

Convera (Offline)

© 2013 by Stephen E. Arnold, www.arnoldit.com

Convera positions itself as a knowledge discovery platform, a marketing angle that vendors have followed. Convera faces challenges delivering its vision to licensees. Sizzle is not the steak in search.

Author's note: This is an unpublished, preliminary draft of a description originally destined for a client report. The information is provided as part of ArnoldIT's archiving project. The information in this draft may not be used without prior written permission. The information in this document was written before Convera went out of business with the sale of its remaining assets to Vertical Search Works.

Convera ushered in the era of selling "everything plus the kitchen sink" search. The firm was among the first to package search as "concept searching," "knowledge management" and "text analytics", thus kicking off an era of calling search something to capture more revenue. The company's contribution to search was to lay out a road map of where information retrieval would go in the next decade. Convera narrowed its focus to vertical search or eCommerce search. Upon its dissolution, Convera professionals moved to consulting, engineering services, or other search vendors.

This information is a rough draft and is frozen.

Introduction


Excalibur Technologies, backed by the low-profile investment firm Allen & Company, was the precursor of Convera. Based in the Washington, DC area, Convera was formed by Excalibur Technologies combined with Intel's Interactive Media Services division. It is a leading provider of content management solutions that unlock the value of digital content. The value of the deal was hundreds of millions of dollars. Convera serves nearly seven hundred customers in twenty-nine countries from its offices throughout the United States and Europe. Convera customers and partners include ABC News, British Telecom, Digital Island, Encyclopedia Britannica, FOX-SPORTS.com, Microsoft, and the National Basketball Association, among others.

Convera offered a universal search system. Video, structured data, unstructured content, and images could be processed, indexed, and retrieved using Convera's technologies. First Excalibur Technologies and then Convera's marketing team and technical demonstrations drew a compelling picture of seamless information access. Content was, asserted Convera, indexed, classified, and delivered automatically to users and other systems. RetrievalWare was a solution to the information problems organizations face. Unlike Fast Search and Verity, Convera was a comprehensive solution with workflow, semantic technology, and proprietary "smart" software. Steady erosion of revenues began in the early 2000s.¹ In 2007, Convera sold its search technology to Fast Search & Transfer and shifted to marketing its technology to publishers.

Excalibur Technologies and then Convera incorporated document scanning and optical character recognition to convert paper content into digital information. The indexing technology was based on a controlled vocabulary technology purchased from ConQuest Software. Excalibur and Convera's marketers, like Autonomy's and Fast Search's sales professionals, assured licensees that the proprietary systems could perform many functions automatically. Convera stressed that RetrievalWare could process different types of content, including video, support an interface with search suggestions and context-relevant results, and extract signals from metadata. In an admittedly prescient push, Excalibur extended its system to index and make searchable video. In the late 1990s, the capability was a key differentiator for the company. With the sale Convera's remaining assets to Firstlight ERA in 2010, the journey of a marketing-oriented company with technology that lagged behind, drew to a close.

¹. The settlement agreement is at <http://contracts.onecle.com/convera/intel.settle.2003.12.23.shtml>.

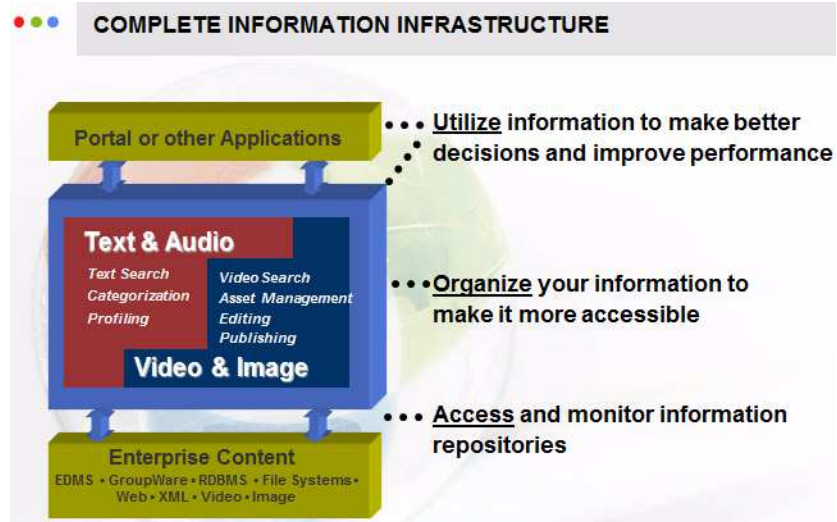
Table 1: Convera: A Bird's Eye View

Product Thumbnail	
1 Search Brand	RetrievalWare. Now incorporated into Vertical Search Works advertiser-oriented services
2 OS Supported	Windows and UNIX
3 Est License Fee	Pricing begins at \$100,000 but special offers and discounts for government agencies can apply
4 Functions	Document imaging support, automatic indexing and document classification, Alleged "real-time content processing and alerting.
5 Claimed Features	APRP or Adaptive Pattern Recognition Processing. Semantic technology to understand content. Ability to handle digital video. Connectors for major file types. Support for structured and unstructured information. Identify subject matter experts from content.
6 Downsides	Performance, particularly for content processing. Need for on-going engineering fixes. Overly complex upgrade processes.
7 Similar To	Autonomy IDOL, Endeca, Fast Search & Transfer
 Product Close Up	Convera's marketing in the early 2000s identified a number of advanced technologies and asserted that these were stable, reliable, and enterprise-ready. For organizations struggling with increasing volumes of digital information, Convera offered a one-stop solution. After installation, the system would address latency, access to multi-language content, integration with other enterprise software and systems, and tools (manual and automated) that would reduce the cost of maintenance and operations. The company had difficulty delivering installations that satisfy the clients' expectations. By 2005, Convera was on a downward trajectory that was difficult to slow.

History

Convera has been one of the leading enterprise search system providers since the company was created in 1995. Like InQuira, today's Convera is the result of combining the former Excalibur Technologies, Inc. with ConQuest Software, Inc. Excalibur Technologies paid \$33 million for ConQuest Software, Inc. In 2002, Convera acquired Semantix Inc., a private software technology development company specializing in cross-lingual processing and computational linguistics technology. Convera's path to growth has been similar to its arch-rival Autonomy's; in fact, Convera and Autonomy have similarly broad enterprise search offerings. Autonomy has differentiated itself by reporting a string of profitable years. Convera has—well, speaking candidly—not performed in an equivalent way. The Convera vision in 2001 was bold and appealed to those who did not understand the cost, complexi-

ties, systems, and staff required to implement what Convera presented as a commercial solution to information woes.



The graphic comes from Convera's senior product marketing manager, John-Henry Gross, circa 2001-2002.

In the height of the Dot Com frenzy, Convera captured headlines when it signed back-to-back deals with two very high-profile companies—Intel in 2000 and the NBA in 2001—for ambitious search-and-retrieval programs. The deals floundered, and the reasons given by those familiar with these now-infamous search programs range from lack of management buy-in to technical shortcomings.

The real reasons are going to be lost in the churn that swirls around many enterprise search initiatives. What remains, however, is a record of sorts. Convera's missteps in 2000 and 2001 cost investors millions of dollars and left Convera saddled with debt.

Today, Allen & Company, a New York investment firm, largely controls the company. In the last 18 months, Convera has made some significant changes. The notable being that the firm hired Claude Vogel, the inventor of Semio, an early visual relationship tool. Convera began an ambitious program to index the Web so that Convera customers could search, text mine, and build custom data sets without having to depend on Google, Microsoft MSN, or Yahoo, or other search engines. Convera's Internet indexing project has been described as a way for a licensee to "create a private-label Internet search system." Early reports are that Convera's index is a way for law enforcement, intelligence, and business research professionals to conduct search and text mining activities without fear that public Internet search vendors can "track" these investigations. Indexing the Internet has broken the financials at most companies in this business.

In fact, only Google, Microsoft, and Yahoo are in the business of “indexing the Internet” in what might be called a “sort of comprehensive” manner. Overlap on text queries across the three services is less than 40%, based on Enterprise Search Reports’ most recent tests. Convera is going to have to invest very large sums to make this a worthwhile information base. Wisely, Convera is said to be concentrating on a small number of “vertical segments,” presumably those of interest to intelligence customers, not the eight billion plus pages of publicly-accessible Web content.

The screenshot displays the Excalibur search engine interface. At the top left is the Excalibur logo. To the right is a search bar with the text "nuclear bomb" and a "Search" button. Above the search bar are options for "Web", ".gov", and ".edu", and a link to "Advanced Search". Below the search bar are options for "Select search type: any, person, date, phone" and "Select media type: text, image, video, audio".

The search results section is titled "Text Results" and shows "Results 1 - 10 of 494,099 for nuclear, bomb (0.83 s)". It includes a "Did you mean:" section with suggestions like "Nuclear attack as a 'Terrorism', bomb as an 'ammunition'", "nuclear weapon as a 'military weapon', bomb as an 'ammunition'", and "Search for all: nuclear bomb".

The main results list includes:

- Effects of a nuclear Bomb**: ...Effects of a nuclear Bomb NAVIGATION MENU Home Software Help Documentation Guest-Book.....or weeks after the bomb. A TEN-KILOTON BOMB DETONATED AT GROUND LEVEL If a bomb in... <http://ram3.chem.sunysb.edu/nucwww/info/nucbomb1.shtml> - 21.1 Kb - Unknown - View Entities Categories
- The Seventh Moon: Nuclear bomb**: ...The Seventh Moon: Nuclear bomb The Seventh Moon This Month's Discussion Topic... ..hill. A nuclear explosion maybe just a symbol of some other dreadful happening Re: Nuclear Bomb Post... <http://tulameen.proboards14.com/index.cgi> - 5.1 Kb - Unknown - View Entities Categories
- nuclear_selden.PDF**: ...nuclear_selden.PDF The Atomic Bomb and the Nuclear Age Mark Selden Cornell University Course Reading The following... http://pawss.hampshire.edu/faculty/curriculum/pdf/nuclear_selden.pdf - 16.5 Kb - Jul 27, 2005 - View Entities Categories
- Nuclear Bomb Pictures**: ...Nuclear Bomb Pictures Nuclear Bomb Pictures Internet Directory | Submit your Site Nuclear Bomb Pictures Nuclear bomb pictures and information... <http://www.sightquest.com/science/nuclear-bomb-pictures-3984.htm> - 14.6 Kb - Unknown - View Entities Categories

On the right side, there are sections for "You may be interested in:", "RELATED PEOPLE AND CONCEPTS" (listing names like Mark Selden, Alan Phillips, etc.), "BLOGS (9281)", and "CONFLICT (3323)".

The newly-redesigned Excalibur interface available in 2006 provides a traditional search box with a number of enhancements. It is the interface used for Convera’s private-label Web search business. These include entity extraction identified as “Related People and Concepts, links to such content as Web logs, related topics such as “Conflict”, and a standard relevance-ranked results list.

In the aftermath of the substantial flow of red ink that washed over the company’s balance sheets in the post-2000 period, Convera today is working to return to profitability. For 2004, Convera reported revenues of \$26.7 million, compared to 2003 revenues of \$29.5. The company reported a net loss of \$19.8 million in 2004, down from the 2003 net loss of \$20.6 million. The 2005 financial data are expected to be equally dismal. The question becomes, “Will Convera survive as a vendor of enterprise search?”

If the company goes out of business, Convera’s aggressive marketing and its willingness to present advanced technologies as commercially hardened are responsible. One contribution Convera made is that the company demonstrated that marketing can sell software which does not work as customers anticipated.

Based on the quarterly reports for 2005, Convera is likely to report revenues in the neighborhood of \$20.0 million with a net loss in fiscal 2005 of about \$5.0 million.

In 2005, Convera cut about 20% of its workforce to focus resources on its Web indexing project. They raised \$29 million in 2005 to hire additional staff and develop a second hosting facility to house its private-label Web indexing service.

The good news is that Convera's management team is making good progress whittling down the debt burden and maintaining the confidence of key investors like Allen & Company. Nevertheless, the financial stability of Convera may be a question to explore before inking a multi-year license for Convera search and text processing technology.

Autonomy, Endeca, and Fast Search have been able to beat Convera in head-to-head competitions in the U.S. Federal government in the last twelve months. Convera's idea to create its own index of the Web was a novel idea when the program kicked into gear in 2005. Autonomy, however, is rumored to be mounting a similar initiative. Fast Search, although not announcing a private-label service, is indexing public Web sites as part of the AllTheWeb.com deal with Yahoo.

Changes in RetrievalWare 2006

Convera provides a full suite of search-and-retrieval products and services. The "guts" of Convera have not changed dramatically since the last major release of the search system in 2005 with Version 8.1.

Enterprise Search Report has been given a glimpse of the enhancements to Convera's core search technology that will be released once beta testing and bug fixes have been completed. Look for Version 8.2 in mid-2006.

The most significant changes for customers of Convera's enterprise search system fall into two categories: overall system enhancements and new utilities.

Key Tweaks and Fixes

Convera's marketing collateral state that the company has made continuous changes and "tweaks" to improve performance of the document processing and query processing subsystems. (Sluggish content processing has been one issue identified by US government entities using RetrievalWare as a search system.) Convera's engineers have recorded some of the document processing subsystem. Other changes have been made to speed up the performance of the user interface, administrative operations, and security subsystem configuration.

One key performance enhancement is implementation of distributed indexing. RetrievalWare now allows a licensee to place indexing servers at differ-

ent content points. If this approach seems like a variant of the Autonomy Topic Server approach, it is similar. Pushing content to a central location for document processing demands huge resources when the volume of content is large. Taking a page from the approach used by Fast Search and Google, Convera now parallelizes the document processing so the load at an indexing point can be distributed across multiple processors. Document processing is disk-intensive, so these parallelized systems need fast access to storage to avoid input-output bottlenecks. These changes are designed to minimize the sluggishness that some RetrievalWare implementations experience.

Second, Convera has expanded its language support. Russian, Korean, and Arabic are supported in Version 8.2. Convera asserts that it can process content in more than fifty languages and permits an English query to retrieve results from content in six widely-used languages.

Convera's clients in the intelligence community need access to scalable systems that can handle the languages "of interest." Convera's objective is to have a way to compete more effectively against Autonomy's language capabilities.

Third, Convera has developed a new spider for content acquisition. One of the notable changes is an improved tool for customizing the behavior of the spider. In addition, Convera additionally provides a code library and sample scripts to help licensees integrate the spider into other third-party applications such as text mining subsystems from Inxight Software, for example.

Finally, Convera has invested considerable time in improving its SDK. The RetrievalWare system has been revised to support Web services. Documentation and sample code have been updated to make it easier for licensees to tune and integrate RetrievalWare functions for specific tasks. A Web Services SDK "bundle" has been created to smooth integration and make customization of browser interfaces easier and faster. A nice touch from Convera was the addition of security extensions to support development of custom document level security. Convera's addition to its security API draws it closer to parity with Autonomy, the leader in search security customization support.

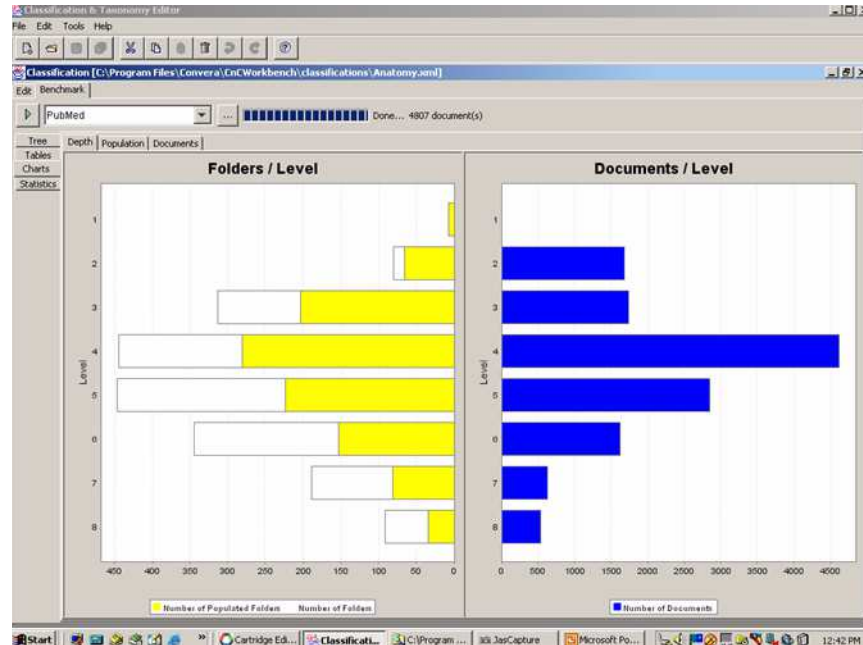
Excalibur Document Management is FileRoom

Convera licenses its records management solution under the FileRoom 2 product name. Because Convera's enterprise search "hooks" directly into FileRoom 2, licensees looking for a fully integrated search and records management solution may want to check out Convera's new version of this product. This software has been optimized and will benefit from the performance improvements that have been a focus of Convera in the last twelve months.

The code base has been rewritten to conform to the J2EE specification. Database support has been modified to enhance stability and compatibility.

Convera enables adaptors to allow any document type housed in FileRoom to be viewed in a Web browser.

From the user's point of view, FileRoom retains the Explorer-like navigation panel, the point-and-click interface, and the search box. However, from the system administrator's point of view, FileRoom 2's functions can be deployed as a thick client, with software running on the user's workstation or a thin client with the code running on a server.



A RetrievalWare graphical administrative tool.

This is an optional document repository module specifically for scanned documents, images and text. FileRoom allows loading, indexing, viewing and managing scanned documents, images and text. Users access FileRoom through a hierarchy consisting of FileRoom documents, where each tier in the hierarchy is a container for storing documents. Users can directly view the scanned image of a retrieved document. Graphs, diagrams, handwritten notations and signatures in the retrieved document are accessible. Document-level security lets organizations control user access at the FileRoom (library), cabinet, drawer, folder and document level.

“Fuzzy” searching capabilities provided by APRP give users some level of confidence that their queries will return the requested information regardless of the quality of Optical Character Recognition (OCR) data. Optical character recognition accuracy is improving. However, errors can make certain data unfindable, however; for example, the misspelling of a person's name or erroneous handling of non-English text within an English language document.

RetrievalWare's Functionality: Reality or Marketing?

Convera's current version of RetrievalWare now matches up better against the product offerings from Autonomy, Endeca, and Fast Search. Mostly under the RetrievalWare brand, the company offers a variety of modules. As always, each comes with its own price tag, and licensees will have to obtain a custom price quotation from Convera to get a specific price for the needed components.

A review of RetrievalWare's core features provides a context for a more in-depth look at how Convera approaches indexing and retrieval, and then some of this complex systems ancillary functions. Most obvious is a guided navigation interface with suggestions for other related content automatically generated by the system once an initial query is launched.

Search-and-Retrieval Services

RetrievalWare offers useful enterprise search technology. Examples of Convera's advanced search features include:

- Indexing functions that allow a user to locate objects of various file types from different repositories, including text, scanned documents, images and video, all from a single search interface.
- Fuzzy search and indexing features that allow a user to obtain relevant search results even when the search terms may be different from, but related to, the original source.
- Document imaging features that support converting legacy paper documents to ASCII that Convera then indexes. A query returns the text "hit" and a link to the source file.
- Built-in categorization service that generates for indexed content a hierarchy of subjects. Users can use this Convera-produced list to browse for information by category in a Yahoo!-style directory.
- A function to identify and associate individual experts on a subject domain within a licensee's organization.
- Alleged "real-time" monitoring, filtering, and profiling tools with a messaging module to notify the appropriate users of new or updated documents on a topic of interest to particular users.
- Image and multimedia processing, search, and retrieval capabilities.
- What the company calls "multi-mode searching," with supports for various types of search, including natural language, Boolean, Yahoo!-style point-and-click listings, and stored queries, as well as combinations of these approaches.

Convera's assertions about semantics, text mining, and automated indexing resonated with many large organizations. The problem was that the computational load these processes impose on infrastructure make them untenable for even large, well-resourced organizations.

RetrievalWare's Advanced Functions

I have become cautious when presented with next-generation search functionality from information retrieval vendors. Convera asserts that it has taken content processing concepts closely associated with university research computing centers and packaged them for enterprise applications. I want to highlight some of these quite sophisticated and little-known capabilities.

Adaptive Pattern Recognition Processing (APRP). According to Convera's marketing collateral, one of RetrievalWare's "core technologies" is APRP or Adaptive Pattern Recognition Processing. APRP makes use of Convera's "semantic network." Incorporating some of the ConQuest Software dictionaries, Convera can perform automated indexing and discern important information from the content processed.

Pattern recognition is a series of recursive algorithms that discover information in a way similar to Autonomy's Bayes-centric IDOL system. RetrievalWare uses it for auto-classification. According to the company, APRP is modeled on the way biological systems use neural networks to process information, acting as "as a self-organizing system that automatically indexes the binary patterns in digital information, creating a pattern-based memory that is self-optimized for the native content of the data." Keep in mind that neural networks and other "smart" technologies require some human intervention to keep the precision and recall without boundaries appropriate to the user's needs.

Once content – including images such as fingerprints, as well as text and databased content – have been indexed, human editorial intervention is theoretically eliminated. Convera's system can automatically generate:

- Topic trees
- Expert rules. Rules must be maintained, which can add to the costs of tuning a Convera system.²
- Sorting and labeling information in database fields

Convera argues that its approach "avoids the inherent subjective biases of categorical indexes." One expects that systems requiring human indexing – such as the Westlaw page annotations for court decisions – make a strong case that systems similar to Convera's generate too many incorrect tags. Which position is correct? Procurement teams will need to test the auto-classification claims of any search system provider. Content volume and type dictate whether an automated system, a manual system, or a hybrid system for building word lists and classification schema is appropriate.

². Rules-based systems like ClearForest work as long as the resources are available to modify or write new rules as content dictates. With too many changes, rules consume programming resources and budget allocations.

Convera presented technical and cost challenges to licensees of the system in the US government. — Stephen E Arnold, consultant to the General Services Administration

Convera's Semantic Network. A “semantic network” is a word list with categories and mapping among words and phrases. (iPhrase makes use of semantic networks, and Microsoft plans to introduce its next-generation search system with an array of mapping tools as well.) The idea is that when a user enters a term, the system can relate that term to other words and phrases. The more sophisticated the mapped links, the richer the query expansion.

A key feature of RetrievalWare is that a user's plain English queries are automatically expanded to include related terms and concepts. Convera argues that RetrievalWare increases the likelihood that relevant content will be returned. An example is that a query for truck would be expanded by the system to include semi and tractor trailer. If the vocabulary in use at an enterprise is not reflected in the semantic dictionaries, these terms can be added manually and mapped to their semantic neighbors. The software recognizes words at the root level, idioms and the multiple meanings of words. This approach can eliminate some of the costs associated with defining keywords, building topic trees, establishing expert rules and sorting/labeling information in database fields.

The baseline semantic network in the English language version was created from dictionaries, thesauri and other reference sources, essentially a built-in knowledge base of approximately 500,000 word meanings, 50,000 language idioms and 1.6 million word associations. These references are based on technology developed in the late 1980s by ConQuest Software that Convera acquired for \$33 million in 1995 when Convera was doing business as Excalibur Technologies. Oracle provides a similar suite of word and phrase resources. Most other search system providers use a version of the word list produced by Princeton University.

For certain applications, semantic networks are pivotal to discovering information that would be missed unless the mappings were used. On the other hand, unnecessary query expansion, particularly when terms have multiple meanings depending on the context, will generate too many hits. When a user wants computer terminals and the system retrieves airline terminals and ship terminals, the semantic network must be constrained to a word list for a vertical market.

RetrievalWare, like Verity, supports domain-specific semantic networks for specific fields of interest. RetrievalWare includes dictionaries for biology, chemistry, computers, electronics, finance, food science, geography, geology, health sciences, information science, law, mathematics, and the MeSH (medical subject headings), military, petroleum, natural gas and petrochemicals, pharmaceuticals, pharmacology, physics, plastics, rubber, and telecommunications.

These lists can be expanded and edited. This sort of manual interaction with these word lists – usually essential – is supported in a manner similar to the process you would use in Verity to refine its dictionaries and word lists.

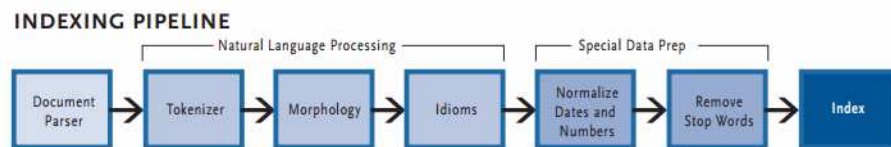
Work Flow Operations. RetrievalWare, according to the company, can perform work flow operations like profiling.

RetrievalWare Profiling, the company asserts, can automatically detect, route and store relevant documents in user-defined profiles, potentially accelerating the timely discovery of relevant information as it enters the RetrievalWare environment. The “hits” matching the profile are provided via e-mail, a Web page, or other means to the interested individual(s).

In addition, a system administrator can configure profiles. The updates or alerts deliver the results of a personalized query transferred to a personal or collaborative folder, along with optional alerts. Automatically providing search results to users is a good idea as long as the users maintain their profile. If users allow profiles to operate without updating, the content becomes less and less useful. Users often undertake new tasks and responsibilities.

One of the reasons for the failure of PointCast, BackWeb, and Desktop Data was the fact that alerts can flood the users’ email in box. The cost of maintaining alerts can be significant if users do not maintain their profiles.

Categorization and Dynamic Classification. RetrievalWare’s Categorization and Dynamic Classification module supports the organization and access of information assets through the use of industry-standard taxonomies. RetrievalWare uses one or more taxonomies to extract concepts and context from information assets. These assets can then be organized into specific views that reflect the personalized knowledge requirements, roles and perspectives of each user.



Convera’s content processing pipeline contains seven functions plus key word tagging. Source: RetrievalWare 8 by Alkis Papadopoulos and Jon Van Winkle, no date.

Cartridges

Convera has embraced the “cartridge” terminology and the engineering concept.³ A Convera licensee can extend the basic RetrievalWare system by licensing special purpose software components (cartridges) for a fee.

Convera’s cartridges include:

³ Illustra, acquired by IBM, allowed licensees to “plug in” software components to add functionality to the database system.

Convera uses dynamic classification to expand individual and intuitive search processing. — Mushtaq Khan, vice president, Convera, June 18, 2003 at the NMCI Industry Forum

Metadata and Text Mining. Convera’s engineers have been leaders in integrating text mining functions with more traditional search and retrieval. Convera’s system generates a traditional index of words and phrases, and it also attaches metadata to each document record. Looking up related entries in word lists generates some of the metadata. Convera provides to licensees “seed lists” or “controlled vocabularies” created for a variety of subjects, i.e. law enforcement, financial services, and pharmaceuticals, for example. With standard indexes and additional metatags, the Convera system can display a standard list of results and “suggest” related topics, generate a visual display of the results using the Semio technology developed by Carl Vogel, Convera’s chief technology officer, and generate a variety of reports about usage, word and phrase frequency, and exceptions in a data set, among others.

Classification Workbench Cartridge. Convera’s Cartridge & Classification Workbench enables the use of manual and automated tools to streamline taxonomy classification development, benchmarking, and deployment. These tools can reduce taxonomy development and deployment times as well as maintenance costs.

Language, Domain and Taxonomy Cartridges. RetrievalWare provides search and categorization results based upon its linguistic processing capability. Through the use of semantic networks – that is, lists of related terms – and taxonomies that cover multiple languages and domain-specific fields of interest, RetrievalWare recognizes and processes words, phrases, and concepts in the context in which they exist. These cartridges are available as pre-packaged optional components to RetrievalWare. Convera also provides development tools that allow customers to customize cartridge content for specific business solutions.

Search Features

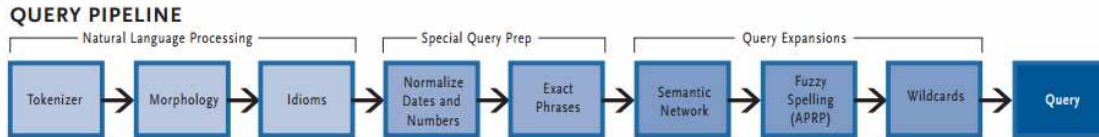
Fuzzy Searching

Convera asserts that its system permits fuzzy searching (relaxing the user’s query to ensure hits in a results list) and morphological operations such as truncation.

Convera’s implementation provides licensees the ability to retrieve approximations of search queries. Like other fuzzy search implementations, Convera accommodates misspellings by users and because of its ability to index binary image files, it can adjust to errors in source documents when indexable text is generated from optical character recognition (OCR) or handwriting recognition programs.

This theoretically reduces the need for OCR clean up (especially useful in applications that handle large volumes of scanned documents). However, selection teams anticipating heavy use of this capability will want to test it

out on their own documents, and there are other valid reasons for wanting OCR clean up beyond search.



Convera's query pipeline includes nine separate operations. The computational load can stretch even the most robust infrastructure. Source: RetrievalWare 8 by Alkis Papadopoulos and Jon Van Winkle, no date.

Convera recognizes words at the root level, which, according to the company, provides “a much more accurate approach than the simple stemming techniques characteristic of other text retrieval software.” Convera links its morphology function with its fuzzy search module to minimize missing words due to irregular or variant spellings.

In addition, Convera recognizes idioms. Like iPhrase and Endeca, the system matches terms against the dictionaries. If the terms are available, the Convera operation delivers useful results. If the terms are not in the dictionaries, subject matter experts must “map” the new terms to the words and phrases in the dictionary.

Cross Language Support –

Like FAST Search & Transfer and Autonomy, RetrievalWare offers cross-lingual options. Users can enter a query in one language and receive conceptually relevant results from documents in other languages. The key word here is “conceptually.” As always, if this is important, test first.

Connectors

Convera asserts that it supports more than 200 document formats. These range from XyWrite (word processing system used by publishers) to Microsoft Word and hundreds of file types from dozens of software products. For example, Convera suggests that it can process content from most popular word processors, e-mail, document and content management systems such as FileNet and Documentum, spreadsheets, Adobe PDF, relational databases, HTML, SGML, ASCII, and more. RetrievalWare synchronizers that recognize when repositories or files have been modified and update the RetrievalWare system manage access to remote document repository and groupware systems.

In addition, Convera provides “adaptors” to make federated searching possible across structured and unstructured information. An adaptor allows RetrievalWare “synchronizers” to provide access multiple native repositories of content from a single point of access. Supported repositories include Lotus Notes, Microsoft Exchange, Documentum, FileNET, Panagon, native

file systems and major relational database management systems including Microsoft SQL Server, Oracle, DB2, Sybase, Informix, Teradata and any ODBC-compliant database. Convera does not explain how the latency of various accessed systems can be overcome to deliver results to a single user in a timely manner. Federated search is a work in progress at companies like Deep Web Technologies and Vivisimo, among others.

Video as a Content Type

Convera is one of a small number of companies asserting that its technology can handle video and provide search and access to a user. Video poses a number of challenges. The principal issue is related to the size of video files. Secondary issues relate to obtaining metadata about a video and indexing the audio track to make full-text searching possible. Early video services cannot provide robust search functionality due to technical constraints and processing costs.



Licensed to the National Association of Broadcasters, Convera's video search "screening room" accepts the users query and displays hits as thumbnail images from digitized video. According to the company, the system "watches" the video and identifies key scenes. The system uses available metadata and close captioning information, if available, to index the system. A mechanism to convert the audio in the video to ASCII text is in development.

Convera offers a specialized system to handle rich media.

Used in conjunction with RetrievalWare Search, it provides for real-time capturing, encoding, analyzing, cataloging, browsing, searching and retrieving of video content, as well as related captured text (closed captions or speech-to-text conversions) and metadata, over corporate intranets/extranets.

For end-users, the product automatically creates a video storyboard, so that searchers can browse, search, and retrieve specific video clips – then play it back in any standard video file formats – without having to watch the video in its entirety.

The product consists of four components:

- Screening Room Capture
- Screening Room Metadata Edit
- Screening Room Explorer
- Screening Room Video Asset Server

Screening Room Capture ingests, analyzes and storyboards analog or digital video assets, including live feeds and extracts. It searches for associated metadata such as captured text (both closed-caption text and spoken audio content converted to text) and key frame images of significant scenes and annotations.

Screening Room Metadata Edit enables users to browse, search, edit and annotate storyboards. In addition, users can select and compile clips from multiple video assets to create new derivative works, export files and metadata in XML format, or output rough-cut edit segments for import into offline editing systems.

Screening Room Explorer allows user access to catalogs of video assets through a standard Web browser.

The Video Asset Server indexes and stores captured video assets for browsing, or search and retrieval via RetrievalWare.

Visual RetrievalWare is a visual retrieval engine, an image processing library, and programmer's toolkit that enables the development of systems that index and retrieve digital images. Users can search for visual information directly from their intranet, a corporate database, the Internet, or other sources using images or video clips as clues.

Visual data is reduced to a searchable index that can be as little as 10 percent of the size of the original image. Users submit queries using examples of visual data or by authoring a visual clue with a graphical product. Based on the shape, color and texture of the visual clue, a list of similar or exact matches is returned.

Excalibur has developed specific RetrievalWare image application demonstrations for fingerprint, face and character recognition.

For example, Convera's Fingerprint Server toolkit provides an environment for creating automated, pattern recognition-based fingerprint filing systems. The fingerprint toolkit includes components for fingerprint image enhancement, feature extraction, indexing and matching, as well as components for associating feature indexes with fingerprint card data. The system includes special algorithms to cope with low-quality images.



Convera's system supported a "more like this" function for image retrieval. The user identifies a face, and the system locates similar faces.

Convera provides a scripting language to define feature extraction functions specifically for identifying fingerprint directional features, minutiae features and focal information.

The company offers a stand alone version of some of its visual search tools. These tools provide a licensee with the ability to log, analyze, encode video, and save the data and video assets in a non-proprietary (XML) format. Screening Room Capture does not require purchase of the entire RetrievalWare or Screening Room system, enabling loading of video assets and meta-data into a third-party database or content management system, or otherwise re-purposing the asset.

Screening Room Capture is also a suitable component for sale to licensees who wish to embed RetrievalWare functionality into third-party software products.

Technical Architecture

The architecture used for RetrievalWare is what the company calls “distributed process architecture.” In simple terms – like Verity – RetrievalWare uses a distributed architecture, avoiding a centralized indexing and document repository scheme. Convera breaks its solution into various modular components that can be distributed across different servers working in parallel when assembling a global search system. The core technology runs as a J2EE application, on either Windows or Unix.

The servers includes a content acquisition subsystem, query processing subsystem, an indexing server, security server, pattern and profiling servers, image server, and optionally, a Web server, as well as separate administrative tools.

The Convera “Text Server” contains a pipeline of indexing, query and display processing modules. These components snap in so that the Convera installation can run on a single machine, on different CPUs in a single server, or on any machine in a network of server computers.

The Pattern Server contains Convera’s build of its APRP, statistical and Boolean searching techniques. The pattern server is language independent, enabling development of multi-language text retrieval applications.

The RetrievalWare Profiling Server is a system for filtering newswires, electronic mail messages, file transfers, and other dynamic information streams in real-time. Its design allows the licensee to integrate retrospective searching and real-time content profiling.

The RetrievalWare Image Server is essentially a suite of tools for licensees who want to build image retrieval applications. RetrievalWare includes components for indexing and retrieving digital images based on their objective content. These components enable pattern recognition-based image retrieval applications that automatically recognize certain types of visual information and provide additional image management capabilities. Qualified programmers can optimize their image indexing and retrieval applications for a variety of specific image data types.

Convera’s optional Web Server is a dedicated front-end server for handling large volumes of user queries. Among other things, it allows licensees to use Convera’s own macro language to easily customize query and results interfaces. Selection teams will want to carefully judge whether Convera’s Web server is an improvement over an existing standard Web server in this regard – the alternative is to use a J2EE web application as the front end for both query and results templates (something Convera does on its own site).

Convera offers security features as well. Convera supports cross-repository security that manages login data behind the scenes to all indexed repositories. Utilizing either library level or document level security, users can only access the files and documents that they are authorized to access.

To tap into the Convera architecture, a licensee uses the RetrievalWare SDK. The RetrievalWare SDK (Software Developer's Kit) is a comprehensive set of tools for building advanced search-based solutions. At its core is a scalable, distributed client-server architecture. Independent server processes help maximize the efficiency and reliability of document loading, indexing and query handling, and support security and encryption/decryption features. Dedicated server processes enable integration of text search and relational database storage capabilities through an open database management system ("DBMS") gateway. The client environment is optimized for the development of graphical interfaces using industry standard tools such as Java and Visual Basic.

RetrievalWare delivers Visual Basic custom controls, remote procedure calls and open server capabilities, as well as engine-level, high-level and client-server application program interfaces ("APIs"). These features help reduce development time for search systems with custom functionality.

The RetrievalWare SDK is an application development environment that includes access to more than 50 APIs. Those APIs include:

- High-level APIs. These are designed specifically for speeding development of user interfaces using GUI building tools such as Visual Basic. Convera provides Visual Basic Custom Controls (VBXs), which perform the graphical functions in the GUI with a single callable module.
- Remote Procedure Calls. RPCs can be used to design bandwidth-conserving communication links with asynchronous processing. The approach makes it possible for RetrievalWare to support thousands of simultaneous users.
- Engine-level APIs. These functions allow integrators to provide customized search functions using RetrievalWare's query structures. A developer can embed RetrievalWare search, indexing, and stored queries into work flow engines, enterprise applications, or content management systems.
- Open server design. Convera provides licensees with tools and sample scripts to allow a developer to add customized features. Convera's server processes enable integration of text search and relational database storage capabilities through an open RDBMS gateway.

ArnoldIT Opinion

On paper, Convera offers search and related functions that few vendors can match. In my experience, the catalog of next-generation functions lays out a road map for enterprise information access. The computing capabilities available in 2006, however, are beyond the reach of most organizations. For those organizations such as the US Department of Defense, the cost of maintaining a Convera system is likely to be high. Convera has landed some big deals based on its search and video capabilities, but only time will tell if the tie ups with Intel and the National Basketball Association will result in successful and financially sound businesses. My view is that most licensees do not understand the technical complexities and costs associated with next-generation information retrieval. I am, therefore, inclined to view some of Convera's assertions with skepticism.

Table 2: Convera Checklist

Attribute	Convera Asserts	ArnoldIT Comment
1 Platform	Windows, UNIX, Linux via a distributed architecture	System is partially distributed
2 Key word search	Supported. Free-text and Boolean	Fuzzy or relaxed search is also supported
3 Text mining	Entity extraction included	Comprehensive text mining is difficult to implement without restricting the volume of content to be processed and trained personnel
4 Automated indexing	Free text indexing. Stop words supported. Bound phrases and synonyms require a dictionary.	The dictionaries require manual maintenance to deal with neologisms, acronyms, and proper nouns
5 Personalization	Certain search operations can make use of a user's profile; e.g., alerts	Profiles require on-going maintenance by the user or other professionals
6 Workflow	Primitive work flow is provided; e.g., routing of content via matching content to a profile	Sophisticated check in, check out, and certain security operations must be coded and integrated via the Convera API
7 Interface	Default interfaces are provided	Customization of the interface is possible. Some functions require original scripting
8 Hosted service	Alleged to be available.	The Intel project is focused on providing Convera functionality from the cloud
9 Administrative interface and tools	Some.	Command line expertise is helpful. Ability to write scripts and code essential.
10 Application programming interface	Yes	
11 Professional services	Available on request	
12 Security	Security controls must be set up via the API	Programming required

Attribute	Convera Asserts	ArnoldIT Comment
13 Connectors	Company asserts that it supports more than 200 file types	If additional connectors are required, third-party tools or custom programming is required
14 Support for structured data	Convera says it supports structured content	
15 Relevance ranking	The licensee can tune relevance via controls or pages containing boosted content	
16 Video	Converse asserts that it can index and manage video content	Video adds another dimension of complexity to a search system
17 Federated search	Convera asserts that it can support federated (metasearch)	Latency is an issue which can be expensive to resolve
18 Fielded search	Convera asserts that its system can search structured data once it is processed	
19 Content crawler	Provided for Intranet and Internet content	
20 Price	Begins at \$150,000	Pricing is comparable to Autonomy's fee for the IDOL system.

Anticipated Benefits

Convera's technology and architecture compare favorably to other enterprise search solutions from TeraTech, Verity, and Autonomy. Specialized applications such as those for law enforcement and intelligence are robust and in many ways set the standard for those user communities.

For broader enterprise search-and-retrieval applications, Convera's system can be tailored to handle departmental or global enterprise search. However, Convera, however, may be better suited for complex, distributed search applications, rather than small-scale search-and-retrieval requirements.

Other benefits of the Convera approach include:

Most standard enterprise search-and-retrieval features are supported by the firm's software.

- The product can be extended to interact with many other systems generating text, database, and – most notably – non-text information such as images and video. Document imaging capabilities are a differentiator.
- Strong financial support from Allen & Company (an investment bank) and from the U.S. Federal government provides some assurance that the firm will not go out of business, although another firm could purchase Convera in the next 12 to 24 months

Possible Drawbacks

Convera has a broad range of sophisticated products and technologies. The company's financial situation has stabilized somewhat, but there have been

reported issues in resolving certain technical issues such as stability, performance, and computing resource requirements. The recent downsizing of the company coupled with the new initiative to index the Internet may tax available technical resources at the firm.

RetrievalWare's performance challenges are significant. Scanning and processing documents consumes computational time and bandwidth. Issues related to permissions to change a document require manually coding of security processes. The technical challenges of providing access to indexes and displaying source documents raised high financial and technical hurdles for licensees. Routine tasks like replication management required manual intervention and work arounds. Replication consistency required human oversight because automated methods were not reliable. Solutions based on in memory caching were expensive and difficult to stabilize.

The drawbacks of the RetrievalWare product line include:

- Financial performance and the high-profile "issues" with Intel and the NBA deal could be viewed as an indication that the firm's marketing hyperbole is not matched by the performance of the RetrievalWare software.
- The product requires a substantial financial, technical, and infrastructure commitment. RetrievalWare is not available in a "lite" version like Verity Ultraseek.
- The costs associated with the product can be a barrier in some commercial enterprises. Government implementations, particularly those entailing national security, have somewhat different standards by which to measure investments in hardware, programming, support, and customizing systems than a for-profit business.

Conclusions

The financial health of Convera remains an issue. The firm's sweeping assertions regarding knowledge management, text mining, and comprehensive video management and search functions are rumored to be expensive to implement and difficult to deploy.

The firm's recent Web indexing initiative can be interpreted in different ways, depending on the point of view of the observer. From a product point of view, creating and index of the Internet to allow a licensee to search and text mine from the Convera Internet index provides an alternative to the use of Google's index and ensures confidential and security spelled out in contract between Convera and its customers. From a financial point of view, indexing the Internet is a very expensive proposition. With Microsoft and Yahoo looking for shortcuts via "social software" and user applied index terms or "tags", Convera is embarking on an indexing project of considerable complexity. More importantly, it shifts Convera from a software com-

pany to an information provider company. The question I ask is, “Is Convera scrambling to make sales in markets different from enterprise search?”

Management has to demonstrate that it can bring a successful product to market and control costs. From the customer perspective, intelligence and law enforcement officials are likely to be early adopters of Convera’s Web index. It is not clear, however, if Convera or another company will benefit from the need to de-duplicate, text mine, and integrate the new Convera Internet content into other information systems. The bigger payoff may be for the services firms providing these data post-processing and consulting services. From the shareholders point of view, Convera must generate profits and grow at a rate comparable to that of other companies in this sector; namely, Autonomy and Fast Search.

Convera has been bypassed in terms of revenue by Autonomy Corporation and FAST Search & Transfer. Although Endeca is a privately-held company, Enterprise Search Report has been informed by those close to Endeca that Convera is smaller than Endeca in terms of revenue and new customer acquisition. If true, Convera is not likely to capture significant new customers unless it can quickly address these major issues:

- 1** Unify its marketing message about its approach to enterprise search. Convera does not have a story to tell comparable to Autonomy’s integration platform IDOL or Fast Search’s pegging search as an application platform.
- 2** Generate significant new revenue. With Autonomy on track to hit \$200 million in annual revenues in the next twelve to eighteen months and Fast Search already generating more than \$100 million in annual revenue, Convera’s top line revenue trails by tens of millions of dollars. Its history of big losses have yet to be erased from its balance sheets.
- 3** Create some buzz. The biggest news from Convera in the last few years has been the reliance of top management on Claude Vogel, a recognized expert in semantic-linguist text retrieval. The “private label Internet” product initiative remains fuzzy. Early customers are reported to be pharmaceutical companies and the U.S. intelligence agencies, but for a project that required investment of tens of millions of dollars in the last twelve months, a broader customer base must be generated.
- 4** Pull ahead of other competitors. Convera must maintain a technical lead over a number of aggressive competitors, including and certainly not limited to Coveo, Endeca, Mondosoft, and

Odyssey ISYS as well as the looming threat of integrated search systems from IBM, Microsoft, and Oracle.

These are, in my opinion, very tough challenges. No search firm has been able to implement the broad range of features in a way that is cost-effective and sufficiently robust to keep pace with the rising tide of digital content. Convera's marketing vision is commendable. Can the company deliver? At this time, Convera's technology may be falling short of client expectations.

It is well to keep in mind that Convera has a number of blue-chip accounts; however, the company's support by the U.S. government and general push for government business may indicate that commercial clients are taking a back seat.

Important questions to ask are:

- What is Convera's track record for delivering on time and on budget solutions?
- Will the firm's assertions that it offers a video search and management solution dilute the firm's ability to invest in its core architecture?
- What is the cost of the manual effort required to maintain dictionaries so that Convera's automated pattern recognition, semantic processes, and multi-lingual functions deliver 80% to 90% accuracy?
- How will the continuing losses Convera and those accumulated in previous years (at least on paper) impact future operations?

Convera appears to be a forward-looking search vendor on track for a financial collapse.

Stephen E Arnold

Minor edits to a rough draft on October 1, 2013