Thesaurus Software and Keyword Constellation Construction

By Jan C. Wright

I recently took on a project to build a set of keywords for a client, a set of terms that would be used to tag Web pages on their Web site. The client and I decided that what would work really well would be to build constellations of keywords; in other words, one keyword when chosen would actually apply several different synonyms and spellings of the term, all in one step. We decided this would be a good approach because the search engine for the site did not do stemming, and upgrades to include that feature would not be available for at least a year. Searching was not working now as well as the client would like, and it needed an immediate fix.

"Stemming" is a capability that allows a search engine to strip a search term down to its root, and then search for all forms of the word. For instance, "caption" might when stemmed actually include words like "captioning" and "captioned." The fact that stemming would not be available meant that we needed to come up with variations of words and include all of them. This particular set of Web pages had a lot of acronyms, each with many variations in spelling and usage, so it would take some time to develop the listing.

Each tagged set of keywords would be filed under a main word or term, and all the variations and acronyms needed to be kept together somehow, so that they could all be tagged at once. To make it more interesting, some terms, such as the concept of "functions," might actually be needed in two or more keyword sets. A function could refer to a programming function, or it could also refer to the functionality of a device.

What we needed was a way of creating and tracking constellations:

> functionality: functionality, functions, functioning, functioned
> captions: captions, captioning, captioned, labels, labelled, labelling

The biggest problem facing me after these decisions were made was to figure out how to actually build the sets without going grayer than I already am. Whatever piece of software I used would need to be fast, to provide output in a variety of formats, and to show which terms were preferred, what terms were in a constellation, and what terms were in multiple constellations. I needed something like this:

> captioned -- Do not use; use captions instead
> captioning -- Do not use; use captions instead
> captions -- USE THIS; it includes "captions, captioning, captioned, labels, labelled, labelling"
> functionality -- USE THIS; it includes "functionality, functions, functioning, functioned"
> functioned -- Do not use; use functionality instead
> functioning -- Do not use; use functionality instead
> functions -- Do not use; use functionality instead
> labelled -- Do not use; use captions instead
> labelling -- Do not use; use captions instead
> labels -- Do not use; use captions instead

I usually build complicated sets of keywords in Excel, but I realized the scope of this project was way beyond my usual methods, and that keeping exact connections and constellation contents up-to-date as I made decisions needed more than Excel could offer. Although I didn't know much about thesaurus software, I realized that perhaps one of these packages might be able to help. I started researching thesaurus software packages, ones that were affordable, and looked at two: MultiTes and TermTree.

MultiTes, the first one I considered, had possibilities. The HTML output of the thesaurus was really nicely done, and would provide a great way for my client to review the constellation sets. But it was so mouse- and dialog-box-intensive that I would soon be forced to seek help for carpal tunnel. I decided to keep looking, as I value my wrists, and I also really detest using dialog boxes. Clicking OK three times for a simple action gets me very crabby.

The second package I tried was TermTree by This to That Pty Ltd., an Australian company. I found that TermTree was less mouse- and dialog-box-intensive, and included some nice drag-and-drop features. Plus, the import capability allowed me to prepare previously existing sets of keywords and indexes for the Web site pages and related materials, and import them into TermTree. This saved a lot of typing. MultiTes also provides a variety of import and export options. But the one export option I felt I needed, one which would give me an Excel-like horizontal layout with column labels, wasn't available.

Both software packages provide downloadable demos of their software,

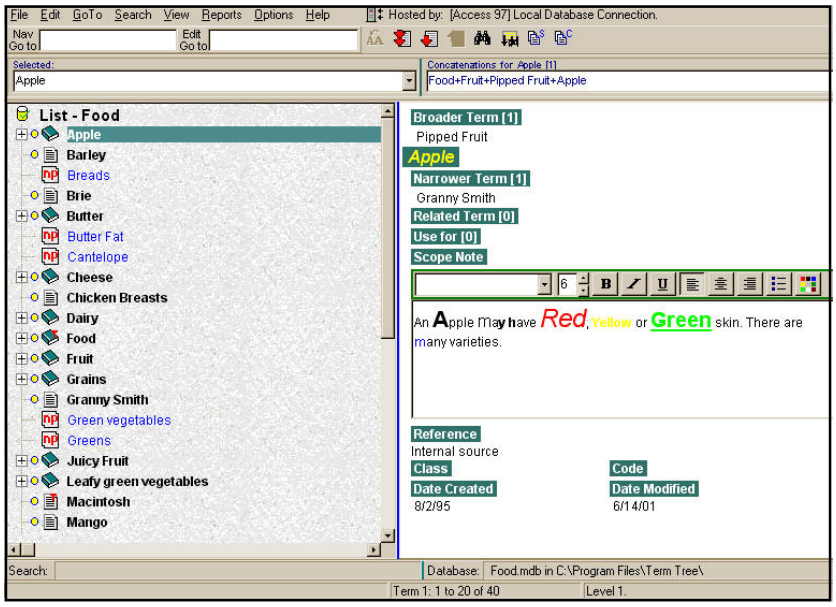

Figure 1: TermTree's display: notice the two-paned approach. + signs in the left pane mean that the term has narrower terms beneath it. The dots in front of terms can be drag-and-dropped into the sections on the right pane. Right-click menus are available everywhere for deleting, renaming, or copying terms.

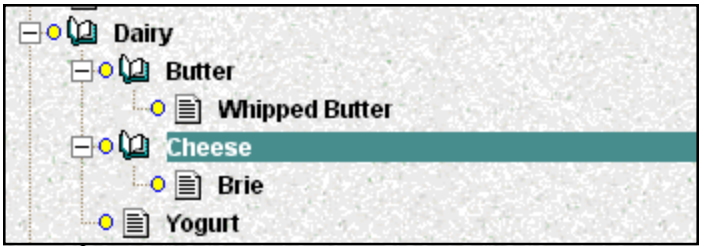

Figure 2: The left pane displays the hierarchy of narrower terms beneath the main preferred term. This display is really useful when building tree structures, but turned out to be one feature not needed in this project.

and both have a nice listing of sources of help in thesaurus construction, allowing me to do some research and get the theories and concepts I needed to consider solidified. I could then adapt what I needed to the building of the keyword constellations.

TermTree also includes support for SQL Server 7 and Oracle 8, and can provide customers with customized DLLs to meet their needs. TermTree even has one DLL that could help with the acronym-and-stemming issues I faced.

As you can see in Figure 1 above, each term is listed alphabetically in the left pane. Unused or "non-preferred" terms are in a different color, making them easily identifiable. You have a variety of display options: you can show all the terms, only preferred terms, or only top terms by using the View menu.

In the right pane, you see the information for a selected term. You can set up relationships between terms, define narrower terms, broader terms, and define scope notes. I imported many acronyms and terms from an existing glossary, and was able to get all the definitions to come in as scope notes, which saved a lot of time and helped me immensely while building sets.

The software tries very hard to keep you from making mistakes, warning you if you try to relate a non-preferred term, or try to use a real term as a non-preferred synonym. Sometimes, a glitch happens. When it does, TermTree provides a database analysis tool that identifies problems and lists them out so you can fix them.

For my purposes, many of these provided fields could be ignored. I decided to include some related terms, but not many. No narrower structures needed to be built, but I did assign a set of about five broader terms to each constellation, letting me know if it had special status in any way, such as being an official title of a specification or internal document.

I found working with TermTree doable, but slow on my older Pentium machine. As I write this, I have upgraded to a newer machine, and it has improved TermTree's performance remarkably. TermTree is based on Access drivers, and you can actually open up the files it creates with Access, but I wouldn't recommend trying to work that way. It's very complex! I opened the file because I accidently typed in a title for a document that was beyond TermTree's field limits, and the entry would not work right any more. Opening the Terms table in Access, I easily found the term, and edited it down to a manageable size. I wouldn't touch anything else in there, as the links and connections between terms are maintained numerically.

TermTree's tech support is wonderful. I imported a bogus character when I imported the massive amounts of data to start the project. Over a series of e-mails, we identified the character, and TermTree rewrote the code and posted a new version the next morning to handle the problem. TermTree is currently

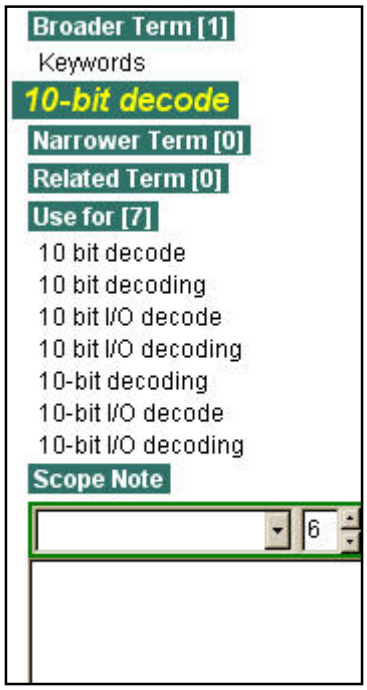

Figure 3: Here's an example of a constellation. "10-bit decode" is the main term for the constellation, and it has been assigned as a simple keyword (rather than also getting a special designation as a standard or specification title). All the alternate spellings and usages are entered in the "Use for" area. These will all wind up applied to the Web page if this term is used. In TermTree's left pane, these non-preferred terms appear in blue, and each says to use "10-bit decode" instead. Designing constellations to work in most situations accurately is a challenge.

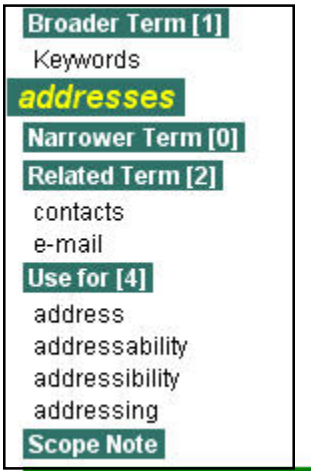

Figure 4: Here's an example of a problem term, one that can belong to different constellations. Rather than place it everywhere it could go, I chose instead to use the Related Terms section to help people to think through whether or not this term is really what they want. This constellation is for bus addressing or device addressing. If you were to look at the "Contacts" term, it also includes "address" and "addressing," but also includes other terms related to maintaining your Address Book. "E-mail" however does not—it is related only in that you might want to consider using it as well if you are applying terms for addressing e-mail.

working on building in more key commands. Until they get some of that functionality in place, I found that QuicKeys allowed me to speed up many processes.

The project isn't over—there are still over 5000 terms to get through. But I found that using different software in a new way has really helped get the project off and on its way. I can now focus on these terms, not on how to do the work.

MultiTes Thesaurus Software: www.multites.com US $295.00

TermTree Thesaurus Software: www.termtree.com.au US $440.00

STC's Telephone Seminars

STC is offering six telephone seminars in 2002. This year, STC is offering online registration at www.stc.org/seminars.html.

In the first seminar, Basil White will discuss "Building a Product, Manual, and Web Site Using Customer-Focused Design." The seminar will be held on January 16th, from 1:00 - 2:30 p.m. Eastern Standard Time

The second January seminar is scheduled for January 30th, from 1:00 - 2:30 p.m. EST. Entitled "Developing a Strategic Framework for Technical Marketing Communication," it will be led by Sandra Harner and Tom Zimmerman.

Other seminars scheduled for 2002 are:

- February 6, "From World-Weary to World-Ready: Usability for International Users." Hans Fenstermacher
- February 20, "Creating Effective Documentation Plans." John Hedtke
- March 6, "Communicating Clarity: Make Your Technical Marketing Matter." Pamela Selker Rak
- March 20, "Creating Usability Goals: Understanding What Usability Means to Your Users." Whitney Quesenbery

The cost for each seminar is US $125.00 and Canadian $140.00. Overseas participants should contact the STC office. An additional $10.00 will be charged for registration received less than five days before the seminar. As cost-effective and time-efficient ways of improving your skills and knowledge, telephone seminars are much like a large conference call, but in a more controlled environment. Simply dial the 800 number from your telephone, enter the provided personal identification number, and you're connected. Then sit back and listen to the presentation and join in the discussions. For one registration, several employees at a company may benefit from the seminar presentation and develop their own interactive discussions.

You may get more information and register online at www.stc.org/seminars.html, or complete the registration form in the December 2001 Intercom. Contact Buffy Bennett (buffy@stc.org or 703.522.4114 ext. 251) at STC if you have any questions.