations and acronyms, capitalization, punctuation, word order in phrases (direct or inverted), and treatment of homographs, as well as listing synonymous and near-synonymous terms and showing term relationships (suggesting additional see and see also references to the indexer).

When indexing texts written by and for specialists in both the UK and the US, the problems of alternative British and American spellings and of different terms used for the same concept need consideration. Some thesauri published for use in the two countries include these alternatives and can thus provide the necessary supplementary vocabulary for the index; in other cases, it may be necessary to locate two thesauri—one from each country—and check one against the other for the terms needed.

The majority of thesauri published are of the monolingual variety but some, for example EUDISED\(^1\), contain terms in several languages; these can be invaluable for checking equivalents when indexing specialized texts containing numerous foreign-language terms which may not be found in general foreign-language dictionaries.

**Standard vocabulary for successive indexes** Users consulting indexes to several volumes of a publication such as a periodical can find the changes in terminology which occur an impediment to easy searching and reliable retrieval. Use of a thesaurus for the appropriate subject field can enable the indexer to supply a consistent indexing vocabulary, while still using the textual terms of the periodical articles in the usual way. For subject fields in which the major information services and retrieval systems use a single standard vocabulary, such provision is a valuable aid to users, easing the transfer from one information source to another and making it possible to search more fluently and effectively.
Index quality

Use of thesauri (as well as other reference tools such as lists of proper names) can help indexers to improve the quality (and therefore performance) of their indexes, with particular regard to the selection and form of terms. One of the virtues of the ideal index looked for by the Society of Indexer’s assessors, as described by Jean Simpkins\(^1\), is that the terms selected for main headings are ‘sensible, suitable to the likely readership, and mutually exclusive’, with ‘every rejected term and synonym . . . covered by a cross-reference’. Indexers who get to know the thesauri and classification schemes in their fields will find this virtue easier to acquire.

References


Protest of a frustrated user

Presently four small men in yellow came into the room stooping and staggering under the weight of an ancient and enormous volume on Sorcery. They set it down on the floor and left the room and the Royal Recorder began to turn the dusty yellow pages, squinting and coughing.

‘There’s nothing under Deer,’ he said at last, ‘except the record and report of maidens changed to deer and back again, all nice and neat and normal, all nice and neat and formal. In every single case the disenchanted lady knew her name. It says so here.’

‘Look under something else then!’ roared the King.

‘Such as?’ the Royal Recorder asked. ‘For instance and example?’

‘Look under Loss of Maidens’ Memory!’ bawled Clode. ‘Look under L!’

‘Loss in this especial case comes under M,’ the Royal Recorder said.

‘How does and can and could it?’ thundered Clode.

The Royal Recorder’s voice was prim and firm, ‘The proper listing, Sire, is, colon, quote, Memory, Maidens’, Loss of, stop, unquote.’

The King closed his left eye and then opened it and closed his right eye. His voice was low and ominous. ‘I do not have the slightest doubt that Pussy-cat comes under Q and Monkey-muddle under R and Donkey-daddle under S. Look up the sorry secret of this nameless child under X or Y or Z, but look it up!’ The shield on the wall trembled.

The Royal Recorder turned to the M’s and ran through Magic, Miracle, Mystery, and Mumps before he found at last what he was looking for, wedged in between Mice and Mountebank.

—from The white deer by James Thurber (1945), quoted by permission of Hamish Hamilton Ltd.

Pat Booth is a freelance information consultant, indexer, lecturer and writer, with previous experience as a special librarian and a teacher of information retrieval studies.