Automated Thesaurus Management in a Networked Environment

Ron Davies
Principal Consultant, Bibliomatics Inc.
48-200 Owl Drive, Ottawa, Ontario K1V 9P7
Internet: rdavies@bibliom.isis.org

Ed Brandon
Senior Program Officer, Information Sciences and Systems Division,
International Development Research Centre, Box 8500,Ottawa, Ontario K1G 3H9
Internet: ebrandon@idrc.ca

While institutional members of an information network may share the use of a common thesaurus, most institutions will need to make additions or changes to the thesaurus based on their particular requirements. Automated thesaurus management systems have not usually taken into account the needs for such local adaptations. While the requirements for networked thesaurus management are complex, it is possible, even in a simple microcomputer-based environment, to meet a significant number of those requirements with a few simple functions. This paper describes basic requirements for managing thesauri in networks, and the means by which these requirements are supported in a specific microcomputer-based multilingual thesaurus system.

Institutions participating in an information network can benefit by adopting a common thesaurus for indexing and information retrieval. Sharing a thesaurus reduces, for each institution in the network, the cost of developing and maintaining a thesaurus of its own; and it also increases the ease with which information can be shared among members of the network (Soergel, 1974). Today, when institutions are looking for ways to improve their effectiveness without increasing costs, information sharing is increasingly important.

Nevertheless, the management and use of a common thesaurus in an information network is complex. Soergel (1974) discusses thesauri in information networks primarily in terms of a "source thesaurus" model (in which institutions extract subsets of terms from a common source), and a "cumulative thesaurus" model (in which a common thesaurus is formed by combining indexing languages in use at different institutions). While these models apply to many projects for thesaurus cooperation, it is our belief that the most common form of network management and use is a more distributed and diffuse model in which a single thesaurus is modified in very different ways at many different institutions. Each organization modifies the common thesaurus by making numerous but relatively minor additions, deletions and changes, in much the same way that many libraries have traditionally made a number of local changes to a standard set of subject headings, such as the Library of Congress Subject Headings. Organizations indexing using thesauri, however, often lack the stricter procedures and longer history of authority file maintenance that libraries have. They often do not have the means, procedures or resources to make these changes in a controlled and effective manner; in many organizations, even large national institutions, the principal means for managing such changes has often been simply to annotate copies of a printed version of the thesaurus.

Users of the OECD Macrothesaurus provide one example of this kind of very distributed network. *Macrothesaurus for Information Processing in the field of Economic and Social Development* (OECD, 1991), as it is officially known, is a trilingual (English/French/Spanish) thesaurus designed for use in indexing and retrieving bibliographic and other information. Directions for this thesaurus are set by an Advisory Committee with responsibility for overall maintenance policy and a Technical Committee responsible for decisions concerning the additions and changes of particular terms. The members of these committees are drawn from the five organizations which share the responsibility for the maintenance programme: the OECD, the OECD Development Centre (which houses the Macrothesaurus Secretariat), International Development Research Centre (IDRC), the United Nations Advisory Committee for the Co-ordination of Information Systems (ACCIS) and the United Nations Department for Policy Coordination and Sustainable Development (UNDPCSD). Now in its fourth edition, and available both in printed and machine-readable form, it is used by many hundreds[1] of institutions throughout the world , including widely-dispersed information networks, such as the International Development Information Network (IDIN).

Experience has shown that users of the OECD Macrothesaurus make local adaptations to the thesaurus primarily for linguistic, geographic or topical reasons.

Changes made for *linguistic* reasons are based on the particular linguistic situation of an institution which belongs to an information network. An institution may not need to use all the languages supplied in a multilingual thesaurus; the institution may need to change spelling of a thesaurus term (*labor* for *labour*) or the form of a term in a given language; in the most extreme case of linguistic change, an institutions may translate the entire thesaurus into another language more common locally.

Geographic changes are changes made to the common thesaurus to reflect the geographic location of the institution and its needs for geographic descriptors. These may take the form of additional descriptors for specific geographic regions or even the replacement of a non-descriptor with a descriptor: a Canadian organization may find the need to add the term Ontario, for example, or a Caribbean organization the need to replace the standard Macrothesaurus non-descriptor *Leeward Islands*, with a specific descriptor in the same form, because the Leeward Islands form an important subregion.

Finally, *topical* changes are made to accommodate new concepts or different terminology in a subject area in which the institution has a particular interest or expertise. These may include both new topics that may later become of general interest to other thesaurus users (and may be included later in the scope of the common thesaurus) as well as topics which will always remain of interest only to one specific institution or a very few institutions, and never graduate to full descriptor status in the common thesaurus.

Given this distributed model of the use of a common thesaurus in information networks, what are the thesaurus management needs of the member institutions, and how can they be met in a microcomputer-based system? First, when initially loading the thesaurus at their local sites, users need a simple and easy procedure to select the language or languages in the thesaurus which they wish to use. Second, when making changes to the thesaurus to reflect local requirements, they need tools to ensure consistency in the thesaurus during and after update, and to assist in recording and publicizing the changes that have been made. Users also need a simple and effective procedure to send proposals to the central site, preferably in electronic form, as well as to track locally proposals for changing the common thesaurus; similarly, the central site needs a means to track and control the proposals that have come in. Fourth, users need a means to assist in exchanging information which has been indexed using the local version of the thesaurus: they need a means to translate local descriptors to the standardized descriptors of the common thesaurus when exchanging records with other institutions belonging to the network. Finally, organizations needs support in updating the locally-adapted thesaurus when a new electronic edition of the standard thesaurus is released.

This list of requirements may appear quite daunting, particularly when considered in the light of the complex, rule-based structure of a thesaurus. Many thesaurus management systems, developed for the management of a thesaurus used at a single organization, or for a thesaurus in which all changes are centrally controlled, do not provide this type of support. Nevertheless, a recent thesaurus management system designed for use by Macrothesaurus users shows that it is possible to provide support for thesaurus management and use in a network even within the constraints of simple software and without tremendous overheads.

The MTM Version 3 software is a multilingual, microcomputer-based thesaurus management system developed by the OECD, in cooperation with the International Development Research Centre (IDRC) and the International Information Centre for Terminology (Infoterm). Designed primarily (but not exclusively) to support the management of the OECD Macrothesaurus, it is used by the Macrothesaurus Secretariat at the OECD Development Centre for the management of the thesaurus. However it is also distributed to institutions who wish to use the Macrothesaurus in electronic form, enabling them to make local changes to the thesaurus and to use the thesaurus in indexing and retrieving information. The MTM software uses Unesco's CDS/ISIS information retrieval system to manage its data and is written in the Pascal-like programming language that forms part of the CDS/ISIS system. The MTM software supports the management of a thesaurus in a networked environment in four different ways: (1) by allowing local configuration of languages; (2) by supporting local additions and changes, including proposals for the modification of the thesaurus as a whole; (3) by providing aids to updating the thesaurus with the distribution of new versions; (4) by converting local terminology into terms from the common, standard version of the thesaurus when exchanging information.

1. Language configuration

Users of the MTM software may configure the software to indicate the dialogue languages which they wish to use, as well as the different languages in which they want to use their thesaurus or load an external thesaurus in machine-readable form. MTM can be configured for up to nine different thesaurus languages. Languages can also be specified as mandatory or optional. A mandatory language is a language in which a descriptor must be present for any descriptor in the thesaurus, thereby enforcing that new terms are entered in all applicable languages. This requirement is not enforced with an optional language, thereby facilitating the work of translating the thesaurus into a new language. The new language can be designated as optional, and the translator can enter terms in the new language, gradually building up the correspondences, and skipping difficult problems if necessary to return to them at a later date. When the translation is complete, the MTM software can be re-configured to indicate the new translation language is a mandatory language; subsequently a batch thesaurus check program (not part of the current release of the MTM software but to be released later in 1994) will then flag any terms for which the equivalent is not found in the new language, so that the consistency of the translation can be confirmed.

2. Management of local terms

A flexible thesaurus management system that supports networked use should support the creation and modification of locally-defined terms. The MTM software allows the user to record new local terms, and to modify existing terms at the local institution. For example, when a user requests the addition of a new term, a menu proposes three different types of terms: descriptors, non-descriptors and local descriptors. Local descriptors are also categorized as belonging to one of two different levels: local descriptors that are specific to one institution only, and descriptors that are shared by other institutions within a localized regional network or subnetwork. The user may also enter, for each language of the thesaurus, the standard thesaurus term closest in meaning to this new local term. This standard term, or replacement term as it is called, is used to replace the local descriptor in bibliographic or other application database records when exchanging information with other institutions as described below.

However it is not enough simply to allow the user to make changes: there must also be a way of tracking and listing the changes made. While complete transaction logging is beyond the capability of MTM software, due to limitations inherent in the CDS/ISIS software and the microcomputer platform on which it runs, the MTM system does provide some simple, but effective ways in which changes can be tracked, depending on the needs and resources available to devote to this task.

First, the MTM software supports the recording of notes, in which the user can record information about the history of a the new term or the change made. Second, the MTM software supports the copying of entire records into a separate log database immediately before the record is modified. These "snapshots" of the record at a given point in time can be cumulated in this log database to provide information on earlier forms of a term. They are especially important in recording deleted terms which would otherwise no longer have a place in the thesaurus.

These same facilities can be used to help users track proposals sent to the central Macrothesaurus Secretariat for possible inclusion in a future edition of the thesaurus. When creating a local descriptor, the user can also indicate if it is also a proposal being forwarded for inclusion in the common thesaurus. Either by recording notes, or by copying the entire record to the log database, a record can be kept of the proposed term, the reasons behind the proposal and any research that went into justifying the addition of the concept. Because data structures are the same in the local and central copies of the database, proposals can be sent in electronic as well as printed form. Some additional fields, for information on the institution proposing a new term, have also been added to the thesaurus data structure to help the central secretariat manage these proposed terms.

3. Updating for new editions

One of the most time-consuming chores in maintaining the local version of the common thesaurus is updating that version when a new edition of the common thesaurus is released. While the MTM software does not provide any automatic update capability, it can assist in making and maintaining local changes from one edition to another, by producing change lists. The facility for producing these change lists, which are simply lists of all the differences between one version of a thesaurus and a later copy, can be used in a number of different ways during this process.

First, it can be used to list differences between the current edition of the standard thesaurus and a new edition, to provide the person responsible for maintenance of the local version of the thesaurus with a clear summary of the changes that have occurred in the common thesaurus. It can be used to list differences between the current edition of the standard thesaurus and the current edition of the local thesaurus, to provide, in effect a list of local additions and changes: using this list, the thesaurus maintainer can review all local changes made, and make the same local changes to the new edition of the common thesaurus. (While this process may take some time, particularly if there are many local changes, it is less time-consuming that having to work from a list of the entire local thesaurus, since the changes have already been flagged.) Finally, when all the local changes have been made to the new edition, the user can produce a list of all the differences between the new edition of the standard thesaurus and the new local thesaurus. This new "list of additions and changes" indicates changes which will affect the work of local indexers and searchers.

4. Information exchange

One of the problems of adapting a common thesaurus to local requirements is that it can be a barrier to information sharing within the network. Institutions in a network share information, usually by sending records or even entire databases created at one site to another site, where they can be loaded and searched locally. However if the originating organization has used local terms to index records, these indexing terms will not be recognized by the receiving institution as valid descriptors in the common thesaurus. Organizations adopting local terms need a method whereby the locally-defined term can be replaced with the standard term drawn from the common thesaurus when searching. The MTM software assists in the exchange of information by allowing users to substitute standard terms for the local descriptors, before exchanging information within a subnetwork or with another institution in the network as a whole.

In practical terms, this facility to replace local descriptors is part of a general facility for translating index terms. For example, an English descriptor entered in the English subject field in a bibliographic record can be automatically translated by looking up the corresponding thesaurus record, and copying the equivalent French descriptor taken from the thesaurus into a different field in the same bibliographic record. Because any descriptors not found in the thesaurus are flagged as invalid, a request for a "translation" of a descriptor in one language into the *same* language serves as a batch validation process. In the MTM system, the translation/validation function has been extended to allow for the replacement of local descriptors. The user can request, by specifying an additional parameter, that when a *local* descriptor is looked up in the thesaurus, the new value is not taken from the descriptor field, but from a field in the thesaurus record which contains the corresponding standard thesaurus term. For example, in the case of the local descriptor *Leeward Islands* mentioned above, the thesaurus record for the local descriptor *Leeward Islands* would have *Caribbean* as its replacement term When the Caribbean organization using *Leeward Islands* as a local descriptor wanted to exchange records with another institution, they would first "translate" the descriptors, asking for replacement of local terms, and writing the "translated" descriptors to a temporary field. When the records were then converted to the exchange format, the data would be transferred from the temporary "translated" descriptor field, instead of from the actual descriptor field which contains the local descriptors.

This simplifies the process of exchanging records: the user has only to run the validation process to create a new subject descriptors field in the source records with replacement terms and send this field, instead of the field with the local descriptor, to other institutions when exchanging records.

Conclusion

The MTM Version 3 software does not meet all of the needs of sophisticated use of thesauri in a networked environment. However it demonstrates that even with a relatively simple microcomputer-based package, it is possible to assist users in managing a thesaurus in a distributed environment, through facilities that assist in configuring languages, in adding local terms, in updating the local version for new editions of the standard thesaurus and in exchanging of indexed records. While the requirements of thesaurus use in a networked environment can be complex, a significant proportion of those requirements can be met in a simple and straightforward fashion.

Notes

[1] A 1988 survey received responses from 386 institutions actively using the Macrothesaurus (Sly 1989); there were undoubtedly many more that did not respond.

References

Aitchison, Jean and Alan Gilchrist. 1987. *Thesaurus construction: a practical manual*. 2nd edition. London: Aslib.

Janik, Sophie and Lise Brunet. 1987. La mise à jour d'un thésaurus. *Documentaliste* 24(6): 215-229.

Organization for Economic Co-operation and Development. 1991. *Macrothesaurus for Information Processing in the Field of Economic and Social Development*. 4th ed. Paris: OECD.

Ritzler, Claus. 1990. Comparative study of PC-supported thesaurus software. *International Classification* 17(3-4): 138-147.

Rohou, Cecile. 1987. La gestion automatisée des thésaurus: Étude comparative des logiciels *Documentaliste* 24(3) : 103-108.

Sly, Maureen. 1989. *Report on the Survey of Macrothesaurus Users*. Ottawa: Information Sciences Division, International Development Research Centre.

Soergel, Dagobert. 1974. *Indexing languages and thesauri: Construction and maintenance*. Los Angeles: Melville.