Stephen E. Arnold
Arnoldit.com

The New Landscape of Enterprise Search

A Critical Review of the Market and Search Systems from Autonomy, Endeca, Exalead, Google, Microsoft, and Vivisimo

Pandia

Stephen E. Arnold
Arnoldit.com

The New Landscape of Enterprise Search

A Critical Review of the Market and Search Systems

Pandia

A Pandia Report

The New Landscape of Enterprise Search: A Critical Review of the Market and Search Systems from Autonomy, Endeca, Exalead, Google, Microsoft, and Vivisimo

by Stephen E Arnold, ArnoldIT.com

Published in Norway by Pandia Search Central, Ruselokkveien 59b, 0251 Oslo, Norway

Email: koch@pandia.com
Web site: http://www.pandia.com

Book design by Maxwell Tyson

ISBN: 978-82-998676-0-3

About Pandia

Pandia Search Central is a site dedicated to search engines, search engine optimization (SEO), and Internet searching. The site offers collections of tools that will help you search the Net better as well as a significant amount of information for Webmasters trying to achieve higher search engine rankings for their sites. There is also a Web log with search engine news, weekly wrap-ups for search engine and search engine marketing news, and in-depth analysis. This work has been in progress since September 1, 1998.

Table of Contents

Preface

More than five years have passed since I wrote the third edition of the Enterprise Search Report. In 2008, I did two monographs about enterprise search. The first was *Beyond Search: What to do When Your Enterprise Search System Doesn't Work*. A few months later, Martin White and I collaborated on *Successful Enterprise Search Management*.

After finishing these monographs, I had grown tired of enterprise search. The systems, regardless of the vendor, were generating significant dissatisfaction among their users. As you will learn, most of the mainstream vendors' search technology works quite well. The caveat is that the licensee must know what the requirements are and have the appropriate resources available.

At the end of 2010, I completed a project that forced me to revisit the 50 vendors I track in my work at ArnoldIT.com and in my blog, Beyond Search. What became clear to me was that the search systems most often discussed in large organizations were not well understood.

Part of the reason was some assumptions made by the search procurement team or procurement manager about search. The widespread familiarity of Google contributed to a perception that enterprise search should work "just like Google." The problem, of course, is that a document in an organization may be accessed once when it is created and then a few times, if ever, when a very specific item of information from it is required. On the Web, search is a game of numbers. Clicks and links determine relevance. In a mobile search, the user's location adds additional context. In an organization, different access methods are needed.

The collision of users' perceptions and the realities of a mainstream enterprise search system from one of the vendors discussed in this report produces significant dissatisfaction with the incumbent search system. No vendor is exempt from this backlash. Users expect a Bing decision engine or Google personalized search result. Enterprise systems deliver information in the form of information injected into a work flow or in another application, or a Boolean keyword result list. Users howl, so the organization begins the hunt for another enterprise search solution.

In many cases, search vendors position themselves to handle this type of rebound or to make a sale when an organization needs to solve a very specific information retrieval problem in customer support, legal matters, or business development.

The result is significant expenditure of effort to figure out which system does what. On close inspection, the mainstream vendors provide platforms upon which almost any type of application can be built.

The scope of this report is narrow. I have focused on the enterprise search technology available from six vendors. Each competes in large organization search procurements. Unlike the 2006 *Enterprise Search Report* which covered 25 vendors and *Beyond Search* which covered 24 vendors, this report focuses on:

- Autonomy Corp. The firm evokes strong reactions from consultants and other vendors. But the fact of the matter is that the firm is likely to break $1.0 billion in sales in the next six to 12 months. Keep in mind that Autonomy has more than 20,000 licensees.
- Endeca. Endeca is an important company. The firm has been in business for more than 12 years, and it has done a good job of marketing its system to high-profile organizations worldwide.
- Exalead. Exalead has one of the most sophisticated information retrieval systems in my opinion. Now the firm is a unit of the highly regarded Dassault Systèmes. With additional sales and technical resources, Exalead is a pioneer in search-based applications and increasingly disruptive.
- Google. The Google Search Appliance generates a warm glow among procurement teams. There are an estimated 35,000 search appliances deployed around the world. More importantly, Google is the "elephant in the room" whenever search is discussed within an organization.
- Microsoft Fast. Microsoft purchased Fast Search & Transfer and now positions the system as the optimal way to search SharePoint content. A close look at the Fast Enterprise Search Platform is needed because many consultants are unaware of or indifferent to the technology of the Fast system.
- Vivisimo. Vivisimo began its commercial life as a vendor of metasearch or federated search. In the last year or so, Vivisimo has been competing as an enterprise search solution.

I track a large number of companies providing search and content processing solutions. I made a conscious decision in this monograph to exclude some search vendors about which there is considerable interest. The principal criticism I heard about my previous studies of search vendors was that the amount of information was too great. Accordingly, I will be discussing some enterprise search applications and their respective vendors in additional "landscape" monographs:

- Open source search solutions.
- eDiscovery content processing system vendors
- Metatagging and semantic search systems

I have included a two-page table that provides brief descriptions of other vendors of enterprise search systems. If one of the vendors in which you are interested is not included, you have enough information in the table to begin your own investigation.

In addition to the writeups about each vendor and the two-page listing of other search vendors, I have included a glossary at the end of this document. The glossary contains entries from my previous monographs, including my Google trilogy. However, I have updated it to reflect some of the new buzzwords; for example, mash up.

If you have comments or criticisms, you may write me: seaky2000 at yahoo dot com.

Stephen E Arnold, June 1, 2011
Harrod's Creek, Kentucky 40027 USA

Chapter 1:

Introduction

Search has become a commodity. Vendors describe features, functions, and services that are like catnip to prospects and promise a solution to a problem traditional information retrieval cannot resolve.

Widely-respected consultants have been hard hit by the economic downturn. As a result, their independent opinions are often little more than Madison Avenue-inspired advertisements.

The Landscape

Enterprise managers shopping for an enterprise search system should feel like kids in a toy store: So much on the shelves, each system fancier and shinier than the next. But the trouble starts as soon as the chosen toy has been unwrapped back home. The glib words on the package meet the harsh complexities of reality in the form of unique organizational requirements, proliferation of repositories, and so on.

This report cannot attempt to address implementation challenges. It is stressed that anyone seeking to purchase a search system should be very aware of the devilish complications that will likely arise throughout the planning and deployment stages.

In describing the landscape of enterprise search, there are two challenges. First, exactly what is meant when a procurement team seeks to license an "enterprise search system"? Search means different things in different organizational contexts. Individual employees often have specific views of search and how an enterprise search system should look, behave, and perform. The notion of "enterprise" is fuzzy. Search is required in government agencies, educational institutions, charities, and businesses large and small. Getting agreement on what search is and what an enterprise search system should do within a very specific organization for users who may have quite particular needs is a very difficult job. The challenge of defin-

ing terms hamstrings the buyer and the seller. Those with deep experience in search and retrieval often make the definition of terms the first step.

Second, it's a challenge to figure out what the system should do. Armed with that information, the potential buyer can then try to match the requirements to vendors' offerings. The vendor, when presented with a list of requirements, works to demonstrate how its system meets the requirements. The problem of course is that words describing search are not the same as using the system in day-to-day business. Not surprisingly, the vendor's system often triggers push back from the licensee. The licensee may not know what is needed until the system is deployed. The vendor then learns what the licensee really meant in the system requirements document.

The landscape of search is full of vendors with technology that can, to some degree, be shaped to the needs of the buyer. Buyers are usually frustrated with incumbent search functions and eager to solve the organization's information access problem. Most organizations have multiple search-and-retrieval systems. A system may be built into Microsoft SharePoint. Enterprise applications like content management or customer support systems typically provide search. Some departments have acquired specialized search systems. In chemical and pharmaceutical companies, research units may license specialized systems to search chemical structure information. In the business development unit, a manager may have leased a Google Search Appliance or installed a department search system from Coveo, ISYS Search Software, or Fabasoft Mindbreeze. If the firm has an information center, there may be a search system providing access to books, monographs, and other third party content. And so on.

The universe of potential candidates begins to look like a large, self-similar blob.

A handful of vendors have relatively well known brands and identities . Other vendors have names that are unfamiliar to most employees in an organization and even some search experts advising that organization.

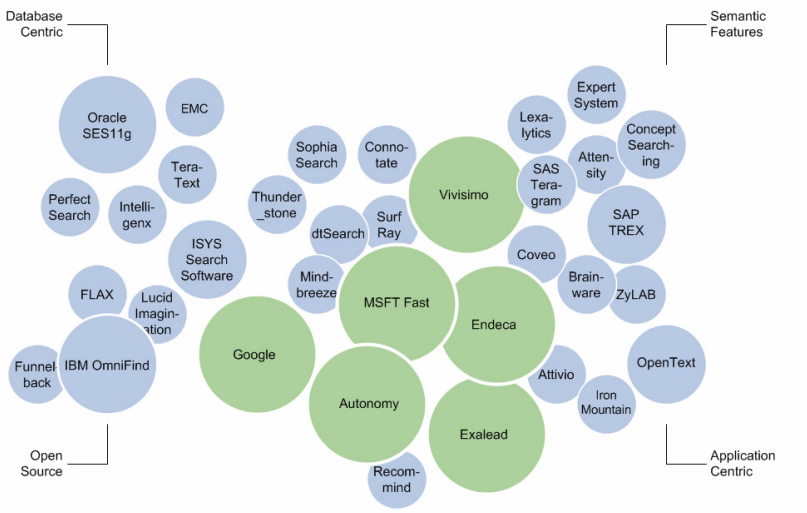
The landscape is rich, large, and unexplored by most information technology professionals and procurement teams charged with finding a solution to an organization's information access or findability problem.

Enterprise Search: The Upside and Downside

Organizations cannot do work unless employees can find information in electronic form. Search, therefore, has become the starting point for millions of workers each day. For the most part, employees can locate needed information. But since the first online systems became available decades ago, a combination of methods are necessary.

Consultants estimate the number of hours employees invest when seeking a fact, figure, document draft, or a PowerPoint presentation. The most common cause of an information foraging exercise is that organizations have a great deal of digital

content. Estimates of the volume are staggering, and it is difficult to know which stupendous estimate of content production to believe.



This diagram provides a representation of the congestion within the information retrieval sector. Each organization competes for enterprise search business. Differentiating among vendors is difficult even for those with experience in enterprise search. The green tint identifies vendors discussed in this report. No accurate representation of enterprise search systems is possible. Vendors reposition themselves to respond to market opportunities. Confusion is the "new normal".

The team working on this report knows that in most of our clients' organizations, storage space must be doubled every four to six months. Some organizations will experience faster data volume growth than others. One fact is incontrovertible. With more data, enterprise search systems are stretched to the limits of the software's capabilities and the infrastructure's capacity.

In our interviews with users of enterprise search systems, we have documented a high level of dissatisfaction, but frustration may be a better word. Documents cannot be found because the indexing cycles were longer than the system's users believed them to be. Users often have to use multiple systems to make certain that the best and final version of the document is retrieved. An even more difficult task is to find a previous version of a document (which may involve retrieving information from a backup system.)

A buttoned-up organization may have a general purpose search system as well as two or more specialized content processing systems. Attorneys have specific requirements and may turn to products that are tailored to the needs of a legal process. Product managers often require systems that provide an overview of textual and numeric data. Vendors use different jargon to explain that a particular system can produce informative, yet eye-catching, outputs. These systems often carry descriptions that suggest "business intelligence" or "analytics" along with the search-and-retrieval features. Employees working in a customer contact or call cen-

ter have search-and-retrieval requirements that are subtly different from the requirements of a user in human resources. On the surface, looking for the answer to a technical question for a customer with an extra-cost support service is similar to a person looking for a resume for a good compensation and benefits manager, but beneath the surface, it's not so simple.

The upside of today's systems is that there is a solution for almost every enterprise search-and-retrieval requirement. Over the last five years, the mainstream enterprise search vendors have improved the usability, performance, and functionality of their respective systems.

When any one of the systems developed by the vendors profiled in this report is set up correctly, it works quite well. Each vendor profiled in this report has a list of reference customers who will back up the vendor's claim for cost savings and improved efficiency. Differences between Web Search and Enterprise Search (Selected)

Feature	Web Search	Enterprise Search	Comment
Architecture	Cloud-based	On-premises, cloud, and hybrid options	An enterprise may require or change architecture arbitrarily
Click volume indicates popularity	High traffic provides a reliable indicator of document importance for most users	Document importance usually unrelated to the number of clicks on a document	Different methods required to relevance rank Intranet content
Comprehensiveness	Service determines comprehensiveness	Comprehensive is defined and a requirement	Certain content must be in the index. If a document is not available, human intervention may be required
Content	Publicly accessible sources	Private content and maybe some public content	The mix of content is spelled out in the requirements. Special licensing with copyright holders may be required
Content types	Text and rich media	Standard file types, enterprise software file types, and possibly rich media	Public Web search systems do not have connectors for Lotus Notes, SAP R/3, and similar content sources
Crawler	Can skip or prioritize certain sources or Web sites	Content must be processed to meet organization's requirement	User expectation for an Intranet is that his / her most recent document is in the index
Database access	No expectation for database content; user takes what is available	Structured and unstructured content required	Users expect access to data in a database or third-party system that uses a database as a repository
Index freshness	User takes what is available from the vendor	Index freshness spelled out as a requirement	Enterprise search indexes must be sufficiently fresh to keep prices, product, and customer information current
Indexing	Web index may exclude content arbitrarily	The enterprise index has content that must be included	A Web index can skip a site that is slow
Interface	Search box and graphic-rich presentations	Customization required	Licensees may require multiple interfaces linked to user roles
Personalization	Driven by advertising considerations	Driven by business requirement	Specific personalization functions may be required for different use cases

Security	Most queries are not secure	System is secure; off site access requires authentication	Enterprise search uses available security flags and may offer additional security features
User expectation	Defined by Bing and Google	Enterprise search will work like Bing and Google	Enterprise search systems must address licensee requirements

The downside of today's systems is that licensees have to know what specific functions the search-and-retrieval system is to perform. Furthermore, licensees have to know what content the system is to process and how frequently the indexes of that content must be updated. In general, more frequent indexing requires a more robust content processing system and sufficient bandwidth to identify, copy, and process the source material. Licensees must provide adequate resources so that the enterprise search system can operate in a satisfactory manner. Starve the search-and-retrieval system for bandwidth, memory, and CPU horsepower, and the search system will not perform at its optimum level.

Most organizations today have experience with search and content processing systems. Users are familiar with the performance, interface, and features of the advertising-supported services from Google, Microsoft, and some promising startups like Blekko.com and DuckDuckGo.com. The vendors profiled in this report have worked to make their enterprise systems more like the free Web search services, but there are important differences of which many decision makers are unaware.

One of the challenges facing organizations looking for a search system and a vendor trying to license its search system is dealing with the confusion between the public Web search systems like Bing and Google and the enterprise search-and-retrieval systems.

Some History Since 2007

Has enterprise search changed significantly since 2007? The answer to this question depends on who is asked. If you query the so-called search experts you will hear many answers, and you will want to evaluate each consultant's viewpoint and verify what the so-called expert says. The team working on this report has extensive experience in enterprise search and content processing, but our admonition applies to our work as well.

Vendor Collapse

Building a search and content processing business is difficult. Since 2007, a number of search vendors have fallen by the wayside. Convera (formerly Excalibur Technologies) went out of business. The company moved from enterprise search to rich media search and then to vertical search. None of the initiatives worked. Entopia, a company that focused on the enterprise, shut its doors. Delphes, a vendor based in Montréal, went dark. Specialists like Siderean Software struggled to close deals. In France, Kartoo shut down. In San Francisco, Groxis turned off the lights.

Because the majority of the companies offering search and content processing are privately held, it is difficult to get a fix on the financial health of some of the vendors. A number of companies have been able to secure venture funding; for example, Coveo (estimated $2 million), Lucid Imagination ($16 million), Vivisimo ($4 million), and Palantir (an astounding $90 million).

Concerns about relying on a startup or a venture-funded search vendor may cause some procurement teams to look for more stable options. Venture funded companies can run short of cash or have management changes imposed on them. As I write this, there have been executive changes in the last few weeks at Lucid Imagination, MarkLogic, Sinequa, and X 1.

At any point in time, the financial health of some enterprise search vendors can be precarious. Not surprisingly, many Fortune 500 procurements go to the vendors with a track record or a vendor providing other enterprise solutions to the organization. The flagship vendors' technology may not be as up-to-date or leading edge as the solutions from smaller firms, but the high-profile search vendors offer the promise of stability. Buying a system from an established, well-known vendor suggests a predictable upgrade cycle and sufficient staff to handle a licensee's specific engineering needs. Most important, the larger search vendors are not likely to go out of business and leave a licensee with an orphan.

Commoditization

Our research indicates that one of the major shifts in the last 36 months is to make search a commodity.

Organizations with recent computer science graduates may find that young engineers want to use free and open source software for enterprise search. Lucene/Solr have become popular systems to perform key word retrieval and provide faceted search. Lucid Imagination offers a one-stop download bundle for Lucene/Solr. And FLAX in Cambridge, England, offers a scalable system. One proprietary search vendor—dtSearch, a Microsoft-centric system—offers a useful search solution for as little as $500. If your organization has standardized on IBM, Microsoft, or Oracle systems, each of these companies offer enterprise search solutions.[1]

Open source search and content processing offer organizations a viable alternative to proprietary commercial systems. There are arguments pro and con for open source. The organization with technical resources and basic search requirements can deploy an open source solution and save the license fees. Open source also makes modifications and extensions easy. There are no restrictions on changing the search system like those encumbering some proprietary systems. However, an organization without technical capabilities is likely to find open source as expensive as

1. The IBM OmniFind 9.x system is based on Lucene and not discussed in this landscape report. Oracle offers its Secure Enterprise Search 11g system. That system is not discussed in this report. SAP offers an enterprise solution as part of NetWeaver and SAP has a search-and-retrieval system called TREX, not covered in this report.

a commercial solution. The reason is that the firm adopting open source will require some third-party engineering support to tap the potential of open source search solutions.

More Familiarity

Another factor transforming the search landscape is the fact that many organizations are now somewhat more search savvy than in 2007. A few years ago, search was not on the radar of many managers. In the last two years, search is a must-have function. Not surprisingly, search is built in or "baked in" with other enterprise applications. Companies like SAS, the provider of analytics tools, acquired Teragram, making text processing a key component of data analysis. Traditional vendors like Autonomy Corp. and Exalead positioned their search and content processing technology as a platform. The idea, which many organizations embraced, was to build on a search and content processing framework enterprise application. Examples range from a postal office tracking system to the analysis of potentially fraudulent transactions.

Repositioning

Enterprise search vendors have been engaged in repositioning. Evident before 2008 and after the financial meltdown in the US in April 2008, search vendors shifted their product positioning to include customer support, business intelligence, and other business unit solutions. Other vendors rolled out industry-specific solutions for financial services, health and medical, and publishing. The result of this repositioning benefited some vendors, but confusion among those looking for search solutions has spiked. Our own work underscores that most procurement teams cannot distinguish the features and functions of the major vendors. When a second-tier or specialist search solution is required, those looking for a niche solution find that the large, medium, and small vendors use the same marketing jargon, assert similar technical methods, and offer equally impressive demonstrations of their products.

Brainware and ZyLAB offer document scanning and work flow systems of which search is a utility function. Startups like Index Engines focused on indexing content sent to back up systems. EMC and Iron Mountain, essentially backup and storage vendors, responded by adding search and retrieval to their products.

In 2011, most search and content processing vendors look essentially the same; but there are differences, often significant ones.

Big Vendors' Words and Deeds

As the financial climate worsened, some vendors let their marketing collateral get in front of the realities of their enterprise search system. At a time when agility was essential, many search vendors have lacked the resources to implement some of the assertions made in the marketing collateral. One telling example is licensees' experience with the Google Search Appliance. Now at Version 6, Google trimmed the GSA product line, a cloud service to index an organization's Web site, and

expanded the scope of Google Apps. In the midst of these shifts, the Google Search Appliance seemed to be taking a back seat. Nevertheless, Google continued to position the GSA as the solution to enterprise information access challenges. The reality is that the GSA works for modest content collections. For large-scale content processing, the GSA may be too expensive and cumbersome.

Other vendors followed similar paths. One example is IBM. The company made a public relations splash with its natural language search system Watson on the television game show Jeopardy. However, IBM markets a version of Lucene as its enterprise search product OmniFind.

On the other hand, Oracle made little or no progress in expanding the footprint of its Secure Enterprise Search 11g product. The description of the system is robust, but the product has some limitations. Microsoft acquired Fast Search & Transfer SA for over a billion dollars; around the time of the deal, Fast Search was dealing with a Norwegian police investigation into its financial dealings, leading to an October 2008 raid on its offices.

None of these large firms has made significant changes in their search and content processing technology. The marketers pulled in front of the engineers and appear to be keeping their lead in 2011.

New Players

New companies continue to enter the enterprise search market. Among the newcomers in the last several years are Funnelback, an open source search system owned by the Australian firm Squiz. Sophia Search made its debut, offering an enterprise search system built on semiotics, a branch of linguistics.

Two companies reinvented themselves with no-cash mergers. Attensity combined with two German firms and shifted from text mining for the US government to analyzing data for advertising and marketing firms. Lexalytics merged with the UK firm Infonics and positioned itself as a provider of social search components.

The graphic *Search Landscape with Selected Vendor Data* provides a simplified scorecard for over two dozen enterprise search vendors. Some of the names will be unfamiliar to readers of this report. The total number of firms offering search and content processing systems to organizations is over 200.

For example, the growing need to manipulate content produced by social networks like Facebook and Twitter permitted some search vendors to reinvent themselves as social search systems. Two of the better known examples are Attensity, a company funded by the US government's In-Q-Tel investment arm, and Lexalytics, a company developed to enhance the tagging and indexing of content in a Microsoft system. Both companies are now best described as vendors of social content processing systems, not search and retrieval, although those functions are available in the firms' systems.

Google entered the enterprise search market in 2002 and in a nine year period has sold tens of thousands of licenses to the Google Search Appliance. However, Google has not made the GSA a viable solution for many organizations. The GSA is among the most costly systems to deploy when large volumes of content must be indexed and made searchable. Google is a cloud computing company, and its search appliance is built by Dell Computers and supported by Google partners. Google talks about enterprise search, but it has not made the necessary commitment to customer support and product usability that licensees have demanded. One competitor said to me, "When I find a company with a Google Search Appliance, that company is a prospect for our search technology. The Google customer knows exactly what is needed to satisfy users."

Rich media applications have begun to have an impact inside organizations. As new employees enter the work force, many bring a familiarity with and a dependence upon rich media as an important way to learn. Organizations are producing podcasts and videos; search systems that were built to process text often require third-party components and time-consuming refitting to handle rich media.

Featuritis

Enterprise search solutions suffer from the same type of feature creep one can see in Microsoft Word and in enterprise resource planning systems. With each release, layers of functionality are added to the core system. Some functions are superfluous; others enhance usability. In the last few years, an enterprise search system vendor is likely to offer such components as:

- Collaborative features. The idea is that when one does a query, the person looking for information can pinpoint who in an organization knows about a particular topic. Armed with that contact information, the searcher can open an instant messaging window or place a phone call. Other vendors add a one-click feature so the user can add index terms or descriptors to a particular result. The search system then uses this "folksonomy" to aid others in finding a particular document.
- Social network functions. The idea is that users are able to create affinity groups around a particular topic. The search system allows a person looking for information to identify these groups or search the content of the groups. Some vendors tout support for external social network content from such sources as Facebook or Twitter, among others.
- Data and text mining capabilities. including content acquisition features that tap into structured data in Excel spreadsheets, databases, accounts receivable, purchase order data, or third-party enterprise applications such as an enterprise resource management or customer relationship mangement system. The "fused" data are available to a user, either in a list of results or in pre-formatted reports. Many of the reports make use of graphics that can be placed directly in a presentation or publishing program.
- Audio and video indexing. A small number of vendors offer options for converting audio or video to text, indexing those speech-to-text files, and including links to the rich media in search results. Most vendors are working to find a way

to meet customer demand for rich media, which often imposes additional work on the search infrastructure.

- Mobile search support. Vendors recognize that enterprise search users may need to access the organization's data from a mobile device. The search functionality is often supplemented with geo-spatial functions, alternative interfaces to accommodate the mobile form factor screens, and on-the-fly results reformatting.
- Understanding the "meaning" of text. Progress is being made in this very difficult component of search. Vendors may use rules (time consuming and expensive to maintain), fully automated systems (statistical, linguistic, and blended methods with or without knowledge bases and controlled term lists), and blended approaches. The shortcut some vendors use is user personalization. The user's department, job function, and past search history provide inputs for the query processing subsystem. Modern systems can therefore deliver results that are more nuanced than the brute force key word approach in the old-fashioned systems.
- Rich indexing or metatagging subsystems. Search vendors and niche software specialists offer components than can assign index terms not contained in a source document. The advantage of systems that assign metatags via "smart software" or via a combination of automatic systems and human subject matter experts is sometimes called "discovery". The idea is that the interface displays related information or suggests documents that may help the user answer a question. Most of the systems discussed in this landscape report support "facets" or rich indexing natively or provide application programming interfaces so third-party metatagging and metadata management systems can interact with the search system's content processing service.

Does this list exhaust the enhancement to enterprise search? No. Organizations often require highly specialized systems. Vendors who create search systems for product management, certain types of electronic discovery and legal research, health and medical tasks such as assigning diagnostic codes to patent medical records, and chemical structure searching are outside the capabilities of the vendors referenced in this report. Each specialized search application demands a high degree of subject matter and business process tailoring. The vendors profiled in this report or mentioned by many of the consulting firms offering advice about enterprise search may say "We do that", but the reality is that although general purpose systems could be customized to handle chemical structures or product manufacturing data, the cost and time required to retrofit a general purpose system or platform are too great. A specialist system makes more sense, costs less, and in general can be deployed more quickly than a retrofit requires.

Battle Lines

Due to the business climate, tension among vendors is increasing. The principal flash points are:

- The Google and Oracle legal matter related to Google's use of code that Oracle claims Google incorporated into its system without permission.
- Google and Microsoft are drawn into conflict in the enterprise market. Google is pressuring Microsoft. The companies are exchanging public barbs, and there are few indicators the tension is decreasing.
- Autonomy defends its 20,000-plus customer base by offering value-added services. When engaged in a direct competition with another vendor, Autonomy is able to demonstrate its wide range of applications.
- Oracle has been particularly aggressive toward MarkLogic Corporation. MarkLogic's technology is an XML repository providing certain features not associated with the Oracle database. Oracle's white paper explaining the deficiencies of the MarkLogic approach shows what a major firm will do to defend its market share.
- IBM, on the other hand, has relied on the Watson publicity stunt to leapfrog Google and other vendors' search and content processing technology. What is interesting is that IBM is licensing its version of Lucene, not its next-generation Watson technology. IBM has demonstrated its ability to use television as an effective publicity mechanism for the firm's next-generation content processing technology.

By 2010, some battle lines formed across the search landscape. Multi-billion dollar competitors are now engaged in harsh combat. Google fights Microsoft Corp. for sales in enterprise software, not just search-and-retrieval sales. IBM struggles with SAP and Oracle for database licenses, mission-critical applications, and services. None of these firms has a robust enterprise search solution. IBM touts its artificial intelligence system on a television game show, yet relies on the open-source technology for its flagship enterprise search system. Oracle struggles with performance and scaling because its search system has roots in technology developed in the 1980s and updated erratically with in-house innovations and technology acquisitions from the little-known Triple Hop.[2] Each year Oracle engages in tire kicking in search of a more robust information retrieval acquisition. But Oracle's management has marginalized search in order to protect its core database and services business. Microsoft purchased the Fast Search & Transfer SA technology and is now working overtime to rationalize the company's multiple search initiatives. Microsoft itself shows little interest in the legal and financial legacy of the Fast property for which the astounding sum of $1.2 billion was paid at the time Fast Search's technology was precipitating a business crisis at the Norwegian firm. As I write this, these enterprise flagship vendors are locked in a life-and-death battle. When a For-

2. The acquisition took place in 2005. Oracle has not made meaningful improvements to Secure Enterprise Search 11g in the last five years. Oracle knows that Oracle database customers will accept SES11g as a solution, but SES11g cannot flourish outside the Oracle database ecosystem. Oracle itself has used the Google Search Appliance to handle certain information retrieval issues for Oracle database licensees.

tune 1000 account is lost, often the only way to replace the lost revenue is to take another firm's client or make an acquisition to obtain revenue and new customers.

The search products and services from major enterprise vendors have created opportunities for enterprise search vendors offering high-value solutions. Search Bing.com, Exalead.com, or Google.com for "enterprise search", and you will be greeted with more than one billion results. An inventory of the vendors engaged in search and retrieval numbers more than 200 firms, ranging from unknown Juru, once the darling of IBM's developers, to the ubiquitous Autonomy Ltd., a vendor that tries not to describe itself with the terms "search" or "business intelligence." Autonomy refers to itself as a purveyor of "meaning-based computing" solutions. The number of startups is remarkable. Entrepreneurs are quick to recognize that the solutions available today share some characteristics:

- Cost. Deploying a system to make information available to employees and contractors in an organization is expensive. The license fee may be one of the lower cost components. The time required to figure out what information to make available and to whom is significant, often measured in months or years. The "sticker shock" of an enterprise search system is one of the major problems a search vendor faces. In an effort to reduce the costs of search, Google, Index Engines, Perfect Search, and Thunderstone offer appliance solutions. The problem, of course, is that appliances must be managed. Even cloud solutions can become problematic. Blossom Software is one of the leaders in hosted search; yet when the licensee has special requests or requires the core system to be fine tuned, Blossom's owner charges for customization.
- Specifications. Many organizations fall for the promises of a search vendor whose pitch is similar to infomercials for products that do everything. Information access needs vary by department, business function, and other factors. Despite the promises of a system that it provides access to "all" information, organizations typically end up with islands of content and multiple search, content processing, and information access solutions. Search vendors have learned that many prospects no longer believe the "we do it all" pitch. As a result, search vendors have shifted to offer what is called a "vertical solution"; for example, electronic discovery for a legal matter, a customer support system to reduce the costs of providing online, outsourced, or on-premises information to customers, or business intelligence systems. This is an interesting category because "business intelligence" is as difficult to define as the older phrase "knowledge management". Because clear thinking rarely accompanies a search system procurement, organizations find themselves faced with user dissatisfaction. Estimates of dissatisfaction range from 50 to 75 percent of a system's users.[3]
- Customization. Search vendors talk about customization and personalization as easy to accomplish. Certain tweaks are easy to make via configuration files, cascading style sheets, and algorithms that recognize a user is a member of a particular department and filter information for that user accordingly. When new

3. This shocking dissatisfaction with search is discussed in Martin White and Stephen E Arnold, *Successful Enterprise Search Management*, Galatea Press, 2010.

features become available on the public Internet, employees may want access to similar services. Enterprise search systems often struggle to accommodate content from such sources as Facebook and Twitter, real-time news feeds, and rich media from such sources as YouTube or Slideshare. Many vendors say their systems deliver these functions, but in actual practice, the implementation may be limited or require that the licensee invest significantly in associated systems and infrastructure. Nevertheless, vendors talk about adding social and collaborative features, often neglecting to spell out exactly what the limitations and costs of the enhancement are.

The Buyer's Conundrum

The question arises, “Which vendor can an organization believe?” The reality is that the nature of information, the inability of a potential licensee to state exactly what information access problem is to be solved, and the nature of the sales and marketing processes are road blocks. The disconnect between potential licensee, users of the proposed system, and the search vendor is difficult to bridge. Every search system vendor has licensees who sing the praises of the search system but not surprisingly, every search system vendor has clients who are deeply dissatisfied. Some consultants talk up the importance of “bake offs” or “head-to-head” competitions, yet the financial climate is such that very few organizations have the time, expertise, or motivation to run objective tests. The result is that almost anyone can set himself or herself up as a search expert and offer help with a procurement. Hundreds of technology services firms assert their ability to work through engineering problems, and most list a number of search systems as part of their service base. The reality is that even engineers working on a search system may not know how to resolve a problem due to the extreme complexities “in the depths”. The only way to remediate some problems is to invest technical resources and time to solve (or more commonly work around) them. The bottom line: the customer pays.

Chapter 2:

Autonomy Corp plc

"Meaning based computing."

"We make technology that reads data and content, understands them and deals with them without human beings needing to be present."—Michael Lynch, Autonomy

Autonomy at a Glance

Autonomy Corp. opened for business in the mid-1990s and has grown to almost $1.0 billion in annual revenues. Analysts have given Autonomy high marks for its vision, technology, market share, and financial success. As competitors for the enterprise market have floundered, Autonomy has innovated, acquired, and marketed successfully, year in and year out.

Today the company's technology platform provides a solution to information retrieval challenges across a spectrum of enterprise requirements. The core upon which Autonomy's success has been built is IDOL or Intelligent Data Layer. IDOL can support eDiscovery, video search, fraud detection, and customer interaction.

Key Developments

In the last 24 months, there have been two crucial developments at Autonomy. The firm has continued to win important accounts, which now number more than 20,000, and the company has continued to add functionality to IDOL- for example, enhanced visualization and advanced analytics, support for social content, and high-performance data fusion capabilities.[4] In addition, Autonomy's remarkable financial performance delights stakeholders. But none of these are as significant as Autonomy's approach to acquisitions and its marketing and sales methods.

Autonomy at a Glance

	Basic Information	Option	Comment
License Fee	Begins at $35,000	On-premises and hosted options	Custom price quotations are available
Search product	IDOL	IDOL is a platform. Autonomy offers options for most enterprise applications	New initiatives include the Aumi-nence health care product line
Technology hook	Meaning-based computing and automated, self-learning methods	Autonomy offers support for rich media, fraud detection, and other specialities	Autonomy's acquisitions are "hooked into" the IDOL platform, so new functions are componentized
Cautions	Accurate specifications and appropriate resources are essential	Resellers and integrators are useful assets to have available	Autonomy IDOL is not a "learn as you go" application. It is a platform with specific methods and operations. Hacking is not advised.
Selected partners	Accenture, Boeing, Capgemini, and hundreds of other highly-regarded organizations	Most information consulting firms have experience with Autonomy	Autonomy provides a search function on its Web site to make it easy to find an Autonomy partner.
Net Net	Autonomy IDOL can integrate structured, semi-structured, and unstructured information. IDOL can process text, voice and video. The system can automatically identify and rank the main concepts within the source documents. IDOL automatically categorizes, links, summarizes, personalizes and delivers information. The system supports collaboration. IDOL can be used to automate operations within enterprise information portals, customer relationship management, knowledge management, business intelligence and e-commerce applications, among others.		

Autonomy acquires companies delivering technology, products, and services that can add to IDOL's capabilities. The company has perfected a strategy in which it acquires companies in high-growth markets, rapidly integrates the IDOL technology into the acquired tools, and as a result transforms the market it enters. To illustrate: in 2005, long before the stampede of search vendors into the customer support sector, Autonomy purchased ETalk Corporation. In late 2005, Autonomy purchased Verity, Inc., which at the time was one of the most prominent search technology vendors. More recently Autonomy acquired Zantaz, Inc. and almost immediately boosted the firm's market presence and at the same time leveraged Zantaz's cloud-based e-mail and storage service into a broader enterprise platform. When Autonomy acquired Interwoven in 2009, some saw the deal as a turning point for Autonomy. It was. Interwoven propelled Autonomy in revenue and opportunity. By integrating features from Zantaz and Interwoven, Autonomy was able to offer e-marketing and collaborative services to existing and new customer segments, including online advertising firms. The point that few emphasize is that Autonomy can manage its acquisitions in a successful manner. The operative word

4. Data fusion is shorthand for such processes as ETL (extract, transform, and load) and adding value to context-free content like Twitter messages and Facebook posts.

is management. Competitors often overlook Autonomy's management strength, which is equal to or better than the highly-regarded technology within IDOL.

Moreover, Autonomy is one of the best sales and marketing operations in the world of information retrieval. There are numerous examples of Autonomy's ability to move into a market before its competitors recognize an opportunity. One was the firm's early push into rich media. The company acquired Virage, a video-management software developer, in 2003. Since that time, Autonomy has become a major player in providing search and retrieval solutions to the broadcast industry. Autonomy then fostered Blinkx, a video search company, which it spun out in 2007. In terms of positioning, Autonomy was among the first information retrieval companies to offer "a portal in a box", an appliance solution to search for an Intranet in 2000. Many search vendors hire former Autonomy professionals because they have been well trained and know the difficult field of search and retrieval.

There are rumors circulating about Autonomy's next major acquisition. The firm will continue to follow the path that has put it at the top of many analysts' league table. But Autonomy's management is the real key to the company's remarkable success and its $6.0 billion market capitalization.

History

Autonomy's technology has its roots in an 18th-century Presbyterian minister's mathematics. Bayes' Theorem sets forth a method by which one can derive inferences about what is analyzed. When the Bayesian numerical recipes are applied to information retrieval, the system "learns"; that is, IDOL automatically (autonomously) forms an understanding of the concepts of the processed content.[5] IDOL makes inferences about the information. The method makes "meaning-based computing" work.

Michael Lynch, one of Autonomy's founders, said:

> Users are inundated with too much irrelevant information on the Internet. Autonomy's Agentware personalized information solutions utilize Neural Network based Intelligent Agents to dynamically understand user preferences which allow service providers to deliver relevant information. This relationship is part of Autonomy's corporate road map to leverage key partner relationships to offer customers the best solutions.[6]

Over the last 15 years, Autonomy has refined, expanded, and integrated operations and functions that fuel Autonomy's market success. There are critics who pooh-pooh Autonomy's methods. However, Autonomy continues to maintain its grip on such customers as law firms, federal agencies, and such firms as Bloomberg, Boeing, Citigroup, Coca Cola, Deutsche Bank, FedEx, Ford, GlaxoSmithKline, Lloyds

5. For more information, see "An Essay Towards Solving a Problem in the Doctrine of Chances" at http://www.stat.ucla.edu/history/essay.pdf
6. Source: http://www.autonomy.com/content/Press/Archives/1997/0428.html

Banking Group, Nestlé, the New York Stock Exchange, and Shell. More than 400 companies offer OEM Autonomy technology, including Symantec, Citrix, HP, Novell, Oracle, Sybase and TIBCO.

Product Lineup

IDOL is Autonomy's main product. However, the company uses its acquisitions and their products to address specific customer segments. Interwoven offers a content management system. Since the Autonomy acquisition of Interwoven, Interwoven licensees get access to the IDOL technology. Autonomy's approach is to solve the customer's problem, harnessing the needed functions to IDOL, thus delivering an integrated, extensible architecture. The Interwoven CMS service is what the customer "sees". The IDOL functionality adds richness to the Interwoven user's experience.

Core enterprise search and content processing products include the IDOL server. The company offers process automation through its Teleform and LiquidOffice business units. The company provides regulatory and compliance products, consolidated archive products and services such as Digital Safe and Arcpliance. The firm provides traditional records management solutions via Meridio Records Manager and content management services via the Interwoven division. For the legal and enterprise sectors, Autonomy offers a range of eDiscovery products and services including LegalHold (eDiscovery notification, preservation, collection, and mapping), cloud-based eDiscovery, and an eDiscovery Appliance. The company offers a specially-tuned version of IDOL search for the legal market as well as the iManage content management product. Autonomy offers specialized versions of its products for the security and surveillance markets, meeting the needs of investigators

involved in fraud detection and other types of intelligence activities. Autonomy offers a full range of products tailored to the media market, including Virage MediaBin (rich media asset management) and ACID (a product that detects copyright infringement).

This list does not exhaust Autonomy's product lineup. The key points to keep in mind are:

- Autonomy's sales team can tailor the product and service line up needed to solve almost any information processing challenge.
- Individual brand and product names are maintained after Autonomy acquires a firm. What is changed is that the acquired company's technology "hooks" into the IDOL platform, thus adding "meaning-based computing" to an established product or service.
- Autonomy offers on-premises, cloud-based, and hybrid options for its products. Consequently, the licensee can tailor a solution that combines on-premises and hosted services. These types of deployments benefit from tight specifications and appropriate engineering resources.

Technology

Autonomy's underlying technology—the Intelligent Data Operating Layer (IDOL)—is founded upon the works of Thomas Bayes and Claude Shannon and now includes more than 170 Autonomy patents. Bayes' work centered on calculating the probabilistic relationship between multiple variables and determining the extent to which one variable impacts another.

Today Autonomy's IDOL server is a "pan-enterprise information access platform". It offers over 500 advanced functions. IDOL, asserts Autonomy, "forms an understanding of all information, whether structured, semistructured or unstructured, and recognizes the relationships that exist within it. This allows computers to harness the full richness of human information, bringing meaning to all data, regardless of what or where it is."

Applying computational power to this concept makes it feasible to calculate the relationships between many variables, allowing software to reveal the context of a piece of unstructured information. Having understood the meaning, Autonomy's approach then relies on Shannon's theory, which states that the less frequently a unit of communication (for example a word or phrase) occurs, the more information it conveys. This is the standard inverse frequency notion now used in virtually all search systems. Ideas, which are more rare within the context of communication, tend to be more indicative of meaning. IDOL's approach is independent of the language of the text and allows the main concepts to be identified and prioritized. Autonomy told me:

> Vendors may claim to perform concept search, but these are not automated solutions and in the end, a more complex approach that is some combination of keyword search and linguistic rules. In terms of solutions we offer from acquisitions, the IDOL platform that underlies these solutions is the main dif-

ferentiator, more so than the solutions themselves. Because no one else can automatically extract meaning from all data types derived from all sources. IDOL is at the heart of innovation.

Autonomy's architecture is well-presented in a diagram that is now several years old. Nevertheless, the basic structural components of IDOL are clearly depicted. Autonomy is correct when it explains that IDOL is a platform, not a single application, although IDOL can be used for a specific function such as deploying a customer support and customer self-service system.

IDOL forms a conceptual and contextual understanding of the content in an enterprise, automatically analyzing any piece of information. IDOL provides over 500 advanced operations to be performed on content. IDOL connectors access more than 400 content repositories, such as file systems, SharePoint, and Documentum. IDOL keyview uniquely supports over 1000 file formats. © Autonomy Corp. 2011

The architecture is distributed, parallel, and scalable. IDOL's architecture delivers unmatched scalability. The system can store over 17 petabytes of data. Other salient features of the IDOL platform are:

- Meaning-based computing. An ability for the system to form a conceptual understanding of content
- Access to all data sources and file types: IDOL supports over 1,000 data types, including rich media, and connects to over 400 content repositories
- Language independence: IDOL's pattern-matching technology is fundamentally language independent
- Compatibility with enterprise operating systems: IDOL is a cross-platform solution
- Compliance support for the US Federal Rules of Civil Procedure. IDOL provides FRCP-compliant search as it searches enterprise content
- Scalability. IDOL delivers linear scalability by use of its distribution model
- High availability. IDOL offers a number of redundancy and fail-over options
- Security. IDOL's "mapped security model" is empirically proven to scale in the enterprise and support most third-party security methods

- Advanced functions. Over 500 advanced functions beyond search; for example, work automation
- Fully customizable and transparent relevance ranking with user-friendly business console style sheets and administrative graphical interfaces

IDOL provides over 400 connectors to acquire web content as well as documents in Microsoft Word, Excel, Oracle database, Documentum, FileNET, Lotus Notes, and other content formats upon installation.

Indexing Highlights

Autonomy's robust platform contains a number of content processing and information retrieval capabilities. Particular functions that are noteworthy include:

Concept Identification and Relationship Matching

The IDOL system ingests an "information object"; for example, e-mail, a Word file, or an audio or video file. IDOL is able to detect the main concepts present in the "information object", and also automatically find conceptually related information. IDOL allows manual and fully automatic linking between related pieces of information, regardless of their format. The concepts in a document can be linked automatically to those in another file. These concepts can also be linked to related concepts within video or voice mail. Hyperlinks are generated in realtime at the moment a document is viewed, removing the need for any manual intervention and ensuring concepts are constantly uptodate.

Precis Function

Interest in document summarization or automated précis generation is increasing. When a user is confronted with a number of relevant hits to longer documents, the system's ability to generate a summary becomes a "must have" function. Document review is facilitated because the précis makes it easy to determine if the source should be reviewed more closely. IDOL provides three types of summaries:

- The basic simple summary comprises a few sentences from the source document
- A conceptual summary displays a few sentences from the document that contain the most salient concepts. These sentences can be from different parts of the source document.
- A contextual summary which relates to the context of the original query. This method allows the most applicable, dynamic summary to be provided within the results of a given query.

Personalization

One feature of the personalization method is that IDOL Server notes what a user does with content. The probabilistic approach eliminates the need to update user profiles manually, although hand tuning is available via IDOL's administrative

interface. One outcome of this combination of context and probability is that IDOL can identify affinity groups within a user community. A pharmaceutical company with worldwide operations can easily identify researchers at different locations sharing a particular interest or line of inquiry.

The method also allows IDOL to perform what Autonomy calls "intent-based ranking". This method personalizes the search experience by customizing the search results according to the users' interests based on their implicitly generated profile.

By understanding the context of all information that a person reads, speaks or writes, whether it be through email communications, Web browsing, or phone conversations, IDOL forms an "implicit profile" of individual users based entirely on a conceptual understanding of that content. Because the IDOL platform is an infrastructure search solution, the system can leverage and consider every action and interaction, regardless of format, across every system, platform and working environment when creating user profiles. The result is an enhanced and targeted end user experience. Relevant content is proactively positioned, easily accessed and ultimately consumed based on the user's interests and current activity at the time of search.

According to Autonomy, "The search experience does not end with personalization.". IDOL continues the dialog with the user to help discover the document with automatic summarization, automatic query guidance (group search results by concept), automatic hyperlinking (deliver related files to each search result), automatic clustering, social search, query suggestion, query relaxation, and ideas cloud (generate prevailing concepts present in the search result), among others.

Automatic Clustering and Categorization

IDOL processes search results and groups them by relatedness. The clustering function operates automatically and licensees can tune the clustering engine to meet the specific needs of a user or particular group of users. Clustering permits point-and-click exploration of a group of hyperlinks to documents that could otherwise be overlooked in a relevance ranked results list.

Organizations can analyze large sets of documents and also the user profiles and automatically identify inherent themes or information clusters. IDOL clusters the unstructured content exchanged in email, telephone conversations and instant messages. Search results can be automatically clustered by concept (e.g. a search for *apple* would result in clusters for the company, the fruit, Gwyneth Paltrow's daughter, etc.). The method is especially useful for keyword-prone users who are unlikely to enter multi-term queries.

IDOL can make use of available controlled term lists or taxonomies. If formal dictionaries are not available, IDOL can generate a list of categories and tag documents with one or more category labels. Categorization is automatic and facilitates routing of particular documents to specific users.

Spotlight Features of IDOL

In a system as comprehensive as Autonomy's, it is difficult to select a particular function or small group of features to spotlight. Based on my experience with IDOL, I want to focus on two features that some of Autonomy's competitors are working to implement in their systems: social and collaborative functions and taxonomy operations.

Social and Collaborative Features

For almost 10 years, Autonomy has enabled collaborative functions within IDOL. The system can identify individuals who share a similar interest or context. However, with the acquisition of Interwoven, Autonomy was able to add real-time social functions to operations anchored in IDOL.

IDOL provides organizations with the ability to listen, measure and engage in conversations across the social Web, including Twitter, Facebook and YouTube. IDOL tracks, clusters and automatically extracts sentiment from any social media site and source and in real time identifies emerging trends based on these conversations. IDOL also provides governance features for social content in order to ensure regulatory compliance.

By analyzing the structures and meaning of language, IDOL determines the positive and negative characteristics of each piece of information and creates relevant classification systems. IDOL can determine the degree to which a sentiment is positive, negative or neutral for the entire content or a segment of the content. In addition, administrators can apply multiple tagging functions and specific threshold cutoffs to determine the sensitivity of sentiment analysis. IDOL is also able to analyze the sentiments contained in audio and video files so that marketers can take advantage of the rich information that resides in multimedia social media assets.

Real-Time Processing

Autonomy's operations occur in a low-latency setting. In practical terms, the content processed by the system becomes available to users and to other IDOL processes within minutes of its entering the IDOL pipeline. Classification, context, and metatagging occur without the hour or day delays that cause bottlenecks in some competitors' systems. Visualizations for business intelligence applications or alerts for fraud detection applications occur seamlessly. Even IDOL's applications for search engine optimization and e-commerce transaction are available without significant latency when appropriately resourced.

Low latency is particularly important when processing audio and video content. The size of the files demands super-computer performance. IDOL takes advantage of caching, parallel processing, and other engineering methods to reduce latency.

Autonomy's success in business intelligence and text mining is one consequence of the system's ability to handle large-scale content streams in near real time. Autonomy's tools like its Application Builder Toolkit allows the licensee to tailor pro-

cessing operations to highly particular content processing requirements such as those in intelligence and financial services deployments.

Operational Views

We think it warrants noting that Autonomy emphasizes how its Meaning Based Computing technology is radically different from what many vendors call "business intelligence".

Content hot spots can be explored with a mouse click. The Autonomy IDOL platform provides a number of visualization options, and the system can be extended with third-party visualization tools if IDOL is used in an environment with an i2 Ltd, Palantir, or similar system.

First, many business intelligence systems force marketers to rely on outdated information. Technicians or specially-trained analysts must code a report and then load the needed information from a data warehouse. Reports then run on this "block" or "cube" of data.

Second, traditional business intelligence typically handles structured data and cannot account for the rich, human friendly information that customers use to communicate today.

Finally, traditional business intelligence systems make assumptions about what to look for. The approach confines the system to pre-defined key performance indicators and previously known pathways. The output looks authoritative but may yield stale, tainted, or incorrect results. Autonomy's approach, which it describes as "a much better way", is to let the data tell you what is important. Oftentimes the most valuable opportunities in business are the "unknown unknowns", the things you did not know were out there.

Autonomy allows businesses to break free from the old business intelligence paradigm. IDOL licensees can listen to and understand all forms of data, see new patterns as they emerge, and act on them in realtime.[7]

The core technology uses pattern recognition to find relevant words and related concepts. Users can express queries without having to know the exact words to snag the information needed. Autonomy's system can be used to make a faceted

search interface available. Its outputs can be manipulated by other Autonomy modules or piped into third-party applications for discovery analysis.

Autonomy's platform can process and understand rich media; for example, voice conversations. IDOL converts speech to text and then processes the output. A human does not have to listen to conversations which is time consuming and expensive. An analyst can use the Autonomy business intelligence interfaces to identify concepts or probe into specific content flagged in a text or graphic display.

Autonomy has created an "investigation manager" module for government investigators. The system permits near real time monitoring and analyzing of person-to-person communications. Autonomy differentiates by offering the IDOL technology as part of a comprehensive information and intelligence platform.

As a student at Cambridge, Mr. Lynch understood that Bayesian statistics could have profound implications for systems attempting to wriggle meaning from unstructured data. Neurodynamics' software and systems - the core technologies in today's Autonomy - are outgrowths of Mr. Lynch's probabilistic modelling and digital signal processing technologies developed by him.

In the early 1990s, commercial search and retrieval systems required that users know exactly how to phrase a query to get information about a topic. Intelligence analysts took one look at the outputs of a Dialog or LexisNexis system and concluded:

1. If we knew what we were looking for, then we would be able to make the Boolean systems provide information. A query on the Dialog system for a recent story on a murder would look like SS (strang* OR chok* OR garrot*) AND (Smith* OR Smyth*) AND UD=9999
2. The information manipulated by intelligence professionals and police was usually not in commercial databases. The data were in the form of ASCII notes typed by a case officer into a terminal, newsfeeds from various services with little formatting in common between Agence France-Presse and Pravda, or from different electronic data obtained from credit card companies, banks, and intercepts.

Mr. Lynch found a ready market, first in the U.K. and then in the U.S. Even today, Autonomy is viewed as the leader in text mining technologies in many intelligence entities.

Strengths

Autonomy has a robust information processing platform and a broad range of applications. Over the course of 15 years, Autonomy has enhanced its core platform and

7. For more information about the Autonomy approach, navigate to http://www.autonomy.com/content/News/Releases/2009/0916.en.html and http://www.autonomy.com/content/News/Releases/2009/1019.en.html

kept in step with new developments in high-performance computing. In addition, Autonomy has been an early participant in social and collaborative content processing.

Although some competitors can complain about Autonomy's position in the market, no other search vendor has generated comparable revenue from enterprise search and content processing. Neither Google nor Microsoft provide financial specifics about their search systems' market success. Autonomy does, and the company's revenue is now approaching $1.0 billion.

Other Autonomy strengths include:

- Financial stability.
- Numerous options and vertical-specific applications available to licensees.
- International footprint, with offices in the USA, Europe, and elsewhere.
- Strong ability to run combined searches against structured and unstructured data sources.

The themes that surface in a review of Autonomy IDOL range from the use of context and conversation to create a superior user experience to advanced rich media analytics. Via the automation of manual processes, Autonomy delivers a boost to an organization's return on investment. IDOL offers what it calls "unified information access" by search via concepts, keywords, metadata, and semantics on intranet, the Internet, and desktops. With more than 400 connectors, Autonomy seamlessly taps into these sources.

The system can handle content regardless of language and offers a comprehensive set of automatic features, functions, and operations.

Companies seeking to apply sophisticated numerical recipes to tasks once handled by trained professionals sought out Autonomy. As early as 2000, Autonomy was applying IDOL to the analysis of customer support call records, warranty cards, answers to open-ended survey questions, and the growing volume of electronic information produced by customers, employees, and publishers. After 2001, the demand for "intelligence" began to grow. Autonomy's meaning-based technology grew rapidly because IDOL could uncover hidden insights within large volumes of electronic information.

The outputs of the Autonomy system can be searched using key words or explored using point-and-click interfaces. The system can also route content (items, concepts, and objects) of interest to a particular user of the system.

IDOL can identify the main idea of a document and find other documents that contain a similar idea. Consider this example: Autonomy IDOL processes documents and identifies groups of words pertaining to ecologically friendly automobiles such as *green* and *hybrid*. The system will know that the ideas are not related to color or engineered crops. When unstructured text is processed, the Autonomy system will map the relationship among concepts. Documents that share concepts or themes are

tagged. The system extracts the concepts and generates a category for them. As noted, the concepts are linked.

Autonomy uses Adaptive Probabilistic Concept Modelling (APCM) to analyze the correlation between features found in documents relevant to an agent profile and then to find new concepts and documents. Autonomy can determine the concepts important to sets of documents. This function allows Autonomy to classify accurately the new documents processed by the system.

When IDOL discerns that one pattern has a preponderance over another pattern in a piece of unstructured information, value-added "tags" are generated for each document or information object. Iterating the mathematical recipes over a flow of unstructured or structured content allows IDOL to extract what Autonomy calls a document's "digital essence." IDOL encodes a "digital fingerprint" of concepts, entities, and meaning in each processed information object.

What is clever is that Autonomy IDOL looks at many traditional statistical arguments in a fresh way. To take a traditional statistical argument, if dice are rolled 1,000 times and generate "snake eyes" 25 percent of the time, Autonomy's system will identify that the dice have been rigged. Autonomy's approach to business intelligence introduces nuance to a process that many competitors boil down to a statistical value without context.

Cautions

The principal consideration with Autonomy IDOL is that it requires appropriate computing resources. An IDOL system starved for CPU cycles, bandwidth, or RAM can become sluggish. When collections contain content on widely varied topics, a subject matter expert may be required to tune the parameters for the IDOL content processing subsystem. The work is not difficult, but IDOL performs optimally when the source content matches the training collection used when the system is first set up.

Some Autonomy customers have not been able to completely eliminate the need for human interaction with IDOL's administrative controls. Prospective licensees will want to test carefully on a corpus that reflects the information objects that will be processed by IDOL when the system is deployed.

Other drawbacks include:

- The Autonomy IDOL system is not an appliance. It is a platform. For organizations looking for a vendor to provide a server that is ready to index minutes after the search appliance is connected will want to look elsewhere. Google, EPI Thunderstone, or Index Engines may be a more prudent alternative.
- Autonomy, compared to some search vendors, does not specialize in providing womb-to-tomb consulting services. Autonomy works with resellers and integrators who perform much of the installation and customization. If you are considering an IDOL system, you will want to involve an Autonomy reseller or integrator. In addition, you will want to identify an Autonomy system administrator and provide appropriate training prior to installation.
- Autonomy's platform, like the platforms of other vendors profiled in this Landscape Report, is composed of a number of moving parts. The system can perform almost any information-related task. Engineering and maintenance are important and on-going processes. Cutting corners on either can undermine the otherwise excellent performance of IDOL.

Autonomy has an effective sales and marketing organization. Presentations about IDOL are compelling and often show IDOL addressing the major information challenge an organization faces. Pilots using the potential customer's own content are perhaps the best way to experience IDOL. Autonomy will sometimes guarantee that IDOL will handle a specific, challenging problem. It will, but there may be some additional effort required. A real-life trial is an important part of the procurement process.

Outlook

The outlook for Autonomy is generally positive. The financial climate is improving in certain sectors and lagging in others. Autonomy has been adept at moving into new markets and making acquisitions that fuel revenues. In the next 12 to 18 months, Autonomy is likely to take these actions:

1. Make another strategic acquisition. With the rollout of Autonomy's health care product/service called Auminence, Autonomy management may seek to strengthen its position in this large, fast-growing sector.
2. Get close to or break through the $1.0 billion in revenue ceiling. Autonomy can achieve this with organic growth and a well-considered acquisition.
3. Escalate marketing pressure on smaller competitors and encroach into customer segments held by the likes of IBM, Oracle, and SAP and similar enterprise-centric vendors.
4. Forge an alliance with an up-and-coming company, possibly in the data management or business intelligence sector. The type of company in which Autonomy may show interest is a data fusion firm with technology that taps open source software and blends open source with next-generation analytics.

Autonomy is likely to benefit from the economic vise clamped on some of its competitors. Particularly vulnerable are mid-stream search and content processing vendors. Autonomy can package its products and services to compete on license fees

and support costs. The Autonomy reputation helps reduce the upsell required to get a potential customer to take a risk on a smaller, less well-known competitor.

Autonomy is also likely to benefit from missteps at some of the Fortune 100 vendors; for example, IBM, Google, Microsoft, Oracle, and SAP. Autonomy's management seems to be more focused on information-centric solutions for the enterprise. Giants like these much larger enterprise software vendors are increasingly diffuse in their quest for revenues. Autonomy can, as it has in the past, bide its time and then make a strategic move that puts the largest competitors off balance.

Barring an unforeseen event, Autonomy is likely to continue on the course that has made it one of the largest, if not the largest, vendor of enterprise search solutions in the world.

Net Net

Autonomy has thrived as other enterprise search vendors have struggled or failed. The company has thousands of satisfied customers. The firm has shown significant management acumen in its acquisitions, the management of acquired companies' technologies and customers, and its financial track record. Most enterprise search vendors struggle to keep pace with the demands of information access among enterprise licensees. As a public company, Autonomy's financial reports provide a yardstick against which to measure the company's performance. Autonomy is now approaching $1.0 billion in sales at a time when most of the more than 200 search and content processing firms are unable to generate sufficient momentum for an initial public offering. Autonomy's Integrated Data Operating Layer has been shaped into a platform that can perform a broad range of search, content processing, information management, and business intelligence functions. Competitors find Autonomy the company to beat in many tender situations. The firm's technology cannot be dismissed or ignored, and we expect Autonomy's "gravitational" pull to increase in the next 12 to 24 months due to its upsell opportunities, its marketing savvy, and its management.

Autonomy Annex 1: Selected OEM Licensees

Autonomy is one of the, if not the most, successful original equipment manufacturers in search and content processing. The company has more than 1,000 OEM relationships. Autonomy's OEM partners use IDOL technology in third-party applications ranging from content management to fraud detection. A licensee of OpenText's RedDot CMS has IDOL's search function embedded in the RedDot system.

Autonomy OEM Licensees (Selected)

OEM Licensee	Key Technology
Adobe	Consumer and commercial software
Cisco	Network hardware and services
Citrix	Enterprise and consumer remote access systems
EMC	Storage
Hewlett Packard	Also used by EDS, an HP subsidiary
IBM	Enterprise solutions, services, and products
Iron Mountain	Archiving and e-discovery services
Kana	Customer support systems
Matrix One	Collaboration solutions
Novell	Enterprise infrastructure
Openwave	Mobile solutions
Oracle	Also used the Oracle subsidiary Hyperion
Support Soft	Enterprise software
Sybase	Database and data management software
Symantec	Enterprise software
Tibco	Enterprise infrastructure
Verdasys	Archiving and security services
Xerox	Enterprise systems, hardware and services

Autonomy Annex 2: Selected Technology Partners

Autonomy has hundreds of partners. Some provide integration; others provide solution environments into which Autonomy fits. You can find an Autonomy partner by searching Autonomy.com for the "find a partner" service. To provide some perspective on the type of partners working with Autonomy, a selected list appears below.

Autonomy Technology Partners (Selected)

Partner	Key Technology
Accenture	Management and technology consulting
Boeing	Manufacturing
CSC (Computer Sciences)	Infrastructure and services
Canon	Industrial and consumer products
Capax Global	Financial services
Capgemini	Management and technology consulting
Captaris (now a unit of OpenText)	Document management
Documentum (now a unit of EMC)	Document management
Dow Jones & Co.	Publishing
Fujitsu	Manufacturing and infrastructure
Hewlett Packard	Hardware, software, and services for consumers, and commercial entities
IBM	Services, hardware, and software
LexisNexis (a unit of Reed Elsevier)	Publishing
Lockheed Martin	Infrastructure services and manufacturing
Logica	Business and technology services firm
Morse (a unit of 2e2)	Information technology services
Northrop Grumman	Infrastructure services and manufacturing
SAIC	Infrastructure and services

Chapter 3:

Endeca

"Search and business intelligence software..."

A company built on Guided Navigation and an innovative data management system.

Endeca at a Glance

Endeca has a well-deserved reputation as one of the leaders in eCommerce search. The company's technology hook is "Guided Navigation." When introduced, the concept was fresh and offered users of an Endeca-based system a way to spot related content. Today the idea of "facets" and "search suggestions" are part of the standard search experience on Web search systems from Google and from such specialist vendors as Autonomy, Exalead, and Microsoft, among others.

To remain fresh, Endeca has moved beyond searching structured product information. The company offers Web search, enterprise search and business intelligence.

Key Developments

The big news from Endeca in early 2011 is the company's launch of the US Census Bureau American FactFinder Web application.[8]. The system makes use of Endeca Latitude, the company's business intelligence engine. The search-centric application shows how a "search" vendor can support database queries. Users enter a zip code, click an option from a drop down list or a map, and the information appears.

8. To explore this Endeca application, navigate to http://factfinder.census.gov/home/saff/main.html?_lang=en

Endeca's system shines when processing the type of data gathered by the US Census Bureau. Is Endeca a business intelligence system, an e-commerce system, or an unstructured content search system? Endeca asserts that the company's technology can handle these three tasks and more. Like other high-profile search vendors, Endeca has morphed from search system to platform over the years.

Endeca at a Glance

	Basic Information	Option	Comment
License Fee	Begins at about $100,000	Consulting and engineering support	Endeca provides an MBA-like consulting service to help licensees tailor the Endeca deployment in an optimal manner
Search product	Information Access Platform with Guided Navigation	Versions of the IAP are available for pushing, mobile commerce, and business intelligence markets	
Technology hook	Guided Navigation, Hybris, and Page Builder, extensible analytics	Third-party enhancements are available from Lexalytics and other firms	Endeca fashions its platform to deliver a product tailored to a market sector. Additional customization is usually required
Cautions	Scaling and performance can be concerns	64-bit architecture with appropriate resources	Appropriate resources are needed to get the most from Endeca
Selected partners	Endeca has an extensive line up of partners, resellers, and integrators	Endeca's core technology is "wrapped" with additional features	Extensive partnering gives Endeca a rich range of functions and features
Net Net	The system may be stretched to accommodate "big data". The system performs well when index updates, query load, and response time are matched. Endeca's core strength is eCommerce with other applications of the technology built on the firm's navigation engine technology enhanced and extended since the late 1990s.		

In the last 12 months, Endeca has taken three important decisions. First, in a talk I heard in 2006, Steve Papas, Endeca's president, said:

> Endeca has added additional capabilities to its Information Access Platform. Endeca is no longer a search company. We are a business intelligence company incorporating our Information Access Platform and a wealth of new functions to make information actionable, not just findable.7In 2010, Endeca stepped up its marketing efforts for the firm's information access platform. One point of emphasis is the use of Endeca for business intelligence functions enabled by MDEX, the semi-structured database at the core of the Endeca platform. MDEX is, according to Endeca, a "a hybrid search and analytic database designed to bring together every piece and type of information required for making critical business decisions, regardless of its original type, format or source."[9] The

business intelligence product and service line was branded as "Latitude." Endeca was adding to its arsenal discovery and data fusion capabilities. The increased marketing gave Endeca more visibility in a hot sector and put the company into direct competition with such vendors as Business Intelligence, Cognos SPSS, Palantir, SAS, SAP Business Objects, and dozens of other companies chasing outside the traditional search-and-retrieval markets.

Second, Endeca has continued to back its developer program, Endeca Developer Network or EDEN.[10] The purpose of the developer network was to position Endeca as an "information platform." EDEN offers knowledge bases, blogs, and a forum function plus software to registered and authorized users. The idea is that a search system can provide a launch pad for enterprise applications that once were available from third-party vendors. Search in third party applications for customer support or inventory was an add-on, not the central feature. Endeca's developer program supports education, and brings together Endeca professionals, partners, and developers who understand that MDEX can be used to deliver a customer relationship management service based on Endeca and its search-and-retrieval features. Endeca's focus on its technology as a platform was taken at a time when other search vendors were making similar repositioning moves in an effort to address markets where search was not a priority.

Third, Endeca has continued to expand its partner and reseller program. Endeca has a certification program and channel and reseller partners and system integration partners. The technology partners provide some of Endeca's sophisticated functions. Examples include LingPipe's natural language processing technology, Basis Technology's multi-lingual capabilities, and Lexalytics' sentiment analysis functionality. In one demonstration project at the Financial Times, Endeca teamed with Lexalytics to create a new information access service called Newssift.[11] Endeca worked with its partners to create the new service. The key point is that Endeca has continued to invest in its partnering programs because these organizations provide sales lead and sales support, technical expertise, and software that adds features and functions to the Endeca platform.

One important point about Endeca in the last year or so is that the company has not been acquired. Microsoft purchased a competitive platform and then a semantic technology company. Dassault Systems acquired Exalead, another company well-positioned in search-enabled applications for the enterprise. No cash mergers took place with content processing companies such as Attensity and Lexalytics to change their corporate make up. Endeca has remained an independent, privately-held vendor. Neither selling itself or an initial public offering were part of Endeca's activities since the firm's infusion of additional venture capital (see below).

9. More information about MDEX is available by searching Google.com for the string "ENDECA Latitude".
10. EDEN information is located at http://eden.endeca.com/web/guest/home
11. The Newssift project was available publicly for a short period of time.

History

Endeca Guided Navigation was an improvement over key word search. The name of the company is derived from the German word *entdecken*, which means "discover." When the company was founded, search usually meant typing key words or a Boolean query into a search box. Endeca changed that approach. By combining the benefits of Yahoo-like tagging with proprietary data management technology, Endeca could "show" users a view of content. A user could type a query and see suggestions or the start page for the search could be configured to show information by categories or topics. A user of an Endeca system in 1999 did not have to type a query. The approach was fresh, even revolutionary.

Since 1999, Endeca has developed a solid client base in eCommerce, enterprise search, and library applications.[12] Although there has been some staff turnover, Endeca has been a stable company. Convera, once a fierce competitor for Endeca, went out of business. Verity, one of the earlier search vendors, sold out to Autonomy.

To jump start growth, Endeca was rumored to be working toward an initial public offering in 2007, and then in early 2008, Endeca announced that it accepted a capital infusion of $15 million from Intel and SAP. Endeca has previously absorbed about $50 million in prior rounds of funding. The company is privately held. Revenues are rumored to be in the $125 million range.

Since the cash infusion, Endeca has stepped up is partnering activities, expanded its developer support programs, and continued its C-level marketing approach. Endeca, unlike other search vendors, works to present the case for its technology at the chief executive officer, chief financial officer, or similar level in an organization. Consequently, some of Endeca's competitors report that Endeca may not participate in head-to-head procurements.

Product Line Up

Endeca's products are configurations of the company's Information Access Platform tailored to solve particular types of problems.

The flagship is the Information Access Platform itself, sometimes called McKinley. The database architecture that makes Endeca's Guided Navigation or "search without search" method possible is called MDEX.[13] Pete Bell, one of Endeca's senior managers, said that MDEX was

12. The John F. Kennedy Memorial Library uses Endeca to search the Kennedy archives and related content at that institution.

13. The MDEX moniker is one that can generate false drops due to the use of the four letters by other companies unrelated to search and retrieval; for example, the Monaco Dealership Exchange and the MDEX Online information service for health professionals.

> a new class of database that has a flexible data model, analogous to XML. That's complemented by a way to query, slice-and-dice, and summarize that XML-like data.[14]

I think of "MDEX" as shorthand for metadata exchange. As Endeca continues to refine its database components, the basic "plumbing" has been reworked in order to permit faster operation. In the last three years, Endeca has redesigned the core data storage architecture of the engine. Endeca exploits 64-bit memory to make it possible to put needed information in memory, thus reducing latency associated with repeated disc reads. MDEX now supports a distributed architecture. With the revamped engine, Endeca asserts that it has taken a standards-based approach. The idea is that customization and integration are easier and less time consuming.

Major products available from Endeca include[15]:

- The IAP or Information Access Platform. This is the MDEX data management component, content acquisition, Guided Navigation, and search system.[16]
- Latitude is Endeca's business intelligence product. Search and analytics are packaged to provide a more user-friendly way to access structured and unstructured information in an organization. Latitude bundles data integration and a discovery framework in the product.
- The Commerce Suite provides a multichannel, online e-commerce operation with search, Guided Navigation, analytics, and search engine optimization. Endeca offers a "Hybris" function that adds such features as enhanced personalization, presentation, and integration of "You may also like" and other merchandising capabilities. The Publishing Suite supports Product Content Management (PCM); that is, information about products, suppliers, and other product-related information.
- The Publishing Suite can process a range of file types and implement search, page layout, search engine optimization, and "Content Spotlighting", which is a link or object triggered by its relationship to the user's query or the user's navigation in an Endeca result set. A user sees and marks relevant documents or segments of a longer document to make repurposing content or research easier. With its Publishing Suite, Endeca pushed into newspaper, book, and magazine publishers where MarkLogic Corp. achieved significant customer wins because databases like Oracle's and Microsoft's SQL Server were not designed for certain content type processing, search, and retrieval.
- Endeca Mobile Commerce is the Information Access Platform tailored to meet the needs of retailers targeting mobile device users. The product provides page building tools called "Merchandising Workbench with Page Builder."

14. Interview published in Search Wizards Speak at http://www.arnoldit.com/search-wizards-speak/endeca.html
15. With each product are fine-grained "solutions"; for example, for government, Endeca offers intelligence and self-service versions.
16. At one time, the Endeca Data Foundry or EDF used technology from a third part, rumored to be ClearForest (a Thomson Reuters company). The resulting tags or markups are used to build the indexes for entities, concepts, and categories.

The product lineup is packaging and marketing. The products do share some common elements: Web site search, Intranet search, administrative components, and various analytics. The underlying platform, the Guided Navigation, clustering service (called Discovery Tags) and the developer tools are tailored to the needs of each market segment.

Technology

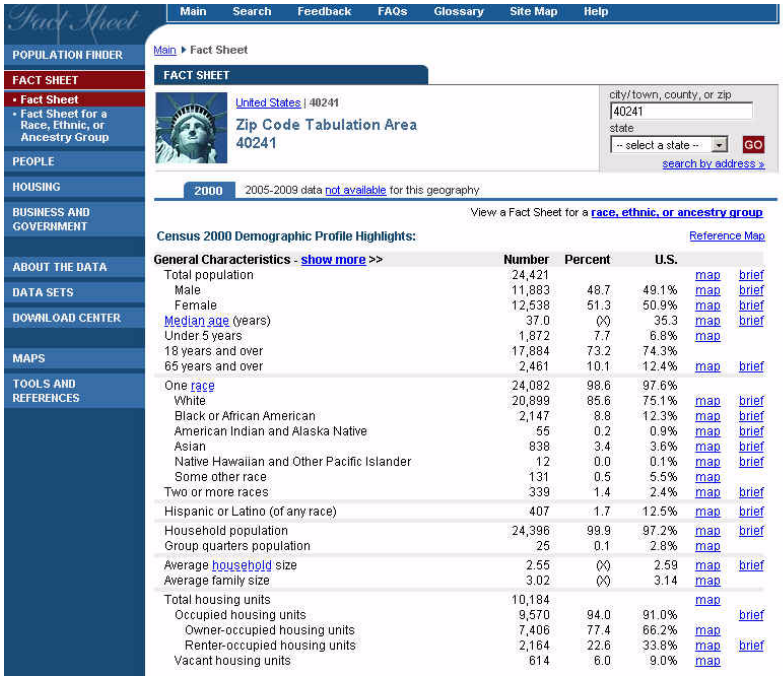
Endeca, like Google, has continued to update and improve its core 1999 technology. Endeca has continues to discuss some of its major technical changes at conferences and on its Web site.

The two major technical shifts were the addition of support for 64-bit architectures and the implementation of an "in memory" method to reduce disk input-output latency. Performance of Endeca's system today is better than it was prior to the implementation of these two technical modifications. The upside of the in-memory approach is that certain functions like computations for advanced analytics execute within minimal latency. The in-memory approach does require that servers have sufficient memory to contain the representations and values. Although the cost of memory has been declining, there are increased hardware investments required as the index grows. Once data must be swapped from disc to memory or from memory to disc, performance is affected, often causing timeouts when under peak load.

Over the last three years, Endeca has successfully enhanced its platform to make it more flexible. For example, analytics functions can be easily upgraded or modified. Third-party systems can be "hooked into" the Endeca system. The Information Access Platform itself can be integrated into other enterprise systems. In short, Endeca has invested time and effort in keeping the platform in step with the needs of its customers and competitive with systems from competitors who have entered the market after 1999.

A glimpse of some of Endeca's technology appears in Endeca's US patents. A useful document to review is US7035864, Hierarchical Data-Driven Navigation System and Method for Information Retrieval, filed May 18, 2000 and awarded on April 25, 2006. The patent document makes clear that Endeca had moved beyond simple key word indexing. The approach can be computationally intensive. Many values are calculated. When new content is processed, some values have to be recalculated. The diagram from US7035864 provides a glimpse of the structure and the various "values" that must be calculated and placed in the index. Endeca's technology incorporates a number of interesting functions. There is a summarizing query mechanism and what Endeca calls a "lexible data model", which refers to meaning-centric nature of the Endeca method.

The "guts" of Endeca's indexing is that content is parsed and tagged (indexed). The Endeca system models data in multiple dimensions, based on "automatic" calculations of relationships. An administrator can control categories and how the data are displayed. The "Navigation Engine" builds the Endeca indexes. Upon processing,



The output from Endeca's system generates a report, not a list of search results. The method works well on fielded data. A link displays a map object and triggers a PDF download or view option for a brochure-type presentation. The queries reflect the Year 2000 census data, not the 2010 census data on February 2, 2011.

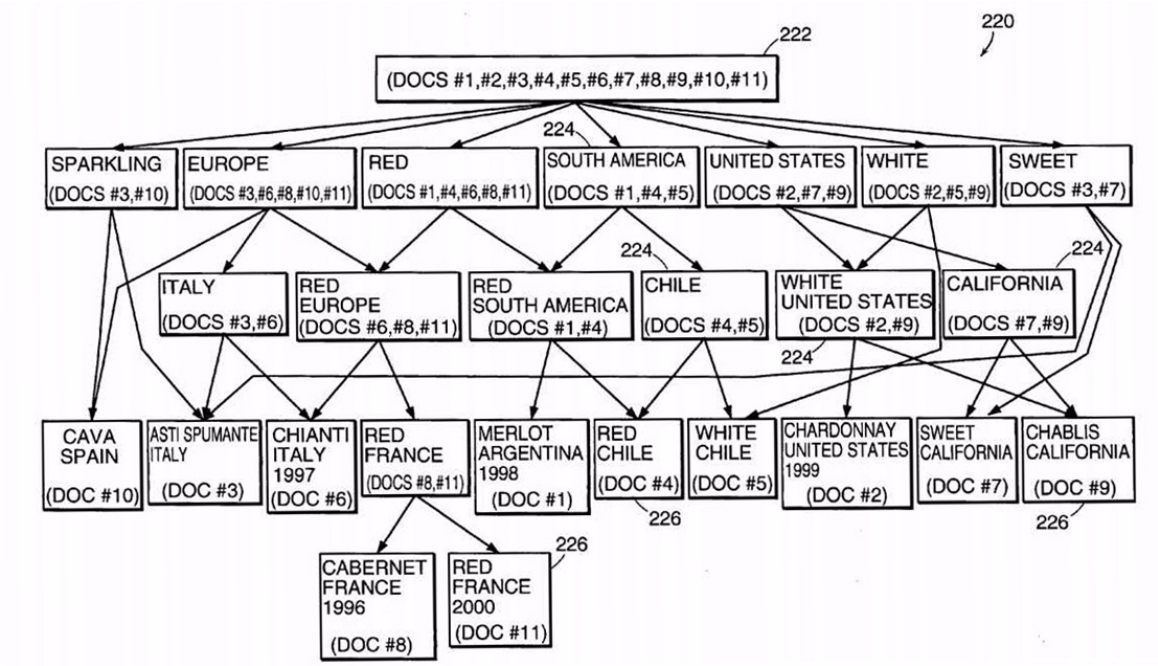
the system calculates the relationship values between objects. The engine takes custom actions for the particular installation from statements in the engine's configuration files. Upon completion of the processes, the system generates the meta-relational index which combines features of traditional databases with technology similar to that used in Google's programmable search engine. .

The database technology embraces a number of technical components in a range of subsystems. There are the Endeca "workbenches". These are software, components, and documentation to make it easy for a licensee to tailor the Endeca platform to suit a particular installation. A Merchandising Workbench permits customization of the Endeca Commerce Suite. For publishers or licensees who require control over presentation of results, Endeca offers its ProFind and InFront administrative tools. These interfaces permit administrators to set certain search options, adjust relevancy ranking and set hit boosting thresholds, and insert and adjust business rules. The work flow component of Endeca distinguishes the company's solutions from some enterprise search products. These settings instruct the content processing subsystem or Data Foundry about the "rules" applied to certain content. Endeca Studio is the interface design component of the system.

Endeca provides filters and adaptors so the content processing subsystem can acquire, process, and tag content. Endeca's system acquires content via crawling, push, file transfer protocol, and traditional ETL (extraction, transformation, and loading). Endeca can process content from content management systems, file systems, SQL databases, enterprise resource planning systems, and traditional office document file types.[17] If a custom connector is required, Endeca provides an appli-

17. Endeca supports such standards as ODBC and JDBC to connect to mainstream relational database systems. The company offers a development framework for creating custom content adaptors.

cation programming interface for developers. The Endeca Content Acquisition System interacts with other Endeca subsystems. One interesting function is provided by the Endeca Data Foundry. A licensee can process XML (Extensible Markup Language). As a result, Endeca can be used in a manner somewhat similar to XML data management systems. The "Endeca markup" is used to build the indexes from words, entities, concepts, and categories. A licensee can use an existing word list if available or rely on Endeca to generate a list of tags and categories.



Modern content processing systems can be computationally demanding. Endeca calculates relationships within and among documents and extracted objects such as entities. The larger the content flows, the more computational resources are required to build the Endeca indexes in MDEX. Updates may require recalculation which may be an issue for some real-time implementations.

Endeca provides APIs that allow a licensee or developer to extend the system via C, C++, scripting languages, and other widely-known methods. With each release of the Endeca system, the company seems to add more customization tools and more graphically-rich presentation components.

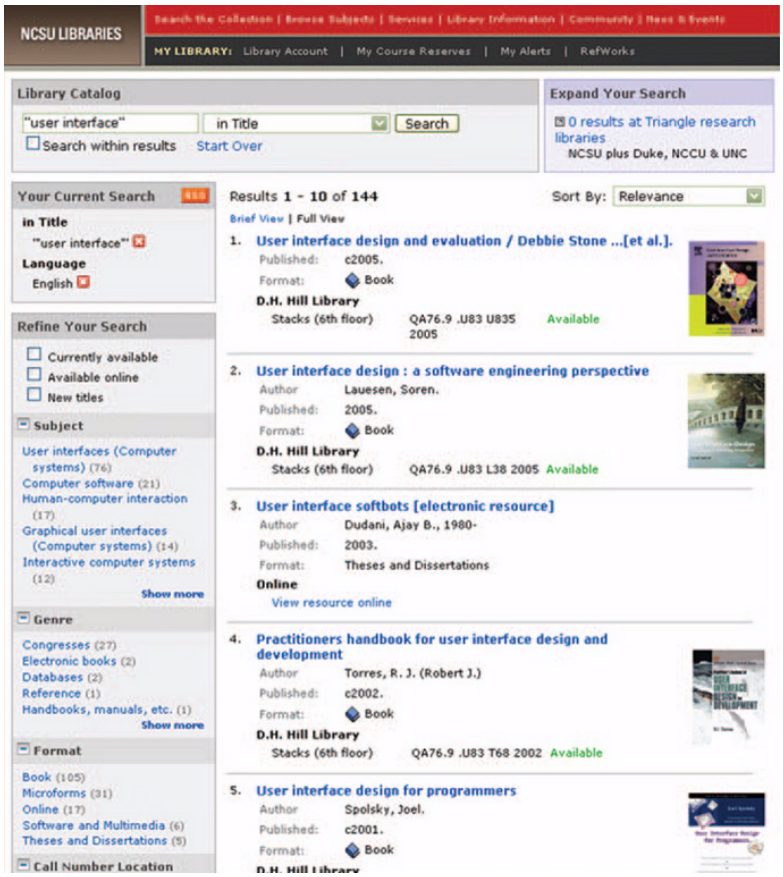
Indexing Highlights

Examples of Guided Navigation characteristics include:

- Orthogonal Dimensions. This means that the Endeca index contains standard index pointers plus special Endeca metadata. The user can view data that are sliced and diced or mixed and matched by concept, term, entity, or other factor such as time. Endeca's engineers describe this as a "meta-relational architecture." For example, a user can navigate by author, then date, then subject, or,

alternatively, by subject, then country, then date. In this way, Guided Navigation dynamically offers multiple paths to every record. More paths means more opportunities to find the specific information required. The order of choice is up to the user. The constraints of navigating down a hierarchy do not exist.

- Hierarchy. The Endeca system does include hierarchies. Instead of being imposed upon data, the Endeca system uses a provided taxonomy such as MeSH (Medical Subject Headings) and models it natively within the Endeca meta-relational architecture. Using this indexing approach, attributes of the dataset can be presented as a hierarchy into which a user can drill down, instead of an undifferentiated list of headings. Peter Bell, one of Endeca's senior managers told Text Mining Review: "This ability to reflect attribute hierarchy is an important element for the user experience. With our system, we eliminate long lists of attributes through the use of logical groupings. We make use of existing taxonomies but bring more flexibility to the user's exploration of content itself.



North Carolina State University relies on Endeca for its catalog search. The main features of the Endeca interface are a search box, results, and point-and-click links to explore suggested content or segment the results by different tags; for example, Subject, Genre, or Format, among others.

Multiple value selections are possible. Endeca's system indexes the discoverable values associated with a record. There is no predetermined data model to limit the number of terms discovered for a document; for example, one record may be tagged with two countries. The next record could be tagged with 22 country names. As a result, any item—category, concept, bound phrase, numeric value, etc.,—can be combined with one or more other metatags.

Metadata used by the Endeca system include:

- Attribute order. The numeric values (date, time, record number, etc.) and other ordering characteristics associated with a record are stored in the MDEX Engine. Accordingly, the values associated with a dimension can be presented to the user automatically by alphabetical or numerical ranges. Alternatively, the order of attribute values can be ranked by prevalence in the current data set. Endeca automatically displays the most likely options available for a particular line of inquiry.
- Precedence rules. Endeca gives licensees precoded rules that can be tailored to control what information is shown to users depending on where they are in the application. For example, when a customer support representative responds to a particular type of inquiry and needs a particular type of information, the Endeca system can prefetch these data, displaying them to the representative as the page refreshes. The search task itself runs automatically, saving time on the call. These rules are important when dealing with sets of heterogeneous records where it is undesirable or impractical for users to browse all attributes at once. For example, an electronic components exchange might have 10 million records that are collectively described by 1,000 dimensions. As a result, application developers may decide to make the most frequently accessed dimensions available in a result set; for example, the component type, manufacturer, and distributor. The secondary attributes like tolerances and temperature ranges might not appear until the user has selected a component type. Application developers can also choose to automate the appearance of the attributes based on ranking rules.

Inherent Metadata

Even though "unstructured" documents like a report prepared in WordPerfect or output as an Adobe PDF file have no structure in themselves, Endeca extracts and stores date, file type, and file size tags to provide useful and manipulable information about a document and add additional structure around the body of a document. These data can be exposed to the user through Endeca's Guided Navigation.

Endeca's text mining processes can process more than 390 different file types. Endeca's system can index content in unstructured repositories holding email or text held in a content management system.

Contextual Metadata

In addition to the document-specific metadata, there is a further class of information that Endeca tags to records during indexing. These metadata include the file structure, including elements of the file path. These objects can be parsed and added to the record as metadata. For example, in an Intranet application, files held in the underlying file system in the "HR" folder can be explicitly tagged as "human resources" and be presented to the user through Guided Navigation. This approach can furnish several layers of metadata in cases in which file structures are hierarchical, ranging from the general source of the record (e.g., the name of the underlying content repository itself) right through to very specific information about the objects in a particular file or directory.

Rules-Based Tagging

During content acquisition from original sources, Endeca can use rules to generate additional tags for a document. The rules can be simple; for example, tagging all documents containing the text “Microsoft” or “MSFT” with an explicit marker such as <Microsoft>. Or the rules can be sophisticated, employing Boolean syntax: for example, <if X AND Y> and <date=June03> add <TAG> for records from June 3 that include both X and Y.

Endeca can use existing thesauri, taxonomies, and controlled vocabularies to increase the speed and accuracy of rules-based tagging.

Statistical Classification

In addition to rules-based tagging, which is itself a type of classification, Endeca employs statistical classification technology through a number of partners who specialize in this area. In the current release of Data Foundry, Endeca relies on statistical techniques based on Bayesian classification to either classify documents based on their similarity to other pre-classified documents or to create a hierarchical taxonomy based on clusters of similar documents.

Classification by Example

When an organization has previously classified documents, or is able to identify documents that can be used as representative examples of a category, these documents can be used as a training set for classifying new documents. The new documents are classified as belonging to one or more existing categories based on their statistical similarity to pre-classified items. The category names are appended as tags in the MDEX Engine and available for Guided Navigation by users.

Automatic Taxonomy Generation

When no previously classified documents exist, the Endeca text mining system can generate a taxonomy that groups related documents in the context of an entire document set. This is a proprietary statistical approach. This Endeca process identifies clusters of documents based on similar word content. The system can create and populate a taxonomy by automatically tagging documents with the prevalent concepts that are significant to their cluster.

Entity Extraction

Entities include names, places, dates, and other words and phrases that establish the meaning of a body of text. Entities are important pieces of information in unstructured documents. Endeca extracts these during its indexing to create a manipulable layer of metadata for a document and a collection. Users can browse or see in a visual display the key content in a particular document or set of documents.

Specifically, Endeca automatically extracts entities such as people, places, and organizations based on a proprietary mix of NLP (Natural Language Processing) and statistical inference. Endeca told Text Mining Report, "Our extraction process is self-training and extensible, so that once an entity is tagged, subsequently all other entities of this type will be identified, extracted, and tagged as metadata." For example, an Endeca customer could tag the chemical compounds mentioned in a number of documents. Then, subsequent documents would be processed so that any referenced chemical compound would be indexed and extracted automatically.

Term Discovery

In addition to entities, Endeca discovers, extracts, and tags terms and phrases. The Endeca system uses a combination of NLP and part-of-speech analysis. In Endeca's Information Transformation Layer, the system identifies noun phrases such as "stock market" and "free market" during document processing. This process is fully automated and requires no human assistance whatsoever.

Dynamic Business Rules

Endeca includes what it calls "Dynamic Business Rules". These are scripts or widgets that can be deployed to highlight specific content or cluster search results. The provided business rules can be modified or used as examples so a licensee can create customized business rules.

The business rules can be used to suggest content or products. The rules can identify records which have specific content or specific records. A retail Web site can use business rules to suggest products for the customer to consider. In an Intranet setting, Dynamic Business Rules can boost specific content so that it appears in the context of a specific query.

The business rules are seamlessly integrated with search and Guided Navigation. These rules can be triggered by the user's profile, time, particular search terms, Guided Navigation choices or more complex combinations of actions.

When a user runs a query, the rules are dynamically selected to provide users with the most relevant content possible.

Endeca's business rules are edited or created in the Endeca Studio tool. The graphical interface allows a developer to create complex rules. By editing Endeca's sample rules, an analyst can create a rule to show items that have been classified as, Marketing documents to any user logged in with a marketing profile.

Strengths

Endeca may be the pre-eminent C-level marketer. Many search vendors focus on procurement teams and mangers in the information technology department. Endeca typically sells at a high level in a prospect organization. Endeca has invested heavily to cultivate the analysts who cover search and content processing. Endeca receives consistently high marks for its system. The company is also quick to modify its positioning to capture new opportunities. The company was one of the first to offer a version of the flagship product for mobile search. The company implements many of the methods taught at top-tier business schools. No other search vendor relies as much on a consultative approach to winning business. When it can get a chunk of real-life data from a prospect, Endeca will build a demonstration of its system using the prospect's own data. Endeca's workbenches and developer network tools are put to use in order to close a deal. Slick marketing and savvy sales professionals have kept Endeca among the top rank of vendors despite the stagnant economic climate.

A checklist of Endeca's strengths would include these items:

- A thought-leader approach to marketing to senior executives[18] backed with user conferences, a useful blog, and trade show appearances[19]
- Proven track record in e-commerce
- A work flow capability that permits routing, stored queries, and specific business actions for reports and content displays under specific conditions
- Growing capability in providing a Guided Navigation approach to answering business information questions positioned as the "simplicity of search with the power of BI"
- Solid line up of resellers, partners, and integrators.

Cautions

Endeca's technology foundation dates from 1998. Despite the significant innovation that Guided Navigation delivered, the internal "plumbing" of Endeca was not designed for the volume of data now routinely flowing through organizations, e-commerce systems, or social content like Twitter messages or Facebook posts. The issues of low-cost scaling and low-latency index updates can become key considerations in the decision between Endeca and another system.

The firm's technology works well as long as the corpus, the index updates, the appropriate latency, and the technical and financial resources are well matched.

18. At one time, Endeca enlisted the support of business guru Michael Porter. At Endeca's user conferences, business thinkers share the spotlight with functional experts from Endeca and its partner lineup.

19. The Endeca blog is at http://facets.endeca.com

Integration can take some time if different partners' technologies are deployed in the Endeca installation. Endeca has made strides in its business intelligence capabilities, but in comparison with solutions from Palantir, i2 Ltd., or SAS, Endeca lacks some of the features and capabilities of these mature business intelligence systems.

Other cautions to research and weigh include:

- The potential for bottlenecks when the volume of content, number of simultaneous users, or near realtime index updates are required. The hardware and infrastructure must be matched to these factors. A mismatch can lead to slowdowns in query processing, results display, and index refreshing.
- Diverse source material—what Endeca calls the "content stew"—can introduce some latency in content processing. Retail companies have ready-made product categories that licensees can map against. A typical company may not have a suitable taxonomy or controlled term list. Automated tagging may require a subject matter expert to interact with the Endeca system.
- Administering the product requires an experienced developer familiar with Endeca's tools and scripting conventions. Endeca has a modern appearance; however, some of the technology dates from 1999. Older technology may be timeless. On the other hand, some older search technologies have significant architectural constraints in certain situations.

Net Net

Endeca is a polished operation. Showcase implementations like Home Depot at http://www.homedepot.com have been funded by organizations willing to invest significant amounts of time, effort, and money.

Endeca has demonstrated that its Guided Navigation invention was an important milestone in search. Endeca's success as a vendor of e-commerce systems has been fueled by the utility of faceted search. Endeca can personalize user experience and showcase products that are hot. These "suggestions" were the first practical demonstration of "search without search;" that is, not making *the user* figure out a query to unlock information.

Endeca has yet to achieve the revenue scale of Autonomy. The company seems well positioned to maintain its present market position. Now more than a decade old, Endeca has built a solid business but has not yet caught fire in the way Fast Search & Transfer garnered a $1.23 billion sale to Microsoft Corp. Nor has Endeca achieved growth similar to that of Autonomy. Endeca could deliver explosive growth if it can expand its share of the e-commerce, search, and business intelligence market.

Endeca Annex 1: Technology Partners

Companies mentioned in the text of the Endeca profile (e.g. Intel, Lexalytics, and SAP) are not included in this table containing representative Endeca technology partners.

Endeca Technology Partners (Selected)

Partner	Key Technology
3Play Media	Transcription of content in rich media
Baynote	Recommendation and social search
Bazaarvoice	Analyze customer feedback
Bee Software	Web crawling, data integration, classification, and management tools
Brightcover	Video platform
Certona	Personalization for eCommerce
ChoiceStream	Product recommendations for eCommerce
Convergence Data Services	Content acquisition and data management for eCommerce and enterprise search
Coremetrics	User analytics
Day Software	Enterprise content management
Demand Media	Content production via contract writers
Hybris	eCommerce platform vendor
i2 Inc.	Supply chain management systems
IBM	Information technology services and hardware
Informatica	Data integration and services
iWay Softare	Enterprise integration
MarketLive	eCommerce platform vendor
MyBuys	Product recommendation system
Nexidia	Audio and video search vendor
Olive Software	Digital publishing solutions
Omniture	Online analytics
PowerReview	Customer-generated reviews technology
PTC	Product life cycle management technology
RichRelevance	Customer personalization and
Searchandise Commerce	Media network for advertising
Semantia	Natural language processing of customer feedback
Silver Creek Systems	Content management for product information
Stibo Systems	Data management technology
Temis	Text analytics
Viewpoints Network	Social technology
Webtrends	Web analytics technology

The lineup of technology partners permits several observations:

1. Endeca has a large number of partners providing platform solutions. This suggests that Endeca integrates into other platforms, not that these partners license Endeca's platform and build their software and systems on Endeca.
2. Endeca uses third parties to handle two increasingly important content types: social media content and video. Endeca, like Autonomy, relies on third party solutions. Autonomy has acquired social and rich media solutions and integrated those into its platform. Exalead has developed and licensed technology for rich media. Exalead developed and integrated its own social media solutions. It appears that Endeca has taken a "Lego blocks" approach to these content types.
3. There are few business intelligence partners. One could use Web Trends and other partners' technology to handle advanced analytics.
4. Endeca is relying on partners to provide natural language processing, semantic methods, and additional classification and indexing tasks. Instead of acquiring a semantic technology company as Microsoft did with Powerset and Google did with Applied Linguistics (formerly Oingo), Endeca has embraced partnerships.

Our view of Endeca's approach is neutral. Endeca's approach reinforces our impression that Endeca tailors a solution to meet the needs of its customers. Instead of relying on one approach, the company can assemble what is required and then in a manner similar to a blue-chip consulting company orchestrate the implementation of the system. The advantage to this approach are numerous. With some Endeca technology dating from 1999, the approach makes it easier to keep Endeca in step with its competition, particularly in ecommerce.

Endeca Annex 2: Reseller and Integration Partners

Companies mentioned in the text of the Endeca profile (e.g.Intel, Lexalytics, and SAP) are not included in this table containing representative Endeca technology partners.

Endeca Reseller and Integration Partners (Selected)

Partner	Market Space
360i	Digital marketing
Accenture	General management consulting
Acquity Groupo	Digital marketing
Arvato Systems	Information technology service
BGT Partners	Interactive marketing
Blue fish Development Group	Documentum and content managements services
Broadstreet Data Solutions	Consulting
Buchanan & Edwards	Information technology professional services
Business Edge Solutions	Consulting
Carahsoft Technology Corp.	US government information technology service
Cognizant	Information technology services
Early & Asscoaites	Indexing consulting
Ecomplexx	Enterprise solution
Enterprise Solution Proviers	Information technology consulting
Envisa	Consulting for “e-business”
Gorilla	Web design
Infosys Technologies	Information technology solution
Ironworks	Consulting
Javelin Group	Euro-centric consulting
Kalypso	Consulting
Keyrus	Enterprise performance management solutions
LBi	Euro-centric full service marketing
Mahindra Satyam	Information technology solutionr
Marubeni Information Systems	Solutions
Molecular	Internet services
NuWave Solutions	Internet business applications
Orckestra	Electronic business solutions
Real Decoy	Consulting services
Resource Interactive	Digital marketing
Rosetta	Web marketing
Sapient Interactive	Web marketing
Sedona Technologies	Endeca implementation services
SkillNet Solutions	Electronic business consulting
Solutions Made Simple	Data management solutions

Sulzer GmbH	Solutions integrator for automotive and financial services
Tackit Knowledge	Enterprise solutions
Tata Consultancy Services	Information technology solutions
Thankx Media	eCommerce consulting and solutions
Virtusa Corporation	Information technology services
Wipro	Information technology solutions

The lineup of selected reseller and integration partners makes it somewhat easier to understand how Endeca sells licenses. The company relies on partners, many of which are general or technology consultants and implementers. Endeca, therefore, gets "baked in" to the solutions that these partners recommend to their clients. Other warranted observations are:

1. Endeca has partners in North America, Europe, and Japan.
2. A sale becomes the primary responsibility of the reseller or partner or a joint effort with Endeca.
3. Endeca's marketing approach makes use of various types of high-level contacts. These contacts via "evangelists" or high-profile MBA-type experts enable "big picture" presentations with the details of a particular solution becoming the responsibility of the reseller or partner who will implement the Endeca system.
4. The firm has a number of "digital marketing" and e-commerce partners which suggests that Endeca has a focus on this particular use of its platform.

Endeca's approach has been effective, but the company has not been able to generate the type of initial public offering interest some other search or content-processing companies did. Endeca, although successful, may be bumping into a "glass ceiling" in terms of growth. As more partners sign up for Endeca, Endeca's sales "touches" increase, but the company has yet to match the revenues of Autonomy.

Endeca Annex 3: Selected Patent Documents

US7035864, Filed May 18, 2000, Granted April 25, 2006, Hierarchical data-driven navigation system and method for information retrieval

US7062483, Filed October 31, 2001, Granted June 13, 2006, as Hierarchical data-driven search and navigation system and method for information retrieval

US7325201, Filed October 16, 2002, Granted January 29, 2008, as System and method for manipulating content in a hierarchical data-driven search and navigation system

US20050038781 Filed September 8, 2003, Published February 17, 2005, as Method and system for interpreting multiple-term queries

US20060053104 Filed November 8, 2005, Published March 9, 2006, as Hierarchical data-driven navigation system and method for information retrieval

US20070106658 Filed November 10, 2005, Published May 10, 2007, as System and method for information retrieval from object collections with complex interrelationships

US20080133479 Filed November 30, 2006, Published June 5, 2008, as Method and system for information retrieval with clustering

US20080134100 Filed October 31, 2007, Published June 5, 2008 as Hierarchical data-driven navigation system and method for information retrieval

Chapter 4:

Exalead (Dassault Systèmes)

At the junction of search and database technologies... from the leader in search-based applications

A vendor with technology that has redefined enterprise information access and applications across business functions...

Exalead at a Glance

Exalead provides high-performance search and semantic processing to organizations worldwide. Exalead specializes in taking a company's data "from virtually any source, in any format" and transforming it into a search-enabled application. The firm's technology, Exalead CloudView, represents the implementation of next-generation computing technology available for on-premises installation and from hosted or cloud services. Petascale content volume and mobile support are two CloudView capabilities. Exalead's architecture makes integration and customization almost friction-free. The reason for the firm's surge in the last two years has been its push into the enterprise with its search-based applications.

The idea of an enterprise application built upon a framework that can seamlessly integrate structured and unstructured data is one of the most important innovations in enterprise search. Only Google, Microsoft, and Exalead can boast commercial books about their search and content processing technology.

Exalead at a Glance

	Basic Information	Option	Comment
License Fee	Begins at $30,000	On-premises and hosted options	Custom price quotations are available
Search product	CloudView is the core platform, CloudView 360 adds semantic processing, info mashups	Quantitative analytics, rich media search, and support for social media are available	Advanced technologies are viewable at http://labs.exalead.com/
Technology hook	Blistering performance, semantic processing and seamless scaling set Exalead apart	The company offers full "cloud" support	Exalead's technical research foundation is comparable to Google's
Cautions	None	Exalead responds to customer needs	The company is based in France with operations in the UK and US
Selected partners	Exalead has partners in major European countries and Asia	Partners can assist with integration and customization of the Exalead platform	The company partners with Dassault Systèmes, Capgemini
Net Net	Exalead is a top-notch engineering firm. The firm's 64-bit architecture delivers excellent performance, customizable semantic processing and painless scaling. Exalead's documented application programming interfaces and solid support for industry standards makes it possible for licensees to customize and integrate the Exalead functions across an organization. The firm's ability to handle billions of documents in Exalead's Web search service is demonstrated at www.exalead.com/search, a Google-style Web index.		

The 2011 deep-dive into Exalead was written by Dr. Gregory Grefenstette and Laura Wilbur. My analysis of Exalead was written without dependence on this authoritative discussion in *Search-Based Applications: At the Confluence of Search and Database Technologies*.[20] The principal difference between the information in the Grefenstette and Wilbur book and my analysis is that the authors do not pinpoint Exalead's innovations as one of the most important innovations in information retrieval as I do.

Based on the information available to the study team, Exalead now offers one of the highest performing and most advanced systems available for enterprise deployment. Now a subsidiary of Dassault Systèmes, Exalead has access to the technical, financial, and human resources of its parent company. The impact will range from expanded operations to commercialization of multi-dimensional information retrieval methods. With comprehensive support for structured and unstructured content, Exalead is certainly one of the leaders in the search-enabled applications it pioneered. IBM, Oracle and SAP treated search and retrieval as an add-on or a utility; Exalead made search the foundation of solutions that reduce costs and improve the leveraging of an organization's information assets.

20. Dr. Gregory Grefenstette and Laura Wilbur, Search-Based Applications: At the Confluence of Search and Database Technologies. Morgan & Claypool, 2011. ISBN: 9781 6084 5507 2.

François Bourdoncle, one of the founders of Exalead, said in April 2011:

> The reality is that "search" comes in two flavors: search technology and search engines. Most of the time, when someone says "search", he or she really means "search engine", which is an end-user application in its own way. But search engines are rarely mission critical applications. On the other hand, search technology is an enabling technology for a wide range of mission-critical applications: CRM, ERP, logistics, procurement, PLM, you name it. And of course, mission-critical applications are significantly more compelling to our customers than the average enterprise search engine used to index an Intranet or standalone SharePoint repository. *Mission-critical applications are where the actual value is created in companies, and where investments are the most productive. That is precisely why we invented Search-Based Applications (SBAs) to use search technology as a way to empower mission-critical enterprise applications with search features and ease-of-use.* (Emphasis added.)

Exalead's combination of engineering innovation and flexible architecture put it in the top tier of companies enabling customer support services, business intelligence, and mobile findability applications.

Key Developments

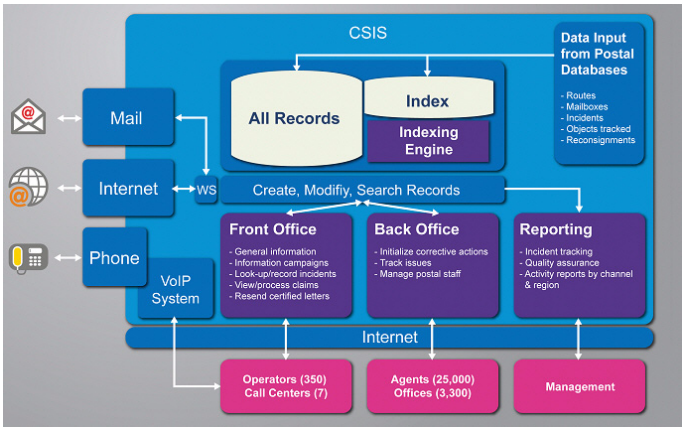
The major development for Exalead was its acquisition by Dassault Systèmes in mid-2010. Terms of the deal were not disclosed, but chatter about one of world's top engineering firms buying Exalead estimates the buyout to be anywhere from $75 million to $125 million. The Exalead technology would continue to be available for licensing, but Dassault took steps to use Exalead's technology across Dassault's various business units. Dassault now divides its brand portfolio into two groups - digital creation and digital aggregation & publishing. This second product group is a strategic move towards broader enterprise solutions, within the manufacturing function and also in the front office. Along with Exalead, that second group includes the 3D SWYM collaboration technology and 3DVia SaaS digital authoring tools. In our view, both firms benefit from this deal as do the licensees of Exalead's and Dassault's technology.

In the last 12 to 18 months, Exalead has made significant strides in building its customer base in Europe, North America, and the Asia Pacific region. Four developments warrant comment.

First, the company hired Dr. Gregory Grefenstette as its Chief Science Officer. Dr. Grefenstette, who received his Ph.D. in Computer Science from the University of Pittsburgh, is a recognized expert in information retrieval and natural language processing fields. He authored the 2010 *Search Based Applications*, a study of search technologies and their use in integrating business processes with search and retrieval.[21] He possesses 15 patents in the U.S. More recently, Dr. Grefenstette worked as a senior expert at CEA (the Atomic Energy Commission). Before that,

21. Gregory Grefenstette and Laura Wilber, *Search-Based Applications: At the Confluence of Search and Database Technologies* (Morgan & Claypool, 2011). ISBN: 9781608455072.

he worked as an engineer at Honeywell Bull and as a principal scientist at Xerox's research and development arm and Clairvoyance Corporation, a spinout from Carnegie-Mellon University. Dr. Grefenstette joins Exalead's large pool of engineers, computer scientists, and information retrieval experts.



A block diagram of a customer service system based on the Exalead CloudView platform. Notice that instead of processing a single source of content, the Exalead approach makes any information available within the user's "view" of a work process. © Exalead S.A., 2011

Second, Exalead announced a significant new variation of its CloudView product line, which is called CloudView 360. Built on the core CloudView search engine, CloudView 360 adds three modules: [a] Exalead's semantic factory, [b] a highly-configurable mash-up builder, and [c] trusted queries (a way to provide search enrichment in an intuitive manner). The semantic factory is modular pipelined data processing that enhances CloudView's semantic processing abilities. Enriched metadata, relationships, classification and sentiment analysis are implemented without bottlenecks by the semantic factory. The mash-up builder is a graphical tool for prototyping and developing search-based application interfaces. Exalead's suggested queries feature is a new navigational facility that provides instant "perhaps-look-at-this" feedback to users based upon underlying data relationships discerned by the Exalead content processing engine. CloudView 360 is now Exalead's most comprehensive platform for building search-enabled applications.

Third, Exalead expanded its Exalabs to showcase specific information access innovations. The company implemented dynamically-updated faceting and dynamic entity mapping across disparate collections. Also new to Exalabs' public demonstrations are Voxalead, Wikifier, and Tweepz.[22] Voxalead processes rich media, indexes the audio content, and makes the source searchable. A query in Voxalead returns a link to the exact point in the video where the relevant content appears. Wikifier demonstrates Exalead's ability to process additional content and map it to an organization's own corpus of content. A named entity such as a company or

22. These technologies may be explored at http://labs.exalead.com/.

product can automatically be enriched with information from the licensee's content and third-party content from the Web or a commercial source. The synthesis is performed automatically and with no slowdown of the Exalead content or query processing subsystems. Tweepz showcases Exalead's ability to process content produced by a social network such as Twitter or Facebook. A query for a person or title returns links to messages and documents about that person, place, or thing. Each of these innovations "snap in" to the Exalead platform. The importance of this approach is two-fold: quick deployment and sharp reductions in the cost of adding new functions and features.

Fourth, Exalead landed a number of high-profile accounts for its search-enabled applications, including the hotly-contested World Bank information retrieval project. The thread running through Exalead's "wins" is that search fits into the licensee's existing enterprise applications. In several cases—for example, for the French postal service and a global shipping company—the CloudView enabled application integrated several standalone enterprise software products. Packages can be traced with near real time updates and inventories with needed information available via a mouse click or standard query.[23] The Exalead implementation assembles the structured and unstructured data and presents the needed information in a report or graphical display. Exalead's platform delivers aggregated and integrated information within a work process context. No training is required to use Exalead's enterprise solutions.

I want to highlight Exalead's push into data fusion, which is rapidly becoming the key function for business intelligence professionals. Exalead's system can normalize a wide range of structured and unstructured content. The system can deliver presentation-grade reports. In addition, the system automatically generates data for mobile devices. A user can obtain a report and display it on a mobile phone. The information in the reports is processed and assembled in near real time. Latency across peta scale flows of information is measured in minutes.

History

The founder of Exalead—François Bourdoncle—was a member of the AltaVista.com development team. The idea for Exalead took shape at the same time Google was hiring AltaVista.com engineers. In 2000, he set up Exalead S.A. to be "a next-generation enterprise search engine." His goals were to achieve greater scalability and higher performance than the AltaVista.com system delivered. His vision was a standardized search infrastructure engineered completely for 64-bit systems. The Exalead system was engineered from its foundation to be unified, scalable, extensible, and real time.[24] Unlike most search systems, Exalead was coded

23. Latency across multiple enterprise applications is measured in minutes with freshness of the index typically below 12 to 15 minutes for a large, multi-location, global organization.

24. Many vendors use the phrase "real time" to describe a search system. There are many points of latency in any modern system. Therefore, the phrase "near real time" is a reminder that latency exists even in super-computer grade installations like Exalead's.

for 64-bit processors. When 64-bit chips became available, Exalead's system delivered customers an immediate benefit as older 32-bit servers were replaced. [25]

Mr. Bourdoncle said in an interview:

> Most people don't know what they are looking for, though they can recognize what they need when they see it. Our system lets users set out on a quest with a simple, less than ideally formed key word search. It then takes them by the hand and helps them accurately locate information, or fruitfully explore related content. What's more, it offers multiple point and click paths to the same information, so users are less likely to miss that golden nugget they're seeking. And it helps them keep their favorite sources a click away.

By mid-2004, Exalead was capturing key accounts in Europe and North America. Exalead rose quickly to the top tier of information retrieval system vendors at a time when Convera, Delphes, Entopia, and Fast Search & Transfer were experiencing increased market friction. Exalead's platform and flexibility were, according to Mr. Bourdoncle, instrumental in Dassault's decision to acquire Exalead. With the Exalead CloudView technology, Dassault has replaced its other third-party enterprise search software and traditional enterprise applications. In addition, Dassault's engineers began in the fall of 2010 to develop next-generation products and services on CloudView. One type of innovation that CloudView makes possible is augmented reality. An application provides the user with layers of information that display on a mobile device or other display mechanism in near real time.

Among the company's hundreds of high-profile customers are PriceWaterhouse-Cooper, Michelin, American Greetings, and Jeffco (a global logistic company).

Product LineUp

Exalead offers a single product line, CloudView. This is the platform upon which Exalead's applications run. There are two hosting options of CloudView.

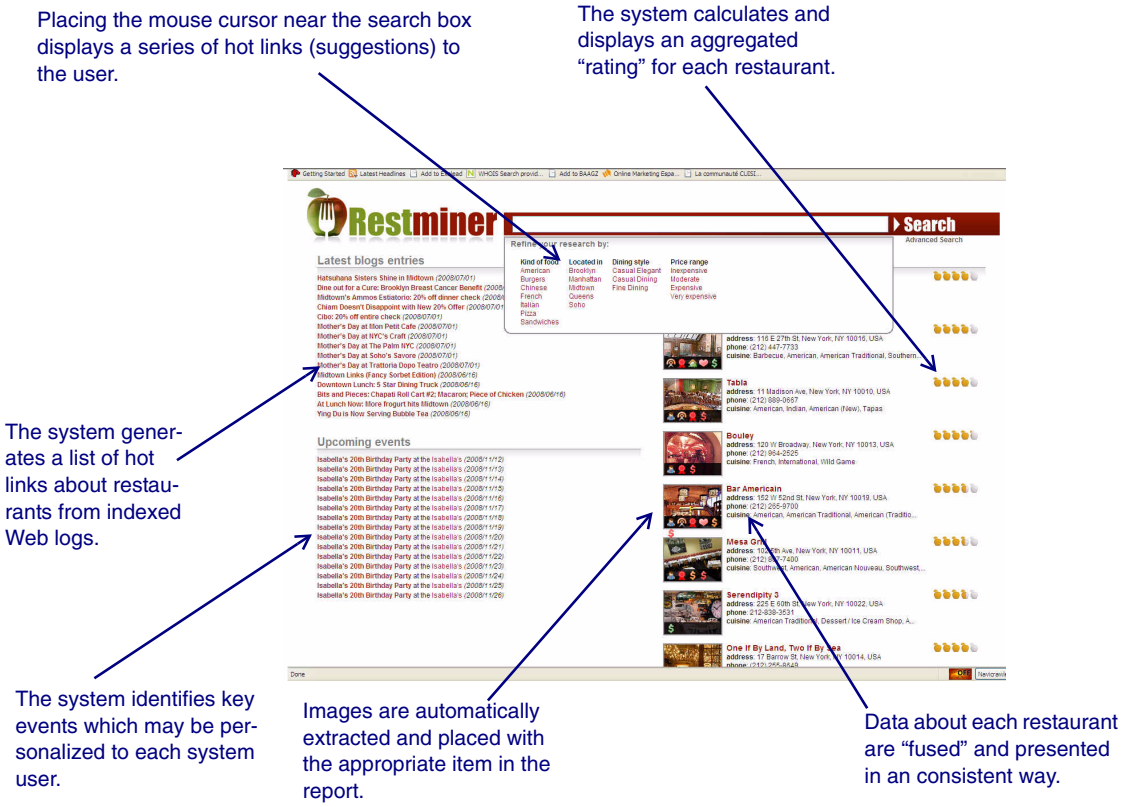
One, is the on-premises installation of the Exalead CloudView platform. Upon installation, the system collects unstructured and structured data from any source, in any format and in any volume, and automatically transforms it into a single structured information resource.[26]

The idea is that this "content representation resource" continually adapts as information and data are processed. The resource can be searched or used in the development of search-based applications (SBAs). Deployment of the on-premises CloudView system requires several days or weeks, depending upon the specific requirements of the licensee. The system makes Web-style search available across structured and unstructured data. The on-premises system features graphical

25. Exalead was one of the first, if not the first, 64-bit, next-generation massively parallel information retrieval systems for the enterprise.

26. CloudView includes graphical administrative interfaces. Administrators can also configure the system by editing configuration files.

administrative interfaces and connectors for standard files and file systems as well as for proprietary systems content such as Lotus Notes. The CloudView platform includes a robust analytics capability for system analysis and for preparing "business intelligence" type reports and outputs for users. CloudView supports Intranet search, Web site search, and e-commerce.



The second is the Exalead On Demand service delivering CloudView functionality from Exalead's data centers. On Demand can be configured as a private label search engine for either general purpose content similar to that indexed by Bing.com or Google.com or vertical content similar to that indexed by travel and hotel and specialty chemical services. On Demand supports social search and user-generated content on public services for blogs and short message systems as well as corporate social content generated by employees for an organization.

Each implementation of CloudView can add components to process audio and video content and petascale flows of social content such as Twitter, and to perform automatic report generation. CloudView can acquire content from Web logs and other social media sources. The user's query can then be answered by automatically generating a report that combines sentiment, numeric data, and third-party content. Michelin and a Canadian mobile operator use CloudView as a research, authoring, and content assembly system. In these types of applications, search is not an add-on or afterthought. Search becomes the defining metaphor of the application and

infuses each action that requires finding an exact item of information needed for a particular task, process, or activity.

Technology

Exalead has developed what it calls “an enterprise-class information processing platform.” Like Autonomy and Fast Search & Transfer, Exalead knows search-and-retrieval is no longer enough to make a sale. The customer needs reassurance backed by a technical architecture that allows search to be an application platform. Search has become the gateway to doing work. A search system, according to Exalead, must allow information to be in one index, and to be instantly findable, and the system must be sufficiently flexible to gracefully incorporate new features and applications.

Content Acquisition and Processing

Exalead CloudView uses named entities automatically extracted from indexed documents and hierarchical metadata, or categories. Past queries by all users can also be provided as suggestions.

Trusted queries expose this metadata in an “assisted navigation” presentation. This assisted navigation system has been patented by Exalead in Europe and the U.S. Licensees can customize the interface to be as simple as a search box or implement a richer interface.

Exalead's approach does not require dictionaries or human intervention (although those approaches are supported). The extracted metatags include file type, author, date, language, and similar document attributes. The system performs on-the-fly categorization. The approach yields folders containing related documents. The effect is somewhat similar to categories generated by Vivisimo's system.

You can see examples of the Exalead content processing outputs at www.exalead.com. The results page provides thumbnails for each document, a feature first introduced in a primary form by Girafa in 1999, and it provides content previews with search term highlighting. The Exalead system suggests related terms providing “assisted navigation” once the original query results list has displayed. Exalead offers one-click filtering for results by, for example, site type (e.g. blogs, forums), language, or file type (e.g. Adobe PDF).

To summarize the content processing functions, the system provides:

- Natural language processing; language detection, sentence boundary recognition, stemming, lemmatization, part of speech tagging and other processes
- Categorization of documents by the available tags. Tag metadata is created by a variety of techniques.
- Dictionaries or word lists aimed at discovery for the corpus; for example, related terms or 'See Also' suggestions.

Exalead points out to the author team that Exalead's public search engine has an index of 8 billion web pages, claiming it is the Web's third largest.

Clustering and Facets

From its inception, Exalead's content processing strategy has been shaped by Mr. Bourdoncle's drive to engineer a scalable infrastructure that could be expanded without the mushrooming costs associated with traditional server architectures and enterprise search.

The infrastructure engineering, based on information available to the author team, shares some broad similarities with the AltaVista.com approach.[27] Google and Exalead appear to have somewhat similar philosophies regarding banks of commodity servers running Linux with some special tweaks. Like Google, Exalead is a mathematics-centric company. There are some linguistic operations, but the core of Exalead CloudView is algorithmic. The Exalead system runs on 64-bit processors and features a "plug in" architecture to allow fast scaling using commodity components. The application workflow is wholly multi-threaded to take full advantage of modern multi-core processors.

Exalead operates its own server farms so customers can use the Exalead system as a managed or hosted service.[28] If you want to have a local installation of Exalead, you can obtain an on-premises license. Mr. Bourdoncle told the author that his engineers have focused on reducing the number of servers typically required to process content in high end applications.

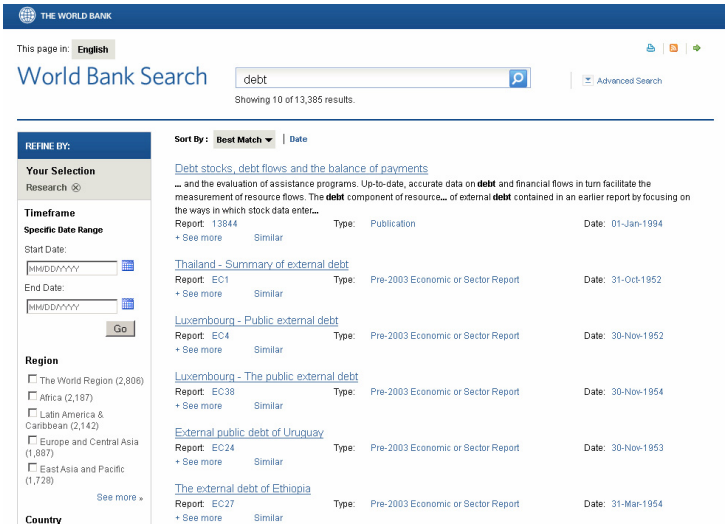
Real Time Processing

Exalead CloudView leverages its computational efficiencies to generate "real time dictionaries"; that is, word stemming, identification of word groups (bound phrases like *White House*), and thesauri or controlled term lists that can be implemented in automated processes. Updates are incremental which further reduces latency and helps ensure the "freshness" of an index or indexes in the Exalead environment. New content is processed in seconds and becomes available to applications or directly to users almost immediately. The Exalead system approaches real-time content processing, although any computing system involves some latency due to network factors external to the core system. The system comes with file processors that can process most common file types. One feature of Exalead's approach is that

27. AltaVista.com was deployed by Digital Equipment Corp. as a demonstration of the firm's next-generation processor and its unique memory management capabilities and multi-threading architecture. After the acquisition of DEC by Compaq and then of Compaq by Hewlett Packard, the AltaVista.com system was effectively an orphan. Google and Exalead sprang from the ground plowed by the DEC engineering team which Mr. Bourdoncle and many Google professionals worked in the 1990s.

28. Exalead offers a cloud-based solution to organizations seeking search enabled applications.

when content is processed, the system also automatically recognizes the language of an information object.



The results display for the Exalead World Bank implementation shows point and click “refinements” to make it easy to narrow content. Hot links display information by region or country. Each result provides one-click access to similar documents as well as the most relevant documents that satisfy the user’s query. Special features like data sorting are one-click operations.

One CloudView operation that is of great value in real time implementations as well as more traditional search-and-retrieval operations is event detection. Exalead technology for entity extraction and categorization enables the system to “type” or “flag” certain information found in unstructured text. The core linguistic functions “recognize” different forms of the same event or action. The system can find syntactically defined structures (such as noun phrases). CloudView then applies entities in order to deliver event detection. Event detection makes it possible for Exalead to answer such questions as *Who did what to whom?* and *What happened where?*

“Snap-In” Design

Exalead has been designed to “snap in” to existing enterprise architectures. Exalead supports most common client-server systems, ranging from branded HP 64-bit servers to commodity Intel and Windows systems and Linux / Unix operating systems.

The core system is written in C, but uses Java wrappers and XML input and output data structures.

At any time, Exalead can expand the functionality of the system. For example, Exalead offers modular components that span social media and analytics, semantic methods and support for existing controlled term lists, automatic indexing and complex workflows. A licensee can process real time video, convert the spoken

words in a rich media object to parsable ASCII, index the content, and return search "hits" or other outputs. The Exalead link to rich media sends the user to the exact point in the audio or video file at which the fact or item germane to the report or user query is located. No serial viewing of the entire program is required.

Indexing Highlights

Exalead CloudView combines natural language processing techniques and web technologies to provide a software platform that exploits business information, both structured and unstructured. The Mining-of-Text open architecture provides flexibility to ensure business applications combine the optimal linguistic features in an optimal sequence. Exalead CloudView provides dynamic and static linguistic resources used by natural language processors (NLP) and semantic processors.

NLP

Natural language processing, as implemented in Exalead, embraces a number of technical processes.

One of Exalead's most interesting capabilities is its NLP tokenization subsystem. Tokenization and normalization converts the information object to a representation that can be manipulated by other Exalead processes. An application programming interface is provided so that the tokenization and normalization functions can be tailored to specific client requirements. While processing a query, Exalead CloudView performs a phonetization of each token based on a set of phonetic rules (for example *vouature* matches *voiture* in French. In addition, at the time of query processing, the system performs an approximation of each token based on a set of approximation rules; for example, *colour* matches *color.*

Other NLP functions "under the hood" of CloudView include:

- Exalead's system performs stemming or lemmatization. The "root" of each word integrates the Exalead language model in order to make enhanced tagging and other operations possible. Exalead supports dozens of languages so that indexing and word sequences are available to the system. At this time, the stemmer supports 15 major languages, including the major Romance languages.
- The NLP system performs part of speech analysis that can perform disambiguation. Exalead supports analysis of sample documents in order to generate automatically a knowledge base so that particular words and other content features are available to the Exalead system.
- In addition to language detection, the system automatically generates word and word co-occurrences dictionaries. These knowledge bases can be used for spell checking to improve the efficiency of search and by other functions within the Exalead system to permit "fuzzy search" or content analytics.
- As the Exalead system processes content, the system gathers statistics about word use for each language. A licensee's linguistic resources are tailored to the

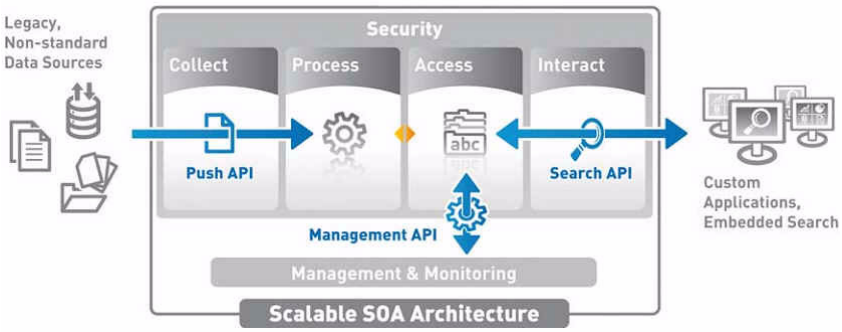
specific business information within an organization. Users and other systems can, therefore, use the "language" and jargon of the organization, which contributes to ease of use and system performance.

Semantic Processors

The semantic processes within the Exalead system permit CloudView to understand the "context" of the processed information. A language makes "sense" when words, sentences, paragraphs, and entire documents are "understood" in terms of certain cues and signals. Exalead makes use of:

- Entity extraction so that a phrase such as "White House" is understood to represent a specific building in Washington, DC and "white house" represents the color of a residence. Native speakers of a language easily make such distinctions without conscious effort. CloudView includes ready-to-use named entity extractors for people's names, geographic locations and company and organization names. Extractors for other kinds of names entities or other languages can be easily plugged into this system module.
- CloudView makes use of what the company calls "ontology matching." While processing a corpus of text, Exalead CloudView supports the extraction of entities or concepts from unstructured data based on an ontology. For example, you can extract the list of company employees and services in all documents.
- Via "fuzzy" ontology matching, CloudView makes it possible for other processes or directly by a user to run a Boolean query between or among concepts. For example, the query string *The Bill & Melissa Foundation* will also return *Bill Gates Foundation* and *Gates Foundation*. Closely related to this "fuzzy" capability is CloudView's detection of related terms. These related terms can be extracted and processed by other Exalead functions. To associate the data with the source document and other documents processed by the CloudView system. These data can be used in a number of ways, including the refinement of the initial query or as items in an interface which permits expanding or narrowing a particular query.
- The CloudView system also performs categorization, based on a user-defined hierarchy or automatically. In the absence of a formal taxonomy or controlled term list, the CloudView system can "learn", create, and apply categories automatically. A training step is recommended to build a library for the automated categorization process.
- Exalead performs clustering of documents. If there is no prior hierarchal structure available to the system, CloudView finds and groups similar documents. The system can also detect different versions of content and perform deduplication operations if required.
- CloudView generates tags that make it easy to provide users with hot links to related or suggested content. This function is often described as "faceted search". Other processes or users can make use of Exalead's facet function.

Among the options are document type, document source, language, and specific topics.



Exalead implements a scalable services oriented architecture. Connections to the CloudView system permit easy integration with on-premises or cloud-centric applications.

The Exalead system uses these different functions in various combinations in order to disambiguate unstructured and semi-structured content.

Strengths

First, the system can be configured to provide at-a-glance dashboards that immediately present relevant information. Users don't need to start with a blank screen through which they're forced to enter a query. The default interface can display information relevant to a user so the “dashboard” shows the user what’s available, what’s current, and what’s important in one or more dataspaces. The benefit of this capability is that CloudView provides operations associated with standalone business intelligence systems. One important plus is that the Exalead system makes use of both structured and unstructured data. As a result, the Exalead system delivers a 360 degree view of the organization based on its information assets in databases, electronic mail, proprietary third party customer relationship management and enterprise resource planning systems and many other data sources and repositories.

Second, the basic, out-of-the-box CloudView system breaks down the walls between structured and unstructured data without compromising security. Access control is used to ensure that individual users see only information to which they have access. CloudView processes, indexes, and makes available answers, not laundry lists of results. The user sees appropriate information in a easily-digested form, on a dashboard, in a report, or snippets of the part of a document germane to the user’s query. Unlike traditional business intelligence systems, no analyst must intervene between the user with a question and the business intelligence system data where the answer resides. An Exalead system permits fully interactive, *ad hoc* queries by users.

Third, the system eliminates the time consuming information discovery sequence that frustrates more than two thirds of information access systems’ users. Laundry

lists of irrelevant results create work. Exalead breaks the expensive, inefficient cycle of formulating a query, scanning a laundry list, opening documents, and hunting for the fact or datum the user needs. CloudView's operational intelligence smooths and streamlines integrating information into the decision making process.

A checklist of Exalead's strengths includes:

- An architecture that delivers high throughput, seamless scalability, and components that "snap in" to the platform
- Ready-to-run components to acquire and index rich media
- A deep bench of experts for customer support and engineering services
- Partners worldwide who are prepared to tailor search-based applications for government entities, commercial organizations, and not-for-profit entities.

Cautions

Several years ago, I reported that Exalead's heavy investment in research and development would require significant resources.[29] Now that Dassault owns Exalead, this consideration is irrelevant. Dassault is one of the premier engineering and technology firms in the world with robust finances and a distinguished track record of technical innovation, research, and development in information systems and other engineering disciplines.

Exalead incorporates a number of technical methods that are proprietary. Although the firm documents its application programming interfaces and supports open standards, Exalead is tight lipped about some of its technical "magic." There is strong support for standards within the CloudView system, but the performance and scalability of the Exalead architecture uses many methods specifically developed to handle exascale flows of content. In today's average organization, the digital content is doubling every three months. Specialized engineering is required at the "bare iron" level to permit low-cost scaling using commodity hardware and today's multi-core processors and high-speed storage methods.

Licensees should expect to devote effort to customization. The default interfaces provide a good springboard for developing e-commerce, customer support, and business intelligence systems. The default screens are clean and functional; licensees typically edit the style sheets and enhance the user-facing screens. The administrative interface is quite comprehensive, certainly ranking with the best blend of power and ease of use. However, these screens are also "clean" and straightforward. Some of the results presentation templates implement a number of Exalead features; for example, date limits, NOT logic, and entity extraction. The licensee will want to implement only the features needed by its users. Turning on every semantic and filtering function can overwhelm some users.

29. See the Exalead analysis in the 3rd edition of *Enterprise Search Report*, 2006, published by CMSWatch.com. The original ESR has been discontinued and reinvented as a more lightweight look at search from a content management viewpoint.

Deduplication functions are good, probably more effective than those available from most vendors. Exalead's system can be extended to handle multiple instances of an information object in a SharePoint repository, but a ready-to-run code shim for SharePoint would be a plus for some organizations. As this report is going into print, we are unaware of any vendor who has cracked this particular SharePoint challenge. Exalead's system permits a solution, however, which is a plus. With more than 100 million SharePoint licenses and an even larger number of SharePoint users, precise deduplication still requires manual inspection. No semantic technology performs with 100 percent accuracy. The context awareness of some Exalead implementations allow some types of semantic false hits to find their way to a report or results list. The study team's tests indicates accuracy in the 80 to 85 percent or higher range. This is a minor issue, but no search and content processing system bats 1,000.

We would also like to see case sensitive search options, which is an important but a rarely used operation.

When Exalead's principal features and core functionality are scrutinized, we find a system that ranks among the best information access systems we have examined, tested, and used.

Net Net

The acquisition of Exalead by Dassault is a good news, bad news situation. On one hand, Dassault has the resources and the market clout to push Exalead even further ahead of other search and content processing vendors. The bad news is that as Exalead moves forward with its search based application solutions, the company will draw the attention of such giants as IBM, Oracle, and SAP. Dassault's own internal needs and the implementation of Exalead within Dassault's products and services for its clients are, in our view, going to stimulate the appetite for search based applications. And with the fast uptake of CloudView, Exalead will undergo continued rapid growth, vying with Autonomy for the role as the world's leading enterprise information and content processing vendor, and finding itself competing against firms with roots in business intelligence, traditional data management, and other enterprise software niches once isolated from a search-based approach. We give Exalead a solid, positive recommendation.

Exalead Annex 1: Technology Partners

Companies mentioned in the text of the Exalead profile for example, Capgeminiare, not included in this table containing representative Exalead technology partners .

Exalead Technology Partners (Selected)

Partner	Key Technology
4D Concept	Document engineering
Acamaya	System integration
Alpha Solutions	Software consulting
Arithnea	E-commerce specialists
Atos Origin	Information technology services providers
ATS	Information technology services
Business & Decision	Technology consulting
CM Inc.	Information technology and services consulting
CMMI Project	Information technology and services consulting
Cohezia	Digital media services
Connect Distribution	Information technology consulting
Contegra Systems	Web development
CSC	Technology consulting and services
Décisionnel	Engineering and consulting services
Deliverance	Technology services
Digirati	Technology engineering and services
DocumMass	Information technology services
Document Text Information (DTI)	Information and technology consulting
e-Spirit	Content management vendor and information technology services
Edifixio	E-commerce and technology consulting
Eficode oy	Technology and engineering consulting and services
ELAR Corporation	Digital content services and integration
EMID Consult	Hosting services
Euriware	Information technology and consulting services
Ever Team	Enterprise content management services
Grupo SIA	Enterprise computer science services
Jouve	Digital content services
Kernpunkt GmbH	Technology services and consulting
Keyrus	Global technology consulting and services
Knowledge Concepts	Information management services
Logica	Information technology services
LTKY	E-commerce consulting services
Mantis Software and Consulting Company	Information technology services
Micropol	Software engineering, consulting, and training

Netbureau	Software services and consulting
Odeon-Ast Ltd.	Information retrieval services and consulting
Redman Solutions	Information technology distribution company
Reply Living Network	Consulting and engineering services
Retis	Information technology services and engineering
SDG	Global management consulting
Search123	Digital marketing services
SEDONA	Information systems consulting services
Sempre Sole	Information technology solutions
Sodifrance	Information technology services for the financial sector
Softline	Information technology and services consulting
ST Groupe	Information technology consulting
Taya IT	Information retrieval consulting and engineering
Terminal Four	Content consulting and engineering services
The Web Fellas	Agile Web and mobile application services
Today Is Now	Information retrieval consulting, services and engineering
VISEO (Homsys and Object Direct)	Integrated management software services, engineering, and consulting

This list of channel partners calls for three observations:

1. Exalead's channel partners cover every continent except Antarctica. The depth and breadth of Exalead's relationships means that professionals with specific cultural expertise can tailor and shape the CloudView platform to meet extremely specific requirements.
2. The number of firms with expertise in the CloudView system means that waits for customer support or engineering services are not required. Services are available without queuing for the only engineer who can answer a particular question a normal situation for some of Exalead's competitors.
3. Exalead is an international system extending beyond its support for most major languages. The company's relationships allow same-day response through the Middle East, the Pacific Rim, and North America.

We believe that Exalead's international scope makes the firm a global player.

Exalead Annex 2: Technology Partners

The CloudView platform makes it possible to "hook" in services from other vendors. Exalead maintains close technology ties to certain companies in order to ensure that licensees can tap into resources that have developed components working seamlessly with CloudView. A selected list of technology partners appears below.

Exalead Technology Partners (Selected)

Technology Partner	Focal Points
DataDirect Technologies	Connectors and middleware
EMC	Content management and storage
Entropy Soft	Connectors and filters
Hewlett Packard	Hardware, management software, and services
IBM	Information technology services and hardware
Lingway	Semantic and linguistic resources
LTU Technologies	Image recognition and search
Microsoft Corporation	Enterprise software, storage, and middleware
Objective Corporation	Content and process management systems
Sciences Po's Medialab	Social solutions
Synapse	Content processing systems
Systran	Language translation systems
Rhe Reuse Company	Content repurposing systems
Vecsys	Speec processing systems
WAND	Taxonomies

Exalead's technology partners provide software and systems that extend the CloudView platform. Exalead has developed a number of next-generation solutions in its own research facilities. Its technology partners develop systems that push beyond the already-robust standard functions of CloudView.

Exalead Annex 3: OEM / VAR / ISV Partners

Exalead has relationships with firms licensing CloudView for use in their products and with specialists who can build full-scale enterprise solutions with CloudView as the foundation platform.

Exalead OEM / VAR / ISV Partners (Selected)

Partner	Key Technology
Atempo	Archiving platform with embedded CloudView
GWAVA	Information security solutions built on CloudView
H&S	Vertical enterprise solutions based on CloudView
Messaging Architects	Compliance solutions
Objective	Content and process solutions
One 2 Team	Collaborative and team solutions
Tera Digital Publishing	The firm's content solutions integrate CloudView
VDR Group	Engineering and manufacturing solutions

At the time of writing (February 2011), Exalead's OEM, VAR, and ISV partners are positioned to serve the needs of closely regulated large-scale manufacturing, high-risk sectors, and on-demand, real-time applications. The fact that these mission-critical applications are built on and with Exalead technology differentiates CloudView from other vendors search solutions.

Chapter 5:

Google Search Appliance

"Fast, relevant search for your Intranet or Web site..."

The Google Search Appliance features collaborative tools and an improved user experience...

Google Search Appliance at a Glance

In 2006, I stated in the 3rd edition of *The Enterprise Search Report*:

> In the last four years, Google has lost its price advantage. The system now is as expensive to deploy for enterprise-wide use as systems from other vendors profiled in this report. The document based licensing model can make a "simple" job cost over seven figures due to the explosion of electronic content inside of organizations. In terms of features and functions, the GSA lags behind Autonomy and Exalead. Endeca, another firm with Silicon Valley roots in the late 1990s, handles structured data in a more effective manner.

The reason for the lost price advantage is not far to seek:

The three-man management team consisted of the two Google founders, Sergey Brin and Larry Page, plus Eric Schmidt. Until he was hired by Google in 2001, Mr. Schmidt had served as the chief executive of Novell Corp. Prior to joining Novell, he was the chief technical officer at Sun Microsystems. In the spring of 2011, Mr. Schmidt was moved "upstairs" to a role on the Google Board of Directors. The new boss at Google is Larry Page.

The shift was positioned as a logical one, but Google seems to be losing its focus on search and has, after more than a decade of trying to diversify its revenues, remains almost totally dependent on search-related online advertising revenue.

Google Search Appliance at a Glance

	Basic Information	Option	Comment
License Fee	Begins at $3,000 and rises as the volume of content processed increases	Buy the GB-9009 and add additional units once the internal expansion is no longer possible	GB-7007s and GB-9009s can handle more documents after an additional fee is paid to the GSA reseller. No additional hardware is required unless the unit's internal expansion is exhausted
Search product	A variant of the Google search system used for Web search	Add-ons from authorized resellers and partners are available	The Google interface delivers basic keyword search which may be configured to include results from other enterprise systems; for example, IBM Cognos
Technology hook	Google's Web search	None	The method of determining relevance may not yield expected documents. Relevance tuning is limited.
Cautions	The GSA has been designed to make deployment simple and quick. Implementing customization for relevance ranking, for example, is difficult	Use third-party software components that add functionality to the GSA. Example: SmartLogic for enriched metadata	Google "locks down" its GSAs. The type of find-tuning and extreme customizations associated with other vendors' search solutions are either prohibited or difficult to implement.
Selected partners	Capgemini, CGI Information Systems and Management Consultants, Northrop Grumman, Onix Networking, SAS	Partners are added. Check Google's partner directory	A partner is required to perform certain GSA customizations
Net Net	The GSA has not kept pace with enterprise search solutions from a number of competitors. The "taxi meter" approach to pricing makes it difficult for licensees to predict the total cost of the system, particularly when text content is growing rapidly in organizations.		

The enterprise revenues have not delivered billions of dollars. One can only point the finger of blame at Google management. The other search vendors profiled in this report have continued to grow, albeit less rapidly since the 2008 financial crash. Google has released new enterprise products and services, but the GSA is not the system that many licensees hope it would be.

Since 2007, Google has made significant improvements in the GSA. However, Google has not yet provided a search solution that makes the GSA the ideal choice for organizations with complex indexing and information retrieval requirements. An "appliance" is, by definition, somewhat constrained. The tradeoff is between ease of use and great complexity. The GSA is designed to be licensed, plugged in, and used to search content. It delivers on quick deployment, easy configuration, and a familiar Google-style search, but it does not permit discovery, robust faceted navigation, or a wide range of customization and relevance tuning options. Other considerations are:

1. The engineering team responsible for the GSA has not be able to match the competition's features quickly or effectively. Integrating the GSA into third-party applications remains a difficult and time-consuming job for experienced engineers. (In contrast, Exalead can deploy a search-enabled application in a matter of weeks.) A GSA device can index with a flip of a switch, but making the results deliver an answer to a business question for a telemarketer takes work by the licensee. The current version of the GSA cannot process audio or video files, a feature available from such competitors as Dassault Exalead, Autonomy, and even Lucene/Solr.
2. Web site search is a commodity service, available from many vendors including Google. Google's own Site Search service provides a satisfactory Web site search and retrieval system. With a free service available, why would an organization buy a GSA to deliver Web site search? For Intranet search, the GSA is easy to set up and can be quickly deployed in a basic security set up such as LDAP. But licensees may not realize that redundancy comes at a price. The GSA GB-7007 with a capacity of five million documents, including the two-year technical support option is about $189,000. To index an additional five million documents, the licensee must pay what amounts to another $189,000. Adding a fail-over GSA requires another six figure investment.[30] The cost undermines the assertion that the GSA is a low-cost solution.
3. The GSA administrative screens are indeed easy to use, almost simplistic. In order to perform certain functions, the licensee must use Google supported methods. Documentation for certain operations is scant, out of date, or incomplete. Even Google partners must "phone home" to Google to get additional input from one of the GSA engineers. Customer support from Google pivots on email, not face-to-face meetings or direct telephone contact. The best Google partners can make the GSA work. Some third-party vendors have successfully developed add-ons to the GSA to deliver such functions as rich metatagging. But for many GSA licensees, the bright yellow devices are a beacon for competitors to pitch a rip and replace, swapping the GSA for a competitive system.
4. Google itself is shifting to the cloud for its enterprise services. The GSA, now at Version 6.8, has been largely unchanged since 2002. The question for some licensees becomes, "Will the GSA become a core component in Google's broader enterprise strategy?" With the current emphasis in the Google enterprise unit on cloud-based services, the GSA may be a stopgap product that could be phased out.

The magnetism of the GSA as an alternative to competitive systems has lost some of it's impact. Google seems intent on displacing enterprise information applications and providing search as a utility. Search is important to the Google enterprise customer, and the GSA creates an opportunity for resellers, third-party developers, and competitors.

30. The prices come from the US government's purchasing Web site, www.gsaadvantage.gov. Run a query for the GB-7007 and you will see the current prices. For a commercial cost estimate, add 12 - 15 percent to the "government" price.

Since the release of Version 6.0, Google has made incremental improvements over the last 20 months. The GSA is, as we prepare this report, at Version 6.8. The last point upgrade took place in June 2009.

Key Developments

Our research suggests that Google has subtly changed the positioning of the Google Search Appliance. When the GSA became available, Google positioned it as a solution for Web site search and enterprise search. "Enterprise search" embraced content behind a firewall. In addition, the GSA could index Web sites "outside" the licensee's firewall and make those results available to users of the GSA.

Google's ability to index content behind a firewall has been described in various Google documents. One of the more complete descriptions appears in a series of patent applications filed in 2007 with the invention attributed to Ramanathan Guha, but other Google documents explain that Google can index content behind a firewall and maintain the confidentiality of that index.

Google now offers a hosted Site Search service. Anyone can specify a Web site or group of Web sites as a source of content for a custom search engine. With this change, the need for a GSA to index a Web site declined.

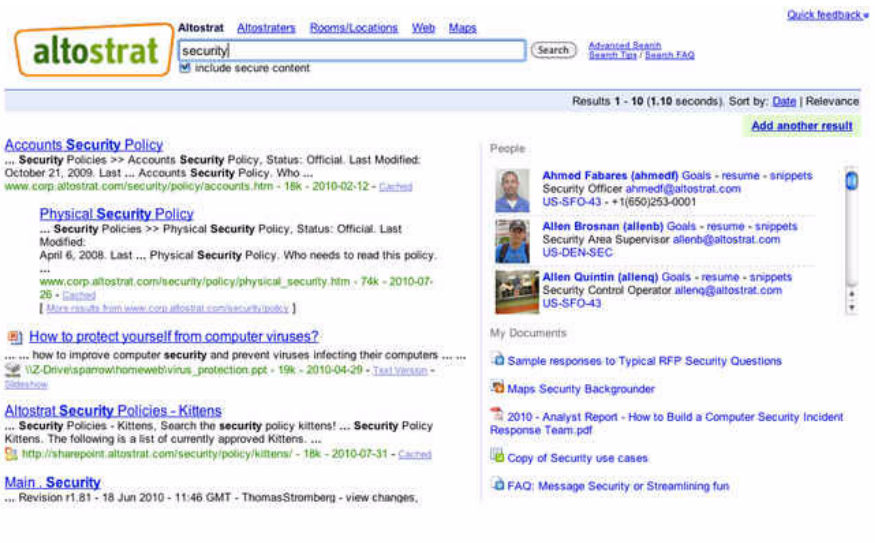
Therefore, with the trimming of the GSA product line to two models, GB 7007 and GB 9009, it seems as though Google is slimming its product line. At some point the GSA may become unnecessary with GSA functions delivered entirely via the cloud (as Google does with its enterprise Gmail and Google Docs services.)

Our discussions with GSA licensees and others familiar with the GSA lineup have revealed several points about the GSAs now available in the market, suggesting that Google has not fully hardened its GSA services.

First, in terms of customer support, Google partners handle inquiries from licensees of the GSA. The partners, in turn, communicate with designated contact points in Google's GSA unit. Most communications are initiated via e-mail, although some partners report having access to engineers familiar with the GSA. Licensees, however, must work through this layered support process. Because the hardware is built by and sold by Dell and other partners, Google's GSA unit is insulated to some degree from the needs of some customers.

Second, one partner (who has asked to remain anonymous) indicates that the search engineering team in the GSA unit is stretched. If true, Google may be shifting resources from the GSA product to other cloud-based product lines. partners have told us that when bugs or anomalies surface, Google urges the partners to develop solutions. Google's interest in developing specific fixes appears to be consistent, but licensees can express frustration that certain features, functions, and services are unstable or can be time consuming to implement. Areas of concern include con-

nectors to enterprise content repositories while matching the security of the licensee's existing systems.



This is a universal search display. The GSA can pull content from different sources and deliver a relevance-ranked list of documents that answer the user's query. Note that this results display incorporates images and links to third-party enterprise applications.

Third, the GSA is an expensive solution for some organizations. The taxi-meter approach to pricing means that a licensee must trim the content processed to remain under the ceiling imposed on each device. If more documents are to be processed, the licensee must contact the partner who is authorized to increase the ceiling on document limits after the licensee upgrades the device. Many organizations have hundreds of millions of document to index with content increasing fourfold every 12 months. With some devices costing $250,000 or more, the GSA quickly becomes cost prohibitive. Few information technology managers will spend a quarter of a million dollars for a computer from Dell just because the organization needs to index more content.

Finally, finding information about GSA is not easy, and that underscores the present state of Google's enterprise offerings, including the GSA. Google assumes that anyone with a GSA will be able to figure out the ins and outs of the product.[31]

The bottom line is that Google has not developed the GSA into a product that can compete against some of the newer solutions available from competitors. In fact, the Dassault Exalead solution uses Google-like principles and incorporates a broad range of features that duplicate the GSA's capabilities and support rich media indexing, social content, and integration with other enterprise software.

The path Google is following with the GSA and its enterprise applications is looking more like Microsoft SharePoint. The GSA adds search as a utility function

31. As a side note, the phrase "universal search" was coined by Fast Search & Transfer and in use as early as 2004. Google appropriated the term and now describes the GSA as delivering "universal search" to licensees.

while the main focus is on cloud services, collaboration, and federated information access. Not to say this is a bad idea, but it is an idea that doesn't match the licensing model. Whereas SharePoint is perceived as “free”, the Google Search Appliance is the most expensive solution that an IT manager can purchase. Should Google offer pricing based on hard disk usage or a flat fee for the appliance since the hardware remains the same, it could then become a more competitive offering.

History

Google rolled out the first version of its appliance in 2002. In the intervening decade, the GSA search system has undergone six major releases with Version 6 the most recent. Upon its debut, the GSA looked like a contender. The product delivers a basic search solution and has been adding features with each version. The system is now at Version 6.8, and it does provide a basic search solution that can be extended. However, the guts of the relevance ranking system remain “locked”. Various workarounds are needed to tune the results to meet the needs of some licensees.

The first versions of the GSA targeted the then industry leaders on three points of weakness that Google identified. First, most enterprise search systems were “kits”, not appliances. Google wanted to deliver a search solution that a licensee could deploy in a matter of hours or even (in large organizations) in less than two or three days. The “simplicity” principle exists to a point in 2011. But over the product’s version upgrades have come additional complexity and changes that provide benefits to organizations relying on Google for other enterprise services, including Gmail and Google Docs. Deploying a basic GSA is as simple today as it was in 2002, but when more sophisticated operations are required, the licensee will need access to a Google partner, expertise in Google application programming interfaces, and Google’s various software components.

The most significant series of changes in the GSA have been the slow but steady improvement in security. The 2002 GSA featured limited security support and was difficult, if not impossible, to integrate into organizations with stringent security and access requirements. Version 6 of the GSA includes support for most popular security models, including forms authentication, Kerberos, LDAP (lightweight directory access protocol), SAML (security assertion markup language), and others.[32] If you plan to implement customized security functions, you will need to write or edit code. Some Google partners can handle this type of customization on a fee basis.

The third change includes direct but not always intuitive support for Google Analytics and Google Sitemaps functionality. In addition, the GSA can integrate with a number of Google’s enterprise services. The complete documentation for the GSA is available online and is current for Version 6.8.

32. You can locate the full GSA security information at http://code.google.com/apis/searchappliance/documentation/62/secure_search/secure_search_overview.html. Verified on February 4, 2011.

Product LineUp

For 2011, the GSA product line up has been trimmed. The search appliance and the cloud-based service are at Version 6.8. For on-premises installations, Google offers:

- Google Mini. The server can process up to 300,000 documents. The system is essentially a demonstration of functionality. Its cost is approximately $3,000.[33]
- GB-7007. The rack-mounted two-unit appliance can process up to 10 million documents. Multiple GB-7007s can be linked together to handle more content. The GB-7007 provides built-in redundancy but Google offers hot spares to ensure minimal downtime. The starting two-year license fee for the GB-7007 is about $30,000 to $400,000 with a three year support plan.
- GB-9009. The server that can process and make accessible up to 30 million documents. To handle larger collections, GB-9009s can be linked. For document collections larger than 15 million, the GB-9009 is the preferred server. The approximate two year license fee for a GB-9009 with a hot backup is about $340,000.

Pricing

Most vendors of enterprise search systems negotiate a price based on the licensee's specific requirements. Google licenses an appliance with a two-year or three-year support package. Prices are available from the US government's public Web site, GSAAdvantage (www.gsaadv.gov).

In other studies of the GSA, we have noted that most analysts quote the price of the Google entry level device, the GSA Mini, and omit the license fees for the GB 7007 and the GB 9009. The table below provides US government prices for the enterprise-grade servers and a few of the options available. In order to determine the license fee for a commercial installation, we recommend adding a minimum 10 percent markup on the prices in the table below. US government prices reflect a government discount that may not be extended to commercial organizations. Educational pricing may also vary.

Selected Google Search Appliance Pricing from gsaadvantage.gov

GSA Model	Configuration	Price	Comment
GB-7007	500,000 documents	$28,387	Two years support. For hot back up, add $6,000
	1 million documents	$66,236	Three years support
	Upgrade from 1 million to 2 million documents	$1,971	

33. The Google Mini is a low-cost trial device. The severe document limits on the device allow a potential licensee to qualify himself or herself as a prospect for a GB 7007 or GB 9009.

GSA Model	Configuration	Price	Comment
	2 million documents	$123,010	Two years support
	3 million documents	$113,548	Two years support
	3 million documents	$158,967	Three years support
	5 million documents	$433,766	Three years support
	10 million documents	$305,066	Two years support
	10 million documents	$423,913	Three years support
GB-9009	15 million documents	$615,053	Two years support
	15 million documents with hot backup	$533,896	Three years support
	30 million documents with hot backup	$993,548	No support listed
	Upgrade from 15 million documents to 30 million documents	$9,856	

A number of interesting observations are supported by the data in this table. First, Google has different resellers, and each reseller offers a different price to the US government. The variance is most likely a result of different royalty or commission deals Google has with its resellers. Therefore, getting multiple Google resellers to provide a price quote seems a prudent step. Second, Google's US government pricing contains a number of inconsistencies. There is no simple way to determine what combination of document capacity and support will result in a particular price. Our suggestion is to identify the specific GSA installation required and then get quotes from authorized resellers. Third, commercial license fees are often higher than the prices quoted on the gsaadvantage.gov Web site. We recommend adding a 10 to 15 percent premium to a gsaadvantage.gov price.

The main point, however, is that the GSA is an expensive enterprise search solution. Consider an organization with a baseline document collection on its Intranet of 10 million Word files, a content management system with one million files, and an enterprise wide Lotus Notes system with 10 million documents accessible via Domino servers. The company would require at a minimum one GB-9009 with a 30 million document capacity for about $900,000. In the span of six months, the company could be bumping up against the limit of the GB-9009 and would have to acquire another unit which could cost another $600,000. If the gsaadvantage.gov prices are accurate, the Google Search Appliance is not the low-cost solution analysts focusing on the low-end devices suggest. A Google Search Appliance solution is as costly as or more expensive than other enterprise search solutions available on the market today. The "taxi meter" approach to pricing means that setting an upper limit on the cost of a Google Search Appliance system is difficult. More GB-7007s and GB-9009s are needed to handle the increase in digital content. Toss in a need to process email, public Web content, or social media. The cost of the Google

Search Appliance may become untenable for many organizations unless they voluntarily remove content from the index.

The pricing communicates one message that is probably desired by Google based on some current limitations: Use the Google Search Appliance for smaller search deployments. The pricing revealed on the gsaadvantage.gov Web site suggests that Google knows its product is not right for most on-premises enterprise search deployments with large amounts of content.

Technology

Google has a large amount of information available without charge. It can be daunting to find the specific information one needs to resolve an issue.[34]

On a feature-by-feature basis Version 6 of the GSA compares favorably with many competitive search systems. The GSA includes connectors for common file types and file systems. If a special connector is required, a partner can create the filter or use other third-party widgets to make source content available to the GSA. However, one should note that this adds cost to an already expensive product that is sold as plug and play.

The GSA includes support for multiple languages. The licensee can activate date search, specifying file date and time information embedded within documents. The GSA allows administrators to customize results presentation. The graphical administrative interface makes it easy to specify what to index and what to include in the indexing crawl.

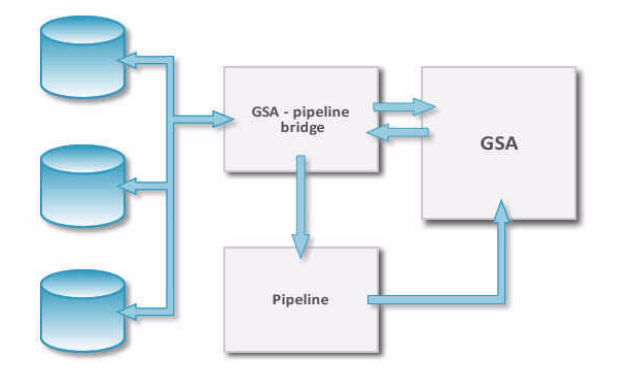
The search technology is locked down, and tuning relevance remains an issue with some GSA licensees. Notable additions in Version 6.8 include:

- Cloud Connect. The GSA can index content in Google Apps, Site Search, and Twitter.
- People Search. The function is not entity extraction. If profile information is available for an organization's staff, the GSA can search it.
- Facets. Google calls "faceted search" Dynamic Navigation. The metadata for a document or other information object can be used to filter search results with specific metadata attributes; for example, documents tagged with a country location such as "Brazil" appear in a result list, if so specified by the user.

Google is adding what we call "soft features". These are functions that are useful but do not make substantive changes to the GSA architecture, now more than a decade old.

34. See http://code.google.com/apis/searchappliance/documentation/68/

The language used to describe the GSA has become increasingly mainstream. Google asserts that the GSA is a "Swiss army knife" and stresses how a licensee can "access all of your business content through one search box." The statement is categorical, and there are exceptions, notably proprietary data stores such as I2 Ltd.'s Analyst Notebook repository. The Cloud Connect service can now index Gmail and Google Docs. Additional work is needed to tap into email and hosted documents that are not directly supported by the GSA. Google also asserts that content is available in "real time." Indexes can be refreshed on an administrator-defined schedule. But in the pursuit of "real time", the GSA's crawler can become overly aggressive and can become a bandwidth hog.



The Findability Blog in Sweden published "Processing Pipeline for the Google Search Appliance." What is important is that the GSA requires the licensee to code a pre-processing component to normalize metadata across sources. SmartLogic, a Google partner, offers a solution for this shortcoming of the GSA. See http://findabilityblog.se/processing-pipeline-for-the-google-search-appliance

The most notable marketing-infused GSA collateral refers to "user experience" or UX. The idea is that a user can access needed information via a range of interface options that make search "easy, useful, and intuitive." UX features built in to Version 6.8 include:

- Endeca-like faceted navigation hot links to content in the user's Gmail or Google Docs account, content from Web sites crawled for competitive intelligence, contacts, or real-time information from Twitter.
- Google's self-learning scoring technology which tunes the behavior of the GSA results to a specific user. Modifications occur automatically.
- Tailoring search results from specific collections of content for particular users; for example, customer support or telemarketing. Administrators can group content and create collections addressing specific user needs; for example, the contracts unit of an organization or the research and development team.
- Implementation of Google's query suggestion feature. Like Google's approach to relevancy, the suggestion feature may be perceived as help by some users and as an annoyance by others.

- The GSA now displays related queries. They define and suggest queries for company-specific acronyms or terminology. The GSA can make use of controlled vocabularies or taxonomies if available.

The GSA allows users to provide input to the locked down relevance ranking system. The "votes" for relevant content will then be used to weight certain values in the relevance ranking system.

Indexing Highlights

Google has published a GSA Data Sheet, with links pointing to useful Google documentation. Here, we focus on a handful of basic features that make or break an enterprise search system.

OneBox and Cloud Connect

Google's "cloud connect" function allows the licensee to index content from Web sites, Google Docs, and behind-the-firewall repositories. The results list displays relevance ranked results from various repositories. With a bit of fiddling, the GSA can index Twitter, blogs, and other social content. The Google "people search" function makes use of the processed content to find colleagues associated with a particular search query. Other vendors describe such features as "federated search". Google sometimes describes its approach as "universal search" or "integrated search". The idea is that the GSA makes an organization's information findable from a Google-style interface.

Results Grouping (or Clustering)

Google added clustering functions to the GSA in 2007. The system automatically groups search results by topics. Google's approach is called "dynamic result clustering" and has been designed to assist users in narrowing a search result set. One of Google's strengths is delivering a laundry list of "hits" in which the user's query is matched. Enterprise users want to find answers, not dig through dozens or hundreds of links. Note that the function does not work for every user's query.

Administrators can control where the clusters appear. For example, if an employee searches for "video" on the company network, the system generates on the fly a set of categories (such as "task") and their location and appearance on the results page can be customized.

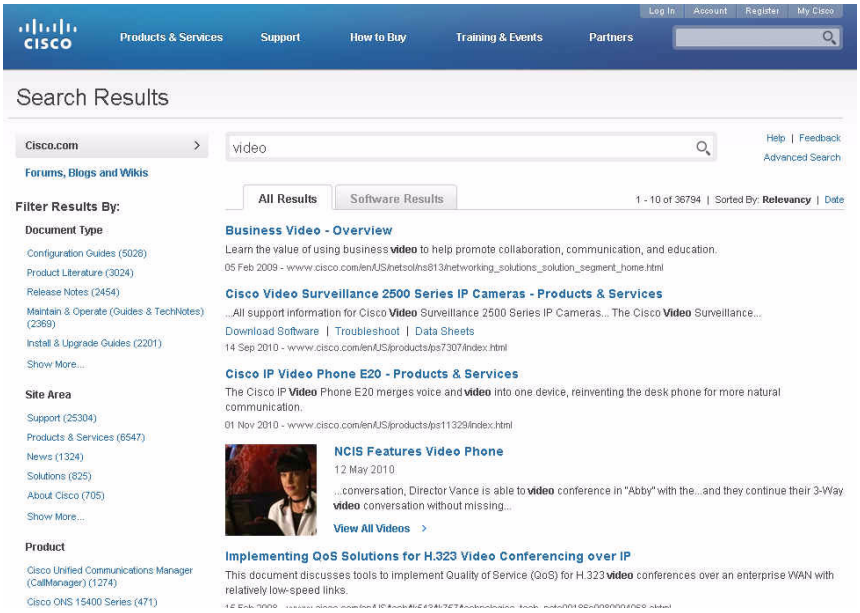
Relevance Controls

Compared to other search vendors' solutions, the GSA offers fewer controls to tune relevance. However, Google did add a function called "source biasing"to make it easier for an administrator to assign a "weight" or an "emphasis value" to a source

or a type of document where licensees have Lotus Notes, Microsoft SharePoint, or an Autonomy Interwoven content management system, the administrator can assign a higher score to e.g. Microsoft SharePoint content, thus boosting that source in a results list. Since 2007, the administrative interface allows source biasing via a graphical interface.

Scaling

The GB 7007 can handle up to 10 million documents. With Google's new architecture, a licensee can connect multiple GSAs. To handle billions of documents. GSA maintains crawl and indexing performance at scale., and query processing times do not deteriorate at scale. GSAs now intercommunicate in a manner similar to that used in Google's own cloud architecture. The integration takes place across appliances, organizational boundaries, and geographic locations.



Cisco uses the Google Search Appliance for its Web site search at www.cisco.com.The clusters appear in the left hand column under the heading "Filter Results By". A click on a cluster narrows almost 37,000 results in this particular query to a more manageable number of hits.

Usability Features

Google, like Microsoft, has embraced the notion of "user experience." Among the more important additions are highlighting for query terms. (A user can spot the most relevant section of a document via the highlights)and a date sort function in the results display. Google's handling of date metadata is less than perfect, but the sort feature is a welcome addition.

The GSA can cache document pages. The problem with caching is that storage requirements can be significant. The GSA can display the cached version of the document if the original source is not accessible at the time the user clicks on a link in the results list.

Security

In 2002, the GSA lacked robust security controls. After nine years of development, the GSAs support secure searching of information protected by basic HTTP authentication and NTLM version 1 and 2. The GSA features a single signon and integrates with most form based single sign on security systems, including Oblix and Netegrity. The GSA includes a secure content API so the users can search across secure content using Google's SAML (Security Assertion Markup Language) and SPI (Google Search Authorization Provider Interface). The GSA supports x509 client certifications via mutually authenticated x509 certificates. The GSA can also integrate with Lotus Notes environments. Google partners can extend the security functionality of the GSA if particular security requirements must be met.

When the licensee activates access protection, the GSA makes a HEAD request on the first two pages of search results for each search request. The system checks to make certain the user has permission to access each document. Access protection places an additional load on the Web server. One workaround is to cache the GSA HEAD request and automatically send a Not Modified response; the result is that the GSA does not go through its verifications.

Additional security issues arise with the fact that the GSA is an appliance. If a customer does not want to use the GSA to serve the search results, for example rendering the XML in another application such as SharePoint, the complexity with security can increase.

Metadata

The GSA can index external metadata repositories and their associated documents so that search across annotated and enhanced content is possible. However, the GSA's native metadata support is not robust, and a third-party solution is desirable if the licensee has an existing taxonomy or controlled list of index terms and wants these used by the GSA.[35]

35. One vendor offering a GSA "snap in" metadata solution is SmartLogic. The software is called Semaphore. Additional information is available at www.smartlogic.com.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE gsafeed PUBLIC "-//Google//DTD GSA Feeds//EN" "">
<gsafeed>
  <header>
    <datasource>sample</datasource>
    <feedtype>full</feedtype>
  </header>
  <group>
    <record url="http://www.corp.enterprise.com/hello01" mimetype="text/plain"
     last-modified="Tue, 15 Nov 1994 12:45:26 GMT">
      <metadata>
        <meta name="author" content="Jones"/>
        <meta name="project" content="hello01"/>
        <meta name="department" content="engineering"/>
      </metadata>
```

Example of code to be edited to activate an XML feed of metadata into the GSA. See http://code.google.com/apis/searchappliance/documentation/50/metadata.html

Collaborative Features

Google has added collaborative features and native support for SharePoint 2010 and Lotus Domino content to the GSA. At first glance, the emphasis on content and collaboration appears to position the GSA as an all-purpose tool along the lines of Autonomy's Interwoven, Microsoft's SharePoint, or OpenText's LiveLink.Yet the GSA, despite Google's enthusiasm, is not yet a platform on a par with other major enterprise search and content processing systems.

In order to establish an instant messaging system using Google's chat technology or to make a voice call from within a Google's results list, additional work is required.

Strengths

The Google Appliance is a solid choice for an organization with content accessible via Web pages. One price delivers a solution that is appropriate for basic search-and-retrieval tasks. Other benefits of the Google Appliance are:

- Out-of-the-box deployment reduces the time, cost, and frustration of enterprise search for organizations with large quantities of Web-centric content.
- System administrator friendly streamlined set-up screens. An included maintenance, support, and hardware replacement warranty (to repair or replace the Appliance if necessary.) Appliances with clusters can withstand multiple hardware failures. The entry-level device relies on RAID support to provide redundancy against drive failures. Google suggests that when using a single Appliance, a licensee deploy two units that are mirrored, which doubles the cost of the system.
- Newer releases of the Appliance allow the licensee to tame the Google crawler. (Prior to Version 3 there were instances of the crawler hammering a server with multiple domains at a hefty pace for hours.)

- Strong potential for user acceptance and no extra education about search for Appliance users (almost everyone knows how to search Google.)

Google's high visibility gives the GSA an advantage no other enterprise search vendor can match in our opinion. Employees at organizations large and small are familiar with the Google Web search system. The idea that the enterprise search system should work "just like Google" is a powerful one. For this reason, even when technical and economic factors argue against the Google Search Appliance, procurement teams may decide to license one or more GB-7007s or GB-9009s. In spite of a number of public relations blunders in the last 12 months, the Google brand evokes strong positive reaction for basic search.

Cautions

Some licensees find the GSA an ideal combination of ease of use, speedy indexing and query processing, and maintenance. Other licensees find the limitations of the GSA overly restrictive. For an appliance-based solution, Google has done a reasonable job of making tradeoffs between customization and the locked-down nature of a pre-built appliance.

Among the issues licensees have identified at conferences and in user groups are complaints about the relevance of the results list. Many essential corporate documents may be accessed very infrequently. Thus Google's traditional PageRank method of determining relevance does not work. Over the nine years of the GSA's life, Google has adopted convention after convention from established enterprise search vendors' systems. In terms of configurability and customization, the GSA is acceptable. It lags behind other systems, however. The GSA is a "black box" and it is not designed to be manipulated in the way one can rework solutions from Lucene/Solr other competitors profiled in this report.

Redundancy requires additional GSAs. Google explains that the GSA is redundant, but the reality is that hot spares are needed. Resellers will explain the options for hot spares, but this information is not put front and center on the Google Search Appliance Web pages.

In order to update the index, the GSA recrawls content. Continuous updating helps ensure a fresh index. But if the GSA is configured to "pull" information from a third-party enterprise system and a large volume of content is generated by the licensing organization, the GSA can be a consumer of bandwidth. The Google crawler interacts with the document limit setting on the GSA. Once document limits of an individual GSA are reached, another GSA must be added to the system, the reseller must be contacted to add headroom to the device, or the number of documents processed must be decreased. For first-time users of the GSA, learning these characteristics can be difficult.

Other issues include:

- The GSA does not accomodate for file sizes in the processing queue. In order to know the size of documents that the GSA will index requires a workaround or a third-party solution.[36]
- Significant customization is often required to make the GSA deliver what the licensee requires from an enterprise search solution.
- The support for metadata is less robust than for other vendors' systems. Google documents its approach to metadata in "External Metadata Indexing Guide."[37] The GSA has not undergone significant changes to its metadata method since 2009. Google provides several different metadata scenarios and each requires the GSA administrator to edit or create scripts to perform the desired function.

Net Net

With effort, the GSA can be used as an enterprise search solution. Although it is an appliance, the customization effort is comparable to that required by other enterprise search solutions.

36. You may want to look at the SmartLogic Semaphore implementation for the Google Search Appliance. SmartLogic is located at www.smartlogic.com.

37. http://code.google.com/apis/search appliance/documentation/50/metadata.html. Verified on February 8, 2011

Google Annex 1: GSA Essential Links

GSA Selected Documents

Document	Location
Documentation for the Google Search Appliance	http://code.google.com/apis/searchappliance/documentation/68/index.html
GSA Connectors	http://code.google.com/apis/searchappliance/documentation/connectors/index.html
GSA Online Help	http://code.google.com/apis/searchappliance/documentation/68/help_gsa/home.html
GSA Search Protocol Reference	http://code.google.com/apis/searchappliance/documentation/68/xml_reference.html
Integrating with Google Apps	http://code.google.com/apis/searchappliance/documentation/68/integrating_apps.html
OneBox for Enterprise Design Principles	http://code.google.com/apis/searchappliance/documentation/68/oneboxstyle.html
OneBox for Enterprise Developer's Guide	http://code.google.com/apis/searchappliance/documentation/68/oneboxguide.html

Google Annex 2: Selected Google GSA Partners

Our research supports our recommendation that licensees of the GSA work with a Google-certified partner. The word "appliance" suggests a toaster- or microwave-like convenience. For basic search, the appliance can be used with few set up headaches. For customized enterprise search deployments, the GSA is as complex as other enterprise search systems.

Google GSA Partners (Selected)

Partner	Selected Resellers and Integrators
Bell Canada	A Canadian Google partner for the GSA
Capgemini	A certified partner focusing on Western Europe
CGI Information systems and Management Consultants	Sales and engineering services in Canada and the US for the GSA
Dell Computer	The PC giant is able to fulfill orders for Google Search Appliances
DMSBT	A Google partner handling the Pacific Rim
Fig Leaf Software	A partner specializing in GSA training
GlobalNet Services, Inc.	Certified partner for the continental US
IBM Cognos	IBM Cognos is a GSA partner
Infosys Technologies	Global technical services firm certified for the GSA
Just Digital	Certified GSA partner handling Latin America and Brazil
Onix Networking	sells and supports the GSAs.
OpenText	Although the company sells a proprietary enterprise search solution, the firm is a Google partner for the GSA product line.
SAP Business Objects	The business intelligence arm of SAP supports the GSA
SAS	The business intelligence firm resells the GSA to SAS customers requiring a Google-style search
Search Technologies	The firm supports and integrates the GSA
SmartLogic	A specialist in taxonomy integration with the GSA
SRA International	A specialist in US government solutions, SRA sells and supports the GSA
Tomorrow Focus Technologies GmbH	A GSA partner in Germany
ZettaServe Pty Ltd.	Australian GSA reseller

The full line up of Google Search Appliance partners is located at http://www.google.com/enterprise/gep/directory.html. Google also maintains a roster of application resellers, add-in products, and developers.

Chapter 6:

Microsoft Fast Search Server

"Enterprise search will change the way people interact with information"

Source: An Update for Microsoft partners, April 2008

The assertion is that Microsoft Fast is the best-in-class enterprise search solution, according to Microsoft...

Microsoft Fast at a Glance

Microsoft is positioning the technology it acquired in 2008 as the "the finest or most superior quality of its kind." Who can argue with Microsoft and the consulting firms identifying Microsoft Fast Search Server as a "market leader"?

Microsoft's Certified Professionals and authorized resellers make the same assertion that Microsoft Fast is the top dog in enterprise search. Purpose built to index Web content, Microsoft Fast Search Server is a complex, massive platform is integrated with Microsoft SharePoint under Microsoft's ownership,. [38]

Key Developments

The key development of interest to readers of this analysis is that Microsoft is marketing Fast Search & Transfer technology as a comprehensive search, content processing, and business intelligence component within Microsoft's ecosystem.

38. An explanation of the Microsoft Fast Search System (MFSS) is available as a Word document at https://partner.microsoft.com/global/40147807. A PDF of this document is referenced elsewhere in this report. They appear to be identical.

Microsoft Fast at a Glance

	Basic Information	Option	Comment
License Fee	Pricing varies based on the licensee's purchases of other Microsoft products and services	Microsoft offers three search solutions. Microsoft Fast is the option for licensees with enterprise requirements	To those without first-hand experience with Microsoft's different search solutions, the Fast option can be confusing
Search product	Microsoft Fast Search Server	Requires SharePoint	Microsoft Fast no longer runs on Unix or Linux systems
Technology hook	A platform, not a search system	Other Microsoft components are required to get maximum value from the system	The Microsoft ecosystem is the environment for Microsoft Fast
Cautions	The product is complex, resource intensive, and includes components dating from before 1997	Test Microsoft Fast Search Server in a head-to-head on identical content	A failure to test a Microsoft Fast Search Server against a competitor's solution may lead to unexpected costs and delays upon deployment
Selected partners	Microsoft has thousands of Certified Professionals and resellers who act as extensions of the Microsoft sales and marketing department	Identify a third party consultant who supports other search solutions, not a "pure" Microsoft consultant	Most Microsoft Fast licensees don't know what they don't know. Accurate, verified information is a must.
Net Net	The Microsoft Fast search solution can be made to work. Success requires three ingredients: [a] dedicated Microsoft Certified Professionals to work on the system on an on-going basis; [b] properly configured and tuned Microsoft SharePoint servers and any other Microsoft components; and [c] appropriate resources to handle spikes, unplanned outages, custom coding, and performance tuning.		

One buzz phrase is that search is part of the Microsoft "business productivity infrastructure" with "powerful user experiences."[39]

A second and related development is that MFSS (Microsoft Fast Search Server) requires other Microsoft components to work. There is no version for Unix/Linux. We find this interesting since the original Fast Search system was built expressly for Unix/Linux. Since the purchase of Fast Search by Microsoft, engineers have been focused on making the Fast Search technology dependent on SharePoint and other Microsoft components.

Other developments since the April 2008 purchase of Fast Search & Transfer include:

- Bundling of MFSS with other Microsoft products. When an organization makes a commitment for reordering or signing up for SharePoint and CALs (client

39. The positioning is located at http://www.microsoft.com/businessproductivity/en/ca/bpi/default.aspx. Verified on January 27, 2011.

access licenses), Microsoft provides incentives or sweeteners to get MFSS into an organization.

- Hard technical information about the performance of MFSS is increasingly difficult to find on the public Web. Even the SharePoint Web logs and the Microsoft Enterprise Search Blog have gone silent or contain marketing information of little use to a person trying to resolve an MFSS issue.[40]
- Certified partners with SharePoint compatible search and content processing systems report strong interest in the MFSS products. Even platform vendors find increasing interest in search and content processing solutions that can deal with the sprawl of SharePoint, Web content, and third-party information.[41]

Microsoft has been extremely successful in getting partners, consulting firms, and analysts to sing the praises of MFSS. Many of these rave reviews were quite surprising to us.[42]

History

Summarizing key developments between 1997 to 2011 is a big job. As there were many significant events. One notable decision was Fast Search & Transfer's decision to withdraw from Web search in 2003, thus ceding the Web search and online advertising market to Overture (later purchased by Yahoo) and Google. Another was Fast Search's aggressive publicizing of:

- Its revenue growth from zero to $36 million in 2002 to the mind-boggling $180 million in 2008. In this reported runup, Fast Search took quite specific aim at Autonomy in an effort to lure away IDOL customers.
- The ability to offer "universal search". As early as 2003, Fast Search's technology was able to ingest structured, unstructured, and semi-structured content. Even today, a system that can process multiple content types is a challenge to set up and deploy. In 2003, the claim was novel and makes evident Fast Search's instinct for offering a solution to a painful problem some other vendors were not addressing.
- The effort prior to 2008 to license, acquire, partner and develop original code. Because Fast Search asserted it could perform data fusion and other advanced processes, Fast Search had to find expedient ways to hook needed functions to

40. The "Microsoft Enterprise Search Blog" is at http://blogs.msdn.com/b/enterprisesearch. The last update was on July 9, 2010. The Microsoft executive responsible for the blog left the company without explanation in mid 2010.

41. See the discussions of Autonomy and Exalead. These companies offer search alternatives for SharePoint licensees.

42. Positive case studies of successful MFSS deployments may be found at http://www.microsoft.com/southafrica/enterprisesearch/success/default.aspx. Like this and other analyst "reviews", it is difficult to determine the accuracy of the examples. For example, one case study describes Dassault Systèmes use of MFSS. Dassault purchased Exalead, a true next-generation search system and, according to our sources, shifting from Fast and other vendors' systems to the Exalead platform.

the Web search foundation upon which its enterprise search rested. The result, not surprisingly, was a big engineering job for most Fast deployments.

To those unfamiliar with the way in which it positioned its technology, Fast Search & Transfer's senior management delivered a "business as usual" message. However, some of the firm's engineering staff were disclosing that Fast Search was in the process of rewriting and re-engineering the core engine as early as 2006. In fact, had analysts been knowledgeable about enterprise search, warnings about Fast Search's precarious technical and financial position could have been sounded as early as mid2007.

Fast Search & Transfer's marketing blended academic and "ivory tower" thinking with science fiction. Examples appear in numerous presentations delivered prior to the acquisition of Fast Search by Microsoft.

A good one is the lecture delivered by Gora Sudindranath, Senior Solutions Consultant, Fast Search & Transfer, in May 2007. [43] The Fast Search system was, according to the presentation materials, able to deliver:

- Community services, including product and content matching
- Discovery, personalization, and navigation plus customer analytics
- Zero-term search, Fast Search's catchphrase for mashups, pushed content, and related automated functions for getting content to a user without the need to enter terms in a search box.

Many of these functions are becoming available today, three years after Mr. Sudindranath's lecture. But Fast Search's system, asserted Mr. Sudindranath, also provided a platform to deliver front office services, connections for answers, people, analytics, and services, plus "business intelligence built on search." The Fast Search presentation stated, "Search IS the Portal." Yet most enterprise search solutions in 2007 were not able to deliver dynamic portals containing rich media, live content, and collaborative functions.

Fast Search's system went event further according to Mr. Sudindranath supporting "intent, aggregation, and monetization."

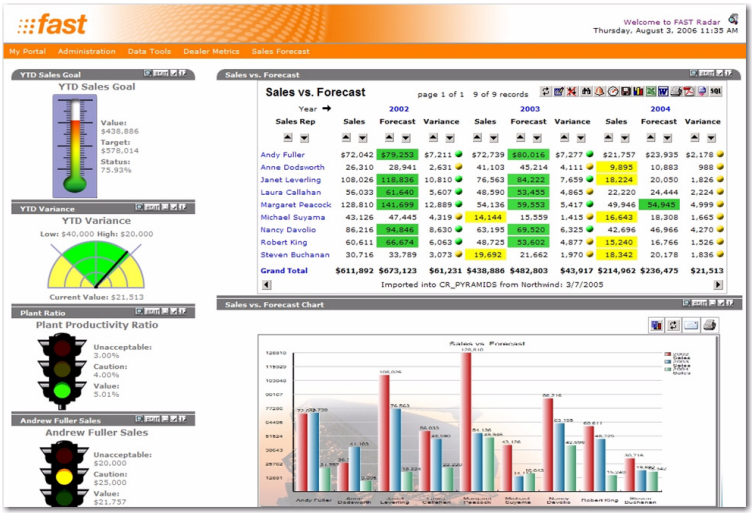
What about mobile? In 2007, Fast Search supported mobile personalization, one click access to such services as those available to Nokia N70 users, feature panels, alerts, and "15 channels based on personal preferences and community recommendations."

Analytics were available through the Fast Radar option, positioned as a business intelligence service. One surprising facet of the lecture was the focus on data qual-

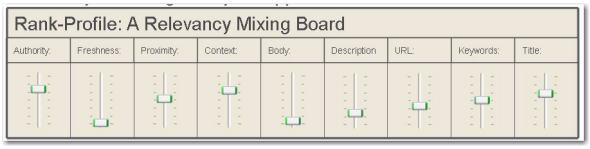
43. This presentation was available online on January 21, 2011, by using Google's advanced search function and searching for this string: "Gora Sudindranath" "May 2007" filetype:ppt. A screenshot from the presentation is included to provide fungible substance to the discussion. The lecture was delivered in May 2007 to the Boston Computer Society Information Retrieval Specialist Group.

ity and Fast Search's ability to normalize source data. Fast Search's term for this function was "cleanse." The Fast Search system then went "beyond search", delivering "flexible relevance."

Slide 14 in Sudindranath's presentation shows a dashboard that compares favorably with the user experiences available from such market leaders as Business Objects and IBM Cognos SPSS. This type of display and the inclusion of the functionality in Fast Search's commercial software suggested that Fast Search had moved beyond other competitors' offerings in 2007.



Mr. Sudindranath's presentation then defined "flexible relevance" in terms of the Fast Search " framework." Unlike Google (described as a black box), the Fast Search approach allowed licensees to adjust more than a dozen "tunable attributes". The idea was to provide an "open and tunable framework to allow you to mix attributes to create the relevancy model right for your application." The method? A series of sliders similar to those for a sound system.[44]



The image appears in "Search 2.0: The Next Chapter of Search," May 2007, by Gora Sudindranath. © 2007 Fast Search & Transfer SA.

How did Fast Search propose to make relevance tuning easy? In 2007, most vendors, including Fast Search, relied on scripts or Web forms displayed in a browser. Fast Search's solution looked like an interface method that was ahead of its time for enterprise search vendors.

The question to consider is, Were Mr. Sudindranath's presentations describing its 'real' technology or a fanciful imagining of what Fast Search could build given time, money, resources, and patient clients? The question has not been answered in the analyst reports, but what we do know is that Fast Search itself was planning to create a new version of its search platform in 2007.

44. See page 17 of the Mr. Sudindranath's presentation.

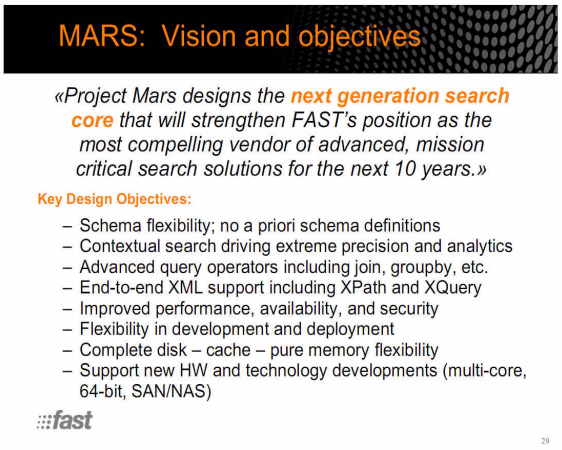
Keep in mind that the financial underpinnings were crumbling in 2007. Microsoft stepped in and purchased Fast Search & Transfer for $1.2 billion dollars. If we assume that Fast Search had revenues of $180 million, Microsoft was paying seven times earnings, but if the "re-evaluated" revenue was closer to $80 million, Microsoft paid 15 times earnings.

Our research revealed that in 2008 Fast Search was in the midst of a massive platform overhaul. In another presentation, Fast Search admitted that it had run into technical problems with the Fast Search technology. You can see the startling information in a Fast Search PDF at the Organisation Européenne pour la Recherche Nucléaire Cern Web site.[45] The briefing reviews Fast Search's "everything including the kitchen sink" approach to search, content processing, and data management. Here is the key passage on page 29 of "Visiting CERN: Taking Search Further":

> Project Mars designs the next generation search core that will strengthen Fast's position as the most compelling vendor of advanced, mission critical search solutions for the next 10 years.

Did Microsoft buy a vision or the version of Fast Search & Transfer's technology that was causing the customers to delay payments?

That question has been ignored by Microsoft, analysts, and so-called search experts.



Slide is from the CERN deck, Fast Search & Transfer SA, 2007. © 2007, Fast Search & Transfer SA. The key point in this slide from the 2007 Fast Search presentation is that prior to implosion and subsequent sale to Microsoft, Fast Search was embarking on a complete overhaul of its system. Was Fast at the end of its useful life? After its purchase by Microsoft, Fast Search would be ported and positioned as the "NexGen" system. The system now offered by Microsoft appears to be a repackaged version of the original ESP which was falling behind competitive solutions in 2006 based on our analyses of vendors conducted for *Enterprise Search Report*, 2003-2006 by Stephen E Arnold. .

Our view is that Microsoft bought Fast Search & Transfer at a very high price. Once in possession of the company, Microsoft had to embark on a process that would chop Unix/Linux from the Fast Search platform and rewire the existing and somewhat problematic system to work in the Microsoft SharePoint ecosystem.

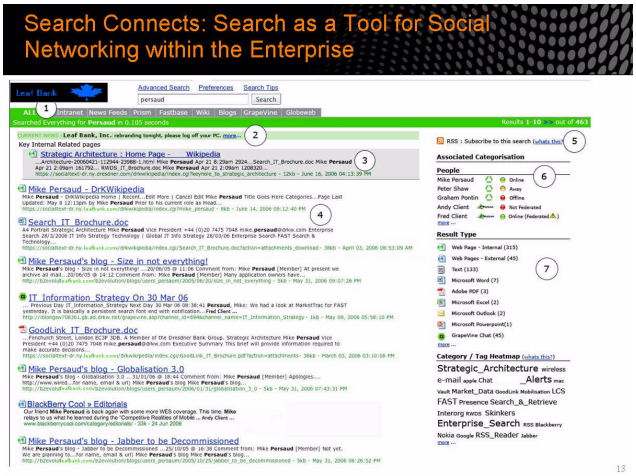
45. Navigate to http://cseminar.web.cern.ch/cseminar/2007/0523/FAST.pdf. The presentation was still available on January 21, 2011.

Microsoft's own documentation for the Microsoft SharePoint Search Server makes clear that there are some known issues with the 2011 version. Microsoft identified some significant issues with its *Microsoft Fast Search Server in Microsoft FAST Search Server 2010 for SharePoint Management Pack Guide for Operations Manager 2007*.[46] The known issues regarding Fast were "Performance collection rules do not collect any data." Just that one? No. Microsoft issued a roundup of known issues from April 2010 to the present.[47] But our research suggests issues with virtualization of the Microsoft Fast Search Server, scaling issues, and integration issues arising because many dependencies are not documented.

How different is today's Microsoft Fast Search Server from the problematic Fast Search Version 5.x system? We can look at a screen shot displayed by Fast Search in 2007 and compare with the screen shot in Microsoft's own presentations.

Screen shot comes from the 2007 Gora Sudindranath presentation:

A 2007 demonstration slide used at the same time Fast Search & Transfer was working on a rebuild of the Version 5.x platform.



Screen shot below comes from a presentation given after the Microsoft purchase of Fast Search & Transfer: The lecture is labeled "Enterprise Search from Microsoft." The presenters are Anna Olsson, Technical Product Specialist, Microsoft Norway, and Torstein Thorsen, VP, Technical Lead, Chief Technical Officer.[48] This illustration appears on page 12 of the handout:

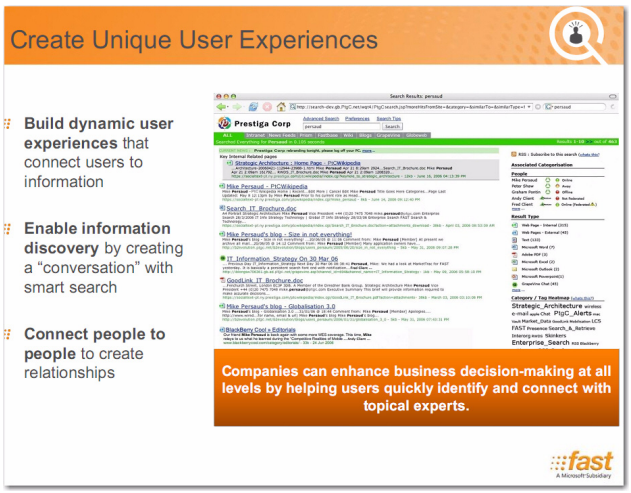
46. This document is available via MSDN. The full title is "Microsoft System Center Operations Manager 2007. Microsoft® FAST™ Search Server 2010 for SharePoint® Management Pack Guide for Operations Manager 2007, April 16, 2010.

47. The summary of known issues appears in the Web page "Microsoft Search Server 2010 and Microsoft Fast Search Server Known Issues/Read Me" at http://office.microsoft.com/en-us/search-server-help/microsoft-search-server-2010-and-microsoft-fast-search-server-2010-known-issues-readme-HA101793221.aspx

48. On January 21, 2011, the presentation "Enterprise Search from Microsoft" was available at www.microsoft.no/portfolio/EPG/EnterpriseSearch.pdf

Our view is that Microsoft has not made as much progress re-engineering the Fast technology as the marketing collateral leads a potential customer to believe. In fact, the assertions about the technical components of Microsoft Fast Search Server have changed little from the assertions Fast itself was making as its business was crumbling and management was working feverishly to keep the ship afloat until a solution could be found to deal with the financial pressures on the company.

Our ongoing monitoring of the "Fast saga" reveals that the academic and marketing influence on Fast Search correctly identified important trends in content and user behavior. Fast Search's professionals were among the first to tout enterprise support for collaborative content and real-time data transformation. However, the Fast Search platform was unable to deliver on these promises. One major newspaper licensee in London told us, "We waited two months, then four months, and finally after seven months we refused to make any further payments and broke our deal with Fast Search & Transfer."[49]



A diagram from a post-acquisition presentation by the Microsoft Fast chief technical officer.

What is Microsoft saying about Microsoft Fast Search Server? A useful summary appears in the document "Microsoft SharePoint 2010. Microsoft Fast Search Server 2010 for SharePoint. Evaluation Guide."[50] A quick scan of this document (after reviewing the 2007 Fast Search presentations referenced in this analysis) is that Microsoft is recycling the descriptions of the Mars project. The "guts" of what Microsoft is shipping as MFSS is a modified version of the ageing Fast Search Version 5.x.

Fast Search & Transfer SA failed because its core platform was expensive to deploy, difficult for licensees to scale, and difficult to have the vast number of het-

49. Statement made in December 2007 to the author in the licensee's office in London, England

50. The "Microsoft Fast Search Server 2010 for SharePoint" document was available on January 21, 2011, at http://www.microsoft.com/downloads/en/details.aspx?familyid=F1E3FB39-6959-4185-8B28-5315300B6E6B&displaylang=eAEF-00104

erogeneous components working in a harmonized manner. Microsoft will want to make certain these missteps are not repeated.

Product Line Up

Microsoft's search product line up is confusing. Most vendors offer a search solution with some options. Not Microsoft: The company provides three search systems to its customers. Before taking a close look at the flagship, MFSS, we look at other Microsoft options.

The first is Free Search Server 2010 Express. The system permits key word search, does not scale, and can only index tens of thousands of documents, not millions or billions.

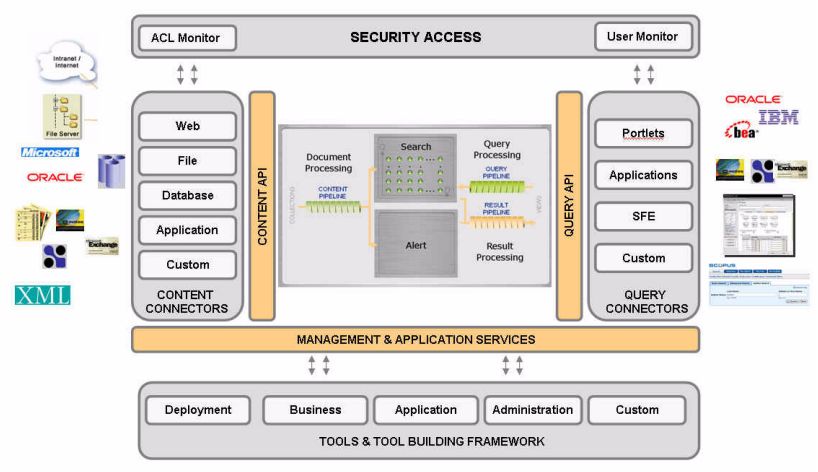
The second is the built-in search function in SharePoint 2010. Prior to the acquisition of Fast Search & Transfer, SharePoint search was positioned as the solution to enterprise information retrieval challenges. That system did not scale, nor was it equipped with the alleged features and functions of a robust enterprise search system like the platform solutions from dozens of companies including vendors included in this analysis.

Technology

Our research suggests that Microsoft has not completed the full rewrite of Fast Search & Transfer's core platform. The Mars version was anchored in Java, and Microsoft seems to have made some changes to the 2007 version of Fast ESP, possibly Version 5.x via software "wrappers". The idea behind "wrappers" is that middleware and interfaces surround the legacy platform. (The ArnoldIT.com team used this approach when we worked with USWest's print Yellow Pages system to output the original USDEX Web site.) The upside of wrapper methods is that the legacy platform is used "as is." New functions can be "hooked in" via application programming interfaces or brute force methods. The downside is that big changes to the underlying plumbing are often expensive. Unknown dependencies can make simple modifications a time consuming exercise in troubleshooting. Another drawback is that appropriate system resources are needed to cope with the latency some wrapper methods impose. Our view is that the MFSS now available from Microsoft is mostly legacy Fast with "wrappers" to make the system work with SharePoint.

What did the core of Fast Search & Transfer's engine look like in the period from 2003 to 2007? This diagram was used by Fast Search's Aleksander Øhrn, Ph.D., in

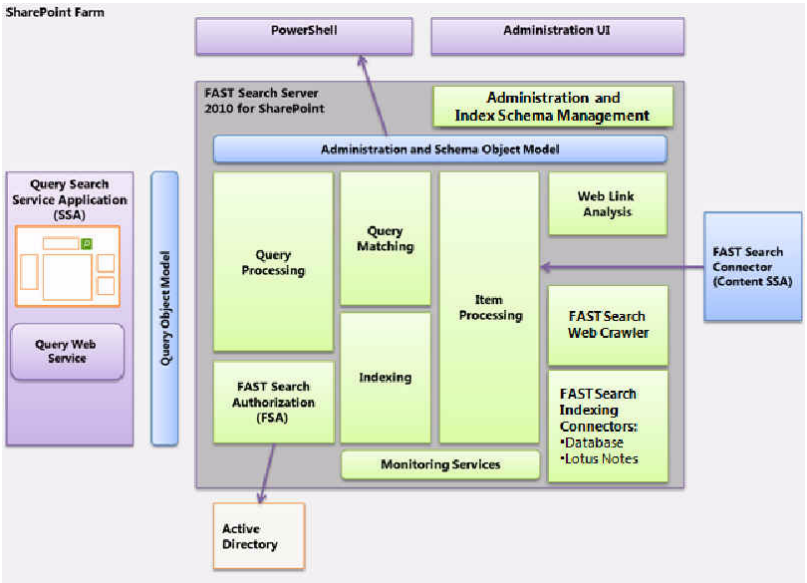
a presentation titled “Contextual Insight in Search Enabling Technologies and Applications.”.



Features to note in this diagram include the support for multiple file types and a wide range of enterprise applications. The system included tools, management and application services, security services, and a number of functions that even in 2011 are found in a handful of enterprise systems. Yet Fast Search would announce that it was rebuilding its system in the Mars initiative.

Our view of this type of diagram is that it certainly identifies needs that many organizations have for a Swiss Army knife content processing solution. The problem was that Fast Search’s Version 5.x system could not deliver in a manner that allowed the firm to operate as a profitable entity. The diagram, therefore, represented a “to be” search system, not a functional “as is” system.

What does the 2010 MFSS search system look like? One can find a useful depiction of the system in the 2010 Microsoft publication *Microsoft Fast Search Server 2010 for SharePoint Evaluation Guide*.[51]



Our comparison of the 2005 block diagram and this 2010 block diagram suggests that the changes made by Microsoft are of the “wrapper” variety. The core engine from Fast Search Version 5.x appears to be unchanged. Scaling, performance, and dependencies are likely to pose some interesting challenges to licensees of MFSS.

Of particular interest are the technical functions of the Fast engine. Like other enterprise search engines, Fast Search can ingest different types of content, index the content, and make it searchable. Where Fast excelled in the period from 1997 to 2007 was in its system's ability to index Web content. On 9/11, for example, the Fast Search AllTheWeb.com news index had near real time content and then-upstart Google did not match the Fast-powered AllTheWeb.com's news search service for timely updates. However, Fast Search exited Web search to focus on enterprise search. Fast Search's engineers then set about reworking the core of Fast Search to deal with the particular requirements of enterprise search. That is when the problems that eventually cut off the company's oxygen supply in 2007 began.

Fast Search went from being a technology leader to a company working hard to make its system solve quite particular enterprise search challenges. Fast Search shifted from licensing a Web search engine to selling enterprise solutions and then trying to refit, extend, and shape the Fast Search core to handle enterprise relevancy, multiple file types including variants of extensible markup language, and outputs that would look like Excel reports and soon.

Long before Google announced "universal search", Fast Search & Transfer was explaining that its system could deliver this Star Trek type of service. Neither Fast Search in 2003 nor Google in 2008 has been able to make good on the promise of "universal search".

Indexing

A quick review of the basic indexing features of the MFSS system is necessary. Key functions that our research has identified as reliable include administrative controls to specify which content should be given priority indexing.

Linguistic Functions

The 2007 core permitted categorization of content via controlled term lists or a taxonomy. The system also supported stemming and lemmatization, a feature which Google (albeit reluctantly) added to its search system years after Fast Search introduced the function. Stemming is particularly useful in clustering and "similar results" operations. The system also provided controls to detect phrases via a lookup table. Phrase detection is now under the umbrella of entity extraction. The original Fast method required a library of phrases; more modern systems automatically identify people, places, and things.

51. You may download this publication at: http://www.microsoft.com/downloads/en/details.aspx?FamilyID=f1e3fb39-6959-4185-8b28-5315300b6e6b&displaylang=en. Verified on January 26, 2011. An easier way to locate the document is to search for the title on Google.com.

As user's query is processed, the MFSS system fuzzifies terms. The idea is to understand the user's query by applying linguistic processing. One can argue that relaxing the user's query generates additional noise or increases the likelihood of more irrelevant records in a results list. But, in an enterprise setting, most users do not know exactly what document is needed, so the "understanding" function (if implemented) can be a help to some enterprise users.

Controlled Term List Support

The Fast core can use controlled term lists to handle synonyms. Hit boosting is also supported so that particular content can appear higher in a results list. In many cases, a stored query can be created to force the system to display specific types of results for certain queries. The Fast core also allowed a system administrator to set up specific rules e.g. to move content after processing into a subsystem that would "push" the results to a user with a specific interest.

Customization

One of the key differences between a brute force Web search and a more refined enterprise search system is the need to "tune" or "customize" certain system operations. There is no doubt that in the period from 2003 to 2007, the Fast core search platform permitted customization. Licensees could access a 500 page manual that documented hundreds of configuration settings. The number of choices was admirable, but the problem many licensees faced was figuring out which control was needed to deliver the desired result. Fast Search's engineers worked hard to integrate third-party code, hook in new functions, and add features. Our work with Fast Search's core system revealed unexpected dependencies. A change made to one option could produce an unexpected effect in another Fast Search operation. Changes are possible via the graphical administrative interface or by editing configuration files that are similar to those found in Vivisimo's enterprise system. With careful study or access to an experienced Fast Search engineer, one could make the system perform in a customized manner.

Microsoft's Additions

To this core, what has Microsoft added in the period between 2008 and December 2010? The major changes are described in *Microsoft Fast Search Server 2010 for SharePoint Evaluation Guide*. Three points warrant comment.

First, Microsoft has added interface enhancements. The use of document thumbnails and a richer graphical presentation layer distinguishes the pre-acquisition Fast Search from the Microsoft Fast Search Server. Microsoft calls these enhancements "additions" to the "user experience" or UX. A more attractive presentation layer is a plus. However, if the underlying system returns content not related to the user's

query or a result set that is missing updated content due to indexing latency, the UX puts a coat of paint over a mildewed wall.

	SHAREPOINT SEARCH	FAST ESP
OVERALL	Solid, out-of-the-box enterprise search tightly integrated with a business productivity infrastructure. Key technical limitations when extending beyond core use cases listed above.	Best-in-class enterprise search with all the knobs and dials needed to build the most demanding applications. Requires significant investment to design, implement, and maintain.
SCALE	Best for content volume less than 20M documents and for performance requirements of less than 5 queries per second.	Designed for extreme scale across 3 dimensions: performance, freshness, and content volume.
INTERACTION MANAGEMENT	Delivers a standard search experience.	Enables a "conversational" search experience.
DOCUMENT PIPELINE	Offers limited document processing capabilities.	Offers sophisticated content retrieval and document processing capabilities.
SEARCH ON STRUCTURED DATA	Enables "no-code" approach to structured data connections using the Business Data Catalog.	Provides an ideal framework for ad-hoc, search-based queries on structured data.
RELEVANCE	Provides strong out-of-the-box relevance for internal applications.	Offers advanced options for tuning relevance.
ADVANCED LINGUISTICS	Includes basic linguistics for major European languages.	Offers sophisticated linguistic capabilities across many languages for improved relevancy.

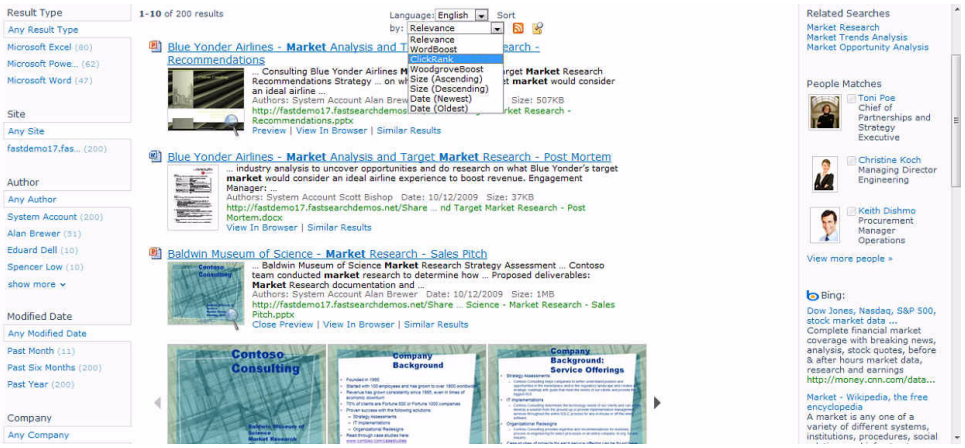
This comparison (page 8) prepared for Microsoft partners in 2008 highlights the key changes in the Fast Search platform. Notice that the underlying engine remains mostly unchanged. Consequently the strengths and weaknesses of the 2007 pre-acquisition Fast Search platform are part of the MFSS "best in class" (page 5) search system. Source: Microsoft Enterprise Search Partner Playbook 2008: Guidelines & resources for Microsoft and Fast Enterprise Search Partners, © Microsoft Corp., 2008.

Second, Microsoft has recast the generation of metadata and the use of file attributes and user information to deliver sorting of results by "managed properties." The sorting function was available in Fast Search, but the Microsoft implementation makes sorting easier, adding date sorting. Google, for example, does not permit meaningful date sorting and lags behind MFSS as well as other vendors with this feature.

Third, Microsoft uses the jargon "deep results refinement" to describe drill down, "more like this", and similar point-and-click operations. The core Fast engine had this capability, but Microsoft has invested effort in making results exploration easier and more useful.

These three features make clear that Microsoft is taking advantage of index content, object properties, and metadata to eliminate the need for a user to know how to cre-

ate a Boolean query or perform certain types of operations on results sets. Most users can figure out that suggested or related content can be accessed in the left hand or right hand columns of the results display. The central results panel provide graphic previews, drop down options, and Google-style relevance ranked results. Administrators can mix and match these elements to create an "experience" appropriate for an individual user or a "group" of users by organizational unit or access control setting.



Spotlight Function

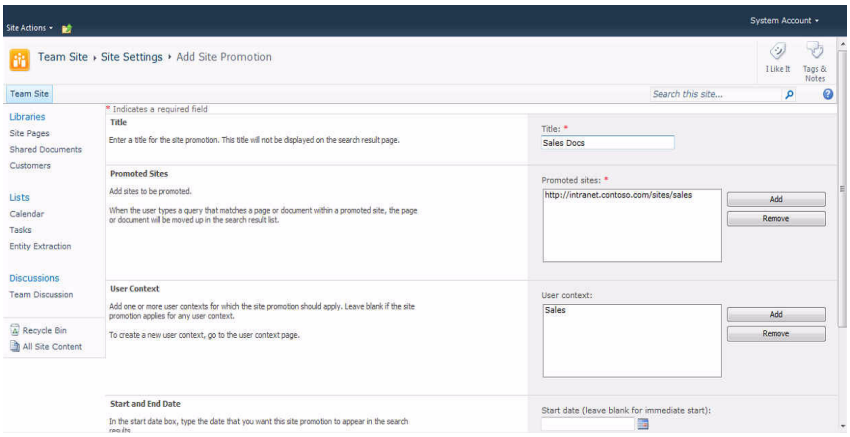
The major change Microsoft has implemented is in the graphical administrative controls. The 2007 Fast Search system had a graphical interface, but Microsoft has expanded what can be done via the graphical administrative interface and improved the presentation. Once again, the change is useful but cosmetic. Implementing these functions is not always easy, but improving an interface is somewhat less daunting than making major changes to the plumbing of a system with design elements dating from a decade ago.

Strengths

The obvious strength of Microsoft and its enterprise search system is that Microsoft "owns" the Fast Search technology. Microsoft's dominance of the enterprise desktop and the broad reach of its ecosystem translate to two advantages:

1. One way or another, Microsoft will be able to make the Fast Search system "work". The cost may be high and the journey arduous, but if the licensee has the stamina, the trip may be rewarding.
2. Microsoft's market presence and its ability to bundle, upsell, and discount its various servers including MFSS means that the Fast Search technology is going to be installed in hundreds of thousands of organizations.[52]

These two “molecules” combine to benefit the consultants who have been Microsoft certified to make MFSS work. The organization that embrace the Microsoft ecosystem pump more money for client access licenses, training, and software into Microsoft itself.



The MFSS hit boosting interface. Content can be weighted to appear at the top of a results list. This function is important in the enterprise when e.g. a change in a policy has to be displayed in certain situations. © 2010 Microsoft Corporation. From *Microsoft Fast Search Server 2010 for SharePoint Evaluation Guide*, page 45.

The phrase “too big to fail” may not apply to Microsoft in general, but the catch-phrase definitely means that the Fast Search technology core is going to be an acceptable choice for many organizations for the foreseeable future.

In addition, Microsoft delivers these payoffs:

- Many organizations see search as a utility, an add on or a nuisance. Microsoft can just upgrade SharePoint licensees when SharePoint’s default system will not do the job.
- Organizations with Microsoft Certified professionals on staff or accessible via a Microsoft Certified partner or services firm will have technical help down the hall or a text message away. Support for other vendors’ search systems may not be as available.
- Organizations are slow to change. Even if there are annoyances with Microsoft’s enterprise solutions, the cost and hassle of a switch are likely to be a significant hurdle. If the MFSS licensee is truly unhappy, then Microsoft has legions of Certified Partners who can sell snap in components that deliver a third-party solution without necessitating a complete platform shift.[53]

Microsoft could print a T shirt that says, “No one gets fired for going Fast.” We have our reservations about MFSS, but the system is likely to be a fixture in many organizations for the foreseeable future.

52. Microsoft offers SharePoint for Internet (MOSS FIS) licenses now for free to Microsoft Certified or Gold Certified Partners to make use of it for their own public Internet facing company Web site or extranet.

53. Snap in search components are available from such companies as Autonomy, Exalead, Fabasoft/Mindbreeze, and SurfRay, among others. For tagging, the Ontolica solution adds useful functionality. See www.surfray.com.

Cautions

We have identified three flashing yellow lights with regard to the Fast Search core and the 2011 MFSS.

First, Fast Search & Transfer failed as a standalone commercial enterprise due to its complexity, lack of scaling, and ad hoc nature. After exiting Web search in 2003, the Oslo- and Boston-based engineering teams raced to build out a full enterprise solution. As Fast Search's sci-fi marketing enticed licensees, the Fast Search engineers had to write code to deliver on marketing's promises. Not surprisingly, the wide range of features and functions in the Fast Search marketing materials proved difficult to get working quickly and economically. The delays and complexities precipitated the financial problems that pushed the company over the edge. Many of those engineering issues remain today.

Second, the Mars rewrite was Java-centric. The Microsoft purchase of the company in 2008 and the present release of the MFSS do not synchronize. We expect that the MFSS search system will continue to undergo fixes and subsystem rewrites. Stripping support of Unix and Linux simplifies the job, but moving the 1997 technology to 2011 and then to parity with next-generation systems like those available from Exalead, to name one alternative, is a very big job. Microsoft has quite a number of challenges at this time, and it is unclear that the engineering team will be given the resources and the time required to make the Fast Search technology live up to the marketing hyperbole.

Third, Microsoft's inclusion of a "free" or "baked in" search system in SharePoint guarantees that many organizations will upgrade. Our experience in search prompts us to observe, "many of these organizations don't know what they don't know." Stated another way, the use of MFSS looks like the "smart" and "logical" step to take to solve enterprise findability problems. Once the project is underway, search becomes caught in the whirlpool of upgrades, fixes, and customization that are part of the Microsoft experience. A demo is one thing; a system that meets user needs is another.

Net Net

Our recommendation is to take a look at the reviews and opinions of the various "experts." Then pick at least two other vendors' systems and run some tests on live content. Without a head-to-head test and a technically and financially anchored analysis, MFSS may snarl an organization at a time when information access is needed for the entity to thrive or survive.

MFSS Annex 1: Technology Partners

A complete list of Microsoft's technology partners is beyond the scope of this report. The table below lists vendors who provide "certified" components for MFSS. Microsoft has a bewildering range of categories; for example, "channel champion." We ignore these because most seem to be labels designed to extend the sales and marketing, not the technical capabilities of MFSS. When you access the partner "solution finder," note that Microsoft does not endorse any of the partners or their software.[54]

Microsoft Technology Partners (Selected)

Partner	Key Technology
Autonomy	Search solution for SharePoint
BA-Insight	Add ons for SharePoint, including a connector to ERP and CRM systems
Concept Searching	A concept classifier that is rules based
Doculex	Document management and search for SharePoint
Fabasoft (Mindbreeze)	Search solution for SharePoint
IO Informatics	Process and search disparate information
ISYS Search Software	Search solution for SharePoint
Knowledge Lake	Content processing solutions for SharePoint
Metafile Information Systems Inc.	Document viewer and search for SharePoint
Raritan Technologies Inc.	Search connectors for MFSS
Recommind	MindServer is integrated with and has a pre-built connector to SharePoint 2007, providing law firms with a highly automated and accurate enterprise search solution
SurfRay	Enhanced metatagging and content processing for SharePoint and MFSS
X1	X1 provides a centralized search window to find, preview and act upon data with patented find-as-you-type searching technology

Finding a specific software solution for a particular SharePoint or MFSS issue is difficult in our experience. The Solution Finder provides some help, but the efforts of search engine optimization "experts" have made precise retrieval time consuming and labor intensive. The situation tells us that a number of individuals are looking for help with Microsoft search systems.

54. Access the directory of partners and solutions providers via the search form at https://solutionfinder.microsoft.com/default.aspx. Verified on January 27, 2011.

MFSS Annex 2: Consultants

Companies mentioned in the text of the Microsoft profile are not included in this table containing representative resellers.

Endeca Technology Partners (Selected)

Resellers and Integrators	Key Feature
Alliance Global Services	MFSS engineering support services
Arke Systems LLC	A SharePoint road map is available from this company
Avalon Consulting	MFSS engineering support services
Cognizant	Portal development and search implementation
DTI Management	An independent solution provider in the field of intelligent information management, enterprise search and retrieval
ESR Consulting, Inc.	Firm builds enterprise search solutions
Hitachi Consulting	eDiscovery specialists
IDV Solutions	Geospatial add in for SharePoint
New Idea Engineering	MFSS engineering support services
R.K. Dixon Company	Can configure SharePoint to conduct effective searches for people
RDA Corp.	Has an online support service for Fast 5.x
Search Technologies	Search Technologies provides comprehensive search solutions and expert professional services to public sector and commercial enterprises
Vorsight Corp.	The company "controls the chaos" that accompanies enterprise search

Finding an "expert" in SharePoint or MFSS is not easy. A number of firms market aggressively, but there is no Consumer Reports type of service to provide an indication of these firms' competence. As a result, many organizations will rely on recommendations from business contacts or meet and interview potential suppliers at a trade show. Finding an individual or organization able to resolve in a satisfactory manner an MFSS problem is a hit-and-miss proposition in our opinion. Our recommendation would be to look for former Fast Search & Transfer engineers and talk with other MFSS licensees to get their recommendations as to potential contractors. The turnover at Microsoft, if it continues, will erode the institutional knowledge about the Fast Search "core" technology. *Caveat emptor* is a phrase to keep in mind.

Chapter 7:

Vivisimo

Vivisimo said in January 2011 that over 100 government agencies have moved from basic search to information optimization...

Vivisimo asserts that it is an enterprise search solution. The system is often evaluated in competitions against Coveo and ISYS Search Software.

Vivisimo at a Glance

Vivisimo Velocity performs "federated" search or "metasearch". *Federated search* and *metasearch* refer to a search system's ability to index diverse content, file types, and repositories which may contain copies, remove the duplicates, relevance rank the results, and display the results list. Results are placed in folders making it easy to explore related hits from the results list.[55] The company offers the Velocity platform and professional services.

The Velocity system takes a user's query and sends that query to different search systems. The results from each search system are presented in a single results list. Vivisimo has technology that performs very good deduplication and on-the-fly clustering of the results.

Licensees can, therefore, implement a type of guided navigation or faceted search while solving some of the more complex challenges associated with enterprise search. Vivisimo's core search and clustering technology has changed in relatively

55. The clustering function is sometimes called "content integration"

minor ways over the last decade. The core of Vivisimo is this original suite of functions.

	Basic Information	Option	Comment
License Fee	Starts at $65,000	Professional services are available	System allows users to access information from one search installation
Search product	Velocity 8	The system can be customized to handle customer support functions	Vivisimo is a key word retrieval system that performs clustering of results into categories
Technology hook	Repositioned as an enterprise search platform	None	Over the last five years, Vivisimo has added a feature to allow users to annotate results to improve relevance
Cautions	The system may require manual configuration	Vivisimo's professionals can handle customization	None
Selected partners	Swets	None	Vivisimo has few technology partners and resellers
Net Net	Vivisimo has transformed itself from a federated search system with clustering to a search-and-retrieval system the company positions as an enterprise search solution. Vivisimo can be used for a range of purposes, but it remains a key word search system with clustering and basic social functions like user-suggested indexing terms. The company took venture funding in 2008. The round injected $4.0 million into the company.		

Key Developments

Since 2006, Vivisimo's major developmental change was the shift from a metasearch system with clustering to an enterprise search solution. Vivisimo's most recent positioning is more grandiose than the firm's original metasearch premise. The firm says:

> Vivisimo Velocity Search Platform 7.0 delivers a new perspective on conceptual search with a new user interface that emphasizes choice in empowering users to express their intent and personalize their search, producing more relevant and more satisfying results. The new user interface for Velocity 7.0 provides easy-to-use check-box options to increase precision in searching for particular documents or to increase recall to find all related information. Configuration is web-based, allowing administrators to easily define default settings.

Investors forced a management change on the company in 2009. Raul Valdes-Perez was given a new role as Executive Chairman. Jerome Pesenti, one of the founders, retained his role as Chief Scientist. The rest of the management team was replaced at the urging of outside investors. Since that change, Vivisimo has repositioned the company to be an "information optimization company", whatever that means. The company has touted social functions that are little more than a feature for an Intra-

net user to suggest tags for a document. Vivisimo is also asserting that the present version of the system performs e-discovery. The approach makes use of on-the-fly clustering. What is clear is that the new management team is more marketing oriented, focused on winning government contracts and featuring better sales professionals than the previous executive line.

In the last 18 months, Vivisimo has worked hard to become a factor in the enterprise search market.

Vivisimo mobile device display. Cluster titles appear at the top of the results list.

New Officers

The key officers are:

- John Kealey, chief executive officer. He is the former President and CEO of iDirect Technologies. He began his career at Coopers & Lybrand and holds an MBA from Washington University in St. Louis. He is a manager, not a search expert.
- Kevin Calderwood, president. He was a founder and managing partner of Amp Capital Partners, a venture capital firm based in Reston, VA and Palo Alto, CA. He is a finance specialist, not a search specialist.
- Patrick Williams, senior vice president, enterprise unit. He is described as a successful executive sales leader. He is running Vivisimo's sales effort. He holds a Bachelor of Science degree in Mechanical Engineering from the University of Texas at Austin. He is not a specialist in information retrieval.
- Tracey Mustacchio, vice president of marketing. She is responsible for the company's global marketing initiatives, including its go-to-market and product strategies. She earned B.A. degrees in Math/Computer Science and Philosophy from Virginia Wesleyan College where she graduated as 1 of 10 Wesleyan Scholars.

Bob Carter, vice president, US government unit. He oversees all business development, sales and operations for Vivisimo in the public sector. He is not a search specialist.

One of the shifts was to pay Gartner Group to promote Vivisimo. The company became a "silver plus" sponsor for a number of Gartner Group events beginning in 2008. This approach seems to have worked. Vivisimo has more visibility among Gartner's clients than it had previously.

Discovery Module

Specific discovery functionality allows for classification of all available data for quick insight into themes and topics without entering a search query. Vivisimo uses clustering capabilities to produce a full auto-classification system on top of a search platform. The new Discovery Module's enhanced collaboration features of express tagging and native document export empowers search users to add their own knowledge and to disseminate content in its native format to colleagues and partners alike. The ability to tag, collect, save, share and archive documents by the thousands or millions through search queries enables knowledge sharing . The new Velocity Discovery Module joins existing platform expansion modules Velocity for Mobile and Velocity for Desktop available separately from Vivisimo.

Mobile Device Option

Vivisimo's system can output results for a mobile device. The approach preserves the clusters (unstructured navigation), pre-defined categories (structured navigation) and licensee-defined tabbed navigation (hit boosted content, for example).

New Buzzwords for SEO

Vivisimo is now using the phrase "information optimization" to describe its search solution. Though meaningless, the phrase makes it easier to find Vivisimo in a Web search. "Information optimization" replaces the abject "search done right."

History

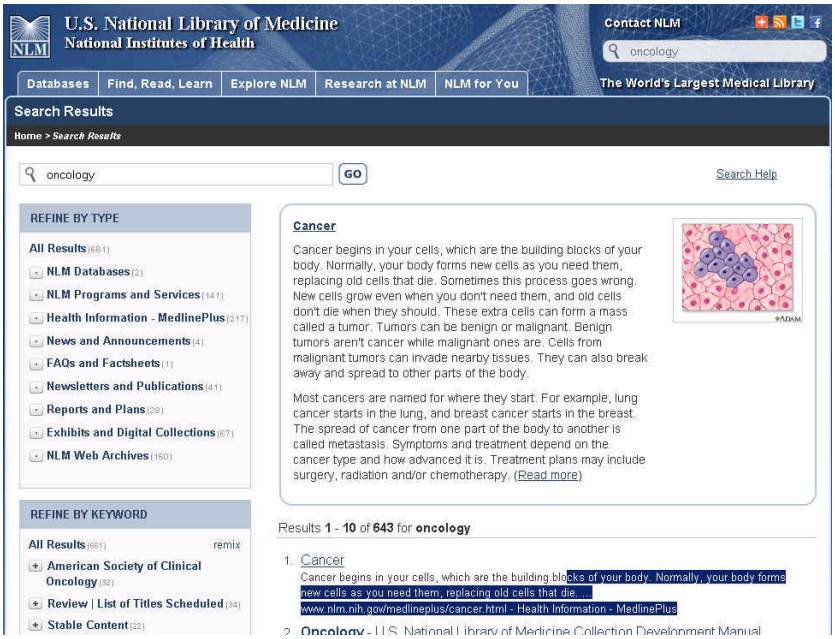
Vivisimo is an enterprise software company founded in June 2000. The technology was developed at Carnegie Mellon University's computer science department. The core differentiator for Vivisimo is its document clustering innovation. Work began in 1998 with an initial National Science Foundation Grant. Vivisimo was essentially a niche player with revenues stuck in the $4.0 to $5.0 million range. Today the company is estimated to have revenues of about $10 million thanks to the sales approach of the new management team. The entrepreneurs behind Velocity were Jerome Pesenti, a computer science graduate student, and his advisor, Dr. Raul Peres-Valdez.

Vivisimo was designed to send a user's query to multiple systems. Vivisimo would then receive the results and cluster them on the fly. The user would see a single list of de-duplicated, federated results. Vivisimo was one of the first vendors to offer a "metasearch" or "federated search" solution for the enterprise.

In 2000, when Vivisimo was founded, an organization with information in a database located at headquarters, a records management system located in the engineering department, and a dedicated server receiving news from Dow Jones had an all-too-familiar problem. An employee looking for information about a particular topic

would have to find someone to run an SQL query to pull the data from the database, then log in to the server with the records management index and run a query on that system, and finally head over to the corporate library to get access to the 30-day news repository.

Running a single query that would "touch" each of these systems and deliver one results list with the duplicates removed was a very difficult and expensive proposition. One vendor—Verity Inc., now a unit of Autonomy Corporation plc—had a system that could provide a similar functionality but Verity required multiple Verity servers. Performance was miserable and the installation and maintenance of the Verity system was labor intensive. Vivisimo knew that the Verity approach was to put specialized computers at each of these information points, index the content, and then allow the user to enter a single query. Verity would pass the query to each of its servers, collect the results, and display to the user a single list of results. Verity worked, and the success of the company was due in part to its ability to have a solution to this common enterprise information problem. Vivisimo targeted the problem but still struggled to generate sales.



The National Library of Medicine implementation displays graphics in results.

The original management team resisted going to the venture capital market. After years of slow growth, the founders did accept injections of capital. Although the firm had received National Science Foundation grants, Messrs. Valdes-Perez and Pesenti were not prepared for the demands of investors. The company took $4 million from North Atlantic Capital in 2008. By 2009, management change was underway.

Today, Vivisimo is a hybrid. The company sells integration services and search solutions packaged as a Swiss Army knife solution, eDiscovery for legal applications, and a customer support platform for call centers and self-service use.

Vivisimo in Action

Because vivisimo is less well known than some of the other vendors profiled in this report, the best way to get a feel for Vivisimo is to look at a showcase installation of the firm's Velocity 8 system.

Vivisimo is now using traditional sales methods to break into the high-profile world of search that Autonomy, Endeca, Exalead, Google, and Microsoft Fast occupy. Vivisimo held the contract for the US government search for two years. The firm lost that contract and now has a showcase site at Airbus and the US National Library of Medicine.[56] Originally the NLM wanted a SharePoint search solution, but Vivisimo won the competition for the account.

The NLM service makes use of several of Vivisimo's features; namely:

- The query is passed against the index of a wide range of sources within the sprawling National Library of Medicine and National Institutes of Health systems. In parallel, the query is also run against the index for a number of other sites, hand picked by the client that are indexed directly by Vivisimo.
- The results of the query are then treated as one result set with any duplicates automatically removed. Vivisimo does a very good job of deduplication.
- Hit boosting. The idea is that a certain piece of information can be placed at the top of a results list.
- Native support for Microsoft SharePoint.
- Tweaks to scaling.
- A query auto complete function.
- The results at the time the duplicate set has been generated are analyzed for conceptual similarity and then grouped or what Vivisimo calls "clustered". The grouped results are labeled using a term or phrase "discovered" by the Vivisimo system via a statistical count, a dictionary, and any terms provided to the system at set up. The number of results in each cluster is shown after the group name.

Not shown in the screen shot is Vivisimo's "key-matching" for common terms. The system can display the most likely answers to be highlighted above the results listings. Key matching is Vivisimo's version of guided navigation. One interface option in Velocity is "tabs" of clustered results. These tabs allow the user to jump between the data sources that most likely contain the answer.

Agency, media, and citizen response to the Vivisimo-powered NLM Web site has been generally positive, but the previous SharePoint and Endeca system were terrible. One NLM official said, "The new system is faster, more user friendly, and more comprehensive than the two previous versions of our NLM site."

56. In 2004 Vivisimo created the Cluster Med service as a demonstration.

Velocity Search Platform

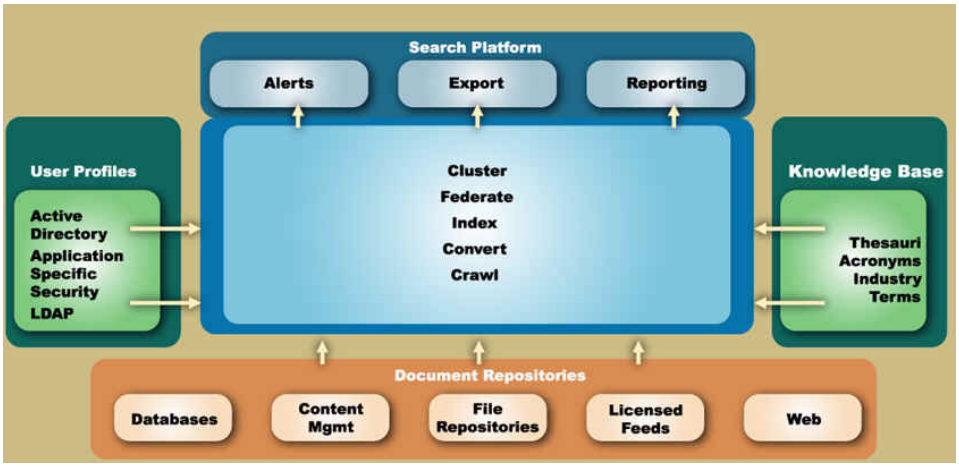
At the outset, Vivisimo was a one-trick pony; that is, it was a federated search engine that featured on-the-fly clustering.

Vivisimo has grown from an invention at Carnegie Mellon to a growing software company with an international reputation.

Velocity has the following components:

- Clustering engine - Automatic categorization of search results without the time and expense of taxonomy building
- Content integrator - System to combine and deduplicate search results from multiple servers, collections, and document repositories on the licensee's Intranet or on the public Internet
- Enterprise search engine - System to allow a user to search for information via a traditional search box.

One of the interesting methods used by Vivisimo is that when the first 500 results are available, the clustering function is performed. These "folders" and their categories are then used to hold the subsequent search results from the user's query.



Block diagram of the general Vivisimo architecture, pre-2008.

Velocity allows licensees to deploy a comprehensive search solution to search simultaneously multiple information sources and present the results in organized folders from a single consistent interface. According to Raul Valdes-Perez, "Velocity meets two major challenges organizations face when deploying enterprise search. First, most search systems do not cluster without training or having a manually-created taxonomy for the clustering system to use. Vivisimo does not need this. So the customer gets the benefit of clustering without the training step and the cost

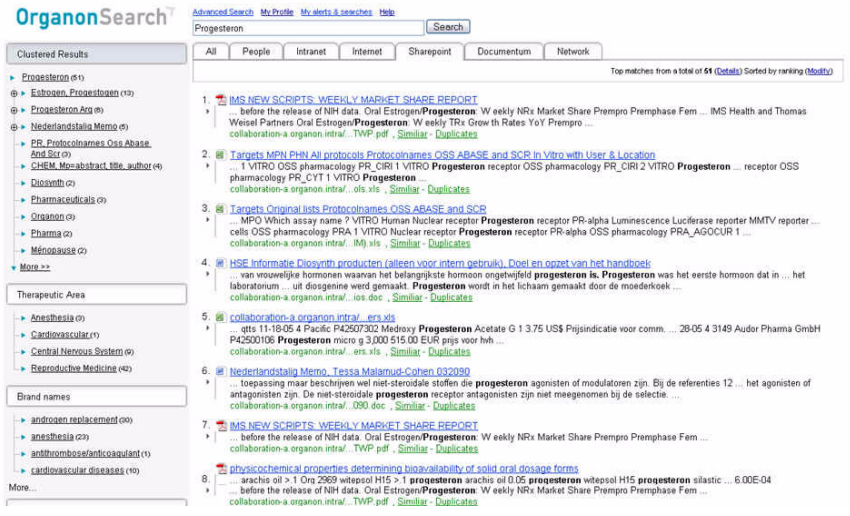
of a custom category list. Second, Velocity can index content from many different sources and make that content available as if everything were sitting in one big server on the user's desk. Because we glue together many different sources, the user needs to use one system to find information. We also deduplicate the results, so the user doesn't have to do this process manually."

In today's environment most enterprises already have information assets and some means of searching them. This creates silos of enterprise content. Vivisimo recognizes this and advocates that enterprises build upon their pre-existing search systems by augmenting them with untapped resources that can be leveraged using Vivisimo Velocity.

Vivisimo built its architecture under the belief that modern search solutions should:

- Leverage existing enterprise search assets by means of content integration of internal and external content sources. Allow administrators to tune the search portal by routing queries to specialized search engines depending on patterns in the user's query.
- Allow users to find all materials from one search box
- Cluster search results into categories, without the costs and complexities of building taxonomies. Provide key-matching for the most common queries and provide an immediate answer to users.

Vivisimo Velocity's search architecture is in line with modern thinking about parallelizing work, scaling, and extracting performance from commodity hardware. The innovation for Velocity is anchored in sending a query to different content sources, obtaining results, de-duplicating the results, and clustering them for the user.



Powered by Vivisimo Velocity, a search on the Organon service from the publications database crawled by Vivisimo's search engine. Results are clustered into topics.

Velocity does not require preprocessing of documents or data. Enterprises have control over how their content will be indexed, never having to reformat documents or change how documents are created and organized. This is important when content has evolved over time and no standards for creation, organization, or manage-

ment were developed. Many other search engines require enterprises to preprocess or reconfigure documents or organizational structures before the crawling can even begin, often requiring dedicated resources and weeks in time. Vivisimo's approach does not force any such preprocessing.

Additionally, Vivisimo Velocity's search engine is unique in that it supports one-to-one, one-to-many, many-to-one, and many-to-many correspondence between documents or search results and matching URLs. Most search engines force correspondence in a manner that one document or search result correlates to a single URL. This often results in inaccuracies and less relevant results and summaries. Vivisimo generates several independent results from a single page, such as a blog's front page in which each entry is a unique result, by parsing the resulting XML feed or HTML output with XSL to provide clustered search results. Velocity 8 permits display of rich media objects in a results list.

The Vivisimo Velocity search is also able to leverage existing metadata. Unlike many other search solutions that require metadata to be embedded within each document, Vivisimo has the ability to attach external metadata to documents automatically. Administrators can easily attach metadata to web URLs or Adobe PDFs even when they do not control those documents.

Velocity makes use of a “staging area” where crawl results are copied and processed. When the index update is complete, the system updates the production index with the refreshed index from the staging area or server.

Representative Customers

Vivisimo’s customers include:

- Airbus
- CB News
- Canon
- Cisco (now shifting to Lucene/Solr)
- Highwire Press (Stanford University)
- Institute of Physics
- Johnson & Johnson
- Eli Lilly & Company
- National Library of Medicine
- Organon
- Procter & Gamble
- Thomson Scientific
- Tyco Electronics

Technology

Vivisimo is one of the vendors positioning search and clustering as a system that can play different roles depending upon the customer's requirements. One role is to "glue" together different collections, servers, and content sources in a "virtual repository." Its other role is to use the Velocity system as an enterprise search engine. Vivisimo has positioned Velocity as a customer support system, touting the firm's clustering as an advanced discovery system.

Clustering

Vivisimo uses a proprietary clustering method which outputs hierarchical clusters.

A major breakthrough was made by Vivisimo in dynamically clustering results and generating labels. Vivisimo still uses a specially developed heuristic algorithm to group - or cluster - text documents. This algorithm is based on an old artificial intelligence idea: a good cluster - or document grouping - is one which possesses a good, readable description. The Vivisimo method for document clustering and meta-search software categorizes search results on-the-fly into hierarchical clusters. Vivisimo Velocity is built on a modern architecture and takes advantage of XML and XSL standards. Configuration can also be done through an extensible set of REST/SOAP APIs. In addition, Vivisimo's solution supports the ability to handle terabytes of information with minimal infrastructure costs compared to other solutions on the market. In practice, there are upper boundaries on the Vivisimo method due to the trade offs between the amount of data retrieved and what must be processed for clustering. Vivisimo typically clusters the first 500 results. As a result, precision and recall can suffer because the full set is not processed, but the outputs are "good enough." Through Web services and SOA protocols, Vivisimo's solutions can easily pull data from existing applications, creating a universal gateway to securely access information among disparate systems within an organization.

Traditional solutions for organizing information like taxonomy building and categorization are complex, time consuming, expensive and difficult to maintain and scale. Vivisimo is trying to change the economics of organizing information by building a solution that is inexpensive and plugs into existing search infrastructures.

Vivisimo Velocity was founded on clustering. Vivisimo's approach allows clustering to be performed "outside" of the search engine. The clustering technology does not need to run on the same platform or server as the search engine. A licensee of another enterprise search product with clustering that slows the indexing process can turn off the enterprise search product's clustering services and "plug in" Vivisimo. The performance of the enterprise search engine indexing goes up and the users of the search system have the benefits of clustering. While others may say that they have clustering capabilities, the performance penalty imposed by other search solutions is high and generates lower quality results.

The content integration provides a single point of access to internal and external content. Vivisimo Velocity can interact with over 600 of the most common data

types inside of an organization. For external content, Vivisimo Velocity works seamlessly with web search engines, licensed feeds, and anything with an HTTP connection.

Vivisimo's functions interact with any search engine through HTTP connections and use XML search engine output or parses its default HTML/Text output, thus avoiding the performance penalty imposed by some search solutions that cluster by performing analyses when contents are indexed.

The clustering is highly configurable and can work in reverse for organizations with a taxonomy. That means that Vivisimo can configure the clustering topics to be based on an organization's existing taxonomy and software will place the relevant search results in the pre-existing topic listing. Organizations can have static topics and the dynamic clustering to ensure that all relevant topics are presented to the end user.

A licensee can integrate the clustering engine into almost any Internet or Intranet search-and-retrieval system. The clustering engine can also be integrated into application software that can make use of the Vivisimo cluster data for data mining or other uses. A CMS with a large number of documents and an embedded search engine from Autonomy, Verity, or another provider can make use of the clustering engine when displaying results. No underlying architectural changes are necessary. However, a licensee will require some knowledge of calling the clustering engine functions and displaying the results on a Web page.

Clustering Minimums

Vivisimo's technology requires a search engine or document index that consistently returns 50 to 500 results. An organization with a small amount of information indexed for search and retrieval will not benefit significantly from the Vivisimo technology. With too few documents for the Vivisimo algorithms to process, the clustering process is not likely to add significant value.

The general characteristics of the Vivisimo search system are a blend of traditional word-and-phrase and faceted or guided navigation search systems. Vivisimo, without any manual taxonomy creation or computationally intense training processes, can cluster results into categories.

We noted these functions:

- Social functions so users can assign tags. The method is little more than the uncontrolled indexing permitted by SharePoint. Vivisimo offers document and keyword tagging, rating and annotation of documents and search terms, use of shared virtual folders, and content mash-ups.
- Clustering requires no maintenance unless a licensee wishes to use dictionaries and chooses to update these manually.
- No pre-processing of documents or collections.
- Simplified installation that uses defaults to eliminate the hand editing of the scripts that make Velocity operate.

- Support for clustering results from standard databases such as Oracle and SQL Server among others. Vivisimo also supports Lotus Notes repositories.
- Additional graphic set up screens which are reminiscent of the Google “fill in the blank” and “choose from a drop down menu” method.
- Vivisimo now has training programs, usually available in Pittsburgh, Pennsylvania, plus online documentation and walkthroughs.
- A flexible alert system.
- Ability to export/save/email results.
- Optimized search results for clustering.

Indexing Highlights

A quick review of the basic indexing features of the Velocity system includes:

Tag Tuning

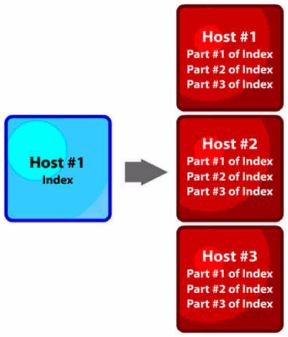
Localization, customization, and tuning are possible via the Vivisimo Velocity API, by specifying stop words and stop phrases, metadata, relative weights for the text fields (for example, title versus abstract), globally important words, lexical stemming, and others. Large sites will want to invest some time in this activity.

For licensees needing customization, a knowledge base module accepts company and industry-specific knowledge such as synonyms, acronyms, spelling variants, taxonomies, and others. A licensee using Oracle's Text search engine as the search system can use the Oracle taxonomies with the Vivisimo clustering software. The clustering processes within Oracle Text can be disabled in order to speed up the indexing process.

Licensees should note carefully that these categories are selected from the words and phrases contained in the search results themselves. This means that categories will be as up to date—or out of date—as the content in the search system, or more specifically, the content in the search engine's result set.

Index Performance

In order to extract good performance from the Vivisimo content processing and index, the company uses two methods. Velocity replicates indexes across servers. Within a server cluster, Vivisimo will segment the index. The idea is that the combined strategy delivers good performance. Scaling can be accomplished by adding more servers to hold replicated indexes or by breaking an index into additional segments on physical servers or virtual machines. Beginning with Velocity 5.5, Vivisimo has provided a graphical interface to add “mirrors” to the index. Once the “mirror” has been set up, Velocity automatically distributes the work load for indexing and/or query processing.



Vivisimo asserts that it can handle more than 1,000 queries per second. New compression methods allow further speed ups. Vivisimo asserts that there is "no hard limit on how many documents can be indexed on a single server – 10M, 50M or even 300M documents can be indexed on a single server (dependent on storage attached and response time needed)."

Language Support

Velocity handles alphabetic languages by including a language specific stop list (list of non-informative words, like English "the", German "nach", Spanish "para", French "toujours") and a stemmer, which recognizes similar meanings among syntactic variants like English "helps", "helpful", and "helping". Vivisimo offers versions of its Velocity components for the major European and Scandinavian languages: Danish, Dutch, French, German, Italian, Norwegian, Portuguese, Russian, Spanish, and Swedish. Globally, nearly 50 languages are supported, including Arabic, Welsh, Japanese, and Chinese. Vivisimo products embed other semantic and syntactic knowledge, but the company declines to provide details.

Customizing Velocity

A licensee can customize most Vivisimo functions through administrative screens and templates.

Vivisimo's clustering subsystem is made up of CGI and PHP scripts as well as XML files. Consequently, a licensee can integrate its functions into almost any third-party application. The company provides API documentation that contains explanations and sample code for integrating the clustering system in programs and system. The current version ships with useful information and examples that document the XML input and output of the system. A system administrator can manipulate these files to further customize the system and its outputs; for example, the change can be as trivial as eliminating the folder metaphor or as sophisticated as modifying the data displayed for each cluster.

Content Integration

Vivisimo supports "content integration", a buzzword that has been replaced by "mashup." The idea is that a user's query will display results from a database and other sources. The de-duplicated results list presents a metasearch to the user complete with folders or clusters to permit result set narrowing. Regardless of where the content is stored (internally or externally) or the number of disparate sources, meta searching will present a single unified view.

Strengths

Vivisimo Velocity offers reasonably rapid deployments into any type of search application. However, large scale deployments still require the personal intervention of Jerome Pesenti, the Chief Scientist for the company.

The benefits of the Vivisimo approach include:

- Ability to leverage pre-existing search and information assets
- Ability to search all content from one search box

Cautions

Vivisimo has fewer drawbacks today than it had three years ago. The company has added some gratuitous features such as support for user tagging. The main change is not the technology; the company has become more marketing oriented since the management shakeup.

Setting up a system that performs multiple functions can be tricky for those without a solid understanding of search, clustering, and script-based configuration. The graphic administration screens put the most important controls in one place. However, a system administrator coming to Vivisimo with modest search experience is likely to need assistance from Vivisimo's technical support engineers, and manual edits of the scripts "under the hood" are often required even though the core system is not a decade old.

Vivisimo provides the information needed to handle relatively simple indexing jobs and the more complicated ones as well. The documentation is useful, but the key to certain configuration settings is in the sample code Vivisimo provides. Finding the half dozen lines needed can be difficult for someone not used to reading code for a solution to a configuration or setup issue. That said, Vivisimo's current documentation is much more thorough and user friendly than the information accompanying earlier versions of the system.

Most systems performing categorizing or clustering functions sometimes put a document in a category that a user might describe as "not intuitive." Vivisimo, to its credit, is clear about the glitches that can occur when software "reads" and categories documents. "There are many reasons," says Raul Valdes-Perez. "First, the mathematics of the algorithms can make a distinction that a human might not. Our algorithms perform quite well, and we are working to enhance them all the time. Second, the document Velocity processes may have an ambiguous title or have some other weaknesses in word choice or writing. Finally, due to the nature of language, a new concept discussed at a conference may not yet be captured in the content in an index. Nevertheless, reviewing documents in either a relevance ranked list or in clearly labeled categories gives the user more flexibility in finding information."

Vivisimo also works best when there are more than 50 hits. It may work better for "discovery" cases rather than for precision retrieval by power users.

Vivisimo Velocity has moved beyond being only a clustering "add on utility" to a complete enterprise search platform. However, the company recognizes that in today's world organizations already have pre-existing search resources and encourages leveraging of these existing resources. Vivisimo assists organizations in identifying their high-value information assets and spotlighting them based on end user queries. This is all done through a combination of the federated search and search solution.

In addition to its core search solution, Vivisimo's technical team has created a unique functionality for users dealing with large numbers of hits and facilitates the implementation of on-the-fly grouping of related information. Vivisimo is pragmatic about clustering. The company provides ways for a licensee to add specific terms to assist the clustering engine in naming folders and enhancing the performance of the automated system. Unlike companies that insist their "intelligent software" can work without human intervention, the Vivisimo team allows a licensee to run the system "hands off" or use human-developed word lists.

Net Net

Vivisimo's approach is unique, and it is almost certain to be emulated by other companies with a specific search value-add that can be used to enhance virtually any search-and-retrieval system. Vivisimo's "no muss, no fuss" approach to a complex information process is a refreshing change from search system developers who want to change the way employees do their work to make search more useful.

Our recommendation is to compare Vivisimo to solutions from companies of a similar size (Coveo, dtSearch, ISYS Search Software, among others.)

Vivisimio Annex 1: Resellers

One indicator of the "impact" of a search vendor is its lineup of resellers and partners. Vivisimo has only two resellers listed on its Web site.

Vivisimo Resellers (Selected)

Partner	Key Technology
Groupnet	Based in Tokyo, Groupnet provides technologies for enterprise search
Swets	A subscription agent which resells Vivisimo in its own SwetsWise Searcher

Vivisimo Annex 2: Technology Partners

Vivisimo has one technology partner listed. It is not clear if this relationship with Microsoft is active or inactive.

Vivisimo's Technology Partner

Partner	Key Function
Microsoft	Microsoft was part of the FirstGov.gov(later USA.gov search) project which has since changed search vendors.

Chapter 8:

Traversing the Landscape

I want to close this report with a few observations about how the landscape of enterprise search may change in the next 12 to 24 months. A number of trends are building like a storm front on the horizon. Marketers and procurement teams are hopeful that one or more of these trends will reshape the market in order to make enterprise search a success.

Landscape Diversity: Upsides

Cloud computing appears to offer organizations a way to shift some costs, sidestep the capital expenditure of old-style search systems, and reduce headcount. Cloud computing is a 21st-century version of time sharing. The present incarnation appears to have some tangible benefits for users who no longer have to be tethered via a real or virtual cable to the organization's in-house computing system. However, for search, cloud computing is a work in progress. Some organizations have made significant progress and are able to offer their customers industrial-strength search with minimal on-site infrastructure. Autonomy and Exalead are, based on my research, ahead of the pack. Other vendors are rushing to harden their cloud services. High-profile cloud outages from the likes of Amazon and Google have made some of the strongest supporters of cloud computing moderate their statements about relying upon a multi-layered network and system infrastructure for certain mission critical tasks.

Open source search solutions are available. Vendors such as Squiz/Funnelback, Lucid Imagination, FLAX, and even IBM provide a community-supported search solution. However, expertise is required to tailor an open source search system to a licensee's requirements. The engineering expertise to shape an open source search solution is not part of the bag of tricks for many information technology professionals. Consequently, the infrastructure, installation, customization, and optimization

expenses can be as high as for certain commercial enterprise search solutions. Some money can be saved, but in many cases, the money is just moved from one budget line item to anther. Enterprise search, whether open source or proprietary, can be an expensive proposition.

Vendors such as IBM and Microsoft, among others, offer a broad range of components that are available to supplement, extend and amplify a basic search installation. The idea is appealing. On a single platform, a licensee can deploy search and then add on such features as on-the-fly processing of multilingual content, automatic indexing and metatagging, and merging structured and unstructured content. The reality for these "kitchen sink" systems is that with each additional component, complexity goes up. The benefit of a common IBM or Microsoft platform often proves illusory. First, there is the issue of lock in. All vendors want to keep a customer. There are many ways to accomplish this, but lock in reduces scope of action and often requires an additional license fee. Second, when new innovations become available, the vendors focused on lock in can take months or years to implement a new feature. The challenges of processing social content in SharePoint or in OmniFind are examples of the "friction" lock in imposes on an organization attempting to make search relevant to its users.

Another trend is the positioning of a key word search system as a way to improve the efficiency and reduce the costs of business intelligence, customer support, or eDiscovery. There are enterprise systems tailored to perform business intelligence functions. Many of these systems include a search component. However, most search systems are not business intelligence systems. Generating canned reports may add eye candy to a PowerPoint slide, but the inner workings may be too limiting for certain applications. Customer support is a specialist area as is eDiscovery. The current vogue is to position any search system as a specialist system. The approach makes marketing sense, but it may not work in the real world.

What is evident from the diagram in the introduction to this report and from the table of enterprise search vendors is that there are many companies offering enterprise search systems. Choice is important, but when there are many choices and no easy way for a generalist to differentiate systems, analysis paralysis is a common result. Enterprise search vendors come and go. In the last two years, Convera and Delphes ceased operations. Other vendors have a very low profile, serving a handful of customers, and teeter on insolvency from quarter to quarter. The financial stability of most enterprise search vendors is difficult for a procurement team to determine. Only a few search vendors are publicly traded and those companies, like Google, do not break out the revenues from the licensing of its enterprise search products and services. The financial picture for most vendors remains bleak. A few have sufficient resources to continue business if the economy falters in the next two years. One reason certain vendors' products find their way into blue-chip accounts is that the firms have significant financial and technical resources. Are these vendors' search systems the best? Maybe not. But the vendors will be in business. The effect is that a few firms dominate and dozens upon dozens of smaller firms fight for specialist contracts. To make the sector more tumultuous, new companies enter the market.Perfect Search and Sophia Search are examples.

The numerous options combined with the challenging financial climate have an upside. An organization seeking an enterprise search system can play different vendors against one another in an effort to craft a good deal. One reason is the emergence of open source options with zero license fees and a promise of reducing the direct and indirect costs associated with proprietary enterprise search solutions. It is difficult to compete with free. The other reason is that certain vendors are bundling or embedding enterprise search within other enterprise applications. Examples may be identified in the packaging and pricing approach of IBM, Microsoft, and Oracle.

Another upside to the enterprise search market is a surprise to many. The technology is sufficiently mature that any of the mainstream systems can be made to work in a satisfactory manner if three conditions are met.

1. Requirements are defined and used in the selecting and licensing process. Without requirements, the likelihood of any enterprise search vendor's system working properly decreases significantly.
2. Appropriate resources are available. Without time, funds, and technical expertise, none of the systems mentioned in the text, its tables, or diagrams of this report will yield a high degree of satisfaction among users. Poor budgeting often hides behind unfair criticism of a particular vendor's technology. The vendor is blamed, but the blame rests squarely on the manager of the search procurement process.
3. Upstream and downstream processes are ignored. A search or content processing system is a dependent service. If content must be transformed or normalized prior to indexing by the search system, the costs and time have to be factored into the deployment. Criticizing a search system for missing content that is not in the index often shifts the problem in an inappropriate way. The outputs of the search system will not meet the needs of certain users in an organization. If the requirements do not reflect those specific needs, many organizations criticize the search vendor for this failure.

These realities point for the need for a pragmatic look at the licensee's responsibilities in the enterprise search process is essential. Most of the churn in Fortune 1000 companies is a direct result of licensee mismanagement, not of technology shortcomings in the enterprise search vendor's system.

New Challenges

Several touchstones will be found on a path through the enterprise search landscape. First, search will be absorbed into other enterprise applications. I have seen a steady increase in the appetite for search-enabled applications and a decrease in users' interest in learning how to hunt for information in several systems. If I am correct, this shift means more consolidation of big accounts among a handful of top vendors. In addition, enterprise search as a stand-alone chunk of software, whether on premises or in the cloud, will diminish. A number of vendors will be directly affected by the "disappearance" or "submersion" of search into other applications.

Second, the sharp increase in rich media will make many existing enterprise search applications warp and buckle. For some organizations video is becoming the primary means of explaining a product or a service. Google itself allows engineers to explain a service in a short video placed on YouTube.com. Indexing and making video content searchable is going to force some vendors to exit the enterprise search sector. Despite the chatter in the trade press and technical journals, text is becoming a smaller part of the content within organizations. Dealing with structured and unstructured data is not, contrary to popular belief, the big challenge. The big challenge is rich media.

Third, social content is not handled in an effective manner by some enterprise search systems. The workarounds required to make SharePoint more like Facebook are complex and interesting from a technical point of view, but licensees do not want "interesting." Licensees want collaboration and other types of interaction to be searchable by entity, information object, date, time, and context. Most enterprise search systems have work to do to deliver a satisfactory experience for a person looking for social information.

Fourth, most enterprise search vendors cannot process in an cost-efficient manner certain types of publicly accessible Web content. In addition, the firehose of short items generated by Twitter, enterprise search systems struggling to cope with blog updates,RSS (really simple syndication) feeds, and key facts that are changed on dynamic pages from competitors or popular airline reservation systems second by second.

Will these challenges be resolved in the foreseeable future?

Not easily. Improvement in enterprise information access are incremental and evolutionary, not revolutionary. Even the impact of new hardware like tablets and new architectures such as cloud computing are coming slowly to search. The jargoin of business intelligence and predictive analytics cannot change quickly the fundamental challenges of information in an organization.

What is certain is that enterprise search systems will improve - and the complexities of enterprise search will increase.

Glossary

Search Terms

Over the years, I have prepared a number of word lists for my different reports and monographs. This report references more than two dozen vendors of enterprise search systems. A broad range of technology is represented among these companies. I have, therefore, assembled a glossary that assists the reader in familiarizing himself or herself with the jargon of search. These definitions are not intended to be definitive.x

Term	Definition
Algorithm	A numerical recipe. A series of specified steps in a mathematical process. Common text mining algorithms are implementations of Bayesian statistics and other manipulations of values derived from text.
Algorithm settings	The settings that specify algorithm-specific behavior in a procedure. Text mining systems include settings that are generated by the system and some that are under an administrator's control. Guessing at a setting can generate unusable outputs. The settings represent ranges for a valid statistical output from a procedure.
Appliance	An instrument, apparatus, or device for a particular purpose or use. Here: Search software and hardware integrated into a server that will plug directly into a standard rack space.
Appliance vendors	Businesses that sell appliances; for example, a search toaster like Google's Search Appliance.
Application Programming Interface (API)	A method for connecting external programs and functions to a search system.

Term	Definition
Assisted navigation	A point-and-click interface for exploring information or performing a search on a topic by clicking on a hyperlink.
Association	A data mining function that shows the relationship among items.
Attribute	A name or a data type. The attribute relates to items in a column in a database table or indexed with a common tag. Attributes can have importance, be collected in a common set, and be used by other processes.
Autocategorisation	Deriving one or more categories that relate to terms found in a retrieved set of documents as a means of guiding the user in refining the search.
Automatic classification	Delivering classification of results without use of a human subject matter expert.
Backup device	A piece of hardware on which copies of software and information are stored for precautionary measures.
Bayesian inference or Bayesian statistics	A statistical approach that calculates the probability of a hypothesis being correct by evaluating the prior probability of the hypothesis and the experimental data supporting the hypothesis.
Behind-the-firewall search	Indexing or searching information that is not publicly available on a company's network behind protective measures. Most organizations prevent unauthorized access to internal information, but the behind-the-firewall search is available for employees who need information for work purposes.
Business intelligence	High-value solutions and information on demand based on information about specific businesses.
Categorical attribute	An attribute where the values directly rate to specific categories. Categorical attributes are either non-ordered (nominal) like age, gender, etc.; or ordered (ordinal) such as high, medium, or low.
Categorization	The process of putting items in categories; for example, items such as IL or Kentucky are recognized, extracted, normalized, and placed in a category tagged 'State.' May be a synonym for classification.
Certification procedure	Verification of hardware or software to make certain it meets specific requirements.
Character	The smallest component of written language with semantic (significant meaning) value.

Term	Definition
Character set	Complete collection of characters for one or more language systems.
Classification	Categorization; that is, a function to distribute things into categories of the same type.
Cloud-based computing	Programs and services running on a third-party server via the Internet. An updated term for time sharing.
Cloud-based service	A service that delivers applications via the Internet.
Cluster	A grouping of items with similarities.
Clustering	Grouping similar documents based on content. Clustering may be hierarchical, probabilistic, overlapping, exclusive, model-based, etc. Used as a synonym for classification or categorization.
Codd RDBMS	A Relational Database Management System that follows general rules as proposed by Edgar F. Codd; for example, MySQL or SQLServer follow the Codd model.
Collection	A set of documents or databases that can be uniquely defined, e.g. a set of documents on a specified shared drive.
Computational linguistics	A discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty.
Concept trees	A graphic representation of the topics and subjects for the content processed by the search system.
Configuration files	Files that govern the operation of search system and its sub-systems.
Connectors	Software scripts that allow two systems to exchange information.
Content processing	The processes that convert a document into a form with index terms and other items in an index to permit a user to perform a search or search-related action.
Controlled vocabulary	A set of official terms developed prior to indexing. May include definitions and tags to map terms to a classification system. Often includes preferred usage terms. Expensive to develop; require continual editorial attention.
Crawler	A software program or script that identifies new/changed content and copies the new/changed content to the search application for indexing.

Term	Definition
Cross-validation	An evaluation technique to determine the accuracy of a classification or other model. When there are too few cases for using separate sets of data for model building and testing, cross-validation is relevant. The data table is divided, with each segment in turn being used to evaluate a model built using the remaining segments. Cross-validation is a feature of Naive Bayes and Adaptive Bayes Networks.
Custom scripts	Programming code lines developed for a specific purpose in running software.
Dashboard	A type of interface that displays a range of information of interest to the user as determined by a personalization component or by the system analyzing the user's query and generating a report.
Data Mining	The process of discovering hidden, previously unknown, and useful information from a large data set. Text mining is a subset of data mining.
Database	A structured array of items (e.g. the license plates of vehicles).
Dictionary lookup	A function that identifies a term and then consults a knowledgebase or word list to obtain semantic (significant meaning) or other data about the term. Many automatic systems reduce computational demand by performing look-ups in word lists.
Disambiguation	The linguistic process of clarifying the meaning in a particular sentence; for example, “he ate the cake on the sofa .” Was the eater on the sofa, or was the cake on the sofa?
Document-term matrix	Values representing the terms in a document, placed in a table. These values can be subjected to additional mathematical procedures.
Drill down	A method of exploring search results or data; for example, a user clicks on a hot link and the system displays underlying data; hence, “to refine a search query.”
Early binding	An approach to maintaining the confidentiality of documents by executing a query only against those sections of the index that do not contain these documents.
Enterprise	A commercial enterprise; hence, enterprise software as distinct from software used on a single user's laptop computer.

Term	Definition
Entities	The names of people, places, companies, products, addresses, email addresses, dates, and similar data. Some text mining systems can separate the elements and tag each item or place each item in an appropriate column in a data table.
Entity extraction	The automatic identification of the names of people, places, dates and other entities through a set of rules.
Epistemology	The study of knowledge by its origin, nature, limits, and methods.
Extraction	Extracting relevant information from a document; for example, identifying the most important sentences in a document in order to generate a summary of the document.
FPGA	Field-programmable Gate Array is a logic chip that can be programmed.
Faceted navigation	A system that allows a user to point and click on suggested terms or topics generated by the search engine.
Facets	A trendy term to describe the categories or headings in which documents, entities or other items have been placed.
Fat client solution	A solution designed to run on the user's own machine to increase performance.
Federated search	Executing a search across more than one data repository simultaneously.
Filter	A function used to specify criteria for selecting or rejecting data.
Graphical editors	An editor interface that allows the display of data in graphical objects and schemas.
Guided navigation	Guides the user to relevant information by keeping the information in context usually using point and click links to drill down further into the data; a synonym is assisted navigation.
Hash function	An algorithm that converts data of any length into short and fixed-length string of characters.
Heuristics	A methodology where the wisest solution is selected using rules. Intelligent systems use heuristics.
Hierarchy	A series of groups inside a system that is organized and ranked.

Term	Definition
High-speed persistent cache	A method for retaining information in a high-speed storage area to increase system performance. A complement to an in memory index. The purpose is to minimize physical disc accesses.
Hit	Jargon for an individual item in a list of results. Sometimes used for the number of occurrences of a search term in a document or a collection of documents
Hosted search	Outsourcing the search function; the index, search system, and technical management are handled by the vendor.
Hybrid display	Combines text, hot links, and graphics on one screen.
Hybrid interface	Synonym for a dashboard interface.
ISO	International Organization for Standardization, a worldwide federation of national standards entities and developer of technical standards. Learn more at http://www.iso.ch/.
Index	A table containing terms and pointers to the documents in which those terms appear. If an index includes tags for categories, pointers allow the text mining system to present items grouped within each category. See also facet.
Index rebuild	Reindexing content after a search system crash or upgrade failure.
Index update	The process of adding new entries to an index.
Information gain	A method to use result lists data to narrow the set of potentially relevant results.
Information retrieval	The science and technology of finding information from a collection of documents.
Infrastructure	The basic, underlying framework or features of hardware or software.
Intelligent search agent	A search system using algorithms to make decisions without human intervention.
Interface	The way in which the user interacts with a system or equipment or programs designed to communicate information from one system of computing devices or programs to another. Now the buzzword user experience or UX refers to a system's interface.
Internal search system	A program that indexes or searches information that is not publicly available on a company's network (behind protective measures.)

Term	Definition
Inverse document frequency	Allocation of a weighting to a search term by number of times a word appears in the document versus the frequency in the corpus. Often referred to as tf-idf for term frequency-inverse document frequency.
Inverted index	An index of text items maps the location of an occurrence to a position in a database and enables queries to be managed without recourse to a sequential string search of each document.
Key word systems	A search system that matches the words in the search box against the words in the index. A key word system may support Boolean logic's AND, OR, and NOT operators.
Key word search	The use of words, usually with Boolean operators (AND, OR, NOT) as a means of constructing a query.
Language identification	A linguistic service that recognizes the language and character set in which the document or text is composed.
Latent semantic indexing (LSI)	An algebraic model of document retrieval based on mathematical techniques that represent a document as a series of values.
Lemmatization	A term used to refer to the process of dropping prefixes and suffixes to obtain the root of a word. A synonym for stemming.
Library of Congress subject headings (LCSH)	A list of standardized subject headings used to index materials by the Library of Congress. The subject headings are arranged in alphabetical order by broad terms, with more precise headings listed underneath.
Linguistic systems	A search system that analyzes language as part of the indexing process.
Linguistic text processing	The functions such as identification of parts of speech used to process the language in which a document is written.
Link analysis	A procedure to calculate link popularity. A text mining system can examine Web pages to identify the network of interactions. Inbound links can be used to determine a Web page's importance and determine search ranking.
Managed search	A third party hosts and operates a search and retrieval system for a licensee.
Managed solution	A synonym for managed search; a version of outsourcing.

Term	Definition
Manual classification	A subject matter expert reads a document and selects and assigns terms by from a taxonomy. These terms are used to index or tag the document.
Mashups	A term used to describe merging two or more sets of data in a single graphical representation such as a map with restaurant telephone numbers displayed.
Metadata or meta-information	Results data that contains information about other sets of data. Metadata can be readily present in the document, such as the title, subject, author and size of a file, or it can be derived, such as its language, genre and usage statistics. Information about data assigned to a document and its components; for example, sorting and displaying documents by date or importance. A synonym for index terms.
Minimum description length principle	A model that describes the best amount of information adequately from the least amount of information encoded and compressed data.
Mining function settings	Determines the type of model to build, its function, and the selected algorithm. See mining function for what it supports. See weighting.
Mining model	The final model built from mining function settings and how it is utilized by the user's specific algorithm.
Mining result (output)	The end product(s) of a mining function.
Missing value	A missing data value because it was not measured, answered, lost or is unknown. The solution: ignore, omit, replace any missing values, or guess from existing values.
Mixture model	Several component functions that are combined to provide a multimodal density. Most of the functions are Gaussian, a mathematical theory of probabilities.
Model	It can be descriptive or predictive and essential in data mining. Underlying processes or behavior come from a descriptive model. Ex.: An association model describes consumer behavior. See also mining model.
Module	The central component of XeLDA (XErox Linguistic Development Architecture), a linguistic engine or service.
Morphological analyzer	A linguistic resource that provides the normalized form and every aspect of speech categories for all tokens identified during tokenization, which is the splitting of a string into token-sets.

Term	Definition
Multi-record case	Each occurrence of a record is stored in a table with columns. Also known as transactional format. See also single-record case.
Natural language processing (NLP)	Use of the rules of native languages to examine the content and meaning of text. Artificial intelligence and a trained rule base for meanings of words are used often. This approach has yet to prove effectiveness as a search technology. Efforts are being made to incorporate this technology into other approaches such as neural network search engines to improve overall performance.
Network feature	A tree-like multi-property structure with component features that are conditionally independent in the network. These features contain at least one attribute (the root attribute) and network features are used in the Adaptive Bayes Network algorithm.
Neural Networks (NN)	A new approach that modifies models from multiple data sources.
Nomenclature	A system of identifying and categorizing things.
Non-Negative Matrix Factorization (NMF)	A tool that uses an extraction algorithm to decompose statistical analysis of multivariate data into two matrices.
Nontransactional format	In a table, every case in the data is stored as one record (row). Also known as single-record case. For multiple records, see transactional format.
Noun phrase extraction	A linguistic resource that identifies words acting uniformly as a noun.
Numerical attribute	A numerical attribute whose value can be an integer or a real number. The values can be skillfully controlled as continuous values. See also categorical attribute.
Online Analytical Processing (OLAP)	An application that summarizes data queries swiftly and is used in business and data mining, such as reporting, analyses, and modeling.
Ontology	A deeper understanding of knowledge in specific areas of systems and their relationship with objects and attributes. The study of the categories of items that exist or may exist in some domain with an emphasis on knowledge representations.
Open source software	Software distributed for free and supported by an informal group of developers.
Ordering	Logical sequence of separate elements.

Term	Definition
Original Equipment Manufacturer (OEM)	A search vendor who licenses software to a third party so that the third party can use the vendor's search system in the licensee's product.
Orthogonal partitioning (O-cluster)	A useful clustering algorithm that locates natural groupings embedded in data. This algorithm technique specifies the data through a hierarchical grid-based clustering model.
Outlier	A datum value that stands out from the central values of the data. They are measured against deviations from the median or the mean.
Parametric search	Searching using attributes defined over one or more knowledge sources. This is relatively straightforward when using structured knowledge. This search is also possible with unstructured knowledge, where intelligent miners might glean concepts represented with the knowledge artifacts. Also see Guided Search.
Part of Speech (POS) disambiguation	The tool that locates the correct grammatical category of a word according to its definition and context.
Pattern matching	Identifying naturally occurring patterns in text, based on the usage and frequency of words, terms, or even letter patterns that correspond to specific ideas or concepts. Usually utilizes probabilistic algorithms such as Bayesian inference or neural networks.
Physical data	Identifies the choices of data to be used as input to data mining through Java interface. This can be used to build a model, model scoring or statistical analysis.
Pigeonhole	A specific, often oversimplified category.
Plain text	Any unenciphered string (i.e. finite sequence of characters) that consists of human-readable characters.
Positive target value	Target values are designated in two classes, positive and negative in binary classification problems. A model then calculates the density of positive target values among a class of objects.
Precision	The percentage of relevant documents in the list of documents retrieved.
Predictive Model Markup Language (PMML)	This XML is a way to define a share models and is supported by SAS, IMB's Intelligent Miner, SPSS Clementine, and many other data and text mining systems. It supports the import and export of PMML models for Naive Bayes and Association Rules models.

Term	Definition
Predictor	A data type used as input to a supervised model or an algorithm used to create a model.
Prior probabilities	The proportionate distribution estimates of a set in a population. Also known as priors.
Production system	A system used to serve content or make functions available to users.
Query processing	Matching the search query against the index.
Ranking	The ordering of a collection of hits into a defined sequence - e.g. relevance, date.
Recall	The percentage of all relevant documents in the index that are presented in the search results.
Relational morphology	A linguistic application that collects a word based upon its derivational family.
Relevance ranking	Among search engines, it is the procedure that sorts the matches to create a set of top match results. An algorithm determines the relevance ranking based on multiple factors, such as location on the page, link analysis, and the proximity of varying words.
Repository	Computer storage of a collection of unstructured text documents.
Resellers	A person or vendor who obtains a product or service, customizes it, and then resells it to a customer.
Response time	The actual or perceived time required for a search system to return a list of results to a user after the query or other instruction has been sent t the search system.
Results	The responses from a search system. A response can be a list of results or a graphic representation of the responses.
Reverse relational morphology	An application that collects all words within a derivational family.
Rule	Instructions a text mining or other process follows when dealing with certain content under certain conditions. (For example, “LGH BV5 is a Canadian postal code.”)
SGML	Standard Generalized Markup Language, of which XML is a subset.

Term	Definition
Schema	Data specification within a database. An entity-relationship diagram.
Score	To generate predictions, one must apply a data mining model to new data. See apply output.
Scripting	Instructions that cause a system or subsystem to perform a specific sequence of actions.
Search analytics	The analysis of the queries used and the outcomes of a search.
Search appliances	A computing device that is pre-loaded with a search-and-retrieval system; the vernacular is search toaster.
Search box	The entry form on a Web page or other interface into which the user types a query in the form of a word, phrase, or other segment of text.
Search system administrator	An individual who is responsible for a search system within an organization.
Search system training	A process to teach a search system licensee how to manage and operate the search system.
Semantic analysis	A technique in information retrieval that relates syntactic structures.
Semantic search	A method of searching for information using concepts that may not be expressed in the text of a document.
Semantics	The study of meaning.
Sentiment analysis	Use of algoritims, dictionary, and other methods to determine if content is trending positive or negative about an entity.
Social search	A method for processing content produced by Facebook, Twitter, and similar services in order to identify entities, trends, or other aspects of information produced by inidividuals using Facebook-like systems.
Software as a Service (SaaS)	A vendor allows licensees or individuals to use software via the Internet without having the application installed on an on-premises computer.
Soundex	An algorithm for encoding a word so that similar sounding words encode in the same way.
Staging system	A testing machine or system used to debug and test software before that software is moved to the production system.

Term	Definition
Statistical search system	A search system that makes use of mathematical routines for processing text. A Bayesian system is a statistical system.
Stemming	The terms in the search index are represented by stems rather than by the original words. This reduces storage and can improve search quality. See Lemmatization.
Stop words	A common word (such as the, of, on, and a) are not indexed; when used in a search query, they are ignored. Stop words are not standardized in all search engines so the same query may vary and yield different results.
Stop list	A set of words, abbreviations and other text elements that are not indexed. Different languages require different stop lists. (In French “the” is the same letter sequence as the drink “tea” in English.)
Structured Text	Information that can be found in fixed fields within a record or file. Relational databases and spreadsheets are examples of structured data.
Structured data	Information that can be stored in a table where e.g. a document is a row and the column heading is a field name.
Structured Query Language (SQL)	An industry-standard programming language for creating, updating and querying relational database management systems.
Tag	A process of marking text as to its function. In HTML, tags allow content-authors to mark up or format a Web document indicating each element.
Taxonomy	Hierarchical in structure, the classification of things or the principles underlying the classification. Almost anything - animate objects, inanimate objects, places, and events - may be classified according to some taxonomic scheme. Taxonomy's first meaning as a strict hierarchy organized as a generalization/specialization relationship among concepts ('is-a' hierarchy) has evolved to a more generic meaning of a scheme for categorization that facilitates browsing of a rich space of content. Web taxonomies often contain cross-links and place a given object in more than one category.
Term Density	Also known as keyword density, the percentage or ratio of a particular search query's terms to all terms on a page.
Term Frequency (TF)	The search metric that describes the number of occurrences of a particular searcher's query term in a Web page. Search engines use the term frequency as a ranking factor in the relevance ranking algorithm for organic search.

Term	Definition
Term position	A measurement of how near at the start of a Web page element a word appears. Words at the start of the body element are more prominent than those later in same element. Position and placement incorporate keyword prominence.
Term proximity	The measurement of two words near one another within a matching page. The closer the words, the better the search ranking result. Also known as keyword proximity.
Term rarity (inverse term frequency)	A frequency measurement of a term used on Web pages. The search engines offer greater weight to pages that contain rare terms compared with common terms when ranking the results. In a multiple-word search query, some words might be very common and others comparatively rare. Also known as keyword rarity.
Term mapping	The process of instructing a search system that “IBM” is equivalent to “ IBM Corporation.”
Text analytics	Software application that employs linguistics and pattern detection methods to credit some greater meaning to the words. Two types of text analytics are entity extraction and document categorization.
Text mining	Extraction of information from different resources and tools to form new facts and analyses. A subset of data mining. Information analysis to discover, tag, and extract facts in a text; for example, the names of people, telephone numbers, and places.
Text parsing	The steps that identify elements of a document beyond the recognition of individual words.
Tokenization	A linguistic operation that divides a sequence or string of characters into words or tokens.
Tokenized	The process of breaking text into its elements; for example, a text can be broken into sentences or paragraphs. Representing tokens as mathematical entities allows them to be manipulated by other processes.
Training data	A collection of content that represents in a statistically-valid way the information the text mining or serach system will eventually process.
UX	A synonym for interface (“user experience”).
Unstructured data	Information contained in e.g. a Microsoft Word file, the message payload in an e-mail message, or the ASCII text generated by an optical character recognition program. No tags or document structure tags like those used in XML mark up are included.

Term	Definition
Visualization	Graphic representation of data so their links are easily identifiable and semantic.
Web site search	A search system that returns results from a single Web site or a group of Web sites that are reached via the Internet.
Weighting (see also scoring)	Assigned a higher value to an item than it would otherwise have had.
XML	Extensible Markup Language. A software program that allows individuals to customize their tags in Web documents using a common recognizable format.

About the Author

Stephen E Arnold has worked in the technology sector since 1971 when he dropped out of the University of Illinois Ph.D. program to take a job at Nuclear Utility Services, a unit of Halliburton Industries. After a stint at Booz, Allen & Hamilton, he joined the Courier Journal & Louisville Times Co. to work in the firm's commercial database division. After the sale of the Courier Journal to Gannett in 1986, Mr. Arnold worked at Bell+Howell. Coincident with his receiving an award from ASIS for his contributions to online information, he joined Ziff Communications. When that firm was sold, he established an independent information practice. Since the early 1990s, he has worked for a wide range of organizations, including the White House and numerous large commercial organizations as a consultant. In 1993, he was one of the founders of The Point (Top 5% of the Internet). Since that push into the Internet, he has worked on a number of start ups. In 2000, he was one of the team working to create the first index of the content available from US government agencies. In the course of his professional career, he was the author of the first three editions of *The Enterprise Search Report* between 2004 and 2007. He has authored three monographs about Google's search technology as well as more than 50 journal articles. He contributes columns and essays to *Enterprise Technology Management*, a UK based publications for chief information officers, as well as columns to *Information Today*, *KMWorld*, *Online Magazine*, *Searcher*, and the *Smart Business Network*. His informal writings appear in his Web log, *Beyond Search*. He also publishes *Inteltrax*, a news service focused on data fusion and analytics. For more information about Mr. Arnold, run a Google search for ArnoldIT or navigate to www.arnoldit.com. Additonal awards Mr. Arnold has received, an archive of his lectures, and other search-related information are available without charge.

Vendor List

In my files are the names of more than 250 vendors of enterprise search and content processing systems. In the original Enterprise Search Report I wrote between 2004 and 2007, I included about two dozen vendors. The resulting report was unwieldy. Vendors share so many similarities and use a bewildering array of jargon in an effort to position themselves and make their systems memorable that I abandoned the project.

I recognize that in today's financial climate vendors are aggressively pursuing organizations with search systems positioned as customer support, sentiment analysis, business intelligence, and eDiscovery "solutions". Some vendors offer products without charge, billing the client for services. Others offer appliances or turn-key systems.

In the list of 32 vendors, I have tried to provide a sampling of the types of systems available as of May 2011. Vendors do come and go. In the last few years, companies such as Convera, Delphes, and Entopia have shuttered their offices. Others have morphs or "changed their spots" in an attempt to find a market for their search and content processing technology. Business intelligence and customer support are two examples of moving from a crowded and often low-margin niche to a less-crowded niches where margins are perceived to be higher.

In an enterprise search procurement, many organizations make a safe choice; for example, using IBM's Lucene-based system which is now packaged with content analytics or using the text search system available to Oracle database licensees. Microsoft SharePoint has become a "safe" choice, not because of a technology advantage, but because Microsoft is the computing standard at an organization. But even in this "safe" decisions, third-party technology is often required in order to obtain specific functionality within the licensee's time and resource budget.

Boldface indicates vendors discussed in this report.

Search Vendors: A Selected List

If you require additoinal information about the vendors in this table, contact the author. For current news about a vendor, navigate to www.arnoldit.com/overflight.

Vendor	Pricing	Positioning	Compares To	Type	url
Attivio	Mid-range	Business intelligence	BA Insight, SAS Teragram	Microsoft-centric	www.attivio.com
Attensity	High-end	Social media	Lexalytics	Proprietary	www.attensity.com
Autonomy	High-end	Platform	Endeca, Exalead	Proprietary	www.autonomy.com
BA Insight	Mid-range	Business intelligence	Attivio	Microsoft-centric	www.bainsight.com
Brainware	Mid-range	Document processing platform	ZyLAB	Proprietary	www.brainware.com
Concept Searching	Mid-range	Tagging utility	SurfRay	Microsoft-centric	www.conceptsearching.com
Connotate	Mid-range	Business intelligence	Attivio, SAS Teragram	Proprietary	www.connotate.com
Coveo	Mid-range	Customer support	ISYS Search Software, Mindbreeze	Proprietary	www.coveo.com
dtSearch	Lower-range	SharePoint search	Coveo, ISYS Search Software	Proprietary	www.dtsearch.com
EMC Kazeon	High-end	Archive search	Iron Mountain	Proprietary	www.emc.com
Endeca	High-end	Search, e-commerce, business intelligence	Autonomy, Exalead	Proprietary	www.endeca.com
Expert System	Mid-range	Mobile search	Coveo, Sophia	Proprietary	www.expertsystem.net
Exalead (Dassault)	Mid-range	Platform	Autonomy, Endeca, Microsoft Fast	Proprietary	www.exalead.com
FLAX	Lower-range	Enterprise search	IBM OmniFind, Lucid Imagination	Open source	www.flax.co.uk
Google Search Appliance	High-end	Platform	Thunderstone, dtSearch	Proprietary	www.google.com/gsa
IBM OmniFind	High-end	Enterprise search	FLAX, Lucid Imagination	Open source	http://www-01.ibm.com/software/data/enterprise-search/omnifind-enterprise/
Intelligenx	Lower-range	Database search	Perfect Search	Proprietary	www.intelligenx.com
Iron Mountain Stratify, Mimosa (dtSearch)	High-end	Archive search	EMC Kazeon	Microsoft centric	www.ironmountain.com
ISYS Search Software	Mid-range	Enterprise search	Coveo, Mindbreeze, Sophia	Microsoft centric	www.isys-search.com
Lexalytics	Mid-range	Social media	Attensity, SAS Teragram	Microsoft centric	www.lexalytics.com
Lucid Imagination	Lower-range	Enterprise search	IBM OmniFind, FLAX	Open source	www.lucidimagination.com
Microsoft Fast	High-end	Platform	Autonomy, Exalead	Microsoft-centric	http://sharepoint.microsoft.com/
Mindbreeze	Mid-range	SharePoint search	Coveo, ISYS Search Software	Microsoft centric	www.mindbreeze.com

Vendor	Pricing	Positioning	Compares To	Type	url
Oracle SES11g	High-end	Database search	Intelligenx, Perfect Search	Proprietary	http://www.oracle.com/technetwork/search/oses/overview/index.html
Perfect Search	Lower-range	Database search	Intelligenx, Oracle SES11g	Proprietary	www.perfectsearch.com
Recommind	Mid-range	E-discovery search	Autonomy, Iron Mountain Stratify	Proprietary	www.recommind.com
SAS Teragram	High-end	Tagging utility and business intelligence	Attivio, BA Insight	Proprietary	www.sas.com
Sophia	Mid-range	Enterprise search	Coveo, ISYS Search Software	Proprietary	www.sophiasearch.com
SurfRay	Mid-range	SharePoint search	Concept Searching, Mindbreeze	Microsoft-centric	www.surfray.com
Teratext	High-end	Enterprise search	Oracle SES11g	Proprietary	www.teratext.com
Thunderstone	Lower-range	Enterprise search	Google Search Appliance, Exalead	Proprietary	www.thunderstone.com
ZyLAB	Lower-range	Document processing	Brainware	Proprietary	www.zylab.com
dtSearch	Lower-range	SharePoint search	Coveo, ISYS Search Software	Proprietary	www.dtsearch.com