Indexing Web Sites and Online Documents

Manjit (Mary) Sahai

Books and printed periodicals remain the primary focus for science editors, but editors are increasingly involved in the publication of documents on Web sites and as online content. As a professional freelance indexer who has prepared indexes for books and online content, I have gained an understanding and an appreciation of what a science editor needs in an index for online documents.

When I began work on my first index for online content, I wasn't sure whether I needed to ask more or different questions compared with what I was accustomed to asking in connection with print-media assignments. If you are a science editor new to online content and in need of an index for it, you too might have wondered whether you have collected and passed on all that is needed by your indexer. This article addresses questions that you might have on the indexing of such content.

An Index Defined

An index is a structured arrangement of keywords that enables readers to find needed information quickly and efficiently. In print media, needed content is located by page numbers. For online content, the location is a hyperlink; when the index entry is clicked, the Web browser or the help engine loads the needed content.

The existence of search engines on Web servers and of index tagging and extraction functions in word-processing and desktop-publishing software might lead one to think that indexes can be generated automatically without human input. Not true. To provide a high-quality and reader-friendly index, an indexer must analyze content and identify and arrange keywords in a useful structured format.

Developing a useful index is an art, not a mechanical or automated process. Although the production of the final formatted index document is and should be automated, the identification of keywords and their arrangement—the elements of a useful index—require an experienced indexer.

A high-quality index

- Helps readers identify needed information quickly.
- Discriminates between useful information and a passing mention in content.
- Excludes information whose location will not help users.
- Indicates relationships between topics by cross-referencing of entries.
- Directs readers to various aspects of a topic by providing "see also" cross-references.
- Groups locations of information on a subject previously scattered in the document.

When readers of a book are unable to find relevant keywords in its index, they will usually scan the table of contents or the headers and footers in the text. Readers of online content are likely to be more frustrated if the index fails to direct them to the information they want, because scanning online documents is generally more difficult than scanning books.

What To Expect of an Online-Content Indexer

Excellent text analysis is the hallmark of good indexers, regardless of the format of the content—print or online. Basic indexing principles apply equally to online documents and book indexes. Most online indexes today look like back-of-book (b-o-b) indexes. If an indexer has a proven track record on print-media indexing projects, it is reasonable to expect him or her to be able to write indexes for online content, although indexers of online content need to be a bit more technology-savvy than is necessary for print media.

It is reasonable to expect an online-content indexer to be familiar with

- Search engines and how they work on Web servers.
- Database-indexing concepts and how dynamic Web-content scripts work with databases.
- Basic HTML tags and how they can be used to create hyperlinks to related Web pages. (Pagination systems differ between electronic and print media—page numbers in print are replaced with hyperlinks from index entries to online content; hypertext links in online documents serve the same purpose as "see" and "see also" references in b-o-b indexes.)

An online-content indexer should also be able to communicate with the Webmaster on technical terms so that the index can be put online as intended.

Selecting Index Keywords

Readability and ease of use are vital in indexes for online documents. An index can help a reader to make good use of the time spent at your site. A reader has a shorter attention span when reading static online content, and a good index can speed the retrieval of desired information.

When assigning an index, be sure to ask the indexer to include keywords that cover chapter headings, table and figure captions, menu options, definitions, acronyms, synonyms, main concepts, main tasks, functions, commands, and parameters.

Most beginners think that keywords or index entries for online indexes must be taken directly from the text. The inclusion of only keywords that appear in the content is a common shortcoming of many online indexes. You should ask the indexer to create keywords for concepts when it is not possible to use the exact words or phrases in the text. Inclusion of synonyms that readers are likely to think of even if they do not appear in the text will improve the value of an index.

Here are some tips to pass on to indexers when creating keywords:

- Create index entries in noun form.

MANJIT (MARY) SAHAI *is a freelance indexer who specializes in creating indexes for medical and scientific books, journals, and Web sites. She is a member of the American Society of Indexers. She can be reached at* manjit@ramcorp.com.

Indexing Web Sites continued

- Pay attention to proper nouns and terms that are case-sensitive—such as some statements, commands, and parameters—and index them as they appear in the text.
- If a primary entry has only one subentry, eliminate the subentry and make it the primary entry with its own locator.
- Do not begin entries with articles (*a, an, the*) or conjunctions (*and, or*).
- Avoid unnecessary articles and prepositions in index entries.
- Arrange all entries alphabetically.
- Ignore articles and prepositions when sorting index entries alphabetically
- If a keyword has more than three links (page references), consider creating subentries.
- Use capitalization, punctuation, and other conventions consistently.

Types of Online Indexes

To help readers find information of interest to them online, Webmasters today use three methods:

- Search engines.
- Back-of-book style indexes.
- Database indexes.

Search engines are relatively easy for Webmasters to implement. These products accept a keyword from a user and retrieve every page that contains it. When content is added, it is automatically included in what is scanned by the search engine. This is a low-maintenance solution.

The major drawback of search engines is that they retrieve pages even if the topic of interest is mentioned only casually and the "hit" offers no information relevant to the user's needs. Search engines can have a high level of retrieval and a low rate of relevance. They can also be of little help when documents use particular words and phrases but the user is searching for a synonymous term that is not in the documents being searched for. For example, a reader browsing an accounting site might be looking for information on sales but use the more generally accepted term *accounts receivable*. The search engine will retrieve documents that contain *accounts receivable*, but not documents that contain sales instead. Needless to say, the reader will not be satisfied with the search experience.

Back-of-book style indexes overcome the problem of irrelevance by being targeted and smaller. They are more labor-intensive and must be updated when content is added to a site. These indexes are usually on a site's navigation bar under the term *site index*. A few sites that have good b-o-b style indexes are

- American Society of Indexers—*www.asindexing.org*.
- US Census Bureau—*www.census.gov*.
- UNIXhelp for Users—*unixhelp.ed.ac.uk*.

A b-o-b style index is a dedicated page (or pages) at a site that looks similar to the index found at the back of a book. This kind of search tool, being familiar to readers, is easy to use. It displays an index with main entries and subentries, and each entry is a hyperlink to a relevant page on the site. (Although dense print documents commonly use three levels of entries, most indexes for online documents use only one or two.) Such indexes provide highly relevant entries and a mechanism, such as a bookmark, to navigate to the desired letter header. For example, to locate a document on dietary fat on a b-o-b style index page, the reader would click the letter D on the index page to move to entries that start with D. Entries are listed alphabetically to allow a quick scan to locate the desired term. Clicking the term loads the relevant document. Book indexes can include multiple locators for a single entry, and online indexes similarly can use multiple subentries.

A few products that can help one to generate such indexes are HTML Indexer, at *www.html-indexer.com*; HTML/Prep, at *www.levtechinc.com*; and WebWacker, at *www.bluesquirrel.com*.

Database indexes require the creation of

- A searchable database that is custom-designed for the site.
- Development of server-side scripts for searching the database and displaying results in a prescribed format.
- Entry of all index terms in the database.

The product delivered by the indexer is a database file that contains index entries and keywords and will be installed and configured on the Web server by the Webmaster. The Webmaster is usually responsible for writing the server-side scripts for query and retrieval. However, some indexers also provide server-side scriptwriting services themselves or through a consultant.

An example of a database index is the MeSH site of controlled medical vocabulary terms maintained by the National Library of Medicine. Another example is the author, title, subject, and keyword search provided by such online retailers as amazon.com and bn.com.

Closing Comments

Publishers of online content increasingly recognize that they need to provide a searchable index on their sites to provide an enjoyable experience to visitors to the sites. As an indexer, I laud them for recognizing the need for online indexes and hope that they recognize the value that professional indexers can add.

Once it has been decided that an online index needs to be created, it is incumbent on the site publisher to decide what type of index— search engine, b-o-b style, or database—would be most relevant. I hope this article has made the various types of online indexes and their strengths, weaknesses, and characteristics clearer.

Regardless of the type of index you choose for your online project, it is important to recognize that a good index is targeted to the user and that its purpose is to help readers identify and locate relevant information quickly and easily. It should analyze concepts treated in the documents, indicate relationships between concepts, provide cross-references, and group information that is scattered throughout a site.

Finally, let me try to shed some light on the costs associated with online indexes and compare them with those of book indexes. A typical rate for book indexes is $3 per page; an index for a 500-page book would cost about $1500. Web-site indexing costs more than b-o-b indexing because it is a more elaborate process. Web-site indexers I know of charge hourly rates because Web sites do not have pages like printed documents. The average rate quoted by indexers for online projects is $50-$60 per hour, and the time it will take to prepare an online index depends on the site content and the type of index desired. Small online-index projects might start at around $3000; larger, more elaborate indexes could cost tens of thousands of dollars (plus the cost of making changes to accommodate the addition of content to the site).