



ASTAWARE SEARCHKEY API

WHITE PAPER

Introduction

ASTAware Technologies INC. ("ASTAware") has developed an expertise in information management systems over the past ten (10) years. Our expertise in search, index and retrieval functionality is embodied within our proprietary, multi-platform technology, ASTAware™ SearchKey.

Over the past decade, we have successfully leveraged our core technology by building applications and products, and delivering customized solutions, to satisfy a diverse range of market segments and customer requirements. We have produced technical manuals, directories, telephone books, aviation regulations, statutes of various jurisdictions, multimedia reference CD-ROMs and data archives, to name but a few.

Our customers are located throughout the world on four (4) continents and represent a varied cross-section of business and government, including Transportation, Telecommunications, Energy, Technology, Manufacturing, Publishing and Financial Institutions. We are extremely proud of the fact that included within the ranks of our blue chip clientele are: Consumers Gas, IBM, Hewlett Packard, Lockheed Martin, Ontario Hydro, O'Reilly & Associates, Qualcomm Inc., Siemens, Silicon Graphics, Sun Microsystems, Thompson Financial Publications, and more.

ASTAware SearchKey API represents our core Java™-based indexing and retrieval technology. SearchKey API offers customized development of search and retrieval functionality assisting businesses in organizing and navigating information in a web-based, e-commerce environment. SearchKey API is readily used by software developers to build search and retrieval capabilities into their own applications and is easily integrated and interfaced with other technologies and products.

ASTAWARE SEARCHKEY API

ARCHITECTURE

Overview

ASTAware SearchKey API's JAVA Developer Kit is a Toolkit and Library to provide component searching and indexing technology. Java's 'write once, run anywhere' concept is illustrated clearly by the fact that our API can be integrated with any technology or operating system. SearchKey API's Information-Retrieval Portal can be used in any computer environment including intranet and Internet server, CD-ROM, DVD, or directly on a desktop. The SearchKey API multi-threaded search engine can distribute search requests from one host, CD-ROM, DVD or desktop to multiple hosts automatically. ASTAware SearchKey API's Information Server consists of three major co-related, yet independent, components: Information Indexing, Searching and Retrieval.



INFORMATION INDEXING

ASTAware SearchKey API Information Server can access information residing on multiple platforms using a file system or TCP protocol to retrieve documents for indexing. SearchKey API is a full text indexing engine that will read every word and number on the document page and add it to the indexes. Since SearchKey API uses an Inverted Index, each found word is stored in the index with a reference back to the source document and a frequency count for both the specific document and the entire index. SearchKey API offers additional support to field indexing of HTML Meta Tags and Text fields used for narrowing information searching. The indexing process is very precise, efficient and incremental. There is no limit to how many documents can be indexed and to how many indexes can be built. SearchKey API multi-index support allows for cataloging information in separate sets and the searching of each set separately.

The Indexing process consists of four steps to complete a full text index:

1. MASTER VECTOR & PARSING

During the Master Vector building process, SearchKey API requires access to the source documents through either the local file system or through an HTTP connection. It opens each document through its URL and reads the first x characters of content, the title of the document if available, and any user-defined summary information. This is stored along with the size of the document and the URL of the document. The extension list provides guidelines for the type of files that are to be indexed and the type of parser that is to be used for those file types. SearchKey API supports the following file formats: HTML, TXT, PDF, PostScript. Keyword parsers for these file types are built into our API. A customization of file extensions is also available, as is the ability of searching audio and video files with the deployment of Meta Tags technology. The Master Vector process creates the FileList.int file and the FileMap.lst file and builds an expanded list file called 'FileList.lst' which is then used to build the index. The parsing process opens each document, reads in the words (stripping the tags if required), and then compares each word to a Stopword list – if it is in the list it is ignored – otherwise the word is stored with its count and location.

2. MERGING

After all the documents have been read and an exhaustive list of keywords has been extracted, the next step is to sort and merge all of these files together. This is a fairly fast process that simply consolidates the lists of all words and records their locations.

3. BUILDING THE ACTUAL INDEX FILE

This phase now takes all of the consolidated words in the word list file and begins to build an index or vector map of the records. During the first phase, a path list is created – an entry for every record or file – this path list is then compressed into the internal SearchKey API structures. SearchKey API uses its own powerful proprietary logic at this point to compress the indexes into very small files, typically less than 15% of the size of the original content.

4. CHARACTER TREES

The final step is responsible for constructing the character trees for the indexes; these, in simple terms, are the compressed vocabularies for the database, and this allows the counts for words to be stored as well. The character tree files are typically much smaller than the indexes.

INFORMATION SEARCHING

The SearchKey API Information Search Service is responsible for handling search requests coming from client programs. The Information Search Service is capable of handling simultaneous requests and passing the results of the search to a results formatter. The search protocol is used to assist the Search Service in understanding the nature and type of query that the user has requested. This protocol allows end-users to use options that help to define a search query more clearly or to narrow the search request to a specific Meta Field, to a named database, or to a specific server.

Software developers can structure queries using Boolean operators (and, or, not, near) that are typically sent in a query string e.g. "computer AND program NOT software" and treats the operators with the precision required. Wildcard searching is also supported. For example, users can enter Java* as a query and their keyword query will include all words that begin with Java and have various extensions.

Search Interface:

Software Developers can design search query interfaces using Java applets, HTML forms or any other programming language. The Information Search Service can be integrated into any application written in Java, or receiving search requests from any other non-Java application using INI. The user's search interface is independent of the Information Server as long as the search queries that are passed to it follow SearchKey API query protocol standards.

Cascading Searches:

The Information Search Server is responsible for providing searches across multiple indexes, whether these indexes are located on one server, many servers, CD, DVD or Desktop. SearchKey API Information Server offers a powerful enterprise search system allowing for independent indexes to be tied together into one single knowledge base. The method of simultaneous searching of independent indexes is called "Cascading Searches". Software Developers can design their own enterprise system allowing for any combination of indexes. For example the custom system can allow for simultaneous searching of CD-ROM/DVD and multiple servers, or Desktop and multiple servers, or CD and Desktop and multiple servers.

INFORMATION RETRIEVAL

The SearchKey API Information Retrieval Service performs an index lookup and the results obtained from the index are passed through a formatting process. This creates a dynamic HTML document set that is controlled by a template. The formatted dynamic HTML pages are returned to the client and viewed in any popular browser. The results pages can contain a combination of page title, URL link, short summary, a content of Meta Tag or frequency. Software Developers can create their own combination of results and can display them within a viewer of their choice, browser, or their own application. The built in Simple and Expanded formats can be used or turned off for direct control of results.

SAMPLE APPLICATION

ASTAware SearchKey API allows you to build any search and retrieval application or integrate a search functionality with other applications. ASTAware Technologies INC. has used the API to develop two of their solutions: ASTAware SearchKey PRO and SearchDisc.

ASTAware SearchKey PRO is a server based full text search engine for multiple servers, domains, indexes and formats. We have used API to develop this commercial application by adding a GUI interface, network communication layer and expanding features of API taking advantage of TCP/IP protocol. Go to our SearchKey PRO demo site at <http://searchkey.com> to see this offspring of SearchKey API in action.

ASTAware SearchDisc is a CD/DVD full text platform independent search engine. We have used SearchKey API to develop this publisher application, by adding Index Administrator, HTML Search Forms. We took advantage of SearchKey API's simultaneous searching to allow for searching for information on CD/DVD and online at the same time. Go to our SearchDisc demo site at <http://astaware.com> to view an online demo or request a runtime version of SearchDisc.

BENEFITS

The ASTAware SearchKey API Information Server offers the following benefits to Software Developers:

1. JAVA Toolkit and Class Libraries;
2. Fully customizable Search Interface and Results processing;
3. Enterprise ready system for multiple indexes such as servers, CD, DVD, Desktop;
4. Built-in platform independence; and
5. Seamless integration with other applications.

CUSTOM SOLUTIONS

ASTAware Technologies INC. offers custom solutions development services to companies looking for the Information Server enterprise system offered by ASTAware SearchKey API. Our development team can tailor make SearchKey API to fit your needs and application. Contact our development team directly at techs@astaware.com to receive a quotation for a turn key solution.