The Emerald Research Register for this journal is available at
www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0024-2535.htm



DIGITAL DIRECTIONS

# Thesauri: practical guidance for construction

Emma McCulloch

*Centre for Digital Library Research, Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK*

Received 29 March 2005
Reviewed 08 April 2005
Revised 25 April 2005
Accepted 27 April 2005

## Abstract

**Purpose** – With the growing recognition that thesauri aid information retrieval, organisations are beginning to adopt, and in many cases, create thesauri. This paper offers some guidance on the construction process.

**Design/methodology/approach** – An opinion piece with a practical focus, based on recent experiences gleaned from consultancy work.

**Findings** – A number of steps can be taken to ensure any thesaurus under construction is fit for purpose. Due consideration is therefore given to aspects such as term selection, structure and notation, thesauri standards, software and Web display issues, thesauri evaluation and maintenance. This paper also notes that creating new subject schemes from scratch, however attractive, contributes to the plethora of terminologies currently in existence and can limit user searching within particular contexts. The decision to create a "new" thesaurus should therefore be taken carefully and observance of standards is paramount.

**Practical implications** – This paper offers advice to assist practitioners in the development of thesauri.

**Originality/value** – Useful guidance for those practitioners new to the area of thesaurus construction is provided, together with an overview of selected key processes involved in the construction of a thesaurus.

**Keywords** Digital libraries, Information retrieval, Controlled language construction

**Paper type** Viewpoint

## Introduction

It is widely recognised that, in many cases, thesauri increase the effectiveness of online information retrieval, resulting in improved levels of precision and recall of user search results (Shiri *et al.*, 2002). Thus, digital libraries are now beginning to incorporate such tools into their services and collections, with the aim of improving both the organisation of information within, and to increase the effectiveness and efficiency with which users can retrieve the information they desire. Indeed, even a decade ago Kosovac (1995) acknowledged that "an increasing number of home-grown, commercial, and research-based systems integrate them in different ways".

How do organisations best select or develop a thesaurus that is fit for purpose? Since no single thesaurus can effectively describe all concepts universally, tailored and subject-specific term sets are required. Much has been documented on the technical processes involved in thesaurus construction and, indeed, on the rationale behind their use (Shearer, 2004; Aitchison *et al.*, 2000; Will, 1997). This column does not intend to duplicate or reiterate such material. Instead, it aims to provide practitioners with some food for thought relating to some of the processes and stages outlined in the existing literature, which will help to simplify the construction process and, ultimately, lead to a more effective end product. Broadly speaking, a number of steps can be taken to ensure 1) the end product adequately describes the collection, 2) it is a manageable and easily

Library Review
Vol. 54 No. 7, 2005
pp. 403-409

0024-2535
DOI 10.1108/00242530510611893

updatable resource, and 3) labour and cost is kept to a minimum. The following ten tips go some way to providing advice on how to meet these criteria.

**Is this work necessary?**

Before embarking on the construction of a "new" thesaurus, conduct an assessment of existing resources to ensure the work is actually necessary. All too often organisations create new subject schemes from scratch, contributing to the huge variety of terminologies in operation and the "associated limitations imposed on user searching" (McCulloch, 2004). It is recommended that where possible an existing thesaurus is adopted to catalogue subject based material. It may be necessary to adapt such thesauri, resulting in a tailored resource, which meets specific needs. Circumstances may make this unfeasible, however. Due to new concepts, perhaps in a technological field, or extremely high levels of specificity within a collection, it may be that existing thesauri are inadequate for particular fields or corpora. In such circumstances, the only feasible option may be to develop a completely "new" thesaurus. (The word "new" is used loosely here since it is unlikely that any resource developed will be completely novel; it is more likely that a large proportion will comprise an amalgamation of recognised and applied terms.)

In the event that a "new" thesaurus is required, first consider the range of established thesauri, taxonomies and terminology sets within the subject field, whether formalised or used by groups or even individual experts within the field. Although no one example may be sufficient to describe a given collection, it may be that a combination of two or more existing resources is perfectly adequate. In such cases, distinct entities should be combined into a single "meta-thesaurus". Without an attempt to create an amalgamated tool, barriers to interoperability will emerge; that is, the use of disparate subject schemes will not lend well to cross searching and browsing by users.

It is also advisable to contact experts in the given field to ask about thesauri or subject schemes they may know of having been adopted within library or online collections. It may be that subject librarians or prolific researchers in the field have been using an "in-house" list of terms, which may be easily adapted into a standard thesaurus. Under such circumstances it is vital that any copyright restrictions are adhered to.

**Term selection**

Having established the need to construct a thesaurus, relevant terms should be identified. One of two starting points may be adopted: firstly, decide on top level headings or sections within the thesaurus and assign individual terms to these as they are gathered; or secondly, collate terms by subject and decide upon appropriate top level headings at a later stage. Depending on the subject area either of these methods may appeal. However, it should be noted that during the development of the thesaurus, as more and more terms are added, the scope of the top level headings, if established early on, may change. It is crucial, therefore, to adopt a flexible approach throughout and not to assume that headings and structure imposed in early drafts will necessarily be reflected in the final product.

Regardless of which of these two approaches is adopted, begin by gathering individual terms from existing resources. In the absence of existing thesauri or taxonomies within the subject field, keywords can be gleaned from a variety of literature such as academic papers, conference proceedings, official information such as government papers, organisational websites (categories used to structure websites

are often useful for identifying top level headings) or mailing list archives, typically used for informal discussion of hot topics – a potential source of current buzzwords.

During this collection phase, note varying forms of the same term and synonyms. These should be included in the thesaurus as non-preferred terms, signified by the *use/use* for notation. To identify synonyms, subject experts should be asked to compose formal definitions for terms thought to express the same concept (this will also be useful for creating scope notes). If the definitions of two or more terms are essentially the same they can be treated as synonyms and the least appropriate or less commonly used in the field should be incorporated as a non-preferred term. Homographs (words which are spelled the same way but differ in meaning) should also be noted at this stage since these must be qualified in order to provide context and distinction.

It is important to pinpoint terms at a suitable level of granularity. Terms should be applicable to a number of items within a collection; however, they should not be overly specific in that they are only relevant to one or two items. A balance must be struck to maximise the benefits for effective and efficient information retrieval. This will largely be determined by the user group, whether perceived or known.

**Structure and notation**

When imposing a hierarchical structure on terms using appropriate relationships, similarly to the case of term selection, the structure imposed on the thesaurus should demonstrate an appropriate level of specificity – too few levels of granularity is insufficient; too many is unnecessary.

A decision must be taken as to whether or not some form of notation (either numerical or alphabetical) is required. If so, this should be one of the final stages in the construction process, as it is highly probable that terms will undergo repositioning within the thesaurus as it develops. It follows that notation should not be imposed until the final version is in place.

The structure and notation should both be kept flexible to allow for easy updating and maintenance. Both should be easily expandable/compressible should the thesaurus grow or diminish.

**Standards**

It is advisable when developing a thesaurus, to adhere to recognised standards. Currently, both the British standards[1] and the American standard[2] are under review. The UK based updating process will result in *BS8723 Structured Vocabularies for Information Retrieval – Guide* (Dextre-Clarke, 2003). Grammatical forms, usage of singular and plural terms, abbreviations, capitalisation, acronyms and punctuation marks are all controlled for by these official guides to standardisation.

Although a thesaurus may be developed for the internal purposes of a relatively small organisation the recommendation to follow standards remains. It cannot be said with certainty that the resource will not be adopted by third parties or indeed, form the basis of another organisations "new" thesaurus in the future. Conformity to standards lends credibility to a resource and may encourage, and even motivate, uptake by other organisations and groups within the field.

**Consultation**

A team of individuals or parties must be established to advise on term selection and structural issues, particularly semantic relationships, and also to provide feedback on

draft versions of the thesaurus. Ideally, this group should comprise specialists from the subject field as well as thesaurus/information science experts. The nature of the feedback required makes it difficult to undertake at a physical meeting or via group discussion. It is likely that, irrespective of the subject area, opinions will vary on the value of individual terms, the form of preferred terms, and the position of terms within a structured hierarchy. In order to limit extensive discussion, e-mail or an online discussion forum may be a more efficient method by which to gather feedback. Since it is unlikely there will be a right or wrong place within the thesaurus for a particular term, consensus may never be reached. It may therefore be practical to assign specific areas of the thesaurus to sub-groups of advisors to limit deliberation, with a single person responsible for overall editorial control. Where difference of opinion exists the editor should make the final decision on issues like which terms to include, which format to adopt, and where to place them within the thesaurus structure.

### Software

Ensure the software with which the thesaurus is implemented is accessible and user friendly, both for cataloguers and end users. Ideally, it should be a familiar, widely used package but, if not, it should be straightforward to learn and training materials should be on hand for the developer, as should support. It may be that different software packages are more suited to the development and final delivery phases of the thesaurus. The package should easily lend itself to updating and modification. A list of potentially useful software packages is provided by Will (2005).

The thesaurus creator will often deliver training to indexers and end users. As such, it would be extremely useful to make notes on aspects of thesaurus usage, navigation and updating procedures during the developmental phase. This should ensure the level of documentation is well targeted and informative for new users, as opposed to documentation created retrospectively, once the creator has become "expert" on the basis and use of the thesaurus. Such notes will form the core of a training manual if required.

### Web display

Web display of the thesaurus is largely dependent on user requirement. A number of questions will help determine how the thesaurus should be presented to ensure maximum effectiveness.

- Do users require to search or browse for terms? Do they want both options?
- Do users want to view the entire thesaurus and all hierarchies simultaneously or do they want to "zoom-in" on particular areas? Should "close-ups" open up an additional browser window so that the area being looked at can still be seen within the context of the overall thesaurus?
- Do users want an alphabetical presentation or index of all terms contained within the thesaurus? That is, do they want to see non-preferred terms displayed?
- Will indexers work from the web based version or will they have a "behind the scenes" interface?

In determining the most appropriate means of web display, accessibility is an important factor to note (Wallis, 2004). The relevant piece of governing UK legislation is the Disability Discrimination Act (Disability Discrimination Act, 1995); most

jurisdictions will have similar laws and it is recommended that WC3 (W3C Web Accessibility Initiative, 2005) guidelines are adhered to which strive to ensure the creation of "websites and applications that people with disabilities can perceive, understand, navigate, and interact with".

## Evaluation

When nearing completion, arrange a testing phase whereby indexers use the thesaurus to catalogue a sample of their material. This will highlight any gaps in relation to particular collections and should also identify any inappropriate entries in the thesaurus.

A trial of the thesaurus from the user perspective would also be valuable. Ask users to run some searches to ascertain how straightforward they find the interface, any search facility provided and interface, and also to consider typical user terminology. Any variation between standard terminology and that commonly employed by users will be highlighted, perhaps resulting in the need to include additional appropriate synonyms.

## Maintenance/updating

Thesauri should be revised regularly, via a well organised and standardised procedure since haphazard maintenance of a thesaurus will lead to a range of problems, both for indexers and end users. Central to the updating process should be an advisory group, approving new term suggestions. The inclusion of terms will therefore be strictly controlled and documented to avoid any potential conflict with existing terms or structure.

Ideally, one or two individuals will be responsible for implementing updates and revisions. This will help to ensure consistent formatting and notation.

## Cost effectiveness?

Some issues already discussed will go some way towards achieving cost effectiveness within the thesaurus construction process. Particularly, advice given on term selection and evaluation should help to minimise costs, since they should reduce the need for widespread research and eliminate the need for any major overhaul once the thesaurus is released and in use.

In addition, Kosovac (1995) proposes that "automatic tools for extraction and clustering of candidate terms" will also reduce overall costs. Such tools may involve high financial investment which may only be justifiable if a number of thesauri are planned. Kosovac also points out that "In situations where fast and efficient access to information is essential savings at the output side can often justify the relatively high input cost". This may well be true in specific situations but an appropriate balance between costs and benefits must be established early on – and stuck to!

## Conclusion

This column has provided some guidance in the form of a range of hints and tips for those constructing thesauri within their specific subject domain. Coverage merely offers some points to note throughout the planning and construction processes. For further reading and a greater insight into the individual steps involved in thesaurus construction, together with practical examples, the author would encourage readers to consult the publications list below, particularly those of Nielsen (2004) and Shearer (2004).

## Notes

1. The British Standards under revision are the British standard guide to establishment and development of monolingual thesauri (BS5723:1987) (BSI, 1987) and the British standard guide to establishment and development of multilingual thesauri (BS6723:1985) (BSI, 1985), of which the international equivalents are ISO2788-1986 (ISO, 1986) and ISO5964-1985 (ISO, 1985) respectively, as produced by the International Organization for Standardization.
2. The American standard under revision is NISO Z39.19-200x Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies (NISO, 2005), of which ISO2788 is an equivalent.

## References

Aitchison, J., Gilchrist, A. and Bawden, D. (2000), *Thesaurus Construction and Use: A Practical Manual*, 4th ed., Aslib, London.

BSI (1985), *British Standard Guide to Establishment and Development of Multilingual Thesauri. BS6723*, British Standards Institution, London.

BSI (1987), *British Standard Guide to Establishment and Development of Monolingual Thesauri. BS5723*, British Standards Institution, London.

Dextre-Clarke, S. (2003), "BS 8723: a new British Standard for structured vocabularies", paper presented at NKOS Workshop, ECDL, Trondheim, 21 August, available at: www.glam.ac.uk/soc/research/hypermedia/NKOS-workshop%20Folder/dextre_clarke.ppt#261,6, (accessed 26 April 2005).

Disability Discrimination Act (1995), Crown Copyright, London, available at: www.legislation.hmso.gov.uk/acts/acts1995/1995050.htm (accessed 26 April 2005).

ISO (1985), *Documentation – Guidelines for the Establishment and Development of Multilingual Thesauri. ISO 5964.* International Organization for Standardization, Geneva.

ISO (1986), *Documentation – Guidelines for the Establishment and Development of Monolingual Thesauri. ISO 5964.* International Organization for Standardization, Geneva.

Kosovac, B. (1995), *Internet/Intranet and Thesauri*, National Research Council Canada, Ottawa, available at: http://irc.nrc-cnrc.gc.ca/thesaurus/roofing/report_b.html (accessed 26 April 2005).

McCulloch, E. (2004), "Multiple terminologies: an obstacle to information retrieval", *Library Review*, Vol. 53 No. 6, pp. 297-300.

Nielsen, M.L. (2004), "Thesaurus construction: key issues and selected readings", in Roe, S.K. and Thomas, A.R. (Eds.), *The Thesaurus: Review, Renaissance and Revision*, Haworth Press, New York, pp. 57-74. Co-published simultaneously in *Cataloging and Classification Quarterly*, Vol. 37 No. 3/4, pp. 57-74.

NISO (2005), *NISO Standards Currently Balloting*, National Information Standards Organization, available at: www.niso.org/standards/standard_detail.cfm?std_id=814 (accessed 26 April 2005).

Shearer, J.R. (2004), "A practical exercise in building a thesaurus", in Roe, S.K. and Thomas, A.R. (Eds.), *The Thesaurus: Review, Renaissance and Revision*, Haworth Press, New York, NY, pp. 35-56. Co-published simultaneously in *Cataloging and Classification Quarterly*, Vol. 37 No. 3/4, pp. 35-56.

Shiri, A., Revie, C. and Chowdhury, G. (2002), "Thesaurus-assisted search term selection and query expansion: a review of user-centred studies", *Knowledge Organization*, Vol. 29 No. 1, pp. 1-19.

W3C (2005), *Web Accessibility Initiative*, available at: www.w3.org/WAI/ (accessed 26 April 2005).

Wallis, J. (2004), "Universal medium", *Information Scotland*, Vol. 2 No. 3, available at: www.slainte.org.uk/publications/serials/infoscot/vol2(3)/vol2(3)article5.html (accessed 26 April 2005).

Will, L. (1997), *Thesaurus Principles and Practice*, Originally presented at Thesauri for museum documentation, Science Museum, London, 24 February 1992, available at: www.wilpower.demon.co.uk/thesprin.htm (accessed 26 April 2005).

Will, L. (2005), *Software for Building and Editing Thesauri*, available at: www.willpower.demon.co.uk/thessoft.htm (accessed 26 April 2005).