

Accurate **search** What a Concept

Technical Discussion of RETRIEVALWARE

A CONVERA White Paper



CONVERATM

Accurate search

What a Concept

Table of Contents



THE NEED FOR A SUPERIOR SEARCH SOLUTION → 2

A NATURAL CHOICE FOR INFORMATION MANAGEMENT → 4

THE VALUE OF ACCURACY → 6

COMPLETE AND ACCURATE RESULTS → 8

CONNECT TO YOUR INFORMATION → 14

THE NEED FOR SECURITY → 18

SUPPORT OF INTERNATIONAL LANGUAGES → 21

THE IMPORTANCE OF SCALABILITY AND MODULARITY → 22

CONCLUSION → 27





THE NEED FOR A SUPERIOR SEARCH SOLUTION



Knowledge—it is the key to success and leveraging it requires making it accessible in a quick, accurate and secure manner. Information and knowledge must seamlessly flow between all processes in an organization and be available to employees, partners, and customers from a user-friendly, single point of access.

Imagine yourself as a knowledge worker in a large organization. You need information on a particular topic quickly in order to solve a problem or make a decision. As you look out at your organization's information landscape, you see a variety of electronic sources. There are file servers, intranet websites, document management systems, groupware systems containing discussions and documents, reports and research your organization has purchased and so forth. To tap into each of these would be a daunting and time-consuming task. But what if you had a single tool that would draw the information you need out of all of those available resources?

This is not a new idea. The concept of a portal—a gateway into an organization's repository of information—has become increasingly popular. A portal can encompass numerous applications such as Information Retrieval, Customer Relationship Management, Help Desk, Enterprise Resource Planning, Business Intelligence, and many more. Portals may be internal resources or part of your public Internet presence, and the individual functions of these portals may only address a subset of all possible requirements. Search and retrieval is a core service required for any portal, supporting many other portal features.

Thus, an information portal is only as effective as the information retrieval technology it incorporates. To be of maximum value, it must provide superior search. Convera's RetrievalWare is an intelligent information retrieval solution with all the features necessary for you to provide superior solutions that meet or exceed your organization's demands.

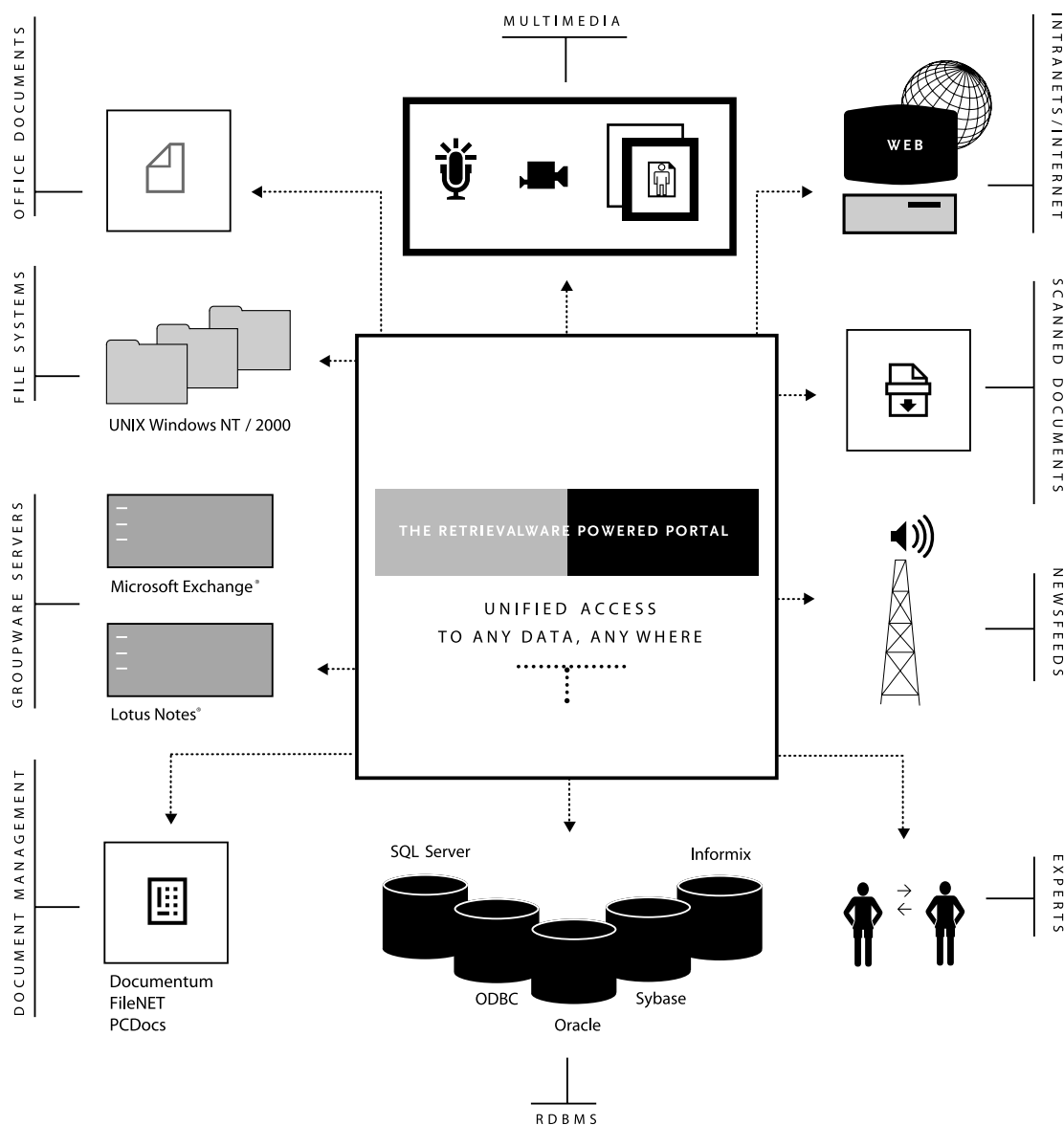


FIGURE 1. INFORMATION LANDSCAPE

How can you achieve superior search results? By employing a technology that enables the retrieval of complete, accurate, relevant information, regardless of file type or language or location, all while securing access. A search technology that meets these criteria reduces unnecessary rework and shortens time spent locating information, which enables your knowledge workers to solve problems faster and meet deadlines. By implementing a less intelligent search technology, you could miss out on these productivity savings and that goes right to your bottom line.





A NATURAL CHOICE FOR INFORMATION MANAGEMENT



RetrievalWare is the industry's most advanced knowledge retrieval solution for

accessing, organizing and utilizing a wide range of distributed assets, all from a common user interface. RetrievalWare incorporates powerful techniques known as Concept Search and Pattern Search, enabled by a mature semantic network™ and our patented Adaptive Pattern Recognition Processing (APRP™), to deliver the most accurate and relevant results. Using RetrievalWare in your solution also gives you the power to securely index and search more than 200 document types stored virtually anywhere in your enterprise.

Concept Search capitalizes on the richness of language, with its multiple term meanings, and transforms it from a problem into an advantage. RetrievalWare performs natural language processing and search term expansion to paraphrase queries, enabling retrieval of documents that contain the specific concepts requested rather than just the words typed during the query while also taking advantage of its semantic richness to rank documents in results lists. RetrievalWare's powerful pattern search abilities overcome common errors in both content and queries, resulting in greater recall and user satisfaction.

“Convera has the richest core technology with both neural networks and text indices that include the most mature semantic network capability.”

→ GARTNER GROUP

This document outlines the value RetrievalWare offers in all phases of information management, focusing on how these features as well as the underlying architecture enable you to implement a successful solution that leverages:

- The semantic network supporting Concept Search
- Adaptive Pattern Recognition Processing for greater recall
- Synchronizers that access content in a wide range of repositories
- Categorization to support browsing and search
- Profiling to alert users to new relevant documents
- A security model that supports secure access to disparate repositories
- International language support through a plug-in architecture
- A component architecture that supports modularity and scalability
- Software development kits (SDKs) and application programming interfaces (APIs) that simplify customization and integration

“In the Portal Tools category of the corporate portal space, suppliers with particularly strong search technology offer component functionality that can make a distinctive contribution to a portal effort. This group includes Convera, with highly developed features in the areas of multi-format search, analysis, and retrieval (including video and other image-based data types) and in semantic network and conceptual search across a wide range of information sources.”



Achieving high accuracy—i.e., retrieving the desired information and presenting it first—is one of the most valuable features of any information retrieval system. Accuracy is the ability of a text search system to find the documents that provide answers, without retrieving documents that provide no useful information. The practical result is that users should have the best documents in the top of the hit list regardless of the actual words used to perform the search. Accuracy is typically measured using two statistics: 1) **recall** (how many of the useful or relevant documents in the system are found); and 2) **precision** (what percentage of the retrieved documents is useful, or relevant to the request). The practical requirement for recall is ensuring that all of the relevant documents in an enterprise be in the result list. It is also important that the proper security functions are in place to ensure that each user sees only the documents that he or she is authorized to view. Only a system that addresses all these points simultaneously should be considered as a potential solution to the problem of information retrieval. With this in mind, it's easy to measure the value of accuracy in terms of time and money to your organization.

“Increasingly customers seek to retrieve all their information – both internal and external – from a single system. This requires that vendors integrate management and retrieval of unstructured and structured text, no mean feat, as well as such things as image files, audio files, streaming video and audio, training materials, and personnel records.”

→ SUSAN FELDMAN, IDC, MAY 2000

The **cost of missing important information** can be very large. While a single instance in which a worker resorts to asking a colleague or information seems like a minimal cost, imagine the expense when you multiply such an instance by many workers numerous times. At the other extreme, a manager missing a vital piece of information may make an unsound decision that could result in the loss of tens of millions of dollars.

Reducing the **time required to browse documents** translates into a very real and measurable productivity improvement. Using typical search technology with only keywords and Boolean logic, a user performing just five queries a day might need to browse many hundreds of documents to find the few containing the right information. This is due to a low precision rate—often below 45 percent. In contrast, RetrievalWare's accurate retrieval and relevancy ranking algorithms often provide a precision rate in excess of 90 percent, cutting the number of documents searched in half. If each document takes 15 seconds to browse, the savings per user per day multiplied over hundreds or thousands of users are enormous.

Even more important, perhaps, is **the cost of customer or user satisfaction**. Most text search systems are under-utilized, simply because user satisfaction with solutions that return inaccurate and incomplete results systems is so low. When users are frustrated with such a system, they stop using it and fall back on slower, manual systems (such as asking colleagues) or they actually re-create the information. As a result, not only is the investment in the system lost, inefficient processes continue and user confidence is diminished.

A Harvard Computing Group study done in 2001 entitled *Portals and Business Functions* addresses the value issue. One of the clients in the study estimates "the functionality available through the portal allows new salespeople to become productive in 2/3 the time. With 5,000 sales staff and an annual 15% staff turnover, this is worth \$27 million in

productivity improvements." Even if search represents only 10% of this effect, that represents almost \$3 million dollars in value!

Another example from the same study is a telecommunications company that conservatively estimates that its employee portal allows its staff to find relevant information more rapidly, thus increasing employee productivity by at least 15 minutes per day—representing a return on investment of 1500 percent.

RetrievalWare's accurate retrieval and relevancy ranking algorithms often provide a precision rate in excess of 90%.





COMPLETE AND ACCURATE RESULTS



RetrievalWare achieves high levels of both recall and precision through advanced search methods and close attention to every detail of the search process. RetrievalWare provides Concept, Pattern and Boolean searching methods that can be used independently or interactively to enable the highest levels of accuracy. RetrievalWare's powerful index and query pipelines use sophisticated formatting and linguistic processing components to achieve these results.

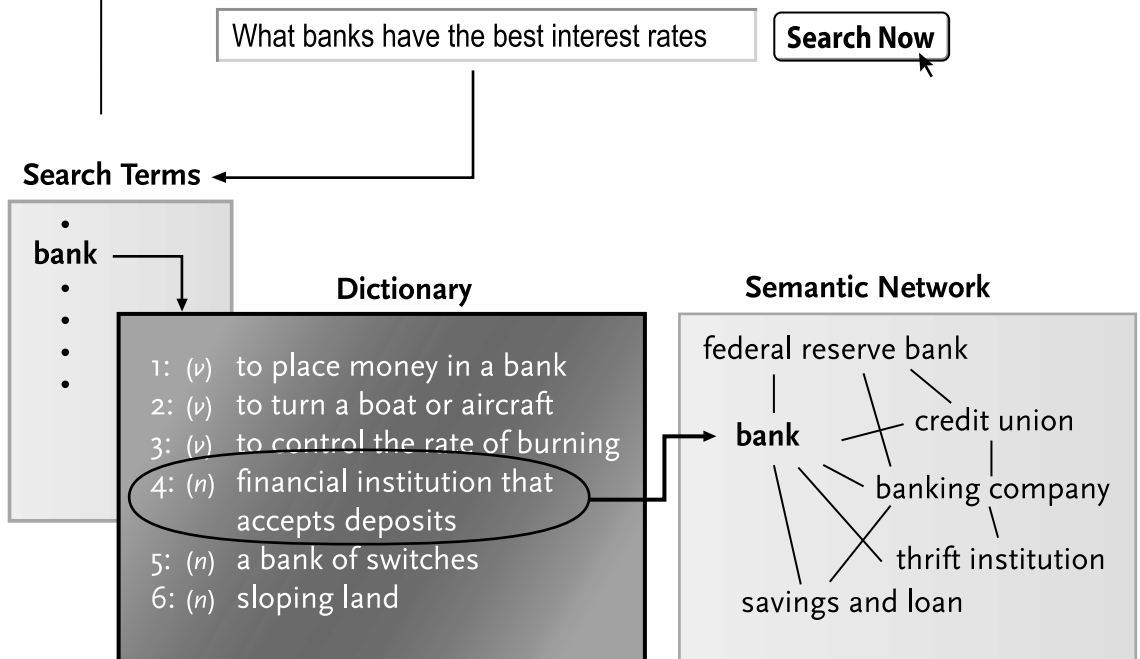


FIGURE 2. EXPANSION OF TERMS THROUGH RETRIEVALWARE'S SEMANTIC NETWORK

→ CONCEPT SEARCH

Concept Search does what we naturally do in conversations with each other, i.e., account for the differences in terms used by individuals to express the same concept. This ability enables paraphrasing of queries (e.g., a search for “international commerce” will find documents that discuss “foreign trade”) as well as clarification of query word meanings through analysis of surrounding words (e.g., the word “tank” when surrounded by words such as a “military” and “vehicle” is more likely to be a fighting vehicle and less likely to be a container for holding fuel).

The core support behind concept searching is the standard RetrievalWare semantic network, a collection of approximately 500,000 English words that expands to over 1.6 million semantic relationships and idioms that are organized by concept. Adding dictionaries in other languages and subject-specific domains can expand the semantic network to address these opportunities.

Although a program can use surrounding words and linguistic analysis to guess at the meaning of a term in a document or query, it is important to provide a user the choice to select the meaning of a term. RetrievalWare’s PowerSearch feature allows users to control word expansion of query terms in the semantic network, which means they choose which, and how many, related terms to include in the search for the most accurate results. Users can also choose specific meanings for their query terms. For example, when searching for banks, a user can easily instruct RetrievalWare to use only those other meanings associated with financial institutions and not those dealing with the bank of a river or a turning aircraft. In addition it is possible to select individual terms within a meaning to sharpen the search even more.

→ PATTERN SEARCH

Concept Search only addresses part of the accuracy issue. Often, words are misspelled due to optical character recognition (OCR) errors during imaging, author errors, foreign transcription errors, errors

made by the searcher and legal spelling variations. Pattern searching as pioneered by Convera uses Adaptive Pattern Recognition Processing (APRP) to overcome these problems. By utilizing a sophisticated internal voting and rating scheme that considers a number of different features of the pattern instead of just character pairs —APRP can find the words and therefore the right documents missed by weaker, “fuzzy” searches based at the character level. RetrievalWare allows the user to view the list of alternate spellings, sorted with the most closely spelled words first, and then select which spelling variations to include in the query.

→ BOOLEAN SEARCH

Although Boolean search is the common denominator among almost all search offerings and is often referred to as basic search, its power and functionality should not be dismissed. The two major drawbacks to Boolean search are the level of difficulty for learning and mastering it and its inability to return rankings between documents. In the hands of a sophisticated power user, Boolean search is an efficient, flexible, and accurate tool. This form of search is also very useful as a filter used in concert with concept and/or pattern search to limit the results and provide ranked results.

→ INDEX AND QUERY PIPELINES OR INDEX & QUERY PROCESSING FOR ACCURACY

During both indexing and querying, words undergo several phases of analysis and tuning, enabling the most accurate and relevant results to be returned. RetrievalWare achieves high accuracy by close attention to every detail of searching large, heterogeneous document repositories. Both documents coming in through the indexing pipeline and search requests traveling out through the query pipeline undergo several phases of analysis and tuning so RetrievalWare can deliver the most accurate results possible.

Search requests entered into RetrievalWare's Query Pipeline are processed through the following steps:

1. Tokenizing identifies strings of characters as words, dates or numbers and determines how to handle special characters.
2. Stop words such as "the", "a", & "and" are removed from the query so their hits don't artificially inflate the document rank.
3. Morphology reduces query words to their root forms, removing suffixes and verifying the existence of the root words in the dictionary.
4. Pattern matching expands the list of query words to include similarly spelled terms in the indexes if pattern mode is enabled.
5. Term grouping identifies words enclosed in parentheses, which allows several alternative words to be treated as a single search term.
6. Exact phrases, words enclosed in quotation marks (" "), indicate the importance of word order and proximity.
7. Idioms (such as "real estate" or "ice cream") are identified so that those phrase hits are ranked higher than occurrences of the individual words that make up the idioms.
8. Numbers or dates, including open-ended ranges such as greater than and less than, are normalized so users can search for them in the body of the document.
9. Query words containing one or more wildcard characters are expanded to include words beginning with or containing a certain string.
10. If the Power Query option is used, users can select specific meanings of query words and/or modify their weight.
11. Using semantic network expansion, words related to the concepts of query words are added to the query word list.
12. Documents are ranked by relevance and displayed to the user.

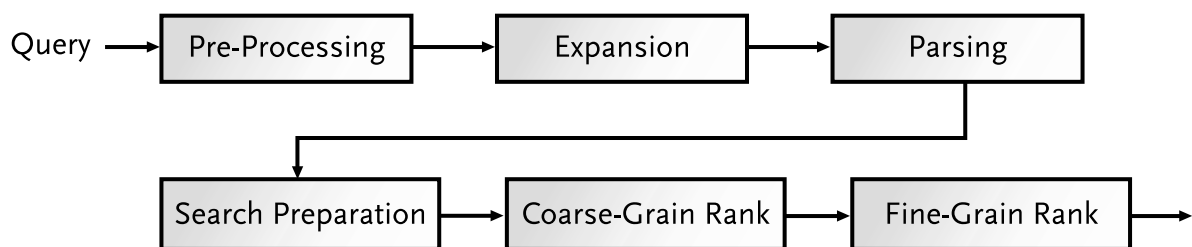


FIGURE 3. QUERY PIPELINE AND RANKING

→ RELEVANCY RANKING

Relevancy ranking is a two-phased process that ensures the return of the most pertinent documents from a search, no matter where or how the documents are stored.

The first phase is coarse-grain ranking, which reduces the search set from the entire database down to a manageable number of documents by using a range of factors such as:

- **Completeness**—the greater the number of query words found in a document, the higher it is ranked
- **Contextual Evidence**—the greater the number of terms related to the query words, the higher the ranking
- **Semantic Distance**—the greater the number of closely related terms to the query words, the higher the ranking

The second phase, fine-grain ranking, uses additional functions to determine the order of the final document list based on:

- **Proximity**—documents containing query words and related terms that occur close together (e.g., in the same sentence or paragraph) receive a higher ranking
- **Hit Density**—the greater the ratio of query words and related words to the total number of words in the document, the higher the ranking

A tremendous advantage of RetrievalWare's relevancy ranking is that it is independent of the individual statistics of the word frequency in each repository, unlike other search technologies that use database statistics. For example, a search solution that uses database statistics and that supports the search of a Lotus Notes repository, Documentum DocBase and a Windows, NT file system makes it virtually impossible to accurately combine the results from a search across these repositories. Instead, it is likely to rank one document that resides in two of the repositories differently for the same query, because the frequency of the occurrences of the query terms differs for each document set. Adding further confusion, it is possible to receive totally different ranking results for a document from

the exact same search across the same repositories if new documents are added and/or existing ones are removed.

In contrast, RetrievalWare provides a consistent relevance value for a document based on the user's query, no matter where the document is or what other documents reside in each repository. While more relevant documents added to the repositories may appear higher on the results list when the same query is again executed, the specific values of each document remain constant. RetrievalWare provides the best method to search across multiple repositories and create a uniform list of ranked documents.

→ DELIVERING SEARCH RESULTS

In addition to recall, precision and close attention to search process, another key determiner of user satisfaction is how search results are delivered.

Even if your solution finds all the right documents and puts the most relevant ones at the beginning, presenting them so they are easy to scan and read is just as critical. Two aspects for delivering search results—pinpointing the relevant hits within the documents and formatting the results for easier reading—are necessary for information retrieval systems.

RetrievalWare provides a comprehensive and customizable method for pinpointing relevant hits within the documents by highlighting them and putting them in rank order. These rankings can be color-coded when delivered to the user with the best hits highlighted in one color, medium in another color and so on. Another way RetrievalWare makes scanning the search results more friendly is with the option, “Go to Best Hit.” For example, one of the documents retrieved with the query, “acquisition of Dogfood.com,” is a 15-page SEC 10K quarterly financial report. It is great that the information retrieval system brought you the document, but you wouldn’t want to scan thousands of words just to find the part on Dogfood.com’s acquisition. Hit highlighting makes it possible to quickly scroll through the document, scan for what’s relevant or to jump directly to the most relevant hits with the “Go to Best Hit” feature.

The other important aspect to delivering search results effectively is how they are displayed. Rarely is the document’s file name sufficient to understand the content and nature of the document. By viewing a short summary of the document, a user can easily scan these brief snippets to determine which documents to open in their entirety. This makes the navigation of search results more efficient and, hence, a more satisfying experience.

The search across structured assets puts an additional demand on delivering search results effectively. For example in a Human Resources

application, resume information for potential candidates might have been stored in a structured database, with many individual cells containing information such as name, address, work experience, education, etc. The challenge is how to present the structured text in a way that is formatted and user friendly. Using the same parser referred to in a way that is formatted and user friendly. Using the same parser referred to in the query pipeline, RetrievalWare formats the content and returns a well-organized, easy to read page, that is customized to fit each user’s specific needs.

→ STRUCTURED DATA AND DOCUMENTS

The subject of structure raises two concepts – structured data and structured documents.

Structured data is stored in a database, where the “structure” comes from the schema of the repository that identifies the meaning of a piece of information based on where it resides in the table. In the case of structured documents each document has “tags” that label each structure and type of information contained within the document. The grandfather of structured documents is SGML (Standard Generalized Markup Language), which laid down the foundations of HTML and XML in all its variations. These structured file formats are the lifeblood of the Web, and provide not only the promise of better machine-to-machine communication and processing, but also more precise search.

Think of going into a bookstore to find a copy of Leo Tolstoy’s War and Peace. This is a relatively easy task; you just go down the aisle, looking at titles and authors until you find it. There’s no need to read the contents of each book you pass by. If the content that an indexing engine provides is just a monolithic stream of text with no tags to identify which part is the title, index, chapter or any other structure of the asset, then you can’t search for a term in a specific structure. In other words, you can only search content by title, author or publisher if the content is “tagged” as such. Now it becomes possible to automatically index both the content and structure and then search based on this structure.

The typical tagging scheme is to have an open tag such as </Author>, which is followed by a string of the author's name and a closing tag such as <Author>. RetrievalWare has the ability, using our powerful parser function, to identify content between tags and store them in fields. This allows a search interface, such as Convera's SmartSearch or a custom search application, to expose these fields for users to search within them. With this capability, a user would simply type "Tolstoy" in a field labeled "Author" to find War and Peace by Leo Tolstoy. When formatting documents for display to the user, RetrievalWare drops the tags for easy reading. Often a particular document set contains many tagged structures and though it is not appropriate to display them all to the user, they are very valuable in certain search applications.

→ THE SEARCHERS AND BROWSERS

There are basically two ways people look for information: searching and browsing. Thus far, this paper has addressed searching—entering in a query in some form and then looking at the results. Browsing is when the assets have been pre-arranged in some structure, which allows a user to navigate through the groupings of content going up or down in level, depending on the need to increase or decrease focus. Support of browsing content requires a deep knowledge of the possible content and how to arrange it in a logical order. These arrangements are often called taxonomies and are the domain of Library Science graduates. Given a good taxonomy, and that is taking a lot for granted, the question is how to populate the categories that it defines? The most precise way is to hire a team of subject matter experts that review all the content and manually place it in a category. You might be surprised at how many well-known Internet sites do just that.

Assuming this is not feasible, the alternative is to employ some additional technology to automatically assign content to a category. This is actually search with saved results! Each category is assigned a query that identifies which assets belong in it and as new content is discovered, it is added. This means that

the accuracy of your search engine is controlling the relevancy (precision) and the completeness (recall) of the content. The quality of your categories will be directly related to the quality and power of your search.

→ CATEGORIZATION

Because RetrievalWare excels in the area of accuracy, it provides the power necessary to support the creation and automatic management of categories. Another very important component for the successful use of categorization, particularly in intranets, is security. In an effort to make the categories as complete as possible, it is easy to post information that should not be seen by everyone. On the other hand, only including information for the lowest common denominator results in categories that are incomplete to your power users. By combining the power of RetrievalWare's concept and pattern search with cross repository and library security, only the documents that a user should see are displayed, thus preserving the total security of the system.

RetrievalWare's broad access to content, precision and recall are also key enablers for high quality profiling. Profiling is the process where the flood of new content entering the enterprise is filtered with rules set by users. The user is immediately notified when content that matches those rules becomes available. Accuracy is critical to this function, so that all the relevant information is found and the users are not disturbed with irrelevant content.

“Through 2002, search techniques will evolve from a focus on indexing words in documents to more complex semantic analysis and pattern matching algorithms.”

→ GARTNER GROUP





CONNECT TO YOUR INFORMATION



Complete and accurate results can only be achieved when your information management solution has access to all the necessary content available.

RetrievalWare Synchronizers let you **extend your search and retrieval applications** to encompass a multitude of repository types. It is the Synchronizer that enables RetrievalWare to securely “reach out” to one or more instances of a source of documents or data, pull their contents into a common index space, so the user can navigate, search, and browse them. A Synchronizer is also responsible for keeping its related index up to date by determining whether anything has changed for the assets or the access rights (ACLs) in a repository.

Convera gives you the flexibility to choose exactly which Synchronizers you need, based on where assets are stored in your organization. Out of the box RetrievalWare comes with Synchronizers for file systems and one for relational database management systems (RDBMS). Additional Synchronizers are available for document management systems (Documentum, FileNET, Panagon), Groupware Servers (Microsoft Exchange, Lotus Notes) and Teradata. If your organization’s solution requires access to a repository or file type not currently supported or not available from a partner, the RetrievalWare SDK includes an Access Filter Module (AFM) Toolkit that allows you or our integrators such as Convera’s Integration Services Group to build the necessary Synchronizers and filters.

→ FILE SYSTEM SYNCHRONIZER

Much of the valuable information in today's enterprise is stored (or in some cases abandoned) in file systems throughout the organization. These are valuable assets and should be included in the enterprise knowledge retrieval solution.

RetrievalWare's File System Synchronizer finds these documents. This Synchronizer can be set up to run automatically at regular intervals to process new, modified and deleted documents in specific file system paths, preserving privacy. There is support for both Windows NT and UNIX file systems.

Synchronization of a Windows NT file system can also extract the access permissions to support secure access. Each instance of the File System Synchronizer handles data for a particular file path. For large dynamic file systems, multiple instances of the Synchronizer can each handle sections of the file server to provide the most up-to-date indexes without sacrificing performance. The Synchronizer remembers the list of files and file properties for all files that were present during the last Synchronizer run and use this information to scan the file path for any new, modified or deleted files, then processing only the appropriate files. Upon completion of the synchronization process, RetrievalWare updates the indices for the new documents, making them available for search and retrieval.

→ RDBMS SYNCHRONIZER

Relational databases have become core technology for storing and processing mission-critical information in almost every organization. The problem is that the mountains of structured and unstructured data in the RDBMS have been islands of information apart from other repositories. The RDBMS Bridge allows RetrievalWare to securely access the contents of any RDBMS to provide powerful search for that application and enables it to act as a conduit that brings together these assets with others distributed throughout the enterprise. Often these RDBMS repositories contain very large numbers of small pieces of content and require a lot of processing. By taking advantage of native calls directly into many RDBMSs, RetrievalWare avoids the overhead of ODBC and JDBC, achieving the highest

performance during indexing and retrieval and providing the broadest range of functions when accessing databases. Once the RDBMS Synchronizer is set up, it only reads from the database, eliminating any conflict with database administration. The RDBMS Synchronizer performs an initial submission of legacy database table data for indexing; then the RDBMS Synchronizer tracks inserts, updates and deletions, which it subsequently uses to submit changes to the indexing servers—which, in turn, update the indices for the RDBMS library. RetrievalWare supports native access to Oracle, Microsoft, SQL Server, Informix, Sybase, and uses ODBC access for Teradata and other databases. Once the data is indexed, RetrievalWare uses the same ability to parse (read and analyze) the data, formatting the data in the fields into a user-friendly presentation for displaying results.

→ DOCUMENT MANAGEMENT SYNCHRONIZERS

Document management systems (DMS) have been important applications for many years, and their importance has increased tremendously with the rapid growth of Web content and other electronic documents. RetrievalWare's Document Management Synchronizers securely access assets stored in the two leading document management systems, Documentum and FileNET Panagon. A custom Synchronizer can also be created for other types of DMS. Each Document Management Synchronizer uses a method appropriate to the specific system for extracting copies of all the documents and determining their access rights. Each Synchronizer then uses a client program on the RetrievalWare server to access and retrieve new, changed or deleted documents at periodic intervals determined by the system administrator. The appropriate documents are passed to the document handler for indexing, cross referencing and—optionally—profiling, before they are made available for searches. The Documentum and FileNET Synchronizers support indexing attachments to documents, as well as indexing access control information for document-level security. Additional RetrievalWare Document Management Synchronizers, such as PC Docs, are available from Convera partners.

→ GROUPWARE SYNCHRONIZERS

Groupware applications and servers have become the lifeblood of every organization, serving as important repositories for content of every kind. RetrievalWare Synchronizers support secure access to both Lotus Notes and Microsoft Exchange content, enabling indexing and searching of these leading Groupware products. Similar to Document Management Synchronizers, Groupware Synchronizers determine which documents require processing using the appropriate method for the product. For new, changed or deleted documents found in Exchange or Lotus Notes, the Synchronizer uses a client program to access the Exchange server and retrieve the documents. The Lotus Notes Synchronizer provides search of Lotus Notes in RetrievalWare and the Lotus Notes Plug-in option provides RetrievalWare search within the Lotus Notes interface. The Groupware Synchronizers can access documents and their attachments and determine their access rights to ensure users see only the documents that they are entitled to see.

→ RETRIEVALWARE FILEROOM

There are many organizations that need to include paper-based documents in their information retrieval solutions. In some cases, these are static legacy information sources, while in other cases they are active content producers. RetrievalWare FileRoom allows scanned paper-based assets to be incorporated into digital repositories by quickly and accurately capturing, loading, indexing and retrieving both image and text documents. FileRoom administrators create special RetrievalWare Libraries that are presented to the user in a familiar hierarchy consisting of file rooms—cabinets, drawers, folders and documents. The containers combine to form an archive of the paper-based knowledge assets that users can expand, browse and view in conjunction with or apart from electronic-source documents. While RetrievalWare FileRoom has been pre-integrated with Kofax® Ascent Capture, which provides scanning and OCR facilities, it can also be easily integrated with many other document capture solutions.

→ RETRIEVING MULTIMEDIA

In recent years the technology for communicating information using video has been more accessible and, hence, more prevalent in the enterprise. How to access, organize and utilize video-based content is the new challenge on the horizon. Convera's multimedia support in RetrievalWare is the first and only solution in the industry. Our multimedia support allows you to ingest, analyze and add metadata to video content via Screening Room Capture. Screening Room Capture is a scalable and multi-faceted video-logging application that oversees the processing of video input and directs the post-capture processing of the results. Video capture handles input from a variety of sources, both analog and digital, and given the appropriate hardware, processing can include any or all of the following:

- Video-event detection (e.g., scene-change detection), which produces a set of "event frame" thumbnail images, along with related information such as the timecode at which the event occurred.
- Embedded text (closed captions or TeleText subtitles) can be extracted, and the audio track can be transcribed by a speech-recognition system.
- Simultaneous encoding of one or more digital copies of the input video, in various formats and at multiple bit rates, for later desktop viewing.
- Standard metadata such as file name, video format, capture date and time and framerate are captured. Also custom metadata such as keywords and framenotes can be added.

Following video processing, post-processing of the results can encompass the following:

- The results of video analysis and text extraction, and all related information—the "CaptureData"—are written to XML files. This captured data can be simultaneously posted to a database server, transmitted to Web servers, copied to storage, and so on.
- The encoded digital video file(s) can be simultaneously transferred to video servers, duplicated, archived, and so on.

Now your video assets can be searched along with the text-based assets. For example a search for “How do I treat high blood pressure” could return not only text articles and reports, but also videos from the FDA and pharmaceutical companies, all in a single ranked results list.

→RETRIEVALWARE’S INTERNET SPIDER
 Web servers using HTTP, SHTTP or SSL, either inside the enterprise firewall or on the Internet, are often the technology of choice for storing and distributing information. The ability to access these resources and search across these and other information repositories in a uniform way is important for knowledge worker productivity. Although the Web started out as a network of text-based documents, it often now leads the way in multimedia technology. Internet Spider is a multimedia, high-performance Web crawler that enables

document delivery to RetrievalWare from any HTTP-enabled server.

Internet Spider navigates Web pages and other document files, collecting their URLs in a database. In subsequent Web crawls, Internet Spider finds any new, modified or deleted files and then passes those pages for updating the database. Like other RetrievalWare functions, the Spider is based on a scalable parallel architecture that allows deployment of multiple instances of discreet components to handle processing-intensive tasks, such as crawling large sites or whole areas of the Internet. Web pages and other contents are then sent to the indexing servers for processing. The content of Web pages can even be stored on the RetrievalWare server for guaranteed and faster display, with the added benefit of hit highlighting.

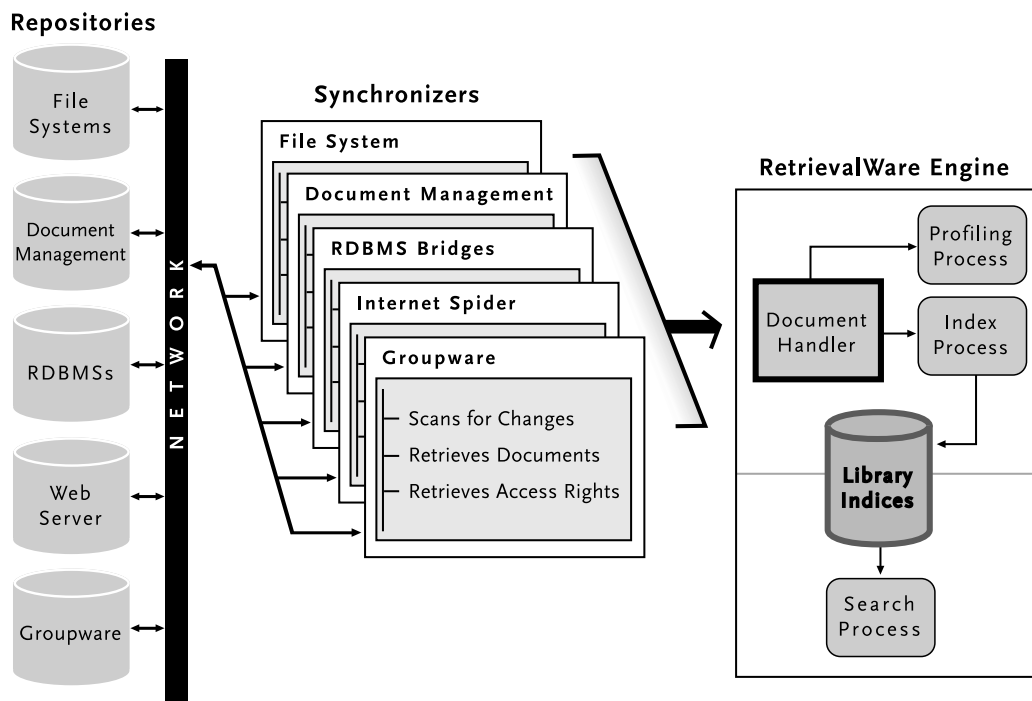


FIGURE 4. SYNCHRONIZERS

“Inability to find information easily on a site is a huge area of frustration for end users, and it will inevitably affect a user’s perception and relationship with an organization.”

→KATHLEEN HALL, GIGA INFORMATION GROUP
 MARKET OVERVIEW: ENTERPRISE SEARCH
 MARCH 23, 2001





THE NEED FOR SECURITY

Now that most modern information retrieval solutions provide easy access to enterprise information, there is an increased need for security. The dynamic nature of matrixed organizations, with teams from both inside and outside the organization, requires a powerful and flexible security system that is easy to manage. RetrievalWare provides a security infrastructure that extends to and leverages the native security of the disparate document repositories in the RetrievalWare knowledge space through a single logon.

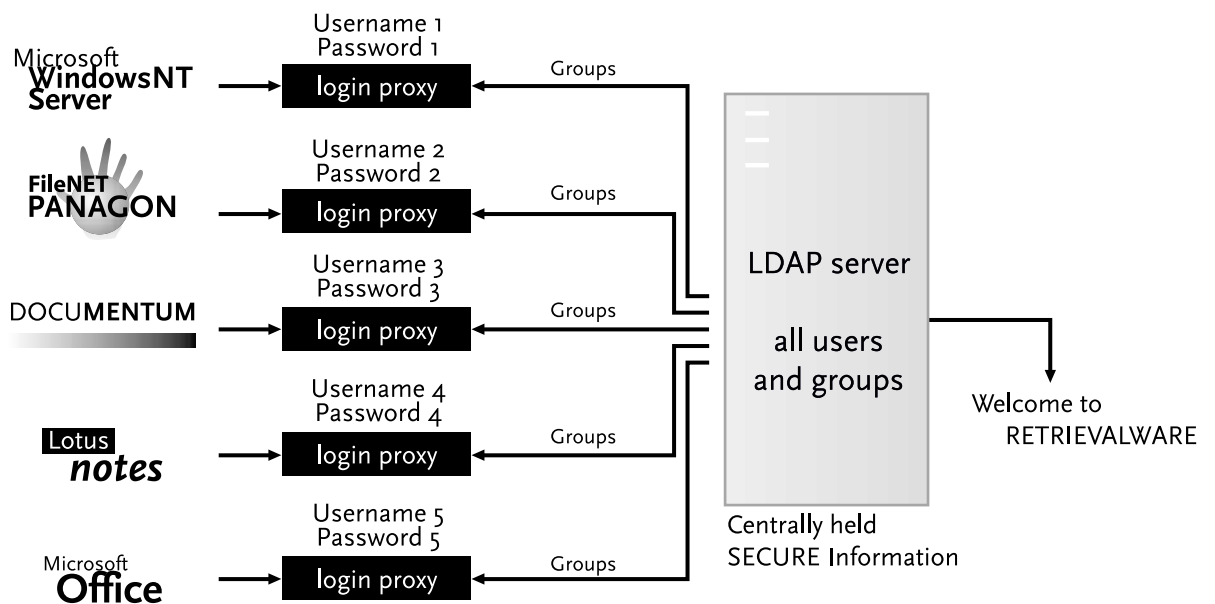


FIGURE 5. CROSS REPOSITORY SEARCH WITH SINGLE LOGIN

→ AUTHENTICATION

The key to any security system is a good front door. When a user logs into RetrievalWare either by the standard client interfaces or any custom interface, the user must enter a valid user name and password. If RetrievalWare is part of a larger solution such as a portal, the login can be automatically performed by that system. Once the user is logged in, the security server generates a unique session key that identifies the session and user information. This session key is used to automatically check the authenticity of every client/server transaction.

The initial authentication can be done against a RetrievalWare held repository or against the security system of one of the supported third-party document repositories through the RetrievalWare Login Proxy Server, which acts as a central point of control for logging clients into a RetrievalWare system. The Login Proxy Server authenticates users attempting access to RetrievalWare by using security information in other systems connected to RetrievalWare. RetrievalWare has built-in login proxies for several common third-party applications, including Windows NT, Microsoft Exchange, Lotus Notes, FileNET Panagon and Documentum. In addition, RetrievalWare provides a Security Toolkit that enables you or Convera's Integration Services Group to develop your own Login Proxy Server to support other sources of authentication such as Enterprise LDAP servers. Once a user is authenticated, a ticket is issued enabling the client to log into RetrievalWare.

With the Login Proxy Server, the RetrievalWare administrator does not need to re-create or maintain current user and group information. Instead, authentication information for RetrievalWare users is obtained from the central third-party domain. The Login Proxy Server also ensures tight security, as client passwords are not stored in local memory and outside processes cannot re-use packets that the servers transmit over the network to gain access to the system, thwarting a common hacker method.

If authentication information is required from an unsupported source, the Security Toolkit within the RetrievalWare SDK creates this service.

→ DISTRIBUTED SECURITY

The architecture of RetrievalWare permits the distribution of processes across multiple systems in order to provide maximum scalability and flexibility. To support this ability, the RetrievalWare security model enables a single logon across multiple instances of RetrievalWare servers wherever they are located. RetrievalWare provides a truly distributed security capability that enables a secure RetrievalWare system to share security group information with other remote, trusted RetrievalWare servers. The user only has to log in once to the local RetrievalWare system and, by proxy, will be logged into other RetrievalWare systems for access to remote libraries. User names and passwords either may be managed from a single "home" RetrievalWare system or managed individually on each distributed server.

→ CROSS-REPOSITORY LOGON

Once a user has logged into RetrievalWare and has been authenticated, there is no need to log into individual secure repositories when searching across one or more of those repositories. During the initial set up of RetrievalWare, an LDAP server is configured that will hold the necessary information to automatically perform the login to each repository. Each user is only required to provide the access information to each repository once during the initial session in order to populate the LDAP server. The information in this LDAP server is encrypted to protect its contents. In addition, all data communications to and from RetrievalWare are encrypted to prevent anyone from monitoring the data or passwords being communicated.

→ DOCUMENT AND LIBRARY LEVEL
SECURITY

During indexing of documents in a secure repository, RetrievalWare can also index the access control list (ACL) information of each document, if available. ACLs are lists of the specific users and groups who may access specific assets. This information is stored in the RetrievalWare index in association with each document. When the user logs in, RetrievalWare tests the user ID and password it has stored in the LDAP server for each of the secure Repositories. Assuming the challenge is successful, RetrievalWare then requests the current group affiliations of the user. Once a user is identified and validated, RetrievalWare allows the user to perform searches and enforces document-level security for assets from those repositories using this up-to-date information it received from the secure repository. This security feature ensures that users see only results that include documents that they are authorized to view and retrieve when browsing those repositories. This prevents the “leak” of information that can occur when summaries, titles, or file names of documents that a user does not have rights to see are made available in a results lists.

This security extends to contents of categories under RetrievalWare’s control; for example, if a subject expert, who often has complete access rights to certain repositories, creates a new category with content from those repositories, users still see only those documents to which they have access. This combination of authentication and access control restricts user and group access to specific libraries, documents and functions so that users get only those results for which they have been granted access by each of the repositories they search.

RetrievalWare provides a security infrastructure that extends to and leverages the native security of the disparate document repositories in the RetrievalWare knowledge space through a single logon.





SUPPORT OF INTERNATIONAL LANGUAGES

The need to manage documents in a variety of languages has increased dramatically in the last few years. The Web and globalization in general are transforming language requirements for even the smallest enterprises. It is critical that any search system allows indexing and retrieval of documents in different languages without setting up a separate system for each language.

RetrievalWare includes modular support for more than 25 languages and, with its Language Plug-in architecture, enables you to easily add other languages to a system. The power of concept search is supported within RetrievalWare through the availability of Language Plug-ins that not only understand the unique character sets, word spacing, morphology, idioms and stop words for basic index processing, but also a semantic network of word meanings that are necessary for concept-based search. Details on the level of support for any particular language are available from Convera upon request.

You can control the language used for indexing manually or you can take advantage of “language tags” in documents, which determine the language-processing module to use for a particular region of text. This architecture allows Convera’s RetrievalWare to not only provide multi-lingual search (the ability to search in more than one language and retrieve results on documents in those languages), but also cross-lingual search (the ability to search in one language and retrieve conceptually relevant documents in many languages). For example, with cross-lingual search, an American financial institution analyst who only speaks English, researching the potential risk for making a loan to a company will get the Italian newspaper article on the defaults by this company to Italian banks last year and the Russian article on their expansion plans. These results are from one search request, written in English.



IDC believes that not only is the global economy a multilingual economy, but that a company that implements a well-designed localization strategy can tap a significantly larger potential market for its products and services.

STEVE McCLURE, IDC

“LANGUAGE MATTERS”

AUGUST 2001





THE IMPORTANCE OF SCALABILITY AND MODULARITY



Because search is a service that is **used both inside and outside enterprises** with a variable number of users accessing diverse and dynamic content, it is very important that the enabling technology be able to grow to support changing needs with a minimal effect on performance. RetrievalWare was developed from its inception on a distributed process architecture, which provides the foundation for building scalable solutions that can maintain concept search performance even as demands on the system rise by orders of magnitude. Unlike search solutions that come up against a “one machine barrier” limiting database size or number of users, RetrievalWare can leverage additional server machines and maintain search performance over large databases and large numbers of users.

→ DISTRIBUTED ARCHITECTURE

RetrievalWare’s architecture provides the ability to build scalable systems that can be distributed across an organization’s computing infrastructure. Because you can distribute single or multiple instances of a process across multiple processors and or servers anywhere on the enterprise’s local or wide area network, your solution can be scaled to meet most requirements. The RetrievalWare architecture is web based, constructed on a client server structure with distributed processes and remote procedure calls (RPC) across a TCP/IP transport layer.

The four main process sectors of RetrievalWare's basic architecture are:

- **Management Processes** (administrative), which route client requests to the appropriate servers and manage server-to-server communication and load balancing.
- **Client Processes**, which process and submit search requests to one or more RetrievalWare libraries and then send the results for display.
- **Search and Retrieval Processes**, which translate client requests into searches against a RetrievalWare index and retrieve documents and hits found as a result of the query.
- **Indexing Processes**, which find new, modified and deleted documents, and create and maintain the indexes appropriately.

No matter how large your solution scales, RetrievalWare's distributed architecture enables you to maintain high performance and reliability, as follows:

- **Response time** between submitting a query and receiving results. Most users will only wait a few seconds at most to get back the results of a transaction. In some situations users will wait longer for very valuable results.
- **Time to index new information** in the source repositories. It is often impractical to have a single indexing process, servicing multiple repositories or in some very large and content dynamic cases a single repository. Our architecture lets you implement multiple indexing processes across multiple servers when required due factors like size or location.
- **High availability**—as organizations depend more and more on their information systems, they will tolerate less downtime. A scalable architecture allows for development of systems with redundancy and system management to support high availability.

→ TOOLKITS TO MODIFY, EMBED AND EXTEND

The RetrievalWare product family is a set of applications that can be used to implement information retrieval solutions right out of the box. For more custom solutions, RetrievalWare offers a series of toolkits to modify, extend or embed these applications in your enterprise. Designed to be used by system integrators, OEMs, software developers, and Convera's Integration Services Group. The RetrievalWare SDK contains toolkits for integrating with RetrievalWare servers, as well as creating or modifying client interfaces.

→ WEB TOOLKIT

Although the standard interfaces provided with RetrievalWare are user-friendly interfaces that provide full access to the power of the product, it is often desirable to modify them or to create completely new interfaces for your environment. The Web Toolkit simplifies these activities. Using customizable HTML forms, the Web Toolkit provides everything needed to enter searches, execute searches, browse document titles and browse document text. The Web Toolkit is far more flexible than traditional Perl approaches, making all HTML documents available to the administrator as

"templates." Using these HTML templates, the administrator can control the entire look and feel of the application, altering the behavior of any interface item simply by modifying the template. The RetrievalWare SmartSearch interface, which is the preferred intranet user interface, was built from the Web Toolkit.

→ JAVA SERVER PAGE TOOLKITS

The emergence of high-volume Internet applications has created the need for increasingly efficient and scalable server-side application architectures. At the same time, the Java programming language has emerged as the language of choice for Web application development. As of May 2000, 57% of Web development projects were based on Java, compared to 37% for C++ and 12% for other languages (source: Cutter Information).

International Data Corporation projects that the number of Java developers worldwide will increase from 1.9 million in 2000 to 3.4 million in 2002. The Java ServerPage Toolkit (JSPTK) addresses the market for search tools for Java developers. The JSPTK allows you to build a high-performance, Web-based interface to the RetrievalWare servers and can support large numbers of concurrent users.

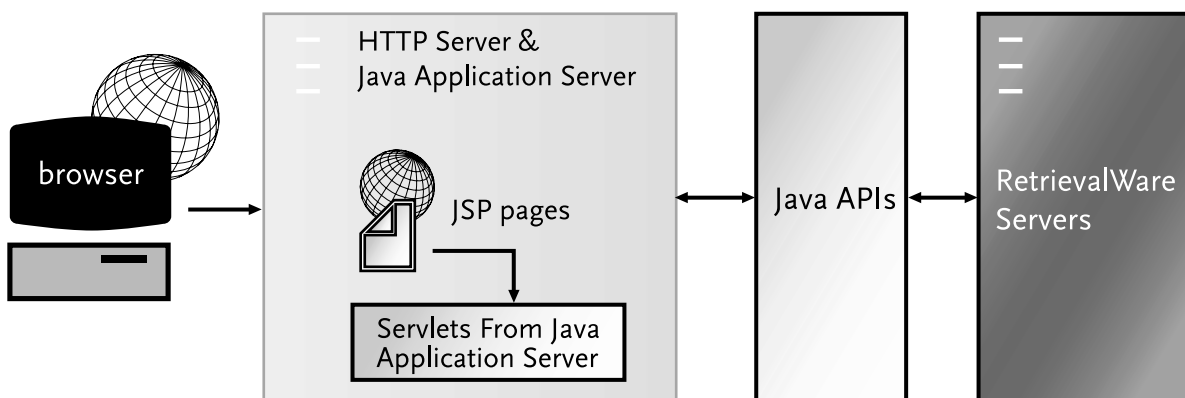


FIGURE 6. SCHEMA FOR RETRIEVALWARE'S JAVA SDK

In addition, because it uses direct interfaces into the RetrievalWare search servers, the JSPTK can handle the highest possible volume of simultaneous users.

Based on Java high-level application program interface (API) wrappers, the JSPTK allows you to build interfaces that use JSP commands to control servlets that are dynamically created and controlled by the Java Application Server and that speak directly to the native interfaces of the search servers. The JSPTK utilizes standard JSP syntax, so it is easy for Java programmers to use.

RetrievalWare makes it possible to perform Java queries that execute several searches, one after the other, so client applications can open several query objects simultaneously and switch between them in asynchronous mode. For example, you can search libraries using different search techniques and then merge the results. Or, you can perform multiple queries that execute over a library of documents simultaneously.

The JSPTK provides access to all RetrievalWare functionality, including storage and re-use of profiles and search results; full RDBMS access; and full search access, including user-feedback expert mode, query by example (QBE), Boolean, concept, pattern, fielded and/or any combination.

→ XML AND COM/ASP TOOLKITS

The RetrievalWare XML and COM/ASP toolkits provide mechanisms for submitting XML-based requests to and receiving XML-based responses from RetrievalWare systems and sending and receiving COM based requests using XML. These capabilities comprise a powerful high-level programmatic interface to RetrievalWare systems and offer an industry standard means for data representation. The XML toolkit enables applications to submit queries and receive results in XML format, for the following functions:

- User Login
- Accessing RetrievalWare library information
- Full Query support including expert searches, recurrent searches and “query by example”
- Return of query results

- Return of resulting documents with hit highlights.
- Query over categories.
- Navigate categories, stored queries, stored profiles, and experts.

The XML API includes a sample JSP client illustrating XML API function calls that is certified on all platforms on which RetrievalWare is certified.

The COM/ASP Toolkit provides a COM object interface that interfaces with the XML API to enable Microsoft IIS/ASP developers to integrate RetrievalWare searching capabilities. This toolkit enables a user to write an ASP application client using the COM object. The COM/ASP toolkit provides a way to send and receive COM requests wrapped in a Java based XML object to RetrievalWare’s XML interface. This capability allows ASP developers to create applications that take advantage of the RetrievalWare powerful functions.

→ LANGUAGE PLUG-IN TOOLKITS

As stated earlier, the need to support multiple languages has increased dramatically in the last few years. The RetrievalWare architecture incorporates Language Plug-In modules that isolate language processing during indexing and search. These plug-in modules support functions such as tokenization, morphology, stop words and idiom recognition, which are language specific. Thus, you can easily add support for a new language by supplementing English processing with additional language plug-ins, without disruption of the existing system. If a language is not already supported by Convera, the Language Plug-in Toolkit has all the tools and information you need to build a new language module and semantic network with corresponding dictionaries to support it.

The Toolkit also provides information for including “language tags” in documents, that tell the system which language-processing module to use for a particular region of text. Similarly, the system supports multi-lingual querying by allowing users to either embed language tags in a query or to select the language in the client user interface. Convera also provides cross-lingual plug-ins that actually enable you to enter a search in one language and find hits in documents in one or more different languages.

→ ACCESS FILTER MODULE TOOLKIT

With the immense number of file storage applications and document formats continuing to grow and the existence of internally developed applications, Convera understands that source documents may be stored in repositories or formats that are not supported by the core RetrievalWare product. That's why RetrievalWare includes the Access Filter Module (AFM) Toolkit, which enables you to incorporate support for any kind of data source or type of data.

With the AFM Toolkit, you or Convera's Integration Services Group can easily customize RetrievalWare's document handling capabilities without recompiling or re-linking RetrievalWare software. You can make simple alterations by changing variable settings or configuration files, modifying scripting rules or launching other programs. For more involved tasks, you can create custom AFM plug-ins. For example, many organizations have home grown document management applications. By building custom AFM plug-ins for these applications, you can access the documents and ensure synchronization with the security and versioning control interfaces.

→ EASE OF CUSTOMIZATION

RetrievalWare's SDK and well-documented APIs make it easy for you to customize and embed RetrievalWare into your interfaces and systems, facilitate system-to-system communication and modify server functions.

High-level APIs

RetrievalWare's high-level APIs are optimized for ease and speed of integration and enable you to create complete custom applications by accessing RetrievalWare directly from within another application. The RetrievalWare SDK provides APIs in C, Java XML and COM (via XML) to support the industry standard integration languages and formats.

The high-level APIs provide you with a significant degree of programming flexibility. Some of the many functions you can implement enable you to:

- Manipulate more than one query at a time through the use of query threads
- Direct how the query will be executed by setting the query properties
- Remove word senses and terms from pattern matching or wildcarding before semantic expansion
- Control the semantic expansion of the query terms by setting the expansion level and word limit
- Access the retrieved document's body text and meta data fields
- Find out about the strength and location of the hits within a document
- Step through hits and move around in the document
- Download a file from a server to a local disk
- Integrate RetrievalWare with a relational database
- Access the retrieved document's text in its original "raw" state (i.e., just as it enters RetrievalWare with all formatting characters, nulls, white space, etc.)
- Access fields and original "raw" text for any document in the library
- Log users into secure systems via username/password or a proxy login ticket

With the RetrievalWare SDK, you have all the tools and information necessary to customize, extend and embed RetrievalWare in any fashion you need.





RetrievalWare provides you with the search accuracy, repository access, security, scalability, modularity and flexibility needed to solve any information retrieval problem. And, it's all backed by Convera's Support and Service Teams. Highly technical personnel are dedicated to ensuring your success by providing you with personalized service and the highest level of technical support available.

Because Convera recognizes that you trust us with a most precious asset of your organization—your knowledge—we make sure that you have access to the experts who can guide you through the process and who will continue to be there for you as long as you remain our partner. Convera will continue to lead the industry with the technology and service that have allowed organizations of all sizes to implement the best knowledge retrieval systems in the world.





ABOUT CONVERA



Convera is a leading provider of software products that access, organize and utilize enterprise data, whether it be text, video, audio or image files. Convera's advanced technologies and products enable organizations to optimize the value of all their content, establishing an information infrastructure that effortlessly scales to provide large numbers of users with fast, accurate, web-enabled access to all relevant information for a broad range of business critical applications. Convera serves over 750 customers in 29 countries from its offices throughout the U.S. and Europe.



For more information
contact Convera

UNITED STATES
800 788 7758
info@convera.com

UNITED KINGDOM
+44 1344 781 800
info@convera.co.uk

www.convera.com





US T 800 788 7758
T 703 760 4085
F 703 748 1255

www.convera.com
info@convera.com

UK T + 44 1344 781 800
F + 44 1344 781 801

info@convera.co.uk

GLOBAL OFFICES
Carlsbad, CA
Columbia, MD
London, UK
Munich, GER
Paris, FRA
San Jose, CA
Vienna, VA