Search Software For SMEs

FultonHistory.com Needed A Low-Cost, High-Function Search Program

by Sue Hildreth

• • •

In 2000, when Tom Tryniski, president and owner of FultonHistory.com, first began developing his Web site—a searchable database of newspaper articles from New York state—he quickly realized he was going to need a very flexible search tool. It would have to be able to index hard-to-read scans of aged and yellowed newspapers and enable users to search on a large number of words for each article.

The FultonHistory.com Web site currently houses more than 5.5 million pages of scanned newspaper articles and photographs from the 1830s through 2005. Many of the articles are yellowed, creased, or torn, making it hard for OCR (optical character recognition) readers to accurately translate all of the text. Yet it was absolutely necessary to glean as many words as possible from the documents so they could be easily found in a search. Manually reading and inputting keywords for each article simply wasn't feasible.

Tryniski also wanted to provide both the scanned text, for search purposes, and the original image so that researchers could see the document and decipher difficult-to-read text themselves. At the same time, however, he needed something inexpensive. Because Tryniski's site is free to all users, he doesn't generate large user fees but depends on voluntary donations to foot the bills.

While searching the Internet, he came across the dtSearch (301/263-0731; www .dtsearch.com) Web application, one of the dtSearch series of search and indexing products. The dtSearch application worked well for Tryniski's needs, as it leveraged both the OCR text and the original scanned images in its indexing and search process. It used the OCR text to build a comprehensive index that includes every identifiable word in a document. That enables a search to return many more documents than would be obtainable with a more traditional indexing process that uses only the title and key words.

While the dtSearch software uses the text for indexing, it returns the scanned image to the user in the list of results.

"If somebody types in 'John Jay Smith' and the OCR scanner has found those words in a document, the site will point them to an image of the page," says Tryniski.

The dtSearch product line includes dtSearch Desktop with Spider for searching files on desktop computers; dtSearch Web with Spider, which can publish a large volume of data to an Internet or intranet site, as well as search content on internal and external HTML sites; dtSearch Network with Spider; dtSearch Text Retrieval Engine for Linux; dtSearch Text Retrieval Engine for Win and .NET; and dtSearch Publish, which lets customers publish a searchable database to a CD, DVD, or portable hard drive.

The dtSearch software can identify a variety of international languages, including Middle Eastern languages, Chinese, Japanese, and Korean. It can search both static and dynamic content and dozens of different file formats.

Easy Implementation & Low Price Are Key For SMEs

Because FultonHistory.com is a one-man enterprise, Tryniski does not have an IT staff to handle a major software installation. He needed a search product that was reasonably easy to install and maintain.

"It was a simple setup [with dtSearch]," he says. "You simply point it to the directory of the files you want to have indexed."

The software can include files at multiple locations in the same index, making it possible to search for and view documents from remote databases in the same set of results. That could be a big benefit if Tryniski ever wanted to share resources with another provider of historical documents.

Price, of course, is also an extremely important feature for an SME. "It was the only product that had all I wanted and more at a cost of only $999," he says, noting that while his content has increased from 300,000 pages to more than 5.5 million since he started, he has not seen the price rise or the performance decrease with dtSearch.

"Look for an indexing product that doesn't have an upper-level cap on the number of pages you can index before

they start to charge more money," he advises.

According to dtSearch, the company is able to keep the price down because it OEMs much of its software to other vendors, relying mainly on word-of-mouth sales to the corporate consumer market. That keeps advertising and marketing costs low.

Scalability & Performance

Performance and scalability is another issue. Large index capacity can mean better performance because the user is spared the need to create multiple small indexes to hold all of the data. Searching across multiple indexes takes longer. The dtSearch index can hold up to 1TB of data—larger than most other indexes.

Currently, FultonHistory.com only takes up 150GB of the available index space. While that has grown considerably during the past several years, Tryniski says that search performance has not slowed as the size has grown. "It scales excellently. It's just as quick with 5 million pages as 5,000 pages," he says.

The main cause of poor performance, he notes, is index fragmentation—a problem that comes from doing many small updates to an index. It's a problem common to all search products and can only be prevented by doing fewer updates—saving the changes for a once-a-month update, for instance.

"It's basically like having a fragmented hard drive. You wind up with little pieces all through the index, and it thrashes the hard drive trying to read this fragmented index. That's more wear and tear and less response time. It's better to wait and write a larger set of files," he advises.

Search Capabilities

The software can also search on synonyms, concepts, phrases, phonic or "sounds like" spellings, wildcards, word stems or roots, and numerical ranges, such as 1901 to 1931.

From Tryniski's point of view, one very important search feature is its ability to do "fuzzy searching," which allows searchers to request the software to look for similarly spelled words. This is useful when searching on names, which can be misspelled or have alternate spellings, and for compensating for smudges and tears in the paper.

dtSearch Web With Spider

Instantly searches terabytes of text across an Internet or intranet site; can perform a variety of different search types, such as concept-based or phonetic searchers, and converts other file types (Word, spreadsheet, database) into HTML; can search multiple sites and return the results in a single view

"Look for an indexing product that doesn't have an upper-level cap on the number of pages you can index before they start to charge more money," says Tom Tryniski, president and owner of FultonHistory.com.

(301) 263-0731
www.dtsearch.com

"The ability to do fuzzy searches is key when you're OCRing old newspapers from microfilm because of the poor quality of the original paper and/or quality of the microfilming," he says.