

## Proceedings of the 7<sup>th</sup> ASIS SIG/CR Classification Research Workshop

### Publishing Thesauri on the World Wide Web

Ron Davies

*The World Wide Web promises to become an important medium for publishing thesauri as electronic publication becomes more common. Thesauri can be published over the World Wide Web in either a static form (where all thesaurus information is written to files at one time) or dynamically (where information is retrieved by a search engine from a database, and formatted into Web pages just at the time that it is delivered to the user). The choice between static and dynamic methods of Web publication influences the format and organization of the thesaurus. This paper discusses some of the aspects that most affect the way in which users consult the thesaurus, in particular the method of locating a term, displaying a term and response time.*

#### **Introduction**

Publishing thesauri electronically has important advantages. It reduces the cost of producing, storing and shipping bulky printed publications, thereby making thesauri more affordable both for publishers and users of thesauri, particularly when users are spread over a wide geographic area. Lowering purchase costs encourages the use of the thesaurus both for indexing and searching. Electronic publishing also allows for more frequent updates, ensuring that users—even occasional searchers—have access to the latest additions and modifications.

While there are a number of different means of publishing a thesaurus electronically, publication over the World Wide Web is particularly attractive. The rapid expansion of the Internet has meant that more and more people have access to Internet services. Many users are already familiar with common Web browsers, particularly those that support a graphical user interface such as Netscape, and do not require any particular browser training. They are increasingly accustomed to using the Web as a source for a wide variety of types of information, and can read about Web sites in popular newspapers and magazines as well as technical publications. Users who search a database through a World Wide Web gateway, either through the Internet or through a corporate intranet, find it particularly convenient to be able to consult the thesaurus used in indexing that database through the same user interface.

In some respects the Web is an ideal medium for publication of thesauri. The markup capabilities of the HyperText Markup Language (HTML) allow a fairly sophisticated presentation of information. The most common browsers use variable-spaced, high quality fonts in a variety of type sizes and weights, and allow the display of graphics to enhance the presentation of the information. The hypertext capability of the Web can be easily applied to the process of navigating through the tree-structures of a thesaurus, allowing users to select a term from a display in a simple way and move directly to additional information on that term. While Version 2.0 of the HTML standard does not allow for the kind of tabular formatting

## Proceedings of the 7<sup>th</sup> ASIS SIG/CR Classification Research Workshop

required for many types of thesaurus display, extensions to HTML to permit the use of tables are already widely supported and will likely be adopted into the standard in the future.

Publishing a thesaurus on the World Wide Web, however, is no different than publishing any complex document in a new medium— there are important decisions to be made in terms of design, format and organization. These choices are at least as complicated as those involved in designing a printed thesaurus, and probably more difficult in the sense that there are not the same number of well-known thesauri already published on the World Wide Web to provide a guide. This paper discusses some of those design issues, and in particular how the mode of publication of the thesaurus – static or dynamic— affects the decisions to be made.

### ***Static and dynamic publication***

A thesaurus can be published on the Web in either a static or dynamic format. In a **static** format (Fig.1), the published thesaurus consists of a series of electronic files, generated at a one particular point in time, which are then delivered without change by the Web server to user's Web browser. The static publication process is similar to the traditional publication of a printed thesaurus, where (usually at the end of a long revision process) the thesaurus is printed on paper or an electronic file is produced and sent to a printer for electronic typesetting. As in the traditional print publication, the set of Web files created may include series of classified, term or permuted term index files with links to more complete term information found in one or more main listings, as well as main term listings in a number of different sequences with complete or abbreviated information. The principal differences are that in publishing a thesaurus on the World Wide Web, data is output from the thesaurus management system to electronic files in HTML, and the HTML files contain the internal hypertext links necessary to allow users to move easily from one part of the thesaurus to another.

**Dynamic** thesaurus publication, on the other hand, relies on an entirely different process (Fig. 2). In a dynamically-published thesaurus, thesaurus records are stored in some form of database (not necessarily the same database used to manage the thesaurus) accessed by a search engine that retrieves records from that database. Communication between the search engine and the Web server takes place through a Common Gateway Interface (CGI) program that translates between the language of the World Wide Web (HTML and the HTTP protocol) and whatever language the specific search engine understands. In other words, when a user submits a request from a Web browser, the request is translated by the gateway into a request to the search engine; the engine performs the search and returns results to the gateway program, which re-formats the response "on the fly" into HTML format and sends it on to the Web server. Each time that a user consults the thesaurus, the response is generated dynamically, and is not stored in the form in which it is delivered. Both full term listings and index listings with abbreviated term information may be generated in this way, depending on the capabilities of the search engine and corresponding gateway.

## Proceedings of the 7<sup>th</sup> ASIS SIG/CR Classification Research Workshop

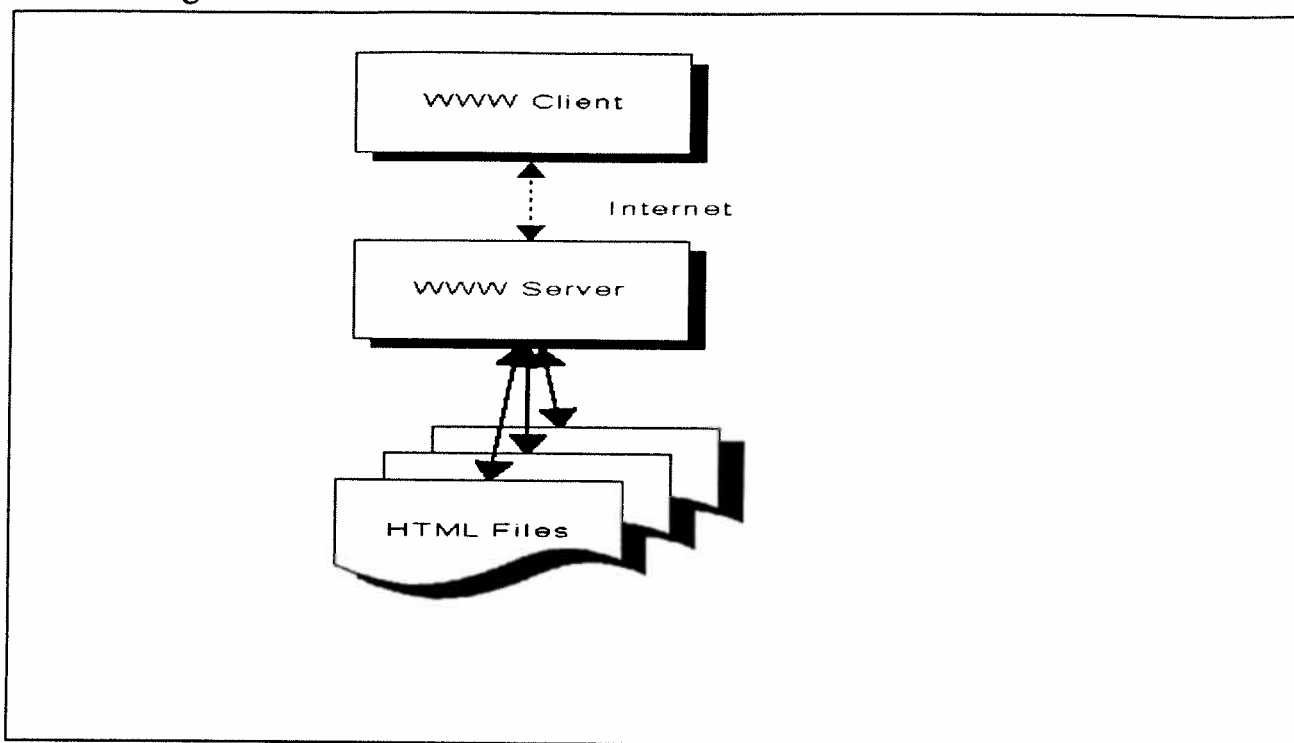


Fig. 1 Schematic representation of static publication

Neither static nor dynamic publication is inherently better: thesaurus developers wishing to publish over the World Wide Web must decide which format is most appropriate to their particular circumstances, depending on a number of different factors. The rest of this paper highlights some of the important design issues that affect how the end user will use thesauri published over the Web, and this ultimately will determine which method is most appropriate for a particular thesaurus. While examples given throughout the discussion are based on a thesaurus with primarily an alphabetical structure, the same points apply to thesauri based on faceted or classificatory principles.

### ***Locating terms***

At the initial stages of consulting a thesaurus, users may already know the descriptor they want to see, or they may want to supply only a partial (i.e. truncated) term, or a keyword that might be found in the descriptor. Regardless of which type of search the user wishes to use, the way in which a term is located differs depending on whether the thesaurus is published statically or dynamically.

With a dynamically-published thesaurus, locating a term is a straightforward process. For an initial entry point into the thesaurus, the user would typically enter the desired term into an HTML form (Fig. 3). The Web gateway then takes the entered term and passes it to the search engine, which would use the value to search the thesaurus database. If only one term is retrieved in response to the search, then the search engine can perform a simple search on the database, retrieve the corresponding term record, and return the result to the gateway. The gateway then re-formats the term record into HTML format and returns the page to the user's Web browser. If more than one term satisfies the request (for example, when the user has supplied a truncated term or keyword) then typically a page with only those terms are returned (Fig. 4), from which the user can select the term that is of greatest interest. In either case, due

## Proceedings of the 7<sup>th</sup> ASIS SIG/CR Classification Research Workshop

to the dynamic nature of the gateway and underlying search engine, the response is directly related to the request of the user.

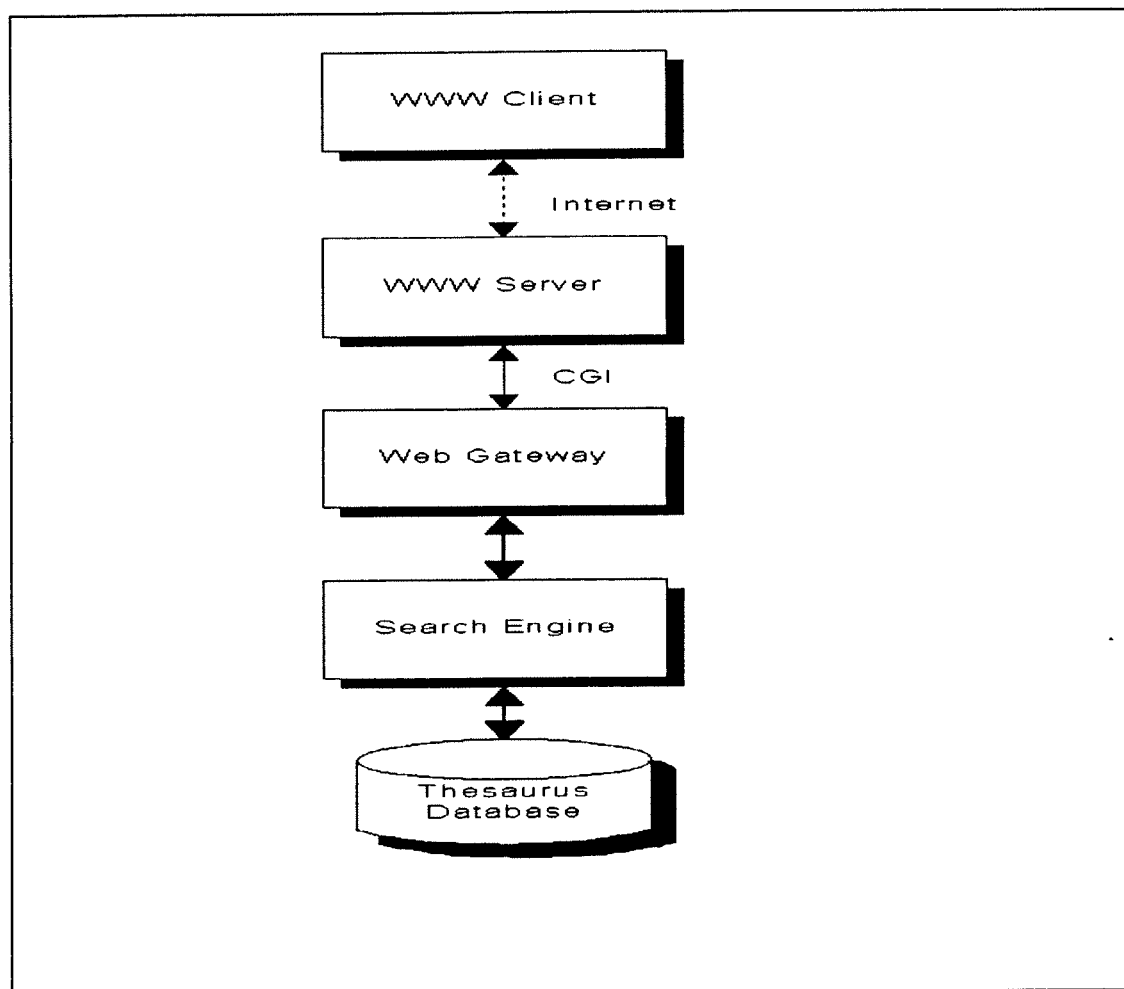
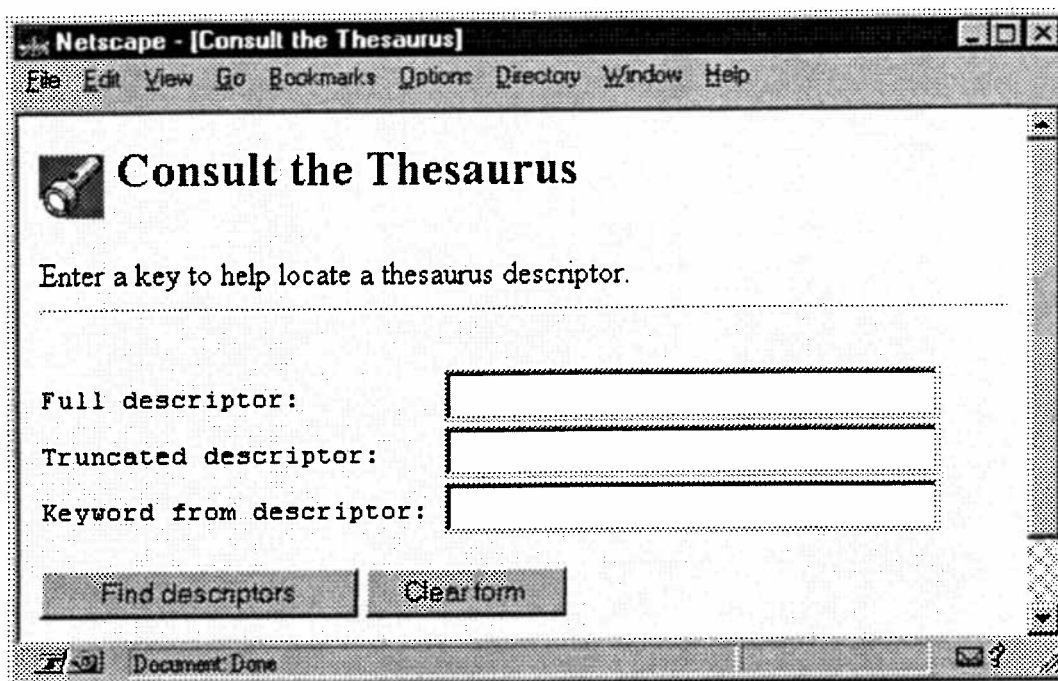


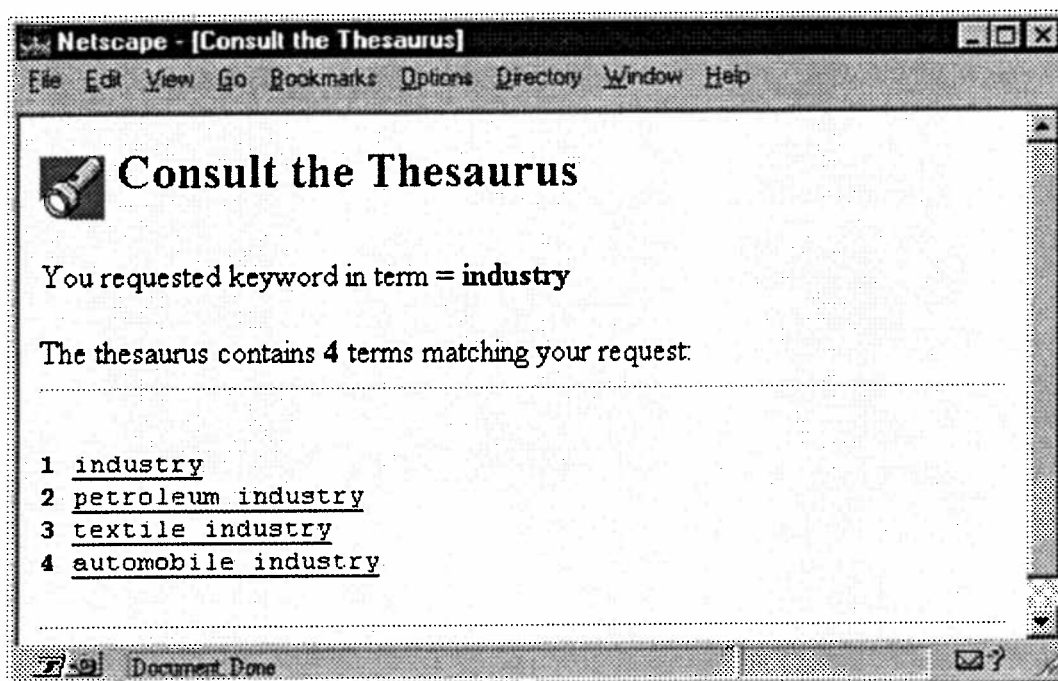
Fig. 2 Schematic representation of dynamic publication

## Proceedings of the 7<sup>th</sup> ASIS SIG/CR Classification Research Workshop



The screenshot shows a Netscape browser window titled "[Consult the Thesaurus]". The menu bar includes File, Edit, View, Go, Bookmarks, Options, Directory, Window, and Help. The main content area has a heading "Consult the Thesaurus" with a small icon of a key. Below the heading is the instruction "Enter a key to help locate a thesaurus descriptor." followed by a dotted line. There are three input fields: "Full descriptor:", "Truncated descriptor:", and "Keyword from descriptor:". Below these fields are two buttons: "Find descriptors" and "Clear form". The status bar at the bottom shows "Document Done".

Fig. 3 Search key input form



The screenshot shows the same Netscape browser window, but now displaying search results. The heading "Consult the Thesaurus" is still present. Below it, the text reads "You requested keyword in term = **industry**". This is followed by "The thesaurus contains 4 terms matching your request:" and a dotted line. A list of four results is shown, each numbered and underlined: "1 industry", "2 petroleum industry", "3 textile industry", and "4 automobile industry". The status bar at the bottom shows "Document Done".

Fig. 4 Search results

Locating a term in a statically-published thesaurus is more involved because the interactivity provided by the search engine underlying the gateway is not available. Typically in a statically-published thesaurus, the user's first view of the thesaurus is a high-level alphabetical or

## Proceedings of the 7<sup>th</sup> ASIS SIG/CR Classification Research Workshop

classified index (Fig. 5). The user must first select a value from this page. He or she will then be presented either with an additional, more specific index, from which another value must be chosen, or a large list of terms, which includes all the terms that share a particular characteristic and through which the user must scroll to find the exact term desired. The process of locating a term in a statically-published thesaurus is necessarily a multi-step process, with several more levels of interaction and less specificity than in the dynamically-published thesaurus.

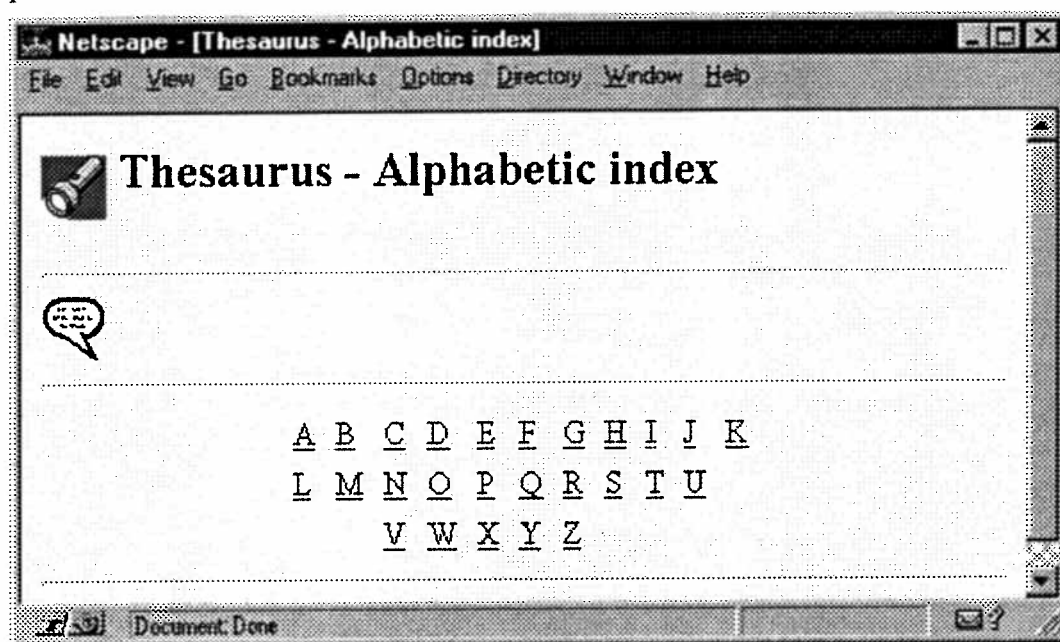


Fig. 5 High-level index

### *Displaying terms*

When a user finally locates a term within the thesaurus, the way in which that information is displayed also differs in statically and dynamically-published thesauri. In a statically-published thesaurus, information on a single term is unlikely to be in a file by itself. While it is certainly possible to publish a thesaurus with each term in a separate file, even a modest thesaurus of one or two thousand terms would result in a very large set of files that would be difficult to manage. Statically-published thesauri are much more likely to group information about a number of terms into a single HTML file. For example, all the terms in a particular facet, all terms in a subject category, or all the terms beginning with a particular letter of the alphabet, might be listed together in one file. Users can navigate directly to individual terms using specific place-marking anchors within that file: for example, they can click on the hypertext link represented by a narrower term to move directly to the full term display for that descriptor. However the term selected is always displayed in the context of the immediately preceding or following terms (Fig. 6).

With a dynamically-published thesaurus, the amount of context that is displayed with a given term depends in large measure on the capabilities of the search engine that underlies the World

## Proceedings of the 7<sup>th</sup> ASIS SIG/CR Classification Research Workshop

Wide Web gateway. In order to present a sequence of terms, the search engine must be able to browse an associated index, collecting terms that fall before or after the target term in the desired sequence.

With some simple search engines (including gateways based on the WAIS search software), it may not be possible to browse an index in this fashion; with other software programs, it may be possible to move forward in a sequence, but not backward to provide a previous term or terms. Software packages without browsing capabilities can present to the user only a single term at a time. While the term can be presented with hypertext links pointing to previous and subsequent terms in a sequence (perhaps represented as "backward" and "forward" navigational buttons as in Fig. 7) to facilitate browsing, the user does not have the ability to see, at a single glance, a wide range of terms. This limitation can be particularly important with faceted or classified thesauri, where the sequence of terms plays an important part in showing associations between terms and elucidating the structure that underlies the thesaurus.

### *Response time*

With statically-published thesauri, and with dynamically-published thesauri with search engines that can browse and collect terms to provide a context for thesaurus information display, the thesaurus publisher must decide how much information to pass on to the user at one particular time. This can be a critical factor in determining how quickly the user can consult the electronically published thesaurus: users with 14.4 or 28.8 kilobyte SLIP or PPP connections to the Internet are only too aware of the difference in response time when accessing an HTML file of ten thousand bytes, and when accessing a similar file of a several hundred thousand bytes. Response time (i.e. the time that elapses between the moment the user requests information and the moment the user receives a response to that request) is particularly complex in an Internet environment where the speed of a network application depends on a number of factors (e.g. network bandwidth, modem speed in the case of dial-up lines, network performance, processing power of the client computer) that are beyond the publisher's control. The most important factor that the publisher can determine is the size of the HTML page that is delivered to the end user.

In a statically-published thesaurus, the publisher determines the size of the HTML files once for all users, at the time when the static HTML pages are generated. The publisher will likely try to come up with a size that both corresponds to the logical internal structure of the thesaurus (e.g. by placing all terms in a certain facet or all terms beginning with a certain letter in one file), and that results in an acceptable response time over a perceived average network connection with a perceived average client workstation. In some cases, if many descriptors begin with the same letter or are found in the same facet, it may be necessary to ignore the logical structure and divide terms into two or more separate files. The fact that the file size cannot change with a statically generated thesaurus also means that this "average" will not do justice to the needs of users on either end of the performance spectrum, i.e. users with slow network connections or small computers, or users with fast network connections or powerful computers.

## Proceedings of the 7<sup>th</sup> ASIS SIG/CR Classification Research Workshop

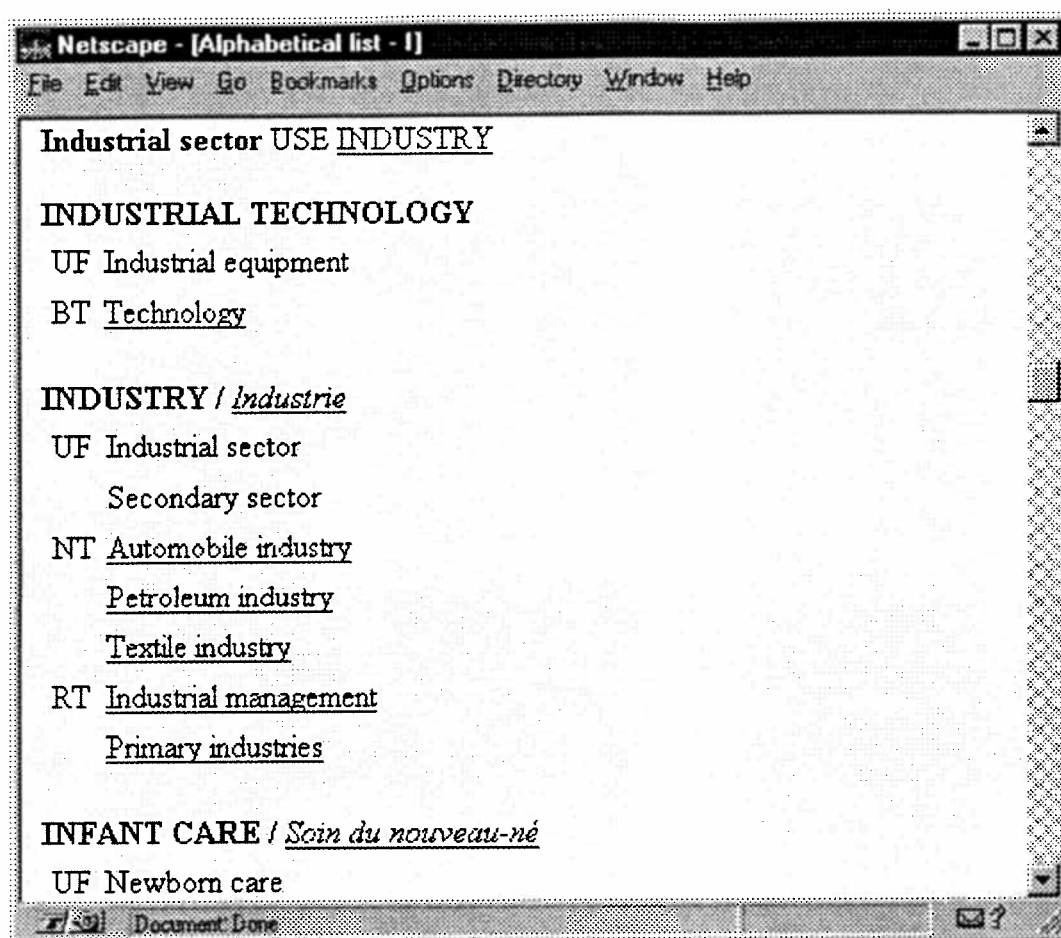


Fig. 6 Term display in a static publication

In a dynamically-published thesaurus, results are generated "on the fly." While it is important to respect the intellectual structure underlying the thesaurus, it is at least theoretically possible to change the behaviour of the Web-published thesaurus to try to suit the user's particular hardware, software or network configuration, or simply the user's preferences at a particular moment in time. If the search engine supports backward and forward browsing of an index file, and can assemble a number of term values into a single HTML page, then the user should be able to specify how far the search engine will browse in either direction to collect terms. A user with a slow network connection might wish to restrict the number of terms fetched in order to provide a faster immediate response. On the other hand, a user with a very fast network connection might want the search engine to assemble a large number of terms, so as to receive more information in the initial response, including perhaps hundreds of terms before or after in an alphabetical or classified sequence, and to several levels up and down in the hierarchy. Though all the terms could not be displayed at one time on the user's screen, the perceived delay in delivering all this information would be small, since most of it would be delivered while the user was looking at the initial screen. The advantage would be that any further movement requested by the user either forward or backward in the sequence would be very quick, since all the information would already have been transferred over the network to the client computer.



## Proceedings of the 7<sup>th</sup> ASIS SIG/CR Classification Research Workshop

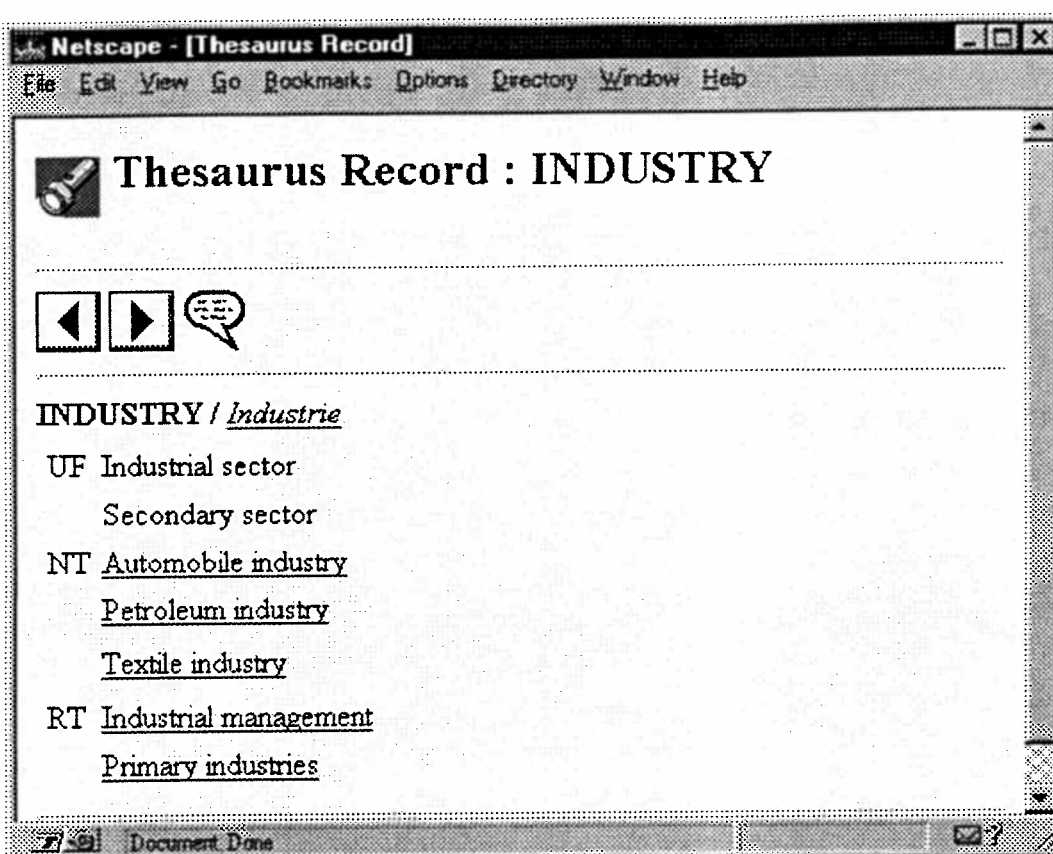


Fig. 7 Single term display in a dynamic publication

### ***Configuration options***

In fact, the flexibility inherent in dynamically-published thesauri allows the possibility for each user to specify other parameters, in addition to file size, that would tailor the thesaurus interaction to his or her particular needs: as long as there is a way of modifying the behaviour of the underlying search engine so that it retrieves and presents information accordingly, these parameters could include the sequence in which terms should be displayed (alphabetical or systematic), the total number of terms to be returned, the number of terms that should precede or follow the target descriptor, and the number of hierarchical levels above and below the target term. However the World Wide Web makes the mechanics of managing this kind of parameter complex. One of the problems with the World Wide Web, at least for sophisticated information access, is that it is a stateless protocol: most Web servers keep no information concerning the identity of a user from one access to another. Either configuration information, such as the preferred size of a Web page, has to be repeated with each access to the thesaurus Web server, or the Web server must be extended to enable it to identify a session and to store and retrieve user preferences using a database. There are several different approaches to extend the World Wide Web to allow for this stateful access, but all of these options have disadvantages and all further complicate the setup and administration of the Web server.

## Proceedings of the 7<sup>th</sup> ASIS SIG/CR Classification Research Workshop

### ***Eliminating the network***

Finally, with a static thesaurus, there is one attractive way to control the problem of response time. Response time is first and foremost a network problem: it can be virtually eliminated if we eliminate the Internet from the delivery process of thesaurus published using HTML. Publication of a thesaurus via static means offers the possibility for a user to transfer all the files that make up the thesaurus onto a local computer, and then to use a Web browser to access those files locally, without having to connect over the Internet to a Web server where that thesaurus is stored. When all the thesaurus files are all found on the local hard drive, the quicker response of the hard drive of the user's desktop computer effectively eliminates any problems relating to access time.

Even if the thesaurus in HTML form is stored on a local hard drive for use during thesaurus consultation, the Internet can still be used to transfer files to the user's computer from a server on the Internet, either through the World Wide Web or File Transfer Protocol (FTP). It is also possible to publish the HTML files of a statically-published thesaurus on diskette or CD-ROM. If the hypertext links in the static HTML pages are coded to refer to other local files, hypertext links within one of the files to another file will still be valid, or will only need minor modification (e.g. through some simple editor utility or through a specialized program run at installation time) to make them work. There are disadvantages to this approach: disk space to store the thesaurus files is required on individual users' machines, and users may continue to use out-of-date versions of the thesaurus, even when more current versions are available for downloading and installation locally (though this is also true of traditional printed publication). However, because no special or sophisticated software (other than the users' Web browser such as Netscape or Mosaic) is required, allowing users to use the electronic thesaurus locally in this format becomes a very attractive and low cost option to other forms of publication.

### ***Conclusion***

The World Wide Web is only a few years old, but it has already profoundly changed our ways of finding and using information. It has provided thesaurus developers with an attractive medium for distributing their publications, but also with a choice to make in terms of the type of publication—static or dynamic—that they will make available. The choice depends not only on the size of the thesaurus and the resources available to install and configure search engines and gateways, but also on the way in the user will locate descriptors, how much information the user will see displayed, and how quickly the system will respond to the user's requests. How thesaurus publishers react to these choices will affect how searchers and indexers consult these tools, and will inevitably have an influence on *de facto* and *de jure* standards for thesaurus organization and display.

## Proceedings of the 7<sup>th</sup> ASIS SIG/CR Classification Research Workshop

### REFERENCES

Cochrane, Pauline Atherton and Eric H. Johnson. "Visual Dewey: DDC in a Hypertextual Browser for the Library User" in Green, Rebecca, ed., **Knowledge Organization and Change, Proceedings of the Fourth International ISKO Conference, July 1996, Washington, D.C.** Frankfurt/Main: INDEKS Verlag, 1996, p. 95-106.

Pollard, Richard. "A hypertext-based thesaurus as a subject browsing aid for bibliographic databases". **Information Processing and Management** 29 (3), p. 345-57.

## Proceedings of the 7<sup>th</sup> ASIS SIG/CR Classification Research Workshop