

# Indexing America Online — Part One

by Seth Maislin, Focus Information Services

Imagine thousands of online documents, whose constantly evolving content is authored by hundreds of people with diverse skills and backgrounds and agendas, and read by millions of subscribers around the world. This is America Online.

Now imagine one self-employed indexer. That's me.

My job, once I chose to accept it, was to apply what I knew as an indexer to the maelstrom known as America Online. It was the most absurdly daunting project I had ever faced. This article is the story of that contract.

I am writing in two parts, presenting the contract as a game. The first part, published here, introduces the obstacles involved and ends with a question: What would you do? In the next issue, I will present the strategy I devised, the implementation of that strategy, the results, and the evaluation.

## The Players

America Online (AOL) is the world's largest commercial Internet access provider, with millions of customers worldwide. Not only do they provide the "standard" services of email and Web access, but AOL also provides a tremendous Web-like database of information. This information is first divided into general topics called *channels*, such as Computing, Personal Finance, and Kids Only (Figure 1). Within each channel are *areas*, which function almost exactly as Web sites. Then, each area can contain any number of documents. Keeping

with the Web analogy, the area is like a home page, and the AOL documents are like pages accessible from the home page.

Songline Studios (www.songline.com) is a team of writers, designers, and developers whose mission is to push the limits of what's possible in Web publishing. AOL contracts Songline to help manage their published content with a multi-channel viewpoint. (AOL, on the other hand, has channel-based departments that often don't communicate with each other.) Songline contracted me to write and consult on AOL's searching and indexing functionality.

## The Game

Although AOL functions almost identically to the World Wide Web, its content is nevertheless separate, managed with a single database. Each record in the database defines an AOL area, including (among many other items) a title, a keyword (much like an alias), and search words. All database data are input manually, generally when an area is first created. Note that the

database records are defined by areas, not by the pages within each area.

There are three methods for a subscriber to locate a particular area. To uniquely find an area, a subscriber can select the **KEYWORD** button and input the area's unique keyword. This is a *known-item search*, much like trying to find a name in a telephone book. If the information matches exactly, the area is called up to the subscriber's monitor. If no exact match is found, the user is notified of the failure.

Subscribers can also browse AOL, starting with the channels page and navigating by category. For example, a subscriber interested in airplane flight information might start by browsing the Travel channel. Browsing involves iterative searching, in that a subscriber can make a selection, view the subsequently provided information (such as a list of areas, pages, or articles), and then choose to make a new selection.

The third method is with the **FIND** button. Pressing this button produces a dialog into which a subscriber inputs text that defines the search goals. This text is compared to the search words in the AOL



Figure 1: Channels page. This page is a good beginning for users who want to browse AOL's content by category.

Reprinted with the permission of American Online, Inc. © America Online, Inc. All Rights Reserved.

(Continued on Page 6)

database, and a successful text-to-search-word match provides the subscriber with a hyperlink to the appropriate area (Figure 2). The functionality provided by the FIND interface is not unlike a subject-based search using a Web search engine. Those who use the FIND feature most likely are familiar with the area but not the keyword, or else they simply do not know precisely what they are looking for. This is precisely the opposite of known-item searching, and in this environment a human-created index will always outperform a search engine.

Unlike the Web's human-made indexes (such as Yahoo!), which tend to neglect a great majority of the Web, AOL's content is managed by a single company and therefore (in theory) could be indexed in full. My role was first to determine which search words would appear in each area, and second to develop a method so that new and changing content could be indexed by others without sacrificing consistency.

Before I could develop a system for choosing the search words, however, I needed to know what I was up against.

### **The Rules**

The most obvious obstacle to overcome is AOL's size. Obvious or not, though, AOL's size proved to be the least important of all obstacles.

Instead, the biggest problem is that information on AOL changes constantly. The fluidity of these changes cannot be underestimated: annual concert dates, seasonal sports data, monthly meetings, weekly contests, daily quotations, hourly news, minute-by-minute flight arrival time updates, and real-

time chats. With AOL's size, the maintenance to continually update the search words would be unacceptable. Instead I needed to understand which area elements would never change, and to index those.

Another issue I quickly became aware of was the database's granularity. Because the search words would enter the database at the level of areas, I could index only

a particular league in that sport, then to a team name, and finally to the player. So using a player's name as a search word would point to the subscriber to a sports news area—a far cry from fingertip-accessible information!

Redundancy of information creates a similar difficulty. All information within an area is indexed by the area. Consider again the athlete example. When a subscriber

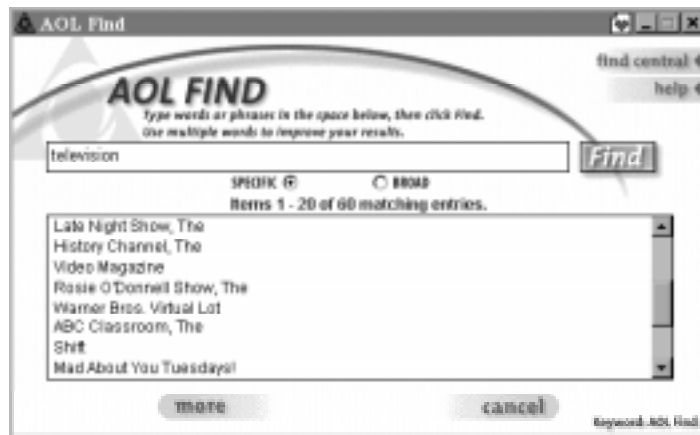


Figure 2: Results from a search for the word "television." Some of the areas with search words that successfully matched the word "television" are listed. The area Video Magazine Online is spotlighted in the next two figures on the following pages.

areas, not the pages within those areas. (See Figures 3 and 4, next page.) For example, an area for a radio station might contain a document for a syndicated radio program. However, if I included the program's name in the search word database, a subscriber who sought information on the program would instead retrieve a link to the radio station. Although the subscriber could then browse from the radio station area to the desired document, this involves an additional navigational step. It is not uncommon for a large area to have a hierarchy of several document levels. For example, to locate statistics on a professional athlete, a subscriber might start at a sports news area and navigate to a document for a particular sport, then to

searches for a player's name, he or she might want statistics, a schedule, salary information, personal background, news of a related criminal trial, or athlete-related chat group information. Even if the sports news area contains all this, I can index the area only once, and the detail gets lost. It's like answering yes or no to a multiple-choice question.

There was one problem I didn't see coming: the area owners and channel managers themselves. Assuming I could select appropriate search words, my decisions might be challenged. The Computers & Software channel, for example, contains over 100 areas. The channel owner wanted the search words "computer" and "software"

(Continued on Page 7)

added to each area. From my point of view, however, an area advertising modems is neither strictly about computers nor about software. Other owners requested common but irrelevant search words, figuring that “accidental” matches would give their area greater visibility.

Finally, errors and user inexperience are problems, too. Sometimes users would select the FIND button

instead of the KEYWORD button by mistake and then expect to uniquely find a particular area. I might accidentally mistype a search word; without any way to discover this, an area could grow unfindable. In addition, naturally, subscribers are unaware of my limitations. They expect the index to contain exact up-to-date information, all spellings of all synonyms, and a rating system.

## The Equipment

Then there are tools difficulties. First of all, even though the database of search words is human-related, the tool ultimately used by subscribers is a search engine. This means that “being close” may not have any value. For example, a search engine may not consider the following equivalent: *woman* and *women*. Worse than that, consider *U.S.*, *U.S.A.*, *US*, *USA*, *United States*, and *United States of America*. To index effectively, the search words should include every syntactical representation of a concept, in addition to synonyms and related terms. At least AOL's searches are case-insensitive!

To some extent, the search technology compensated for this with a *stemming algorithm*. This means that common word suffixes are ignored, such that only root words are compared for exact matches. Plurals such as *tickets* and *buses* are replaced by their singular forms, and gerunds like *teaching* and *voting* are treated as infinitives (*teach* and *vote*). This feature is extremely useful—consider *clothing* and *clothes*—but there are downfalls. Stemming creates contextual discrepancies, such as when a search for *fishing* locates a pet store that sells *fish*. (Then there's *vegetable* and *vegetarian*!) Stemming can also be inappropriate, as when the word *news* is shortened to *new*. Truly, this last example is a serious problem, since any search for “news” would turn up areas for “New York” and “New Zealand.”) Finally, stemming simply won't do a complete job: *mouse* and *mice*, *tracking* and *trackball*, *hamburger* and *burger*.

Figure 3: Home page for the Video Magazine Online area. Any matched search word for this area will provide the user with a hyperlink to this page.

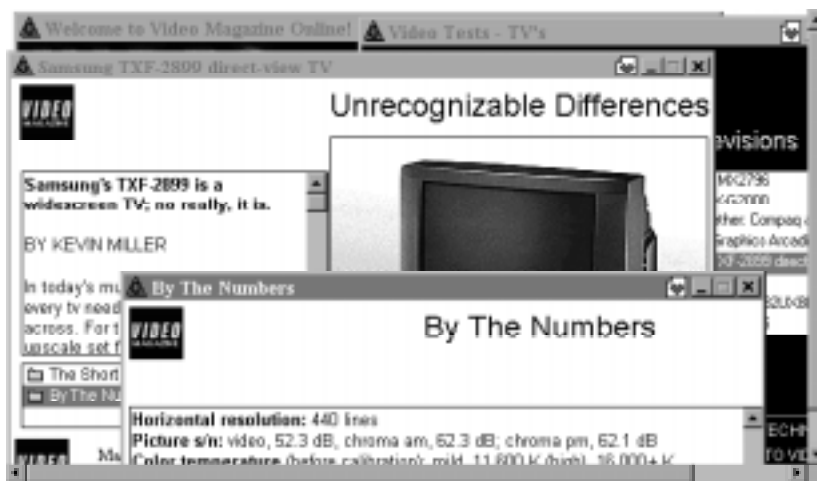


Figure 4: Several subpages of the Video Magazine Online area. There are four open windows, identifiable by their title bars. To navigate from the home page to the “By The Numbers” page, select the Televisions menu item on the home page (not the Video Tests button), and then on subsequent pages select the Samsung TXF-2899 direct-view TV menu item and the By The Numbers item. Here is where database granularity becomes a problem. Although the resolution information for Samsung's TXF-2899 might be indexable, successful matches on this information would not bring subscribers to “By The Windows.” Instead, they would get only Figure 3.

(Continued on Page 8)

The search engine also utilized *stop words*: common words such as articles and prepositions that are ignored in search queries because potentially they could clutter up search results. For example, the word “and” cannot be searched. This is a beneficial feature when text is searched, but not nearly as useful when an indexer has control over the search words.

Figure 5: 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 (see Figure 4) are completely ignored.



Whereas stemming and stop words are behind the scenes, subscribers also have tools of their own. Subscribers can use the boolean operators AND, OR, and NOT. Searching with AND means that searches must match more than one word at a time (e.g., “new AND york”), and OR is used when looking for areas with at least one matching search word (e.g., “milk OR dairy”). NOT is used when you want the search to disqualify particular matches (e.g., “news NOT york NOT zealand”). Subscribers can also use wildcards, although it should come as little surprise that fewer than 2% of all searches contain them.

When a search proves very successful, such that the user is presented with many choices, AOL does nothing to rate them by importance. (This functionality didn’t exist in the database design.) The database is searched from top to bottom; search results are listed in

database order. In fact, the newest areas, which are added to the end of the database, will always appear last in the search results.

Finally, in what would prove to be the most frustrating limitation of all, the search words can be accessed only through the database, one area at a time. Global changes are impossible. Printing the search words is difficult. Thus there existed

no effective means of comparing similar areas, or even reading what had been done so far.

### Intermission

The second half of this article will be published in the next issue of *A to Z*. In my conclusion, I will explain how I addressed the above challenges, and explore the viability of my choices. For now, I challenge readers to develop their own solutions. Try to answer the following questions:

- Should any special consideration be made because of AOL’s size?
- Because content is always changing, how (and when) does one decide what gets indexed? Consider both the removal of outdated material and the additional of future data.
- If only areas can be indexed, how does one address an area’s subpages?

- How much influence should others be given when it comes to indexing individual areas?
- What is the best way to respond to user inexperience and errors?
- How should one address the limitations of text matching and the stemming algorithm?
- Can (and should) the index compensate for the lack of a rating system?

## Announcements



**Articles Wanted:** The newsletter is looking for good articles on indexing to help our mem-

bers learn more about specific techniques, access method theory, software packages with indexing modules, etc. If you have an article idea, please contact Pilar Wyman, pilarw@aol.com, or Jan Wright, jancw@mindspring.com, for more information.

Contact the editors for submission requirements and file formats. The next deadline for articles is August 1, 1998.



**Questions for Q & A Wanted:** If you have a technical indexing problem that you

would like answered, consider submitting it for our Q & A section! We will find an experienced indexer to publish an answer, helping others who may run into the same problem. Send your questions to Pilar Wyman, pilarw@aol.com, or Jan Wright, jancw@mindspring.com.