Comparative evaluation of thesaurus creation software

Heather Hedden

This article provides a comparative evaluation of three software programs used for the creation and editing of controlled vocabularies, thesauri, or taxonomies: MultiTes, Term Tree 2000, and WebChoir TCS-10. These thesaurus tools are comparable in their features and are all available in single-user desktop versions affordable for the freelance taxonomist.

Indexers have the skills, and sometimes the need, to go beyond creating indexes and create controlled vocabularies or thesauri used for indexing. But understanding the principles of thesaurus creation is not enough for one to efficiently complete a project. As in the field of indexing, having a good software tool is also important.

The size and scope of a thesaurus project can vary greatly. It could be a simple term list that an individual indexer creates to help with the indexing of a longer than average book. At the other extreme the project could involve a large collection of inter-related authority files of tens of thousands of terms managed by multiple controlled vocabulary editors in a large organization. Similarly, the software tools available vary in their capabilities, number of users supported, and price. At the high end, a number of thesaurus products (also called taxonomy tools) integrate thesaurus creation and maintenance features along with the capabilities of auto-categorization, human indexing, and/or search to provide more complex full-featured solutions. Examples include Early & Associate's Wordmap, Data Harmony's MAIstro (which includes both Thesaurus Master and the indexing tool M.A.I.), and Dow Jones' Synaptica. This article, however, looks at products that facilitate only thesaurus creation and maintenance and have affordable single-user versions.

A thesaurus can be considered as a type of taxonomy which supports not only hierarchical relationships but also associated term relationships, cross-references from nonpreferred ('used for') terms, and the additional option of notes for each term. ('Controlled vocabularies, thesauri, and taxonomies' (Hedden, 2007) explains the definitions of thesaurus and taxonomy in greater detail.) Published national and international standards provide specific guidelines for the creation of information retrieval thesauri. The products evaluated here all support the characteristics of a thesaurus in accordance with these standards. They are indeed marketed as 'thesaurus construction tools,' although they are sometimes also called taxonomy tools. For simplicity, though, we will refer to the person using the software to create a thesaurus as the 'taxonomist' or simply the 'user.'

The basic requirement of a thesaurus tool is to maintain terms and their associated relationships and other attributes. All these relationships are reciprocal between pairs of terms. By using a thesaurus tool, the taxonomist only needs to create or edit the relationship in one place. If the user decides to rename or delete a term, the change is reflected in all its relationships. Merging and subsuming terms may also be supported. Features for optional scope notes and user-defined classification categories for each term are also expected in a thesaurus tool.

The products covered in this evaluation all meet the above basic requirements and share several additional features. These include designating candidate and approved terms, indicating term creation date and modification date, permitting a term to belong to multiple hierarchies (polyhierarchies), and disallowing illegal relationships (e.g. circular relationships). They all run only on Windows and include online Help. They can also export taxonomies in platform-neutral formats for use in other systems. Free demo versions for each are available for download.

The following topics will be evaluated and compared for each product: thesaurus display, term display and editing, thesaurus searching, user-defined relationships and attributes, rules enforcement, importing and exporting or reports, and online help and tutorials.

Thesaurus display

A thesaurus is typically displayed alphabetically, with relationships and attributes listed at each term. A simple taxonomy, on the other hand, with the emphasis on broader–narrower (parent–child) relationships is typically displayed as a hierarchy. Thus, a thesaurus/ taxonomy tool could conceivably present the list of terms more than one way, and indeed the different tools do offer different displays. Thesaurus tools also differ in whether they create single isolated thesaurus files or named hierarchies that can be related to each other.

Term display and editing

Each thesaurus term has relationships and other details that can individually be displayed and edited, although the various thesaurus tools differ in how this is done and in what attributes are supported. Creating terms and their relationships can be performed through various means: main menu selections, context menu selections, shortcut key combinations, other keyboard commands, toolbar buttons, and mouse drag and drop.

Thesaurus searching

Thesaurus navigation is important not merely for the end-user of the implemented thesaurus, but also for the creator of the thesaurus, who needs to know, for example, whether a given term has already been created.

User-defined relationships and attributes

The ability to define relationships, types of notes, and categories of terms is an important set of features for making a thesaurus tool versatile and extensible. Although standard thesaurus relationships are limited to broader term, narrower term, related term, and use/used for, more specific user-defined relationships might be desired. More complex types of relationships are what distinguish an ordinary thesaurus from what is called an ontology. Relationship types between terms that give more meaning, such as 'produced by,' 'owned by,' 'purchased by,' and 'utilized for,' are known as semantic relationships. With the growing interest in the Semantic Web, it is becoming increasingly important that a thesaurus tool also support user-defined relationships.

Categories are often used in taxonomies to classify terms for end-use, by source, or for any purpose the thesaurus developer may have. Categorizing terms makes them easier to batch edit for the taxonomist and also possible to designate for certain end-use search interface characteristics. Such user-defined categories could also be used to distinguish generic terms from named entities, terms that the user decides may be permitted to have certain semantic relationships, and terms that belong within a facet for a faceted taxonomy. There could be a master taxonomy, for example, and certain subsections categorized for certain audiences or services.

Rules enforcement

A thesaurus should follow certain rules, such as not having terms that duplicate within the same hierarchy or category, terms that lack hierarchical relationships with any other terms (orphan terms), or terms that have a broader term which is narrower than a narrower term of the first term (circular references). Not necessarily 'rules', but controls may also be desired over additional areas, such as whether to allow multiple 'use' references.

Importing, exporting, and reports

In addition to building a thesaurus from scratch, i.e. manually typing in each term, a taxonomist may want to take advantage of external lists of terms to import and incorporate into the thesaurus. Most often these might be lists of names, organizations, or places, but it could also be the incorporation of legacy taxonomies, along with their relationships, into a new system. Thus, importing or batch loading of data into a thesaurus tool is an important feature.

Exporting the thesaurus into formats that can then be imported into other systems is crucial for a thesaurus tool. A thesaurus is not just to look at but is also to be used in the indexing/tagging of documents or web pages and then by a final end-user for search and retrieval of those documents and pages, which is done in one or more other software systems.

Generating interim reports of various kinds can aid the taxonomist in the task of building a thesaurus. Occasionally a thesaurus is published for third party use, in which case various outputs, including a printed document, might be desired.

MultiTes Pro

Product name: MultiTes Pro, version 2007.02.01
Product vendor: Multisystems (Miami, Florida, USA)
Price: US$295 single user; US$1,295 for 5 users; US$2,495 for 10 users; US$3,950 enterprise deployment.
www.multites.com

Thesaurus display

MultiTes displays the thesaurus in an alphabetical list of all terms, interspersing nonpreferred terms (in italics) among the preferred terms. There are no options to change the user interface display. To view a hierarchy of the thesaurus it is necessary to select the 'Hierarchical' or 'Top term' options from the Report menu, which can generate simple text files to the screen. The 'Hierarchical' report displays the full parent/child hierarchy path for all broader and narrower relationships for each term. The 'Top term' report displays a simply hierarchy of the path of narrower terms, if any, under each term.

MultiTes' display, while limited to alphabetical, does include some additional information for each term which the other tools do not include in their alphabetical displays. Making use of the full width of the screen in a tabular format, MultiTes includes columns for term status, types and number of relationships, note types, category names, and language. The user can customize the display to show only the desired columns of information. While it is a minor inconvenience to have to go through the steps of generating a report to obtain a hierarchical view in MultiTes, the user has the options of restricting a hierarchical display to a certain subsection or category of the entire thesaurus. While the hierarchy cannot be expanded in the main view, more terms fit into the MultiTes display, about 38, compared with about 28 in Term Tree and TCS-10.

MultiTes creates single thesaurus files with the extension .th2. Relationships cannot be created across files. To create multiple authority files or facets with interrelationships, these each have to be created within the same file and designated by user-defined categories.

Term display and editing

In MultiTes, since the thesaurus list takes up the entire window width, to view or edit a term's details involves clicking method of creating a new term, An existing term must first be selected from the right-hand thesaurus pane. Once selected, Insert Term can be selected from the Edit menu. If a term is to have more than one broader term, then Add Additional BT must be selected from the Edit menu. While creating terms first as narrower terms to existing terms is the most common method by which taxonomists work, it should not be the only method permitted. Sometimes a taxonomist might want to create a new term first as a related term to an existing term.

TCS-10 has a limited drag-and-drop feature. Unlike Term Tree, a term cannot be dropped to add any relationship, rather it can only be moved from having one broader term to having a different broader term. This is done through the thesaurus tree display pane. If the destination broader term is not visible in the display, the procedure cannot be done, so in a large hierarchy many broader terms will need to have their hierarchies collapsed first so that the target term will appear in the window display. The practicality of the drag-and-drop feature is therefore somewhat limited.

Thesaurus searching

As with MultiTes and Term Tree, TCS-10 features both a search box at the top of the scrollable thesaurus list and an advanced search feature from the menu. The search box, unlike MultiTes and Term Tree, does not support truncation. On the other hand, text within this search box does not merely match the start of a term, but also can match a word within a term. As this kind of search can yield multiple results, a window pops up with all matching results, a feature found only in the advanced search of MultiTes, from which any matched term can be selected and jumped to.

The advanced search (Ctrl+F) feature from the Edit menu, and also via an icon button, does not have as many limiting options as the advanced search in MultiTes or Term Tree. The searchable fields are only the term name (descriptor) and scope note. However, TCS-10 uniquely has the advantage of offering guided Boolean search. Options include search Within Results (AND), Add to Results (OR), and Not in Results.

User-defined relationships and attributes

TCS-10 supports unrestricted user-defined types of use/used for relationships and types of related-term relationships (through the Maintenance menu's Lookup Table) but restricts the broader/narrower relationships. Instead, it provides the option to choose from three additional types of broader/narrower relationships, all of which are in accordance with ISO standards. They are defined in the system with the abbreviations BTG, BTP and BTI, which can be selected from a drop-down list. BTG stands for broader term – generic, which is when a narrower term that 'is a' member or type of broader term. BTI stands for broader term – instance, when a narrower term is a specific (proper noun) instance of a generic broader term. BTP stands for broader term – partitive, which is when a narrower term is part of a broader term whole. If the user has more specific thesaurus relationship requirements, the limited choice of standard relationship types can still be useful. For example, a geographic hierarchy is actually a kind of whole-part relationship, so the BTP relationship could be used to relate cities to their states and states to their countries. I consider the BTI relationship particularly valuable, since named entities have their own particular attributes and may need to be set off differently from generic broader/narrower term relationships.

TCS-10 also supports user-defined notes and other attributes. The note types are in addition, not instead of, the system-supplied Scope Note field. Unlimited user-defined attributes can be created for terms. Called ambiguously 'User Relation Type,' these attributes could be utilized, for example, for location or contact information of an organization or person term. A feature unique to TCS-10 is that one can even create authority control for a relation type, such as a lookup table of industry codes or state abbreviations.

Rules enforcement

TCS-10 does not permit the creation of a relationship to a term that does not yet exist. TCS-10 also includes user-defined systems options (from the Edit menu) to allow/disallow duplicate terms and to allow/disallow deletions of terms with narrower terms. The user can also specify whether narrower terms should display under their parent term in an order other than alphabetical.

Importing, exporting, and reports

TCS-10 supports the greatest number of formats for importing data. From the menu, Tools > Import, the choices are ASCII, XML, and MARC. For the import of ASCII text files there are additional submenu options for importing a simple descriptor alphabetical list, a standard thesaurus list with notes and relationships for each term, and a simple hierarchical list using periods for indents. There is also an ASCII import wizard to aid the process. For importing XML, a tag mapping table needs to be completed.

TCS-10 supports exporting as XML, ASCII, MARC, HTML (with hyperlinks for the web), and a proprietary format for Webchoir's other indexing and search products. Additionally, user-created look-up tables can also be exported. HTML export options include web pages with or without frames.

Report options from TCS-10 are comparable with the other thesaurus tools and include alphabetical, hierarchical, parallel hierarchy, hierarchical portion, rotated, hierarchy list, and descriptor by category. Terms can be included/ excluded by date, approval status, or alphabetic range. For the alphabetic report, all relationships, notes, categories, attributes, etc. can be included or excluded in the output. Additional kinds of reports include top term list, used-from (UF) list, URL list, image list, category list, and candidate term list. The reports are output as either ASCII text files or HTML. Unlike the output of TCS-10's export feature, the reports generated in HTML are simple HTML files lacking term hyperlinks.

Online help and tutorials

TCS-10's Help, while comprising a reasonable amount of information, is not easily searchable. The bulk of it is on merely eight pages, one for each menu item. So, when searching on Help, the user is returned to one of these pages and must scroll through the long page to find the highlighted search term. Furthermore, the online Help file has only Contents and Search, and lacks an index. The Help text also displays too wide for easy reading. In addition to the online Help file, the same Help documentation is also in an HTML file.

Summary

There is no clear-cut best or worst thesaurus software. Each tool fulfills the basic requirements. Different user preferences and different taxonomy projects will determine which product to choose. In building a thesaurus, speed and ease are important. Each tool permits adding successive relationships of the same type (nonpreferred terms, narrower terms, or broader terms) simply by hitting the enter key.

MultiTes is a good-value thesaurus creation tool. Its basic features are easy to use, but some of the reporting and batch methods of making changes are not so intuitive. I especially like the way it is easy to define unlimited types of relationships and is simple to import terms with relationships. Compared with the other tools, MultiTes lends itself most easily to offline text file importing and exporting. Its major drawbacks are the inability to subsume or merge terms in a single step and its lack of a spell-check capability.

Term Tree is graphically not so clean and attractive. I feel that there are too many unnecessary icons, but the software is very functional. The taxonomist can add relationships to a term selected from the browser in just one step/click, unlike the other two packages. Term Tree serves efficient term entering and editing by making use of keyboard shortcuts and single-action toolbar buttons. Whether the feature of drag and drop is really that important is another matter.

TCS has the advantage of supporting the dual level of thesaurus databases and named hierarchies within those thesaurus database. This is a useful feature for creating multiple facets or top-level terms for a single thesaurus project. Terms can have relationships across different hierarchies, including broader terms in more than one hierarchy. My main complaint about TCS is that the broader/narrower term relationships are not part of the editable relationships of a term and that the entire thesaurus must be created from the top down.

Reference

Hedden, H. (2007) Controlled vocabularies, thesauri, and taxonomies. *Indexer* **26**(33) (March), 33–4.

Heather Hedden *is a taxonomist with Viziant Corporation, a continuing education instructor with Simmons College Graduate School of Library and Information Science, and the manager of the Taxonomies & Controlled Vocabularies SIG of the American Society for Indexing. E-mail:* heather@hedden.net

Some indexer-authors

SI Scottish Group member Margaret Christie had her first volume of poetry published as a chapbook in autumn 2007. (Margaret Christie, **The oboist's bedside book.** Glenrothes: HappenStance, 2007. 36pp. ISBN 978-1-905939-11-4 £4.00) The poems provide a fascinating insight into the production of sound, as it is shaped into music by the artist's technique, endless practice and the sheer joy of playing. Some poems have featured in previous collections, but this book heralds the start of solo volumes.

Margaret summarizes her activities thus: 'Margaret Christie indexes, copy edits and proofreads to support her oboe habit . . . .' She even includes references to indexing in her poems!

Another distinguished member of the Scottish Group, Colin Will, is also a published poet.

Hazel Bell, an advocate and practitioner of fiction indexing, has published cumulative indexes to A. S. Byatt's novels (Hazel K. Bell, **The Frederica Indexes: Cumulative Indexes to A.S. Byatt's Novels 'The Virgin in the Garden'/'Still Life' and 'Babel Tower'/A Whistling Woman'.** HKB Press, 2007. 21 pp. ISBN-13: 978-0955250361.£6.00); and see the advert on this page for another of her publications.