True and tested products: thesauri on the Web

Michèlle Hudon

Hundreds of thesauri are now available on the Web in either static or dynamic HTML formats, most as standalone products. Even though they have moved into a different environment, their nature and structure, and even their main functions, have remained the same. The article discusses the issue of whether the thesaurus can play a significant role in web-based information retrieval without evolving into a more flexible tool.

Database and book indexers are familiar with the indexing and retrieval thesaurus, a tool they have used primarily as a source of controlled vocabulary to represent the subject content of information resources. The thesaurus was introduced as an indexing aid at about the same time as computer-assisted information retrieval became more common, but well before the advent of computer-driven information searching and full-text availability.

Many predicted the demise of the thesaurus in the short term, along with that of other controlled indexing and searching languages, as powerful search engines appeared on the Internet and on the World Wide Web, and information resource description and retrieval suddenly became everybody’s business. Yet, ten years or so later, thesauri are still around, having made their way onto the Web itself, and acquired new functions in the process. Thesauri can assist with learning and assimilating information, provide knowledge-based support for end-user searching, improve information display, provide indexing assistance and facilitate multi-database searching. Remarkably, thesauri are still being developed in accordance with sets of national and international guidelines formalized in the 1960s, to which few significant modifications have since been made.

The first 40 years: 1959–1999

Following a ‘lengthy and confused intellectual’ pre-history (Roberts, 1984: 282), the first modern thesaurus emerged in 1959 as a secondary, suggestive rather than prescriptive, indexing aid used by indexers to enlarge their vocabulary. The thesaurus had developed over a period of 15 years under the combined pressures of rapidly growing, entirely new subject areas and collections, of new patterns in the use of information, and of spreading applications of computers in information storage, processing and retrieval. While a first group of information specialists chose to abandon the orthodox retrieval principles applied by libraries, particularly through bibliographic classification schemes, a second group led by Calvin Mooers and Charles Bernier asserted the value of terminological control and conceptual structuring; it is, of course, the second group’s thinking and work that led to the development of the thesaurus as we have come to know it.

The Cranfield studies considered all indexing languages as fundamentally similar, consisting of an indexing vocabulary together with means of showing semantic relationships to help improve recall, and with syntactic devices to help improve relevance (Cleverdon, 1967). The traditional content of a thesaurus conforms to this description. Its usual components consist of a lexicon (or entry vocabulary), scope notes and a network of relationships among terms.

The third component of the thesaurus – its network of paradigmatic relationships of equivalence, hierarchy and associations between terms and/or between the concepts represented by these terms – is its focal element. The relationships allow thesaurus users to perceive the underlying logical structure of a domain as it is represented in its literature. The integration of relationships into the then developing thesaurus was suggested by Bernier in 1957. From that time on, much emphasis was placed on this component of the thesaurus, which was to become a critical structural element distinguishing thesauri from other types of controlled indexing languages.

Holm and Rasmussen believed that the thesaurus being developed at Dupont de Nemours offered the level of conceptual and terminological control needed ‘to eliminate the educated guesses . . . made by both indexers and retrievers in order to find information stored in the system’ (1961: 187). Since then, the main functions of the thesaurus have been defined as follows:

. . . to promote consistency in the indexing of documents, predominantly for post-coordinated information storage and retrieval systems, and to facilitate searching by linking entry terms with descriptors. (National Information Standards Organization, 1994: 38)

Reich and Biever (1991: 336) have described these functions as the indexing consistency function, a thesaurus introducing ‘order and language standards into indexing terminology’, and the retrieval function, a thesaurus serving as a ‘source for searching vocabulary’.

At the input end of an information system, it has always been assumed that the availability of a thesaurus facilitates the indexing process and contributes to the maintenance of indexing quality. Over the years, the thesaurus became more of an indexing aid than a searching one, and indexing quality and consistency came to be seen as a guarantee of better retrieval. Although a natural function of the thesaurus would be to remind them of related ideas and terms that might be valuable in searching, information searches seldom took advantage of the full lexical suggestive power and navigational capabilities of the thesaurus.
Several reasons have been given to explain information searchers' lack of interest in an instrument that has the power to increase the efficiency of a search, reducing the time required to recall a suitable and satisfactory number of relevant information sources. Unfortunately, thesauri used for indexing the contents of databases have not always been made available to searchers in a format that would be convenient and appropriate to the environment. By 1990, even if thesauri were already an integral part of most computerized information-retrieval systems, most were still available only in print format (Pollard, 1990: 158), and print thesauri tended not to be used by online searchers who did not understand their format and structure, and had too little time and patience to learn about them. As for thesauri available online, they were criticized for being incomplete, confusing and difficult to use, with displays that were judged at times inferior to those of their printed counterparts (Weinberg and Cunningham, 1988: 417).

The display of thesaural data has always been a sore point with thesaurus designers and users. Thesaurus users – indexers and searchers alike – have repeatedly requested flexibility in presentation and modes of access, suggesting that various types of listings and graphic displays could help them move about in the thesaurus and would allow them to bring into focus the portion of the structure required at a given time (Bertrand-Gastaldy and Davidson, 1986). Graphic displays of thesaural data, which might be particularly useful to non-specialist users who do not, and need not, understand the conventions used by the indexer nor the subtleties of the indexing language (Bertrand-Gastaldy and Davidson, 1986), were never popular amongst thesaurus designers and managers. Bertrand-Gastaldy and Davidson (1986: 241) remarked on 'the conservatism that characterizes the elaboration and presentation of thesauri', suggesting that thesaurus designers develop a better understanding of how information is communicated, and try to assimilate the new technologies so that they could make more efficient use of them. Nevertheless, matters relating to display consistently received little attention in the scientific and professional literature, as well as in the standards and textbooks relating to thesaurus construction.

Because thesaurus users obviously had no choice but to adapt themselves to the tool, suggestions were made for the integration in information operations of another type of thesaurus, the end-user thesaurus, designed to guide and facilitate end-user searching of textual databases. End-user thesauri would function as part of the interface to link the vocabulary of the searcher and the vocabulary of the database. Sources for terms included user queries, texts containing potential answers to queries, expert human indexing and existing linguistic reference tools, such as other thesauri, dictionaries, glossaries, etc. The end-user thesaurus never acquired any kind of official status; on the web field, it has quickly lost more ground than its more traditional competitor.

Thesauri on the Web

The first discussions associating thesauri and the Web took place in the early 1990s. The main advantages of publishing thesauri electronically were then defined as a marked reduction in the costs of producing, storing and shipping bulky print products, a predicted increase in the use of thesauri because of lower purchase costs and extended availability, and an ease of updating ensuring that users would always have access to the latest additions and modifications (Davies, 1995: 37). What was seen as an attractive option in 1995 rapidly became the only option if thesauri were to play a role in the new information world. By the end of the decade, hundreds of them had made their way onto the Web. To date, thesauri in the English language still dominate the offerings, with an increasing number of multilingual thesauri also becoming available.

Interestingly enough, thesauri are now marketed as searching aids in an environment in which full-text searching is prominent. It seems obvious that the semantic structures provided by thesauri can play an important role in both organizing and retrieving web-based information sources. Subject access through a knowledge structure such as a thesaurus or a classification system is a significant feature of good-quality subject gateways (Shiri and Crawford, 2000: 277). Gateways use thesauri to index web pages and provide structured and more consistent subject access for browsing and searching, much as was the case in the earlier era of human-based indexing and computer-assisted retrieval. Unfortunately, web browsers are not currently thesaurus-aware, and existing metadata formats actually make little use of thesauri.

Finding thesauri on the Web also remains a problem. Lists of online thesauri, such as the High-level Thesaurus project (HILT) A–Z of thesauri (http://hilt.cdirl.strath.ac.uk/Sources/thesauri.html) and the Databases of thesauri set up at the Queensland University of Technology (http://sky.fit.qut.edu.au/~middletm/cont_voc.html#Databases) are at best incomplete. Like any other web resource, full thesauri have been known to appear, relocate and disappear for unclear reasons. If thesauri are to become a standard feature of web-based information transfer activities, it is imperative that an online registry be established and maintained to replace the traditional thesaurus clearing-houses where thesauri were collected, described and indexed, and from where information on them could be disseminated.

Structure and display

Traditional thesauri can be difficult to navigate, but contemporary browsers and navigators should make them easier to use, thanks to web-related technologies. Software producers have, indeed, been quick to produce new versions of their traditional thesaurus management applications, offering more user-friendly web interfaces, together with navigation and browsing capabilities. Thesauri available on the Web at this time, most of which were at some point traditional print thesauri, exhibit varying levels of sophistication. A diminishing few are still offered in simple static text format or as flat files – exact replicas of their printed counterparts. Static thesauri consist of a series of files generated at a particular point in time, which are then delivered as a whole by the web server to the user's web browser. The user's first access to the thesaurus is through a
The screen display theoretically offers possibilities for flexible screens to replace the limitations of the printed page. The limitations of alphabetical and hierarchical displays, the limitations of the flat-file thesaurus, even if it does make use of hyperlinks and allows for relatively easy navigation between its various sections.

Most thesauri are presented in HTML format, with more or less effective use of hyperlink technology. A few still use hyperlinks to connect the various sections of a thesaurus, but not individual term records, while the majority have fortunately adopted a dynamic HTML format with fully navigable hyperlinks. In the truly dynamic thesaurus, term records are stored in a database accessed by a search engine that retrieves records on request (Davies, 1995). The AGROVOC thesaurus (http://www.fao.org/agrovoc/), the EUROVOC thesaurus (http://europa.eu.int/celex/eurovoc/) and the OECD macrothesaurus (http://info.uibk.ac.at/info/oecd-macroth/) are three examples of tools that make good use of hyperlinks.

Schneiderman (1989) has suggested that hypertext is appropriate in situations where ‘there is a large body of information organized into numerous fragments, the fragments relate to each other, and the user needs only a small fraction at any time’. Pollard suggests, in turn, that thesauri fit this pattern exceptionally well:

A typical thesaurus contains many thousands of individual terms that are explicitly related to one another in a well-defined manner [and] at any time the user of a thesaurus is interested in identifying only a small subset of the available terms. (Pollard, 1990: 160)

Hyperlinks appear, indeed, as a most appropriate technology for exploiting to its full extent the semantic structure of a thesaurus, its most important and significant component.

Of course, a hypertextual presentation of thesaural data does not solve all access problems. The simple alphabetical display, traditionally the main display in a thesaurus, can be easily mapped onto a hypertext system, but it has the inherent disadvantage of not showing a full hierarchy at the entry point for a term; the same was also true in the print format. Despite the importance of the concept, and especially in view of problems already identified with the display of thesaurus data, little work on hypertextual presentations of thesauri has been reported in the literature. As for graphic displays, they are as uncommon now as they used to be.

While the electronic display of thesaurus data can remedy some of the limitations of the print format, making it possible, for example, to switch back and forth between alphabetical and hierarchical displays, the limitations of the screen have replaced the limitations of the printed page. The screen display theoretically offers possibilities for flexibility and customizability that one could only dream of in the print environment. But in fact some web-based thesauri are currently less rich in access than the print versions of the same product (Milstead, 1998).

In the mid-1980s, Bertrand-Gastaldy and Davidson (1986: 246) had suggested the following rules for the design of thesaural displays and interactive interfaces to the data:

- maximize mnemonic devices (colour, typefaces, etc.);
- provide the displays rapidly, with rapid movement within the thesaurus;
- allow simple consultation, with simple instructions;
- permit different approaches to consultation – for the specialist or non-specialist user, and for indexing and retrieval;
- allow different starting points – from a specific descriptor in context or from an overview with zoom;
- order the information;
- offer a choice of graphic or textual displays without leaving the program.

The transition of the traditional thesaurus to a web environment was most likely the perfect time to put these rules to the test. But, on the whole, ways of presenting data and of interacting with them have changed very little. Interfaces are sober, functional and effective (see for example the Seattle city clerk thesaurus at http://clerk.ci.seattle.wa.us/~public/newtoc.htm), but little imagination has gone into making thesauri attractive to non-specialist searchers; experiences provided by such innovative tools as the Plumb Design Visual Thesaurus (http://www.plumbdesign.com/thesaurus/index.html) remain more intriguing than truly useful. Is it that the traditional structures and interfaces are so good and are working so well that there is no reason to re-examine them? Or could this be due to the fact that standards for the construction, management and display of thesauri have yet to be revised, that they do not yet account for the amazing capabilities of new processing and interfacing technologies, and that they do not yet account for the fact that information indexing and searching are no longer the preserve of trained specialists?

Functions and uses

As was also the case in the traditional environment, thesauri vary as to their functionalities and usability. Most web-based thesauri are not fully embedded as search and browse aids in information-retrieval systems and web search engines; presumably, as standalone products, they continue to be used in the development of search strategies or as reference tools for learning more about the structure of a domain. Loading the thesaurus onto the Web is the easy part; it is impossible to know how often, by whom, and for what exact purpose, the thesaurus is being consulted.

Thesauri that are fully integrated into information-retrieval systems are of more immediate use for web-based information retrieval. Thesauri that operate in connection with a database, such as ERIC (http://searcheric.org/) or HASSSET (http://155.245.254.46/services/zhasset.html), offer advanced support by providing information searchers with a suitable starting point and by using thesaurial relationships for query expansion. These thesauri also help searchers identify and select alternative search terms and use them for query reformulation purposes (Shiri and Crawford, 2000). Searchers do appreciate this painless way of using thesauri, but they are not necessarily
aware of the potential impacts on recall, precision and general quality of results.

**Potential developments: what does the future hold for thesauri?**

Through its journey to a new information environment, the thesaurus has remained pretty much the same in nature, structure and functions. It remains popular, thanks to the vigilance of trained indexers and librarians who can be quite convincing when talking about what is, no doubt, a very useful indexing and searching aid in any environment. But can the thesaurus survive in the medium and long term in such an interactive and dynamic environment as the Web, where its competitors – ontologies, taxonomies, semantic networks and even traditional classification schemes – continue to break new ground?

As early as 1990, Schmitz-Esser was issuing his first call for a new round of standardization that would lead to the second-generation thesaurus. This would be used in expert systems where its role would be to ‘keep track of the real-world terminology and to relate it to the terms of the pre-established thesaurial structure’ (Schmitz-Esser, 1991: 144), converting users’ requests into query statements as required by the information system. Ten years later Schmitz-Esser (1999) again suggested that the survival of the thesaurus is guaranteed only if it evolves into a tool useful for linguistic engineering. Everybody agrees on the principle, but there is as yet no consensus as to what exactly should be changed and what the thesaurus should become. Moves towards a richer structure with more information of a definitional nature are discussed, but to date little has been done in that direction.

Expansion and specification of the set of relationships that structure the thesaurus are recurring proposals. The limitations to three types of semantic relationships have always been criticized, and the ill-defined and all-encompassing associative relationship was never considered appropriate for supporting non-hierarchical query expansion, particularly in an automated environment. Not only has it been suggested that relationships should be specified and named (e.g. cause–effect, agent–action, process–product, etc.), but discussions as to whether desirable relationships may be different for different categories of users, from scientific and technology-orientated domains to humanities, from culture to culture, and so on, have also been launched. Those who still use thesaurus to index admit that more specification could indeed be useful, but they are wary of the complexity that such specification would introduce in a tool already considered difficult for untrained people to use. Where thesauri are used by search engines, it remains to be demonstrated that a more precise but more complex structure could be more useful than what currently exists. The desirability for all searchable to browse the Web using their own personal thesauri has also been discussed; although theoretically interesting and eventually feasible, the idea is not very practical, the actual number of users displaying an explicit need for a thesaurus to navigate the Web being quite small.

Even if thesauri continue to exist for a while as separate tools to which both indexers and searchers refer when tagging documents and queries, their survival cannot be guaranteed if this continues to be seen as their main function. The thesaurus must move behind indexing and searching interfaces in order to play its role as a vocabulary control tool that does not require users to interact with it as a separate operation unless they really want to. As indexing aids, well-structured thesauri combined with language processing techniques should be capable of ‘bestowing a second wind to automatic indexing’ (Miller: 333), a next step being the indexing of ‘Internet massives’ (Miller: 334). As searching aids, thesauri can continue to be used for query expansion or limitation, and as filters to improve the relevance of retrieval.

Another threat to the long-term survival of the thesaurus may actually be linked to its popularity and the fact that so many different thesauri currently exist and continue to be developed. In an environment where interoperability, reusability and shareability are dominant values, the proliferation of more-or-less standardized and more-or-less specialized thesauri overlapping with several others is probably not the best move. Modified tools, such as superthesauri, meta- and macro-thesauri, have moved beyond the proposal stage and are currently being tested in organizations that distribute information not restricted to a subject domain or discipline. Around the world, governments are developing such thesauri: examples are the Government of Canada Core subject thesaurus (www.thesaurus.gc.ca) and the US GILS topic tree (www.fidocat.com/GILS/two.pdf). These types of general/overarching thesauri could eventually find their way to other, non-governmental, areas of the Web where they could be most useful.

The participants at a joint NISO/APA/ASI/ALCTS workshop held in 1999 concluded that there was a need for an updated standard that would put the emphasis on interoperability and provide for a broader group of controlled vocabularies than just the thesaurus. It is not clear whether any more work has been done on this front, and whether, three years later, the original recommendations remain relevant. Despite the major changes that have affected the environment in which thesauri are used, the standards committees have remained surprisingly quiet on matters relating to thesaurus construction. This is not really surprising given that standards tend to officiate existing practices rather than to lead the way into new territories. A new standard should be submitted to the community as quickly as possible, however, if the goals of conceptual and technological compatibility are to be kept within reach.

**Conclusion**

Few people would deny any more that the more full text there is, the more help users need in navigating it, even where a powerful search engine is available (Milstead, 1998). This is why thesauri will survive and should even have a bright future on the Web. But we are now at a crossroads, not unlike Mooers and Bernier in the 1950s, who concluded that existing subject representation tools – classification schemes and subject headings – were not entirely appropriate for use in the developing computer-assisted or
computer-supported information systems. They selected the best parts of the old tools and added new elements to shape the whole into a new form that would behave well for uses then current (Mooers, 1985: 249). The first-generation thesaurus was born, an efficient indexing and searching aid that became immensely popular.

A second-generation thesaurus is now, indeed, needed. The new thesaurus will be developed with the information-searchers’ needs and habits in mind and be structured so that it can be used most efficiently in search-engine-driven information environments. Much time is currently being spent trying to convince various categories of web managers and users that a slightly modified version of the traditional thesaurus, a tool that remains very expensive to develop and maintain, is something they absolutely need. Even more time should probably be devoted to the design and testing of truly new, richer and more versatile models. On the semantic level, the thesaurus of the future should offer better-defined relationships, specifying the nature of the linkages between concepts and between terms. On the practical level, the second-generation thesaurus should be designed and offered in a more interactive fashion, along the lines of what Bertrand-Gastaldy and Davidson suggested even before anybody could predict that thesauri would eventually be used on global networks, by non-specialists and for purposes other than information indexing and searching.

Notes
1. Although there is evidence that a few indexing and retrieval languages similar in form and content to the thesaurus had been used during the previous decade in experimental situations (Roberts 1984), the first thesaurus of the modern era is considered to be that developed in 1959 for the indexers at E.I. Dupont de Nemours Engineering Department. This thesaurus, however, was neither published nor widely distributed. The Chemical engineering thesaurus, a direct derivative of the Dupont tool, was published in 1961.
2. All websites referred to in the text were accessed on 13 March 2003.

References

Michele Hudon is Assistant professor at the École de bibliothéconomie et des sciences de l’information, Université de Montréal, Canada. Her areas of expertise include indexing procedures and evaluation, and thesaurus design and evaluation. She is the author of a textbook on thesaurus construction (published in French by Les Éditions ASTED in 1995) and of several articles on multilingual thesaurus development. She is a former President of the Indexing and Abstracting Society of Canada (IASC/SCAD). Email: michele.hudon@umontreal.ca

User manual index praised!
Given the quality of user manuals these days, the [Nikon digital camera] 4500’s is outstanding. It has an excellent index, which includes a list of the icons used by the camera and the page references for them. I can’t count the number of times I’ve forgotten the meaning of an icon and had to pore over the pages of a manual searching for the key.

Website testing is an easy way to make money if you have a few spare minutes and a phone or computer with an internet connection. Testbirds is conducting tests on the following products: Websites and portals. Mobile apps. Internet of things (IoT). Wearables. Games. You will earn points by testing websites and apps that can be redeemed for cash and other rewards. Summary. Website testing is a great way to make a few extra dollars each month since most tests only take 20 minutes or less. And, since you most likely have a laptop or smartphone with a built-in camera and microphone, all you have to do is apply to become a tester and immediately begin testing apps and websites. Do you test websites when you’re not taking paid opinion surveys? What’s your favorite site? Truest definition, being in accordance with the actual state or conditions; conforming to reality or fact; not false: a true story. See more. Test your memory on these verbal firecrackers from the week of June 29 to July 5! Question 1 of 7. anchorite. These objects were praiseworthy, yet they were not the highest aims of the truest and purest ambition. Captain Cook|W.H.G. Kingston. In such a union as the author quoted has here described men and women find life's deepest and truest joys and satisfactions. Analyzing Character|Katherine M. H. Blackford and Arthur Newcomb. See more examples. Tested and True. 2 February 2018 Â· How To Make Money Testing Free Products from Home. Did you know you could make money testing free products from home? There are TONS of ways to test out free products and get paid for it. realonlinesurveys.com. Get Paid To Test Free Products Online: 10 Legit Product Testing Sites. Did you know you could make money testing free products from home? There are TONS of ways to test out free products and get paid for it. Did you know you could make money testing free products from home? There are TONS of ways to test out free products and get paid for it. Tested a... Since 2000 MyOffers has been giving away some of the best prizes on the