

Indexing America Online — Part Two

by Seth Maislin, Focus Information Services

Editorial Note: This is the second half of an article published in the May 1998 issue. To obtain the first half, please check our website at <http://www.stc.org/pics/indexing/> for the May issue.

Instant Replay

In the first part to this article, I introduced my contract with Songline Studios to index the online content of America Online. In that article, I presented a list of obstacles I needed to overcome to fulfil my contract. Here are most of them:

- America Online's size
- Irregular and constant fluctuations in online content
- Database granularity (i.e., inability to hyperlink to pages other than area home pages)
- Political pressure from America Online managers and area authors
- Errors and user inexperience
- Text matching limitations (e.g., ignorance of synonyms, stemming algorithm inaccuracy)
- No rating functionality for successful matches

The Challengers

I am an indexer, headquartered in Massachusetts, who now writes 60 to 80 indexes annually. I provide professional development training both privately and to organizations such as the American Society of Indexers. My specialties are indexing and information theory, particularly with online tools.

To deal with AOL's size, I contracted others to help me with "the dirty

work" of typing search words in the database records. At the height of the project, four people worked in a large room at open desks. Although I was in charge of the project, most of the nitty gritty was accomplished by two other individuals. Following my lead, they reviewed each of the AOL pages, decided the important concepts of each area, and chose representative vocabulary. With their assistance and the involvement of several Songline employees, the project took under one year, from the initial planning stages through testing, review, and summary. In theory, then, AOL's size could be conquered by allowing enough time to complete the project and contracting enough indexers to do the work.

The Score

The project did get completed: AOL now has a more consistent and more accurate Find system than ever before, including a methodology for adding future areas to the database. Here is that methodology, summarized for this article.

- Information is indexable if it is important. Here is a good rule of thumb: If a subscriber ends up at an area, you ought to be confident that the subscriber isn't going to get angry for being there.
- Include area keywords and important title words as search words.
- Search words should either (a) match the document text; (b) be synonymous with document text; (c) represent ideas presented in the text; or (d) repre-

sent slightly broader categories of function or ideas than those presented (e.g., using the search word "pets" to describe a page about cats).

- If the information is not on the main area page, index it only if (a) the material can be intuitively found using well-labeled navigational tools, or (b) the information is only one click away from the main area page.
- Use verb infinitives.
- Use all lowercase, except for proper nouns, for better search-word readability.
- If a search word has an irregular syntactical variation, include that variation as a new search word. (For example, always pair *woman* with *women*.)
- Do not use stop words (articles and prepositions) as search words unless necessary to make sense to *you*. For example, don't use "cats and dogs," but do use "United States of America." Articles and prepositions aren't searched.

Many of these suggestions address head-on the text matching limitations. For example, synonyms and irregular syntactical variations are explicitly requested. During the indexing process, style sheets were generated, listing both obvious and unusual combinations (e.g., *woman/women/womyn/girl/gal/female*). These lists were then consulted before inputting search words, since not all may be appropriate.

Other guidelines are to improve readability within the database, such as requesting stop words for readability, initial-uppercasing

(Continued on Page 7)

capitalization, and infinitive verbs. Although these choices do not change the software's performance, they do reflect my personal preference. I believe in having a consistent interface for other indexers and writers.

One tool that proved helpful was the (proprietary) records of failed search queries. Many of the failed searches were legitimate — subscribers sought information that wasn't there. Other failures were caused by misspellings, particularly when proper nouns were involved.

Figure 5: (Editor's note: Figures 1-4 appeared in the first half of the article in the May 1998 issue.) The search words for the Video Magazine Online area. Notice how many of the concepts are listed repetitively to assist the search engine, but also how critical details on the area's subpages are completely ignored.

However, the inefficiencies inherent in the AOL index showed most plainly when it came to the two greatest obstacles, currency and depth of information. These are worth another look.

When an event occurred "in the world," a flurry of online activity would follow. For example, the bombing of the 1996 Olympics likely inspired many users to search for the words "Olympics" and "bomb." Though these stories were thoroughly covered in the News Channel, the word "bomb" was never added to the index. This is because news usually stays current for only a short time before disappearing. In addition, there is so much daily news that maintaining a current index is unfeasible. Instead, users are expected to browse the News Channel. So here is how I address changing information: If I

know it is going to change soon, don't index it. Instead, index by category. Although users may find the search functionality incomplete, I want to guarantee that the results — no matter how limited — are trustworthy.

Regarding database granularity, often users would search for information that we knew existed online, but that would not have been easily accessible from the area home page. Again, I chose having an incomplete index over an untrustworthy one: Why should a user be sent to a page and not



know why? (See Figure 5.) There is much disagreement about this. Some believe that subscribers like the challenge and excitement of finding something new and that they will be inspired to explore. Others believe that exploration is an iterative search (browsing-oriented), but that the Find button is designed for people who are not interested in false hits, misdirection, and "red herrings." I instead implored area managers to develop better internal navigation. One controversial suggestion would be to design the database such that individual pages have search words. However, because AOL is so huge, and is run by so many people, trying to introduce such a monstrous, rewritten database — without interrupting subscriber access — is a job I wouldn't wish on anybody.

This brings up another difficulty: the pressure from channel managers and area authors to add certain search words to specified areas. There are advantages to keeping them involved. The index could be kept more current, with each area under independent management. Also, search-word choice might be more educated. I think the disadvantages are greater. First, such micromanagement would be a logistical nightmare of communication and authority battles. Second, any consistency among areas would be lost with the introduction of hundreds of indexers. Finally,

with dozens of indexers trying to access and update the database simultaneously, performance would be compromised on many levels.

To address user inexperience and error, I included keywords and area title words as search words. (Many users accidentally start a keyword search with the Find button instead of the Keyword button.) In addition, a new, one-page area called "Refine Your Search" was written, always to be presented in the list of successful matches. This page is simply a help screen, consisting of a list of user-oriented tips regarding the Find system. Suggestions included "Check your spelling" and "You can make your search more precise by using the AND operator."

Finally, I want to say that although AOL's size was not a major obstacle, it couldn't easily be ignored. Consistency among areas was a major challenge, especially without a global search-and-replace ability. Redundancy of information (that is,

(Continued on Page 8)

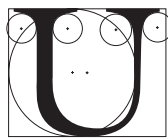
numerous areas with the same basic content or audience) would present bulky search results. Shortcomings in the index, such as not indexing certain subpages, seemed much more pronounced. Query sheets of synonyms and syntactical variations became more difficult to maintain. And areas could be created, subdivided, redesigned, rewritten, or deleted without warning or notification, challenging the indexers to rigorously reevaluate which areas remained to be indexed, or needed to be indexed again.

Demanding a Rematch

In less than one year, the goal of this project was fulfilled: to make AOL an easier place to search. I am satisfied with the final product, although I can imagine several AOL subscribers who are not. Although the information industry is always asking new questions of me, this was the only time I had been forced to break the tenet of a good index: enabling readers to find every valuable statement. I would like the opportunity to address the challenges again, perhaps with greater authority. I want to investigate redesigning the database, scheduling "guerrilla" team indexing updates, and working with area authors.

Working with America Online was a great opportunity for me to learn and help develop the theory behind handling volumes of online information for millions of people. I look forward to working on this project again. In fact, I suspect that I will ... once AOL again appreciates the need for a good, current index.

Upcoming Events



Upcoming Events lists information on indexing events.

To list your event in the next issue of *A to Z*, contact Bill Graham, BillGrah@bhip.infi.net, 941-739-4218, or Jan Wright, jancw@mindspring.com by December 15, 1998.

➔ September 26, 1998: **Indexing Workshop** (San Diego STC) Lori Lathrop will help indexers sharpen their skills from 9:00 A.M. to 5:00 P.M. Registration takes place at 8:30 A.M. The cost of the workshop is \$79 for STC members and \$89 for non-members. A continental breakfast and lunch is included in the price. Contact Sue Heim, Phone (619) 546-2400 Ext 323, Fax (619) 546-1285, E-mail Sue_Heim@msn.com or Sue@centraxcorp.com.

➔ September 26, 1998, 1:00 P.M. to 3:00 P.M.: **Philadelphia Group of ASI**, Marshallton United Methodist Church, 1282 West Strasburg Road, Route 162, West Chester, PA. Kamm Schreiner will demonstrate SKY Index. Contact Nancy Guenther, 610-436-4049, nanguent@chesco.com.

➔ October 9-11, 1998: **1998 Annual Conference of the Society of Indexers**, Grand Hotel, Grand Parade, Tynemouth, Tyne and Wear, NE30 4ER, England. Nancy Mulvany will deliver the keynote address. According to the organizers, the Tynemouth conference will be a practical indexing conference with sessions run by indexers for indexers. For more information, write to the Society of Indexers

by Bill Graham, Upcoming Events Editor

at Mermaid House, 1 Mermaid Court, London, SE1 1HR, United Kingdom, Phone +44 (0)171-4034947, E-mail 10624.1745@compuserve.com, Website: <http://www.socind.demon.co.uk>.

➔ October 10, 1998: **Learning Indexing: What are your Choices?**, Pacific Northwest Chapter, at Portland State University. The meeting will take place from 11:30 A.M. to 3:30 P.M. It will include a networking lunch with active indexers, a business meeting to plan for the rest of the year, and an opportunity to work through an actual lesson with other indexers. Short presentations, information, and course materials from several courses will be available for browsing. Contact Jan Wright, 206-784-2895, jancw@mindspring.com. For more information, see the chapter's website: <http://www.mindspring.com/~indexwest/ASI>.

➔ October 10, 1998: **Fourth Rocky Mountain Indexing Conference**, Colorado Chapter of the ASI, Meadows Library in Boulder, Colorado. Do Mi Stauber will present a workshop (9:00 A.M. to 5:30 P.M.) titled, "**Facing the Text: Content Analysis and Entry Selection in Social Sciences and Humanities Indexing**." This is a professional development seminar of the ASI. The cost for ASI members is \$35 after 9/15. For non-ASI members, the cost is \$35 after 9/15. The cost of registration includes lunch and snacks. The workshop is open to ASI members and