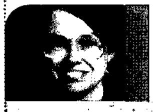

the **dollar** sign

Marydee Ojala
Editor, *ONLINE*

Finding and Using the Magic Words: Keywords, Thesauri, and Free Text Search

A major characteristic that distinguishes information professionals from casual Web searchers is our determination to find the best, most appropriate search terms. We sometimes construct elaborate search strategies to obtain highly relevant, timely, and accurate information. We don't put a couple of words in a search box and trust that the top 10 results will fulfill our information need. In fact, information professionals consider the "magic words" of controlled vocabularies to be our ace in the hole. We rely on these to find information that our clients can't, and we revel in their astonished looks when we succeed where they failed.

We understand the arcane language of thesauri, taxonomies, and ontologies. We recognize the limitations and the advantages of free text searching and know when to combine those with thesauri terms. Who else, when perusing a library catalog, would automatically search for `cookery` rather than `cook books`? The Library of Congress Subject Headings have come in for their share of ridicule—and many of the truly strange headings have been rectified. My first professional job was as a cataloger for a major multinational commercial bank, and I still remember hating the subject heading `interest and usury`. It's not there anymore, since cooler heads realized the terms are not synonymous. (EBSCO's Business Source Premier has a "see" reference from `interest & usury` to `interest`.) In 1998, it was big news that LC changed `moving-pictures` to `motion pictures`. But today's researcher on the industry is much more likely to use `films` or `film industry` as a search term. LC semi-accommodates this, having added terms such as `feature films` and `adventure films` to its vocabulary.

> Information professionals consider the "magic words" of controlled vocabularies to be our ace in the hole. We rely on these to find information that our clients can't and we revel in their astonished looks when we succeed where they failed.

Perhaps a library catalog search using the prescribed term `motion pictures` will return highly relevant books. But take that to the Web and things change. It's common language, not controlled vocabulary, that rules the day. Statistically, there is probably much more on Web pages that utilize some form of the word `film` that will return relevant results. Sundance is a "film festival" not a "motion picture festival"—although ABI/INFORM uses `motion picture festivals` as its thesaurus term. Trying to fit old words to new concepts is like trying to squeeze a size 8 foot into a size 6 shoe.

CONJURING CONTROLLED VOCABULARIES

In business literature, terminology changes in an almost faddish fashion at times. A thesaurus becomes a moving target (not to be confused with a moving picture). New technologies affecting industries are hard to pin down, terminologically speaking, when they first appear. What, for example, to make of YouTube? Suppose your research project was determining corporate use of YouTube, perhaps precipitated by JetBlue's CEO filming an apology after the airline stranded passengers last February, which the company uploaded to YouTube. Given the unique product name, you don't need controlled vocabulary to search for YouTube; you can simply search the name.

The good researcher will recognize that competitors to YouTube exist and should be added to the strategy for your research project to obtain comprehensive retrieval. To do this, you either OR together the competitor names (assuming you can determine them) or look for a controlled vocabulary term. At LC's Web site (www.loc.gov), there's one book with YouTube in the title that's been assigned subject headings. Actually, it's only been assigned one, Internet videos. Search that term in the subject heading field, and only the one book appears. This tautology is reminiscent of the distinction early databases made between descriptors and identifiers. The former were controlled vocabulary; the latter were uncontrolled vocabulary. Identifiers, in my mind, were a precursor to the current tagging phenomenon.

Blogs are another example of potential perils in choosing thesauri terms for new technologies. When they first appeared, only a few years ago, the common name was "Weblogs." Although it's still used, "Weblogs" has been supplanted by "blogs" in common parlance. If you'd chosen Weblogs as a controlled vocabulary term for your database's thesaurus (or if LC had chosen it as a preferred subject heading), you'd look foolish and need to change it. However, in checking LC's Web site, there is one book that uses Weblogs as a subject heading, while 44 have blogs.

KEYWORD SLEIGHT OF HAND

When bibliographic databases containing periodical articles began, producers published a hard copy of their thesauri, sometimes only annually. Today the databases are online and can be updated on a much more frequent basis. This allows the vocabulary to fit flexibly with changing circumstances. However, the possibilities of inconsistent application of controlled vocabulary still exist, causing experienced searchers to rely on their instincts as well as thesauri terms when selecting keywords for a search strategy.

In the financial world, special purpose acquisition companies are being formed. These so-called "blank check" companies incorporate for the sole purpose of buying other companies. ABI/INFORM actually created the thesaurus term special purpose acquisition companies, but it had only four hits at the end of May. A free text search on the acronym SPAC shows the variety of terms used. Here are three examples. Note that the first two use the thesaurus term, but the third, which is certainly relevant to the research request, does not.

"Should You Sell To a SPAC?" uses the descriptor terms Special purpose acquisition companies, Acquisitions & mergers, and Initial public offerings. It has the classification codes 9190 (CN=United States), 2330 (CN=Acquisitions & mergers), and 3400 (CN=Investment analysis & personal finance).

"Buyout managers learn new uses for SPACs from hedge funds" uses the descriptor terms Corporate finance, Fads, Special purpose acquisition companies, Trends, Buyouts, Hedge funds, Business models, and Investors. It has the classification codes 9175 (CN=Western Europe), 8130 (CN=Investment services), 2330 (CN=Acquisitions & mergers), and 3400 (CN=Investment analysis & personal finance).

> Blogs are another example of potential perils in choosing thesauri terms for new technologies.

"Market Awaits SPAC Flood as Deal Deadlines Approach: After a rush of blank-check IPOs in 2005 and 2006, some of these vehicles are approaching the cutoff date to find a deal with the threat of liquidation looming" uses the descriptor terms Investment companies, Investment banking, Acquisitions & mergers, Public companies, and Equity stake. It has the classification codes 8130 (CN=Investment services), 2330 (CN=Acquisitions & mergers), 3400 (CN=Investment analysis & personal finance), and 9190 (CN=United States).

MAGIC MOMENTS

How best to find the most appropriate terms to use in a search strategy depends upon whether you're in a database that incorporates controlled vocabulary or whether you're in a free-form environment such as the Web. Looking at the former first, in Dialog, you can use the EXPAND command to see descriptor terms. If you're using ABI/INFORM on the ProQuest CSA platform, use the Topic Guide. For EBSCO's Business Source Premier, go to the Advanced Search Option and check the "Suggest Subject Terms" box (if it's not already checked, which is normally the default option). Dialog's EXPAND lists items in alphabetical order, but ProQuest CSA and EBSCO*host* will show descriptor phrases with your term occurring anywhere within the phrase.

At Factiva Search, click on one of the Intelligent Indexing options (source, company, subject, industry, region, or language) to determine relevant index terms. Remember that most of Factiva's assignment of index terms is done using an automated, rule-based program. With Factiva Search 2.0, look at the Discovery Pane to the right of your screen of search results. Here you'll find a tag cloud of descriptor terms, which Factiva calls its News Cluster. Graphs for companies, sources, subjects, and industries show the most popular ones mentioned in the retrieved articles. Below those is a clustered group of keywords.

Another trick is to click on a descriptor term you like once you've found a highly relevant article—not only when using Factiva, but with other search systems as well. This runs a search on that term. Scrutinize the descriptor terms in the results set for additional terms either to modify your original search or to start a brand new one. Experienced researchers will identify this "trick" as the "Pearl Growing" search technique, which dates from the early 1980s. A classic article describing this, along with the Building Block and Successive Fractions techniques, is "Online Bibliographic Search Strategy Development" by Robert Wagers and Donald T. Hawkins in *ONLINE*, Vol. 6, No. 3 (1982): pp. 12–19.

Industry coding systems can suggest alternative terms for your search strategy. Suppose your research topic is finding assisted living facilities in a particular geographic area. Regardless of the political correctness or noncorrectness of the terminology, you might choose NAICS code 623110, "Nursing Care Facilities." Some of the subsets use phrases such as "convalescent homes," "convalescent hospitals," "nursing homes," and "skilled nursing facilities." NAICS code 623111 is also pertinent and uses the phrases "continuing care," "retirement communities," and "assisted-living." Given their prominence within the industrial coding system, they are excellent candidates for a free text search.

AD WORD WIZARDRY

Sometimes information professionals can take advantage of the advertising bent of Web search engines. Search engine optimizers go to great lengths to identify words most searched upon so they can use these words to drive traffic to specific Web sites. If the theory of the wisdom of crowds has any validity, then information professionals can incorporate words commonly searched into their search strategies to improve retrieval.

You can find keyword suggestions at Overture's Keyword Selector Tool (http://inventory.overture.com). It tracks actual searches (one assumes on Yahoo! since Overture is owned by Yahoo!) to suggest search terms on which to bid. You can only enter one term in the search box, although the results are frequently phrases. A search for `retirement`, for example, retrieved retirement community, retirement plan, retirement planning, retirement calculator, senior citizen retirement, active adult retirement community, retirement living, retirement gift, saving for retirement, social security retirement, and military retirement pay, among others. Note that the search term appears in all of these; it's the surrounding words that are of interest to the SEO community.

Google has a similar site for its AdWords (https://adwords.google.com/select/KeywordToolExternal). Unlike Overture's version, you can enter both words and phrases. The one word search `retirement` yielded similar phrases to Overture's, but added retirement jokes, 401K retirement, retirement age, retirement tax, early retirement, and medicare retirement. There seemed to be more searches on Google focusing on retirement parties than at Overture, and, if you scroll down to the end of the list, terms begin to appear that include assisted living but not retirement. Both Google and Overture show results by the volume of searches done on the word or phrase. However, at Google you can also display results by cost and position estimates, volume trends, and possible negative keywords.

KeywordDiscovery (www.keyworddiscovery.com/search.html) is another place to look. It collects searches from multiple search engines. There is a free word lookup component to this site, but hardcore SEO people pay $49.95 per month for a more robust version. The free version allows only for one-word searches. A recent search for `biofuels` suggested alternative terms such as `biomass`, `alternative energy`, and `manure`. You may have to do multiple searches at KeywordDiscovery, using synonyms you've already thought up, to truly mine the value of this site.

SORCERY OF SEARCH TERMS

The notion of index terms predates any of the standard business databases, since it began with printed periodical indexes and library catalog cards. The need then was obvious. The catalog and the index were guides to, respectively, journal articles and books. No full text was available. Index terms were necessary to describe the concepts in a book or article, since no one could actually apply free text search terms to all the words in the book or article. The thesauri terms acted as a surrogate for the actual items' words.

Today you can search the full text of articles in most business databases, the full text of books on Amazon and Google Books, and the full text of Web pages using any Web search engine, making the value of indexing somewhat questionable. Additionally, as the Web search engines improve their relevancy algorithms and augment their content sources with premium content, the importance of controlled vocabulary decreases.

Most information professionals point to the value-add of thesauri to justify the cost of traditional databases. In fact, Factiva published a white paper on this very topic, written by Jan Sykes, back in 2001. This is a point we need to reexamine as the online searching world changes and evolves. In the words of Old Lodge Skins, played by Chief Dan George in the 1970 movie (film? motion picture?) *Little Big Man*, "Sometimes the magic works, and sometimes it does not." If it works less often than it did in the past, how much of our expertise in using keywords, indexing, and thesauri will be valuable in the future? Our magic words could lose their potency, diminishing our sorcery skills and weakening our conjuring capabilities.

Marydee Ojala *(marydee@xmission.com) is the editor of* ONLINE: The Leading Magazine for Information Professionals.