What makes a good Web index?

Maureen Henninger

Schemes, indexing, site search tools, the use of meta tags (identifying metadata and descriptive metadata), and navigation devices and aids

Last year I was asked to be one of the judges of the 1998 Web Indexing Prize offered by the Australian Society of Indexers. I assume that this request came to me since I am a member of the team which teaches the short course Indexing Web Documents through the Continuing Education Program of the School of Information Systems, Technology and Management, University of New South Wales. However, I would like to think that the request came because of my expertise in finding information on the Internet (the 2nd edition of *Don't just surf: effective research strategies for the Net* has just been published by UNSW Press).

The creation of information and the identification and discovery of information are the two sides of the coin. The first activity should not be carried out without a methodology for the second. Indexers and publishers understand that the creation of a book is a poor effort without consideration given to chapter headings, tables of content, footnotes, bibliographies and concept indexes: back-of-book indexes. It should be a logical step to transfer this understanding to the 'new' publishing medium, the World Wide Web, whether the published 'document' is an article, an individual web site or a reference tool such as a bibliography.

Classification schemes

Classification is a process whereby similar material is placed together, either physically or conceptually. It is a process that enables the discovery of the existence of information by browsing. In the example of a traditional book, the information is classified by dividing the corpus of the work into chapters, often with further hierarchical divisions, represented by subheadings. On the Web this process is almost always carried out with individual 'documents', generally for the very simple reason that the document generally exists first as a word-processed document that has all these built-in features.

Other types of Web publications, bibliographies and individual web sites have generally adopted the classification process, either home-grown schemes or internationally recognized schemes such as Dewey Decimal or MeSH (Medical Subject Headings):

- General subject directories sort Web sites into hierarchical subject categories, generally reflecting broad subject disciplines for example Britannica Internet Guide, Yahoo (homegrown), BUBL (Dewey Decimal), and AustLII (adaptation of Moys).
- Subject gateways (bibliographies, 'webliographies') often use home-grown classification schemes; others use standard schemes, for example OMNI (MeSH).
- Web sites generally employ classification devices such as broad subject categories, site maps and alphabetical listings.

Indexing

Indexing is the process that allows the retrieval of specific items of data from the entire corpus of information. Indexing is of two general types, keyword indexing and concept indexing. The traditional back-of-book index contains both types; the first, keyword, can be created by computer programs as easily as by humans, and generally more inexpensively. Concept indexing is most effectively created by human indexers, although there have been many experiments in computerized concept indexing.

In Web publications both keyword and concept indexing are used, although the latter rarely. These are achieved by the use of:

- Web site search tools;
- meta tags; and
- alphabetical lists of both keywords and concepts.

The most effective strategy for finding information is to go to the site you know has the information. As an individual site is focused on a particular area or topic, a site search tool (generally a search engine) allows the information to be found quickly and directly.

The consulting team which maintains www.searchtools.com <http://www.searchtools.com/> gives the following criteria for sites which should be indexed with an automated search tool:

- Sites with valuable data in many pages. The exact number of pages is hard to define, as it depends on the density of the data. If you have more than 50 book reviews, for example, visitors will wish to search for other books by the same author, or other books they've heard about.
- Sites which get many visitors arriving from search engines at pages deep within the site hierarchy.
- Growing sites that are adding new and valuable information.

The technology of search engine indexing is still relatively primitive. It is therefore important that documents are indexed with good metadata, as these automated indexers will become more sophisticated in the collection and searching of metadata elements.

Meta tags and other metadata

Identifying metadata

All metadata aid the identification, description and location of information and are generally specified as elements or attributes. Many of the current descriptive metadata schemes list these attributes as 'core elements', for example the Dublin Core, AGLS, and ANZLIC metadata schema. Some elements facilitate the discovery of a known document such as title, creator/author, date, etc. — these attributes are identifying metadata. They describe the document as a unique object.

Descriptive metadata

In order to facilitate the discovery of information about a specific subject (as in the case of the information seeker who does not know of an existing document) further metadata attributes need to be used. The attributes, variously named subject and/or description, give some indication of the data content held within the document — its 'aboutness'.

HTML provides the meta tag content but at this time very few documents on the Web use metadata at all. When there is metadata it is generally a string of keywords, or a phrase. Dublin Core and AGLS provide for exhaustivity and specificity by allowing subject terms from a recognized scheme such as Dewey or MeSH to be added. Such authorities of course can be used without using the Dublin Core elements.

Annotations

An annotation is not 'indexing' but a long-standing bibliographic tool — a type of descriptive metadata. It is even more valuable if the content is evaluative in nature and as such is an excellent addition to subject gateways.

Alphabetical lists of keywords and concepts

The types of metadata discussed so far facilitate the discovery of potentially satisfactory documents; however they do not provide direct access to the exact information held within those documents. This access requires 'rich discovery tools', which may be a simple alphabetical arrangement of single document titles held on a Web sever or a professionally created back-of-book-type index which provides access not only to specific words, but to subject concepts. Such an index can be more effective than the automated indexing of a site, but it is expensive and should only be considered for important sites or for compelling reasons.

Navigation devices and aids

Whatever tools are used to give access to Web documents, in the online environment more help is required. The familiar aids of 'see' and 'see also' references are translated to hyperlinks, but must be logically organized and judiciously applied. And always at the end of a list suggestions of further possibilities are very welcome and useful.

Finally in the Web environment there can be a sense of disorientation; the familiarity of the logical sequence of a print document is missing. A good Web index should include good information design elements which provide such orientation. There should be navigation devices which provide immediate access to all the major 'chunks' of information in the 'document' or site. Most importantly, these navigation devices should be visible, though not intrusive, at all times.

Maureen Henninger is Coordinator for Continuing Education, School of Information Systems, Technology and Management (SISTM), Sydney 2052, Australia.
Email: m.henninger@unsw.edu.au