



Research Report

Beyond Search:

What to do when Your Enterprise Search System Doesn't Work

April 2, 2008

by Stephen E. Arnold



Gilbane Group Inc.

763 Massachusetts Avenue
Cambridge, MA 02139 USA

Tel: 617.497.9443

Fax: 617.497.5256

info@gilbane.com

<http://gilbane.com>

Table of Contents

List of Figures	vi
List of Tables	viii
Preface	x
Executive Summary.....	1
Fixing a Broken Search System.....	1
How Do You Avoid Common Mistakes?.....	1
Beyond Key Words	2
The Next-Generation Search Market.....	2
The Market Layout	4
Beyond Search at a Glance	5
I. Introduction: Setting the Stage.....	9
Search 2008	11
Does Your Organization Have Search Flu?	11
Maintaining Legacy Search Systems	14
Fixing a Search System	14
Avoiding Some Common Pitfalls	18
Getting to More than Key Words.....	32
The Symbiosis of an Interface and Search Technology	33
Identifiable Trends.....	39
Payoffs and Liabilities of Rich Text Processing.....	46
Market Context.....	51
The Superplatforms	52
Autonomy, Endeca, and Fast Search (The Big Three)	55
Up-and-Coming Vendors.....	63
More Choices, More Functionality	68
What's Next?	70
Google and Dataspaces	73
Semantic Technology at Google.....	73
Transformic: A Meta-Meta Approach to Content	75
Possible Impact of Google's Dataspace Technology	81
II. Tracking the Players	83
A Snapshot of the Beyond-Search Market.....	83
The Beyond Search Market Map.....	86
III. Vendor Profiles	91
1. Access Innovations.....	92
MAIstro Suite	93
Technology	95
Customers.....	97
Benefits.....	98
Downside	99

Net-Net	99
2. Attensity Corporation	100
The System in Action	102
Technology	103
The SDK - Attensity Integration	107
New Features.....	107
Upside.....	107
Downside	108
Net-Net.....	108
3. Bitext SA	110
The Bitext Data Suite	111
Technology	113
Upside.....	114
Downside.....	114
Net-Net.....	114
4. Brainware, Inc.	115
How Brainware Works.....	116
Associative Memory: The Brainware Innovation	117
Features	118
The System in Use.....	120
Upside.....	120
Downside	120
Net-Net.....	121
5. Cognition Technologies, Inc.	122
Examples of the System in Use.....	124
The Knowledgebases	125
Key Features.....	126
Upside.....	127
Downside	128
Net-Net.....	128
6. Connotate Technologies	129
A Content Processing Riff	129
Connotate's Agent Approach	131
Professional Services.....	133
The System in Action	133
Upside.....	134
Downside	135
Net-Net.....	135
7. Dieselpoint Inc.	137
Key Features.....	138
Open Pipeline	139
Technology	142
Dieselpoint in Use	143
Upside.....	143
Downside	144

Net-Net	144
8. Exalead	145
The Company	147
The Technology	147
Product Line Up	149
Customers.....	151
Upside.....	151
Downside	152
Net-Net	152
9. Exegy	153
Hardware+Software	153
The Company	153
Technology	156
Product Line Up	159
Upside.....	161
Downside	161
Net-Net	161
10. IBM Corporation	162
IBM Content Processing Products.....	164
IBM Partners/Developers	168
Upside.....	169
Downside	169
Net-Net	170
11. Information Builders Inc.	171
The System in Action	172
Technology	173
Search and Rich Text Processing.....	176
New Features.....	177
Upside.....	178
Downside	178
Net-Net	178
12. Intelligenx	180
The Technology	182
Intelligenx Features	184
Discovery Engine in Action.....	185
Upside.....	186
Downside	186
Net-Net	187
13. IntelliSearch Inc.	188
Technology	189
Upside.....	192
Downside	193
Net-Net	193
14. ISYS Search Software	194
The Company	194

ISYS Product Line Up	195
Rich Text Processing	195
Other ISYS Features	199
Technology	201
Examples of the System in Use	201
Upside	201
Downside	202
Net-Net	202
15. Lexalytics Inc.....	203
Technology	204
Examples of the System in Use	205
Key Features	206
New Features	207
Upside	208
Downside	208
Net-Net	208
16. Linguamatics Ltd.....	210
Examples of the System in Use	210
Key Features	211
Technology	212
New Features	214
Upside	214
Downside	214
Net-Net	214
17. Microsoft Corp.	216
SharePoint Server Search	217
Fast ESP	219
Tools, Not Toasters	222
The Company	225
Upside	225
Downside	225
Net-Net	226
18. PolySpot SAS.....	228
Hybrid Technology	229
Rich Text Processing	230
Big News: Collaboration	232
Examples of the System in Use	233
Upside	233
Downside	233
Net-Net	234
19. Recommind.....	235
Customers	236
Technology	236
Upside	239
Downside	240

Net-Net.....	240
20. SchemaLogic Inc.....	241
The Company	241
What SchemaLogic Does.....	243
Technology	244
Product Line Up	246
The System in Use	246
Upside.....	247
Downside.....	248
Net-Net.....	248
21. Siderean Software Inc.	249
How Seamark Navigator Works	251
Customers.....	253
Upside.....	253
Downside.....	254
Net-Net.....	255
22. Thetus Corporation.....	256
Semantic Search.....	257
Thetus in Use.....	258
Thetus Publisher	259
Fusion	261
Searching with Thetus.....	262
Upside.....	262
Downside.....	262
Net-Net.....	262
23. Vivisimo Corp.	263
A “New Breed” of Search System	264
Technology	265
Velocity Search Platform.....	267
Upside.....	269
Downside.....	270
Net-Net.....	270
24. ZyLAB.....	271
Technology	271
Examples of the System in Use	273
Basic Functions	273
New Features.....	276
Upside.....	276
Downside.....	276
Net-Net.....	276
Glossary	277

List of Figures

Figure 1: Google's Ubiquitous Search Box	32
Figure 2: Endeca's Guided Navigation.....	33
Figure 3: Siderean "Snapped Into" Oracle SES 11g	34
Figure 4: Autonomy IDOL Hot Links	36
Figure 5: Inxight Entity Extraction.....	37
Figure 6: Vivisimo Automatic Classification	38
Figure 7: Temis Luxid Interface Options	39
Figure 8: Hyperbolic Relationship Map	41
Figure 9: Conoco's Dashboard Display	46
Figure 10: The Squeeze on Autonomy, Endeca and Fast.....	62
Figure 11: Coveo and SharePoint	64
Figure 12: Exalead's Interface	66
Figure 13: The ISYS Customizable Interface	67
Figure 14: Siderean Graphic Interface	68
Figure 15: Interactions not Captured in Most Content Processing Systems.....	78
Figure 16: Semex: Mining for Personal Information Integration	79
Figure 17: Managing Dataspaces.....	81
Figure 18: Google Universal Search - Multiple Content Objects.....	82
Figure 19: "Beyond Search" Market Sector Functionality	87
Figure 20: "Beyond Search" Market Sector Vendors.....	88
Figure 21: Access Innovations' MAIstro Interface.....	93
Figure 22: MAIstro Rule Building Interface	94
Figure 23: Attensity product stack.....	101
Figure 24: Attensity unstructured data transformation process.....	103
Figure 25: Bitext NLP Adds Functionality to Microsoft's Live Search.....	111
Figure 26: The Bitext data flow is straightforward.	113
Figure 27: Brainware's Search Interface	116
Figure 28: Cognition's Search Interface	124
Figure 29: Cognition Highlights Multiple Search Concepts.....	125
Figure 30: Connotate's Agent Library.....	131
Figure 31: Connotate's Flow	132
Figure 32: HVM's Dieselpoint Interface	139
Figure 33: Dieselpoint's "Open Pipeline Architecture"	140
Figure 34: Exalead's Panels.....	146
Figure 35: Exalead Assisted Navigation.....	149
Figure 36: The Exegy System's Data Flow	156
Figure 37: The Exegy Appliance.....	158
Figure 38: The IBM OmniFind Architecture	163
Figure 39: The IBM UIMA Annotation Viewer	167

Figure 40: Information Builder's Platform.....	172
Figure 41: Information Builder's Search Result Screen	174
Figure 42: Intelligenx Discovery Engine.....	181
Figure 43: Paginas Amarillas' use of the Intelligenx Interface.....	183
Figure 44: The IntelliSearch Interface.....	189
Figure 45: The IntelliSearch Alerts Manager.....	192
Figure 46: ISYS' Default Interface	196
Figure 47: The ISYS API.....	199
Figure 48: ISYS Reports.....	200
Figure 49: Lexalytics Report	204
Figure 50: FAST Marketrac Report Using Lexalytics.....	206
Figure 51: Linguamatics Search Term Report	211
Figure 52: Linguamatics Express.....	213
Figure 53: SharePoint Search Architecture	217
Figure 54: FAST ESP Components	220
Figure 55: Extracting "Emotion" with FAST.....	222
Figure 56: The PolySpot Interface	229
Figure 57: PolySpot Search Term Highlighting.....	232
Figure 58: Recommind's Advanced Query Interface.....	239
Figure 59: SchemaLogic's Metadata Management.....	242
Figure 60: SchemaLogic's Architecture	243
Figure 61: Oracle's Use of Siderean.....	250
Figure 62: Seamark Navigator	251
Figure 63: Thetus Search and Data Display.....	257
Figure 64: The Thetus SDK.....	259
Figure 65: USA.gov's Use of Vivisimo.....	265
Figure 66: ZyLAB's Search Result Interface.....	272
Figure 67: ZyLAB's Administrative Tools	273

List of Tables

Table 1: If You Need Additional Support	15
Table 2: The Most Common Pitfalls.....	19
Table 3: The Most Significant Differences	57
Table 4. Quick Look at Access Innovations	92
Table 5: Technical Highlights for MAIstro	98
Table 6: Quick Look at Attensity Corporation	100
Table 7. Technical Highlights for Attensity	106
Table 8: Quick Look at Bitext SA	110
Table 9: Technical Highlights for Bitext	112
Table 10: Quick Look at Brainware, Inc.....	115
Table 11: Technical Highlights for Brainware Inc.....	118
Table 12: Quick Look at Cognition Technologies, Inc.	122
Table 13: Technical Highlights for CognitionSearch	127
Table 14: Quick Look at Connotate Technologies.....	129
Table 15: Technical Highlights for Agent Community Gen2	134
Table 16: Quick Look at Dieselpoint	137
Table 17: Technical Highlights for Dieselpoint.....	138
Table 18: Quick Look at Exalead.....	145
Table 19: Technical Highlights for Exalead	151
Table 20: Quick Look at Exegy.....	153
Table 21: Technical Highlights for Exegy.....	155
Table 22: Quick Look at IBM	162
Table 23: Technical Highlights for IBM.....	166
Table 24: Quick Look at Information Builders.....	171
Table 25: Technical Highlights for Information Builders	176
Table 26: Quick Look at Intelligenx.....	180
Table 27: Technical Highlights for Intelligenx	185
Table 28: Quick Look at IntelliSearch Inc.	188
Table 29: Technical Highlights for IntelliSearch Enterprise Search Platform	190
Table 30: Quick Look at ISYS Search.....	194
Table 31: Technical Highlights for ISYS Search	198
Table 32: Quick Look at Lexalytics Inc.	203
Table 33: Technical Highlights for Saliency 3.2	208
Table 34: Quick Look at Linguamatics Ltd.	210
Table 35: Technical Highlights for Linguamatics.....	212
Table 36: Quick Look at Microsoft and Fast Search.....	216
Table 37: Technical Highlights of Microsoft SharePoint Search and Fast ESP	223
Table 38: Quick Look at PolySpot SAS	228

Table 39: Technical Highlights for PolySpot Enterprise Search	231
Table 40: Quick Look at Recommind	235
Table 41: Technical Highlights for MindServer 5.2	237
Table 42: Quick Look at SchemaLogic Inc.....	241
Table 43: Technical Highlights for SchemaLogic Inc.	247
Table 44: Quick Look at Siderean Software Inc.....	249
Table 45: Technical Highlights for Siderean Seamark Navigator	253
Table 46: Quick Look at Thetus Corporation.....	256
Table 47: Technical Highlights for Thetus.....	261
Table 48: Quick Look at Vivisimo Corp.	263
Table 49: Technical Highlights for Velocity.....	266
Table 50: Quick Look at ZyLAB	271
Table 51: Technical Highlights for ZyFind.....	275

Preface

Search is part of every professional's life. Little work can be done unless we can find the information, document, or data we need. When search systems return too many or irrelevant results, we have to fall back on manual methods. These are less useful today than they were a decade ago. The paper filing cabinets are either jumbled or chaotic. On deadline, we have to conduct a frantic investigation involving telephone calls, e-mails to co-workers, and flipping through paper and digital files looking for what we need.

I heard the phrase "beyond search" in a talk given by Sue Feldman, the search expert at IDC, in 2003. Ms. Feldman correctly identified a trend driven by companies using technology to overcome the straightjacket imposed by key word indexing.

Almost five years later, I seized upon the phrase as a way to provide some guidance to organizations confronting user dissatisfaction with search tools for finding content on an Intranet, what I call behind-the-firewall search. I added a subtitle to help narrow the field of focus for this monograph. The phrase "what to do when your enterprise search system won't work" allows me to provide some practical tips and profiles of companies whose technology may address some common search issues.

I have tried to provide guidelines for repairing, augmenting, or replacing an incumbent search system. In addition, I wanted to give the reader a sense of the alternatives to key word search now available. From a compiled a list of more than 200 organizations offering fixes, alternatives, or utilities designed to "fix" broken search systems, I have selected a wide range of commercial products, representing different technologies, countries of origins, and functions. My intent is to give the reader a brief overview of issues, options for resolving them, and a basic grounding in what's on offer in 2008.

This study is a reaction to the increasing frustration that my work in search and retrieval has made clear. In December 2007, I conducted a research study of America's largest scientific and technical organizations. Our interviews and Web survey revealed that 60 percent of the users of the organization's search system were dissatisfied with it. These data are interesting because when we conducted a similar study in 2006 for the third edition of the "Enterprise Search Report," dissatisfaction was in the 50 percent range.

Jane McConnell, Paris-based head of Net Strategy, approached me after my remarks at the *International Online Meeting* in London, England. She said, "Your data are identical to those we have gathered. I was startled to learn that dissatisfaction with Intranet search systems was the same in the U.S. as it is in Europe."

Vendors rarely provide information about how their customers' users perceive their system. One of the reasons is that most vendors license software and then move on to the next sale. Licensees work with integrators or partners. The developers of the software are not usually making post-sales visits unless the contract stipulates this continuing involvement. Another reason is that most licensees don't do user

satisfaction surveys. Search system administrators are overwhelmed with work. The procurement team is usually disbanded once the vendor decision has been made. In most organizations, search becomes an orphan. With users grouching about poor performance or erratic relevance, few volunteers venture forth to address these issues. Senior management has other, presumably higher priorities.

My work reveals that most Fortune 1000 organizations are involved in fixing, procuring, or changing their search systems throughout the year, year in and year out. Search is a white noise problem; that is, it is ever present and difficult to separate from other problems. Enterprise software almost always includes a search function. Employees quickly learn that there is not one search system. There are typically five or more. The Content Management System may have a search system embedded in the document creation and management tools. The CMS vendor, however, usually licenses search-and-retrieval technology from a third-party vendor. Autonomy Ltd. and Fast Search & Transfer are two prominent licensees of their technology as Original Equipment Manufacturer (OEM) deals. Database systems come with search systems as well. Some vendors like Oracle Corp. offer optional search technology. In the case of Oracle and IBM Corp., these companies bundle a command-line tool, offer home-grown technology, technology purchased from another firm, or search technology from a partner. Imagine the surprise of a user who discovers a version of Verity technology inside a CMS system, a database search system running on Oracle's Secure Enterprise Search technology, and a Web search system provided by IBM using Endeca technology. No wonder users are annoyed. Different systems and different implementations of search get in the employee's way. In these cases, search is more than annoying; search is a problem.

The need for a different approach to Intranet search is evident at conferences. The attention is no longer exclusively focused on finding information on the Internet or World Wide Web. Conference organizers are offering sessions about public search systems like Google's, Microsoft's, and Yahoo!'s. In 2006 and 2007, conference programs were peppered with presentations on semantics, search appliances, taxonomies, dashboards, work flow alerts, automatic classification of content, and assisted navigation. Each of these buzzwords is explained in the Glossary that accompanies this study. The point is that conference organizers are responding to a growing demand to move beyond search.

This study addresses some of the main features of this beyond search trend. I am not an academic, and I am attempting to put into a business context very complex technologies and processes. Mistakes are inevitable, and I am responsible for any errors in this work. My goal is to provide information to a professional involved in procuring a search system or search system enhancement, a system administrator working to improve an existing search system, an entrepreneur looking for an overview of what's on offer from companies around the world, and investors determined to find a way to capitalize on the opportunities in this market sector.

Unlike the first three editions of the *Enterprise Search Report* which grew by the third edition (CMSWatch.com, 2006) into a search encyclopedia of more than 650 pages, I

don't go into great technical detail, provide a return-on-investment model, or spell out the pitfalls of a hasty search procurement. Please, buy a copy of the present edition of *ESR* if you want this information. Similarly, I don't rehash the technical information about Google that appears in *The Google Legacy* (Infonortics Ltd., 2005) or *Google Version 2.0* (Infonortics Ltd., 2007). Instead I focus on one of Google's newer "beyond search" initiatives, referencing my earlier technical analyses of Google, not repeating that information. Readers of this study will find new information, so my previous studies' information has not been duplicated. In some cases, you will find that my more recent work has altered my assessment of certain systems; for example, the "upstarts" now are worthy alternatives to better-known vendors. In the case of Google, the information is based on research I conducted for a large software firm in October, November, and December 2007. To my knowledge, the enterprise implications of the Google dataspace technology has not appeared in an overview study before. I have included a glossary, and as in my other monographs, I have tried to present definitions for a professional who is not an academic or content processing expert.

Beyond Search has become a longer study than originally intended. I've tried to answer the questions I've been asked in the last few months. One major change is that the profiles of selected vendors are brief. I have intentionally selected vendors that represent some of the more innovative approaches to search, content processing, and metatagging. You will find that some of the vendors are almost unknown in the United States. Innovations in search are no longer "made in America". In fact, IBM's semantic e-mail technology comes from engineers and scientists whose roots are more global than Silicon Valley. You won't find in-depth discussions of Autonomy, Endeca, and other high-profile systems. That information is now readily available in the fourth edition of the *Enterprise Search Report* and elsewhere.

Stephen E. Arnold
ArnoldIT.com
Harrod's Creek, Kentucky
April 2, 2008

Executive Summary

The central premise of this study is that key word search and retrieval usefulness is slipping. In behind-the-firewall scenarios—what some people call Intranet search or enterprise search—key word search is the principal way an employee finds information in digital form.

For some behind-the-wall queries, key word search is useful. When the user looks for a unique name or term, key word search can match the query to the document containing the word.

As the volume of information increases, the search box in which the user types words and queries has become a test of the user's skill in figuring out exactly what words and phrases are needed to unlock the combination to the system holding their information. The cost of ineffective information retrieval is difficult to calculate. Without effective search-and-retrieval many business processes don't work. Search, therefore, must work.

Fixing a Broken Search System

How does an organization in today's challenging financial environment fix a search system that doesn't work?

The choices are stark. You can patch the existing system. Search and content processing systems today are sufficiently complex that slapping a bandage on an ailing system may not work. And once a system begins to degrade, a quick fix will not often be an enduring one. In mission-critical failures, you have to get up and running quickly with full knowledge that another problem lies in the future. Your solution, therefore, is to budget to react and move from problem to problem in order to deliver "good enough" search.

You may also reinstall the search system, apply patches and upgrades, and re-index the source material. In some cases, a re installation will remediate the problem. Many organizations find that the nettlesome problems can be resolved by starting with a clean slate. There are some penalties associated with starting over, and you will have to weigh downtime and cost to determine if this path is right for you.

The increasingly attractive solution is to "rip and replace". You look at some of the options available and replace your existing system with a different one. Many options exist, and you will find that some of the systems described in this study may deliver a better search experience at a lower cost than your existing system. However, time and expense play a large role in a "rip and replace" solution.

How Do You Avoid Common Mistakes?

You will need to go back to business basics and make certain you know what your requirements are. You need to have a plan. And you need to have a budget and a way to track your costs.

So armed, you will be equipped to deal with the five most common pitfalls my research identified; namely, inadequate infrastructure, lack of work processes and policies for vendor system updates, overconfidence in your knowledge of search, poor coding, and inadequate planning. There are many tips and tricks for reducing the risk from management and technical error. For example, you can make certain you have budgeted for skilled people to interact with the system's tagging in order to keep egregious mistakes from taking place. Your organization has its own jargon, and it's important to make certain that those terms are mapped to documents pertinent to that concept. This is an editorial resource issue, and it is as important as performing real-life testing before deploying a system.

Beyond Key Words

To move beyond indexing the words in a document, a system must have some way to figure out some of the implicit nuances in the document and identifying explicit information about the document such as its date of creation, the file type, and other useful facts.

Most users identify a “beyond search” system by its interface. There is a search box and other information as well. Most of the systems profiled in this study can generate an interface with suggestions for related content. These systems show a list of concepts and ideas in order to alert the user to what's available. Maps or other graphic devices provide a visual cue to the information.

But the interface is a signal that the system goes beyond key word queries. The “plumbing” makes the interface possible. A range of algorithmic techniques, linguistic methods, and statistical procedures can assign a document to a category, extract the names of people, places, and things in a document, and classify documents and individual paragraphs by their subject.

Many modern systems use the power of today's computers to apply linguistic techniques to a document. In effect, the software “reads” a documents and makes an effort to “understand” its meaning. Some systems can produce abstracts of longer documents so a user can get the “gist” of a document without having to scan the entire piece.

The benefit of these next-generation systems is that users have ways to find information in different ways. The tyranny of the search box is broken. Users report that finding information is easier and more enjoyable. If the information is not in the system, the user learns that quickly so alternatives can be pursued right away—saving even more time. But these “beyond search” systems are complex, and like their “dinosaur” predecessors can be difficult to manage.

The Next-Generation Search Market

Search has become one of the “next big things”. The market is in turmoil. New product announcements flow daily from the more than 150 vendors active in the behind-the-firewall search sector. Acquisitions create new opportunities for companies like

Microsoft (buyer of Fast Search & Transfer SA) and SAS Institute (buyer of Teragram). For the customers of Fast Search and Teragram, there's uncertainty.

The vendors themselves are like chameleons. Their marketing lingo can change overnight. The use of jargon makes it difficult to figure out exactly what a vendor's system does. It's even more challenging to determine if the system works in a real-world situation like the one at your company.

Today you have a choice of buying a behind-the-firewall search system from giant companies like IBM, Microsoft, and Oracle. I call these firms super platforms because these vendors provide many mission-critical enterprise systems. Search becomes an "add on", often available at a very attractive price. Search may be bundled with an accounting system, for example. Search, in effect, is free because it is included with the larger license.

You can license a system of a mid-tier of high-profile search vendors. In this category are Autonomy, Endeca, and Fast Search & Transfer. These companies have created a good business for themselves in behind-the-firewall search, but now find themselves under pressure from newcomers and the focal point of acquisition efforts.

There are a number of up-and-coming vendors. These companies offer systems that compare favorably to those available from the super platforms and the blue-chip vendors like Autonomy. Each of these companies is enjoying fast growth and have a number of satisfied customers. You will want to pay attention to these companies—Coveo, Exalead, ISYS Search Software, and Siderean. One or more of these companies will "move up" to blue-chip status, filling the position of Fast Search & Transfer which is now part of the Microsoft super platform offerings.

You can also "go open source." Lucene is a viable option for some organizations. You can also license a search appliance from Google or another vendor. Some vendors, not discussed in this study, offer a "hosted" option and handle your search system from a remote data center.

Not surprisingly, there are some significant forces at work in the search "space". These include commoditization of search, helped in part by the no-cost Lucene open source system. The market, in general, is moving away from key words into more sophisticated business processes, including business intelligence. Marketing, not technology, is more important than ever. Search vendors are becoming more skilled in the ways of Madison Avenue, so "buyer beware" becomes an important catchphrase.

Google has its search appliance or GSA. But the company has other search technologies as well. One of the most interesting is Google's dataspace technology. A combination of search and advanced content processing, dataspace could leap-frog the solutions discussed in this study. Google, true to its desire to keep a low profile with regards to its technology, won't comment about dataspace. It's important to recognize that search and retrieval is a "problem". A quantum leap forward by Google or some other company could reshape the entire market without much warning.

The Market Layout

To help you keep track of the players, this study contains a market map and a market scorecard. You will be able to see which vendors are in the search tools business. You will learn which vendors provide building blocks upon which you can set up a customized system. There are vendors who provide deep analysis using technology developed for the intelligence community. Other vendors provide a data management or database approach to make manipulation of information useful and informative. Other vendors use advanced mathematics to discern patterns in data and information. Others provide an enhanced search function that blends key word search with specific advanced content processing operations such as on-the-fly classification.

This study contains profiles of 24 systems from several different countries. Dozens more asked me to include them in this first edition, and I had to make some decisions about whom to include and exclude. The 24 profiles provide a representative view of the different methods, approaches, strengths and weaknesses of the systems, and a useful orientation to the dynamic field of behind-the-firewall search. The study provides six to nine page discussions of these companies' systems:

Companies Profiled in Beyond Search

Access Innovations	IntelliSearch Inc
Attensity	ISYS Search Software
Bitext SA	Lexalytics Inc.
Brainware, Inc.	Linguamatics, Ltd.
Cognition Technologies, Inc.	Microsoft Corp.
Connotate Technologies	PolySpot SAS
Dieselpoint, Inc.	Recommind
Exalead	SchemaLogic Inc.
Exegy	Siderean Software Inc.
IBM Corp.	Thetus Corp.
Information Builders Inc.	Vivisimo Corp.
Intelligenx	ZyLab

Search terminology is fast-moving. I have prepared a glossary and made an effort not to lapse into arcane and specialized jargon.

In closing, search is a complex and increasingly important function in an organization. Approach it with a commitment to excellence. The effectiveness of your organization's work processes and decision making are at stake.

Beyond Search at a Glance

This chart provides a quick reference to the companies profiled in *Beyond Search*.

Vendor	Technical Approach	Upside	Downside	Comment
Access Innovations	Systematic approach to controlled term and taxonomy management	One of a very few systems able to generate ANSI-standard vocabularies and taxonomies	System requires a knowledge of and commitment to vocabulary and taxonomy construction	Strongly recommended for controller vocabulary and taxonomy building and maintenance
Attensity	Rich content processing for metagging and content discovery and analysis	A sophisticated, hybrid system that squeezes content for meaning via iterative processing	A hybrid system with strong intelligence and discovery functions makes this more of a special purpose system	System has a strong reputation from parts of the U.S. intelligence community
Bitext	Works at the lexical level to perform and support automatic synonym and query expressions using NLP queries	Built in knowledgebase lexicon and semantic mappings	Company has a very low profile outside of Spain and the European Community	Provides ability to add NLP functions to an existing search system.
Brainware	Statistical engine with knowledgebase support for search and content discovery	Allows identification of relevant documents even when the content processed is unknown to the user	Technology is different from that of most other vendors. Head-to-head testing useful to grasp the system	An effective discovery and analysis tool
Cognition Technologies	Rich content processing using a proprietary linguistic knowledgebase	Allows an organization with a need for intelligence-agency grade text processing to implement an advanced system	Does not support non-text content	Company works hard to implement the controls associated with information science precepts

Vendor	Technical Approach	Upside	Downside	Comment
Connotate	Rich metatagging for selective dissemination of information	Unique “search without search” approach; easy-to-use agent technology	The “report” or “pushed information” approach is well suited for competitive intelligence and monitoring	Licensees will find the system useful for monitoring and competitive intelligence
Dieselpoint	XML database with advanced metatagging operations	Flexible, high-performance content processing and data management system	Company has a low profile in search, content processing, and XML data management	A versatile system suitable for ecommerce and content processing
Exalead	Behind-the-firewall search with advanced content processing features	Advanced features and customization combine with high-performance search and retrieval	Company is European centric and stepping up its marketing in the U.S.	Strong engineering makes this a contender for behind-the-firewall search and content processing
Exegy	High-speed appliance solution for information discovery and analysis	Can process terabytes of text, identifying items of interest for analysis	An appliance solution suited for monitoring; an academic spin out	Good choice for high volume intelligence and financial monitoring
IBM	A tool kit approach that permits construction of a specialized content processing system	IBM will be able to make any system work and scale	IBM is a mindset with its own procedures and methodologies; can be expensive	No matter what feature or function you want to implement, you will be able to deliver that to your users
Information Builders	Rich content processing of structured and unstructured data	Flexible architecture, which handles BI, search, and discovery requirements	Requires commitment to an entire framework	This approach points to a future in which search is a component of an enterprise system
ISYS Search Software	Behind-the-firewall search with entity extraction, classification, and document preview from the results list	High-speed, feature-rich next-generation search system	Company lacks the profile of more aggressive marketers of search and content processing	Strong product with a commitment to customer service

Vendor	Technical Approach	Upside	Downside	Comment
Intelligenx	A scalable database solution that supports text analysis, search, and metatagging	Ability to handle large volumes of content at high loads	Low profile for this company	Very high performance with strong analytic capabilities
Intellisearch	A semantic system that can support key word search supplemented with metatagging	A modest service footprint that holds down hardware costs	Company is a newcomer to the U.S. market	Strong following in Europe and building in North America
Lexalytics	Ability to measure sentiment or tone at the document, summary, and entity levels	Combines text mining and using a query passed against another index to resolve ambiguities	Self funding limits financial resources for marketing.	Consider if you want to provide users with reports instead of result lists.
Linguamatics	Advanced linguistic processing based on real-time NLP-based querying	Strong knowledgebase support, tabular results and rich configuration options	Quality of knowledgebase content has a significant impact on processing	Best suited to well-formed XML and structured content
MSFT Fast	Bundled search plus the Fast Search toolkit for building customized text solutions	Financial resources and market reach of Microsoft	Complications of an acquisition for end users including license fees and terms	A seismic shift in the behind-the-firewall search market; future impact unclear in Feb 2008
Polyspot	Key word search supplemented with metatagging for behind-the-firewall applications	Many features presented in a pleasing UI with SDK and API available for extending the system	Low profile in North America, technical support comes from Europe	Polyspot delivers on a number of rich text processing features.
Recommind	Bayesian "engine" refined for behind-the-firewall search and legal discovery	Engineering enhancements to speed system performance	Works best when licensee understands information discipline	Flexible tool for search and eDiscovery; useful in the enterprise as well

Vendor	Technical Approach	Upside	Downside	Comment
SchemaLogic	A content management system for metadata	Master metadata framework allows simpler access, integration and delivery of information	Long sales cycle, requires dedicated hardware and a careful configuration and deployment.	One of the founders left the company in early 2008
Siderean Software	Rich semantic indexing to support assisted navigation	Sophisticated content processing for search and assisted navigation; XML is a core competency	Focus on semantic technology	Combines the type of interface made popular by Endeca with robust content processing at an attractive price point
Thetus	Federated search of different content types	Access provided to disparate data types and their lineage	Core component is among most complex content processing engines available	One of a handful of military-grade content processing systems available
Vivisimo	On-the-fly classification plus comprehensive search and retrieval functionality	Rapid deployment; simple interface	Categories may confuse some users	Now an effective, full-enterprise search and content processing solution
ZyLab	Statistical and semantic engine with knowledge base support for search and content discovery	Can be extended for repository services and text mining; includes third-party visualization tools	Low profile; terabyte systems require a dedicated administrator	ZyLAB is solid, multi-capability vendor in content processing with customers throughout the world and double-digit growth

I. Introduction: Setting the Stage

Confusion about technology and the cost of search and content processing are familiar companions in many organizations today. Industry consolidation –Autonomy buying Verity and Microsoft acquiring Fast Search & Transfer – are two examples.

There is a tendency for procurement teams to blur the distinctions among three types of content processing and search systems. The ubiquity of free Internet search systems from Google, Exalead, Microsoft, and Yahoo! inform our belief that one- or two-word queries or a single click in a personalized Web page will deliver relevant results in milliseconds. These systems work brilliantly. Because someone else foots the bill, you and I have access to them for free. The temptation is to ask, “Why can’t our search system work like Google’s, Exalead’s, Microsoft’s, or Yahoo!’s?”

The second type of search is Web site search. Most people form an impression of an organization by its public-facing Web site. The Web site search system can be a junior version of the Google, Exalead, Microsoft, or Yahoo! “regular” Web search system. Google provides a free service called Google Custom Search Engine so anyone can use the Google Web search technology on a single Web site. The confusion about Web site search and other types of search comes about when an employee can “find” information on an Internet search engine but not on the company’s own, internal search system.

The third type of search engine is one that indexes information on the company’s network. The idea is that information needed by employees for work purposes should be searchable. Most organizations prevent unauthorized access to their internal information. These content processing and search systems, therefore, operate behind the organization’s firewall. The firewall protects sensitive information and creates an internal/external information distinction.

Behind-the-firewall search is difficult because of security, access restrictions, and government regulations about how certain information must be implemented and kept operational. Another twist is that many organizations want their employees to have access to both the internal information and information on the public Internet. Furthermore, behind-the-firewall systems ideally have to integrate with other enterprise software such as an accounts-payable or human resources system. A firm’s attorneys may need special types of functions for legal matters. The chemists want to search using graphic chemical structure diagrams. The public relations and marketing departments want to search product technical specifications and PowerPoints.

The behind-the-firewall search, therefore, is a complicated beast. Most people call behind-the-firewall search *enterprise search*. But that term is now somewhat devalued. I don’t think it conveys what is required for a system that processes content and makes it findable to the employees of an organization.

Behind-the-firewall search systems share surface similarities with public or ad-supported Internet search engines. However, behind-the-firewall systems are sufficiently different from these Internet search systems in another important way, the

people operating these systems must operate within specific budget, technical, and infrastructure constraints. Dissatisfaction with search is often directly related to the organization's budget for content processing. Trying to operate one of today's behind-the-firewall systems without adequate technical, financial, and management resources is source of many complaints about search. Even a search "toaster" appliance takes care and feeding. Cut corners and any system from any vendor will become a major problem very quickly.

If a text or content processing system makes it impossible for accounting to issue a payroll, you know which system gets cut back. Not surprisingly, in most organizations there is considerable misunderstanding about search in general and content processing specifics.

For an employee trying to close a deal, search is not just entering a phrase like *Roberts Plumbing* and getting a list of documents in which the word *Roberts* and the word *plumbing* appear. In a work setting, finding information is not an option. Search is the raw material of work. An employee needs information about a specific company like Roberts Plumbing in order to do work, usually under some sort of pressure.

Key word systems into which an employee types some words, and maybe an operator like *AND* are viewed by many users as slot machines that don't offer very good payoffs. The user must "guess" the magic word. If correct, the system provides the needed information. When rushed or on a deadline, most users avoid systems that don't provide the needed information.

When a search system generates a long list of results, some employees don't have time to open and conduct a manual inspection for the information. Search often doesn't help. It creates more work. That's one big reason there is so much interest in systems that show what's available through a system content map and systems that generate answers without the user having to type a complex query, semantic search. Sure signs of an under-resourced and poor search system are yellow sticky notes and piles of paper. The search system forces users to find a way to locate needed information when the search system cannot.

With the volume of digital information increasing sharply, organizations have to have systems that make finding the "right" information quickly. Search or some variant such as text mining are on the hot seat. With the costs and risks of not finding information needed to close a big deal or fight a legal claim, the demand for better search and findability is rising in step with the volume of electronic information flooding networks.

Search – specifically enterprise search or behind-the-firewall search – find it very easy to find potential customers. Any organization with a search system is likely to be looking to buy another one. In fact, most Fortune 1000 companies, if my research is correct, have a minimum of five search systems. How many does your organization have?

Overworked search system administrators find themselves on the firing line. The most-asked question I hear is "What can I do to fix our search system?" without expertise, money, infrastructure, and time. There are more options today than at any other time.

That's the good news. The bad news is that the number of options makes deciding on a specific solution more difficult.

Confusion and cost are the handmaidens of behind-the-firewall search.

Search 2008

We're now almost a decade into the 21st Century, and behind-the-firewall search is a familiar topic in trade journals, at conferences, and on Web logs.

Vendors of less expensive and "intelligent" systems have reported strong sales. The companies benefiting from the crisis in behind-the-firewall search include Coveo (Montréal, Canada), Exalead (Paris, France), ISYS Search Software (Sydney, Australia), and Siderean Software (El Segundo, California), among others. Some companies have embraced open-source search, electing to use search technology supported by volunteers who work for the community. Tesuji, a little-known company founded in Hungary, is a Lucene-based vendor. On the other end of the spectrum, the IBM OmniFind search system blends Lucene with its home-grown technologies.

To shake up the market, Microsoft first started giving away its search solution with its SharePoint server. A few weeks later Microsoft announced that it acquired Fast Search & Transfer, one of the high-profile search vendors. No one is certain how the Microsoft-Fast deal will affect licensees and competitors.

Search and content processing vendors must sort out confusing signals from the market. Mixed messages abound. Organizations want more features and more intelligent systems. At the same time, budgets are tight. Technical resources are constrained. Unlike 2007, 2008 is fraught with uncertainties about economics and technology.

There's a growing realization that when it comes to behind-the-firewall search, one size does not fit all. In 2008, the market for behind-the-firewall search is unsettled. Consolidation seems inevitable in today's market. Simultaneously, new search and content processing solutions enter the competitive fray at what seems like a quickening pace.

There is no single, perfect search solution. Every behind-the-firewall search "solution" is a compromise.

Does Your Organization Have Search Flu?

Here's a quick temperature check. These are hot spots in behind-the-firewall search identified in our 2007 research. How many of these challenges have been satisfactorily resolved in your organization?

Access Challenge

Users want more choices than a search box. Do your users want a system that "shows" what's available? One that suggests possibly useful documents? A system that displays a *dashboard*, essentially a point-and-click interface with certain information displayed

for a specific user? Instead of typing queries, users can mouse-click a topic or hot link to get information. Does your organization “push” information to its users?

The Laundry List Challenge

A list of results – a laundry list of documents that match the query – creates work. The user has to click on each item in the list and hunt through the document for the needed information. It's no surprise that users want systems to reduce work, not create more. Rank-and-file employees want their search system to alert them when an important event occurs and automatically display or “push” the information to their computer or mobile device. The notion of automating certain queries and the system sending the needed information is in sharp contrast to a system that makes the user do the work. Key word search is a “pull” approach, and many users don't want to key certain queries over and over again. Do you have to perform tedious, repetitive tasks when using your organization's search system by searching for the right content within your search results? If so, you have a laundry list challenge to address.

Many Sources, Many Servers Challenge

Many organizations want to give their employees access to both internal and Internet information. To make the situation even more difficult, the internal sources of information are scattered across different repositories in different locations, often in different file formats and languages. The term used to describe systems that can find, index, and make searchable information in such an environment is called a federated search system. The idea is that a user accesses one interface for information, not a different interface and system for each type of information. Content in different formats must be made consistent. Correct handling of different versions of a document is needed. Duplicate content must be winnowed automatically so the user doesn't have to perform this task. Does your organization need a federated search system?

Structured – Unstructured Challenge

You know that information resides in databases. Much of that information is available from interfaces specifically designed to process the data in rows and columns. Structured data is where payroll records, purchase orders, and factual information such as prices are kept. Unstructured data, on the other hand, refers to the information in standard documents, e-mail, PowerPoints, and PDF files. Does your organization have structured and unstructured information integrated and available in a search system?

Performance Challenge

Most behind-the-firewall search systems return results less quickly than a free, Web search system. Does your organization's search system return results in less than a second, less than 30 seconds, or longer? Sluggish systems can force users to create a manual work-around. Systems that users ignore increase an organization's information usage costs.

The Freshness Challenge

Yahoo! updates its news every few seconds. Each news story carries a message about when the information was updated. Behind-the-firewall search systems are often criticized for not having current information. Users don't understand why a particular document is "the old version." Does your organization have a system that makes available the most recent information, or do users have to talk to colleagues to locate the most recent version?

How did you do? If you found yourself agreeing with two or more of these statements, you may be a candidate for some search and content processing reengineering. Most of these challenges can be successfully resolved. You will need to select the appropriate vendor and have the resources available to fuel the engine of change once you establish your search priorities.

Maintaining Legacy Search Systems

Fixing a Search System

Let's assume a search storm strikes your organization's search system dead. Also, let's imagine that you are the person given the job of "fixing" the problem. In most organizations this means diving into the problem, identifying the problem, and getting back online as quickly as possible. Cost is usually not the main concern at a time of crisis. What can you do to get the search system up and working quickly?

Repair

The options for fixing a search system problem are limited. There are three. Let's look at each and then review some rules of thumb for determining which option may be appropriate for specific situations.

First, your search system vendor offers some technical and engineering support. If you have a service agreement, you need to contact the vendor with the specifics of the problem, as you understand them,. Your search system vendor may offer you some initial telephone troubleshooting, and then, if the problem is not resolved, will offer you some options. The options will vary due to the particulars of your license agreement.

One option is a telephone walk through, followed in some cases by remote diagnostics. If your search system is running on servers at a third-party location such as Rackspace or a similar operation, you will be working through layers of technical personnel.

Typical Trouble Spots

In many cases, rebooting the search system corrects the problem. Routine troubleshooting can look for these problems:

- Hardware failure in a server or storage unit
- Network outage due to a hardware or configuration change
- Network saturation with search-related tasks
- Insufficient storage which causes a search subsystem to terminate, typically in the content processing or query processing subsystems
- Changes to the thresholds for relevance or activating additional content processing functions via the search system's administrative interface
- Index won't update or has lost pointers to certain information
- Interruption of power or a cooling issue that caused a device to shut down unexpectedly

Once you have determined that the problem seems to be related to one of these causes, you can take remedial action yourself or contact your search system vendor. Vendors offer a wide range of technical and engineering support. Some provide their own engineers. Others have relationships with resellers who provide service. For example, Autonomy has expanded its consulting and support arm in the last 18 months. Endeca,

Fast Search & Transfer, and other high profile vendors have also ramped up their professionals services' activities.

Problem	Actions to Consider
Hardware failure	Replace the hardware and test. If storage devices fail, restore the indexes or other software components, and test. Rebuilding the indexes may be necessary.
Network outage	If the problem occurs when a network device is upgraded or replaced, configuration may be the issue.
Network overload	To reduce load on your network, you can [a] reduce the number of users and queries [b] reduce the aggressiveness of content acquisition [c] check process priorities to ensure that search functions are not consuming too much bandwidth [d] expand bandwidth
Storage issues	Insufficient storage or storage device malfunctions can cause problems of many types. Faulty storage devices should be replaced. Fault tolerant storage reduces some risks, but a poorly performing device may create bottlenecks that slow another process; for example, writing temporary files during index updates and requires replacement.
Index incorrect	Corrupt indexes may be due to either hardware or software errors. If restoring an index does not resolve the problem, you may have to rebuild the index, staging the rebuild to allow you to determine if the system is working correctly
Device failure	Some search systems are plagued with hardware failures. The cause is [a] low-cost devices that are not reliable; [b] inadequate cooling for the devices; [c] a power problem
Unexpected time outs	Intermittent problems are difficult to resolve. Look for [a] hot spots and bottlenecks in the log files; [b] errors in custom scripts, particularly when search is interacting with a separate third-party system; and [c] down-line issues caused by hardware failures elsewhere or large blocks of data residing on a sluggish subsystem
Vendor unresponsive	Vendors can be short-staffed or unable to resolve problems due to other hardware and software. You will need to work with your existing technical resources and advisors with system expertise

Table 1: If You Need Additional Support

Let's assume your search system vendor is unable to assist you. You can turn to specialists such as New Idea Engineering in Silicon Valley.¹ Your vendor may have authorized resellers or partners. If you are in a major city in the United States, you may be asked to contact a third party. The "gotcha" is that the authorized, certified partners and resellers may have to re-contact the search system vendor. These reseller-partner relationships vary.

If your troubleshooting identifies an interaction with another enterprise software system, you may be caught between vendors who politely suggest that the "other" party

¹ <http://www.ideaeng.com/>

has responsibility. This finger-pointing may be best mediated by a third party specialist with a working relationship with one or both vendors.

Because of the complexity of modern networks, it is possible that you will have to involve other firms' engineers or turn the matter over to a specialist in your area. You know the drill. Search appliances can be easier to troubleshoot because the vendor or an authorized reseller delivered a server designed to perform within specific parameters. Keep in mind that if your colleagues created custom scripts, fiddled with various system settings via the administrative interface, or hacked the system, you will be required to perform methodical troubleshooting.

Restoring a Search System

If you must restore the system to a previous state, there are several points to keep in mind.

First, restoring the search system may require you to rebuild the index. Some systems have no provision for restoring an index short of operating two search systems in tandem. When one fails, you then fall back on the backup system.

Second, some search systems cannot be restored from a backup device. The basic system must be reinstalled and then specific procedures followed to restore your customized settings. None of the vendors whose systems we have tested intentionally makes reinstallation harder. Installation systems vary from vendor to vendor.

Third, a restored system may require updates before you can reinstall the index or launch an index rebuild. As a result, the reinstallation can take time. Depending on the hardware used, you may encounter unanticipated steps. For example, in some hardware configurations operating under Solaris, various digital signatures are used for security purposes. Keep codes, user names, passwords, and other security-related items at hand.

Fourth, replacement of some hardware devices may entail a specific certification procedure. Some high-end equipment cannot be moved into production until a certification sequence has been followed. The problem may be as minor as entering codes and allowing an IBM Serveraid device to reinitialize. The more complex procedures may involve a specially-trained engineer who must okay the system.

You may not be aware of the complexities associated with getting back online a large-scale, behind-the-firewall system. There is no easy way to circumvent the technical and procedural steps you must follow. If you take a short cut, the system may have an unknown weakness that can bring about another unexpected problem.

Replace

A couple of bouts with search system failure is often enough to trigger a project to find an alternative to an existing system.

In most organizations, *rip and replace* means turning off the existing system. You then deploy a new system. Unfortunately even the most dysfunctional search systems have

their loyal, often vocal adherents. Many organizations find it easier today to replace a search system without removing the incumbent system. This study profiles a number of companies whose software snaps into a search system or provides a specialized function that works better than the incumbent's subsystem for a particular process.

Today, rip and replace is almost as easy as a telephone call.

One approach is that you can contact one of the vendors of search appliances. You provide the estimated number of documents you want to index, and the vendor arranges for you to receive one or more servers. The vendors offering this solution today range from well-known companies such as Google to some less familiar such as EPI Thunderstone, Index Engines, and Planet Technologies. New appliance vendors enter the market frequently, so you will need to do a check to make sure you have a current list from which to work.

A second approach is to download Lucene, an open-source search system, and use it as a search engine. Lucene is reasonably easy to set up. If you are wary of open source software, you can license a version of Lucene from a commercial vendor like Tesuji. This company offers engineering and technical support services to its customers. Keep in mind that some of IBM's search solutions incorporate Lucene.

A third approach is to tap vendors who offer a hosted search or managed search service. The idea is that you open a port on your firewall, provide a list of targets for the service to index, and provide a search box to your users. The index, the search system, and the technical management are handled by the hosted search or managed search vendor. Your options for this type of solution are surprisingly broad. Blossom Software has a customer list of more than 400 commercial and governmental entities that provide enterprise search technology. A number of vendors offer hosted or managed services for search and content processing. Another variation is available from Fast Search & Transfer. The company will assume responsibility for your Fast Enterprise Search Platform (ESP) and provide a full-time Fast-certified engineer to operate the system.²

A fourth approach is to license another on-premises search system. You have many choices in search and content processing systems. Some of the search systems "snap in" or "bolt on" to your existing search system. You can operate these in parallel, shifting certain processes that are bottlenecks to the new system. In effect, you patch your incumbent search system with special purpose medicine.

Alternatively, you can freeze your incumbent system and bring online a replacement system. You may find that the interesting new companies offer a lower licensing cost, have introduced the specific features you require, and have designed the system for commodity hardware running Linux. No search system is without its flaws, but you

² Prior to Microsoft's buying Fast Search & Transfer, Microsoft was allowing certain Microsoft Certified Partners to host Microsoft search on the partners' servers. The Certified Partners could provide hosted search to their customers. It is not clear as of Jan. 23, 2008, how Fast Search's system will be meshed with the Certified Partners' hosted search option.

may be able to reduce certain costs and get back online more quickly than you thought possible.

Several vendors profiled in this report emphasize their ability to get up and running quickly, ISYS Search Software and Siderean Software to name two.

In short, rip-and-replace may not be as difficult as it would have been two or three years ago. Please keep in mind that incumbent vendors will lobby to keep you as a client.

Which Remediation Path Is Best?

No single best way exists. You may find yourself working through the routine troubleshooting procedures, solving the problem, and sticking with your incumbent system.

You may want to enhance your incumbent system or replace it. You have several different approaches to explore: hosted and managed solutions, appliances, and quick deployment on-premises systems.

Any of these approaches will work. Your specific situation determines your options and your scope of action. Based on our work with dozens of organizations wrestling with behind-the-firewall search, you will want to do thorough troubleshooting; for example:

- Gather data about the problem quickly. Assess the information, seeking advice and counsel from your colleagues. If the problem is a hardware failure, replace the defective device. Test. Deploy.
- Coordinate with the search system's technical support team. If that's not available, turn to your search vendor's reseller or partner. Explain the problem and seek suggestions. If your experience with the vendor is positive, you may want to have remote diagnostic run, if possible. If the system is offline, a site visit by a vendor's technical specialist may be warranted. Avoid random experimentation with configuration files or reinstalling the system.
- Listen carefully to the recommendations of the vendor. Before taking action, get the vendor to provide an e-mail or fax explaining what must be done, why, and the fees. If you and your technical resources are confident in their ability to perform a system restore or similar action, make the fix. If you are uncertain about any of the procedures recommended by the vendor, get the vendor or an authorized reseller to make the fix.

Avoiding Some Common Pitfalls

Based on our experience with behind-the-firewall search and content processing systems, several common problems appear again and again. Let's look at the five major pitfalls with older or legacy search systems and what to do to avoid them.

Pitfall	Actions to Consider
Inadequate infrastructure	Implement a quarterly upgrade cycle for the search system. You may be over-resourced for a short time, but the system will consume available resources
Vendor system updates/upgrades	Implement a tandem system or at the very least a development, staging, and production server. If the production server fails, you can use the staging server as a hot spare.
Overconfidence	Participate in vendor training and document any changes made to search system configuration files
Poor coding	Implement testing procedures and use configuration management systems to manage your code
Planning	Create a plan that includes tasks and then seek the advice of a consultant, make modifications, and work in a methodical way

Table 2: The Most Common Pitfalls

Inadequate Infrastructure

The servers, storage, bandwidth, and RAM are inadequate for your search and content processing workload. When your system was first installed, you were at ground zero. Over time, the volume of content processed has increased, usage typically has increased, query processing ramped up, and additional features such as entity extraction are sucking up CPU cycles and consume disc space. Typically you are finding that multiple points in the hardware infrastructure must be beefed up. Adding servers and other devices takes time and costs money. Most information technology budgets are limited. You may have to limit the volume of content processed or put a cap on the number of system users. Neither solution is ideal, but you face some stark choices. When costs or engineering are intertwined, you need to document the situation and seek support from management. If you can, budget for quarterly infrastructure upgrades. Avoid bottlenecks by expanding the system before the problems occur.

Updates/Upgrades

Vendors “push” updates to their licensees. In most cases, the updates have no material effect on the search system. But sometimes an operating system update coincides with a search system update and unexpected events occur. Many software vendors have limited testing resources, and your particular configuration may be one not thoroughly tested. There are two different ways to avoid this pitfall. Let’s look at each.

The rollback is the process used to restore your search system to an earlier state. The “gotcha” with search system rollbacks and data restores is that indexes may have to be rebuilt. For small document collections, reindexing is not an issue. For a system with tens of millions of documents, the index rebuild may take days or longer. The good news is that the problem can be easily addressed. The bad news is that users may expect the system to be up and running quickly. To deal with this problem, communicate what you are doing and the time required to get back online.

An alternative is to create a redundant search system. The idea is to operate two systems in tandem. If one fails, the system rolls over to the backup system. You can,

then, update the failed system. When the update proves stable, roll over to it and update the second, fail-over system. The problem with fail-over is that it adds to the cost of the search infrastructure. Nevertheless, tandem systems provide a mechanism to be up and running quickly after a failure.

Overconfidence

Some engineers love to tinker. They fiddle with settings or try tweaks to learn about the system. Search systems, however, are complicated beasts, and it is not easy to grasp the interdependencies within these complex systems. Even a simple indexing system for a single machine is a complicated program and often behaves in surprising ways. Larger-scale systems consist of many subsystems, often carefully balanced to juggle such tricky variables as query processing and response time, index updates, and certain content processing functions like entity extraction and classification. Many search and content processing systems come with plain text configuration files or graphical editors. Fiddling is easy. Unfortunately, a single configuration file change can bring the search system to its knees. Troubleshooting a random change is time consuming and therefore expensive. The good news is that configuration files can be restored. The bad news is that data corruption may have occurred. Rebuilding the indexes may be necessary.

Sloppy Programming and Inadequate Testing

Time is a scarce resource in many organizations. Short cuts are the norm. Most search systems allow programmers to interact with the indexing and content processing subsystems via application programming interfaces (APIs). Some programmers ignore documentation, preferring to dive in and learn by trial and error. Scripts or code with errors can create unexpected behavior in the search system. The fix is simple: Remove the offending code. If possible, limit the number of cooks stirring the search broth. Procedures and policies are the best protection against scripting errors.

Planning to Avoid Problems

No one wants a search system or content processing system to suffer problems. But problems occur. To make troubleshooting easier, consider planning to avoid certain pitfalls. Putting procedures in place before the search system is acquired is a prudent step. Also, make certain that the search system has adequate hardware resources and a planned upgrade cycle. By doing this, you can avoid some of the hardware and system issues discussed so far. Planning a year or more ahead is useful.

You will want to work out a specific procedure with your search system vendor about updates and upgrades to the search system. Vendors' marketers are eager to push new functionality to licensees. You certainly should license new features needed by your users. However, the installation procedures must be designed to keep the system up and running. A production system and a staging system are needed. Most organizations only use a production system for search. When a failure occurs, there is no fail over. The phrase *adequately resourced* means having a well-architected, redundant system and trained professionals on your team.

Finally, you will want to spell out who may make changes and under what circumstances; establish the quality assurance process enforced when a configuration

change is made. This will prevent random changes to the configuration files before the system is installed.

Some Tips and Tricks

Over the years, I've gathered a dozen tips and tricks to help smooth the shift from key word indexing to content processing. I've also included some recommendations on how to handle common problems that crop up when working with sophisticated search and text processing systems. We've tried to highlight problems that in our experience bedevil organizations, regardless of size.

1 Pricing

The vendor cannot or will not provide you with a price quotation. There are several parts to the issue of price. Let's break out the pieces of the price puzzle and tackle each briefly.

First, vendors of search, text processing, business intelligence and other systems often have a published price. Some call it a price list; others call it a floor price or the price below which the vendors cannot go. You'll find that most content processing systems have a minimum licensing fee in the \$50,000 range, but the first-year costs are often higher, sometimes hitting a \$1 million or more. Remember, this cost is for the right to use the software system. The reason the floor price is "low" and the first-year installed price is 20 times higher is because of:

- **Customizing.** Complex systems – and content processing is no exception – don't work like Microsoft Excel. Complex systems must be installed, debugged, tuned, and deployed, sometimes in phases. To get these systems operational, vendors have to change configuration scripts and sometimes write new code to get the systems working "as advertised." Marketing and sales people are not responsible for anything related to the system but selling and marketing.
- **Technical support.** Once the system is up and running, you may want to change how the system performs. The most common change is an adjustment to the indexing subsystem. Aggressive indexing and index updating can slow query response to unacceptable response times. Another common change is to force the system to boost certain content to the start page. These types of changes are trivial to an engineer familiar with the system. To someone not equipped with this knowledge, a minor tweak can wreak havoc. To get the system back on track, you may have to purchase technical support. When a system is licensed, information technology staff may assume "we can fix anything." The reality is that not even a search vendor's inexperienced engineers can fix their system. You will have to pay to get the informed support you need.
- **Troubleshooting.** You install the new system. Unexpected behaviors surface. Most often relevance is poor or the assisted navigation subsystem suggests useful sites that have nothing to do with the displayed category. What's the problem? The knee jerk reaction is to call the content processing system vendor. You find that the content processing system is not the problem. You have to work through the larger computing infrastructure to find the culprit. In short,

you have to investigate or pay someone to investigate the behavior. Problems of relevance or classification may be traced to initial system training, so new training data must be assembled and the system retrained. A lack of computing resources can generate erroneous results because certain processes fail to run to completion. The index does not get newer entries so users see incorrect or incomplete results.

The second aspect of system cost is infrastructure. Most search and content processing system upgrades take place without a thorough engineering review. The assumption made by most of the licensees is “We have enough hardware and bandwidth to handle this system.” The reality is that not only does the organization not have the hardware, storage, RAM, and CPU cycles to handle some advanced text processing functions but the engineers do not know the weaknesses of their current infrastructure. The reason is that turnover in many organizations creates knowledge voids. The current IT team may not know the problems because those engineers have not worked with the content processing system long enough to understand its true capacity. When a content processing system crashes or brings the internal network to a halt, the fix costs money. Most IT organizations don’t have the resources to do significant reengineering and upgrades. Any expenditure sets off a budget brush fire. The result is that the root problem is not “fixed.” A temporary patch is applied until the next problem.

The third aspect of cost is that most organizations assume that content processing is a standalone system. It is not. When a content processing system (CPS) works well, it is because the licensee has invested time and money in broader enterprise publishing work. File and data transformations are reduced or eliminated, thus saving as much as 30 percent of an IT department’s operating costs. Duplicate content is prevented from entering the content processing subsystem. Nothing fouls relevance ranking subsystems faster than two dozen variants of the same document.

When you try to estimate these three costs – vendor fees, infrastructure, and work flow modifications – you have almost no substantive data on which to base your analysis. Rough estimates are the best you can do.

The fix is to work up a budget, track direct and indirect costs, and keep management informed about the actual and anticipated expenditures. After you have installed and managed several behind-the-firewall search systems, you will have a better sense of the costs involved. Many organizations now have this deeper-cost understanding. If you lack these data, invest time doing some research and data gathering. Consultants may be helpful, but, as always, choose your expert carefully. There is no certifying authority for search expertise.

2 Vendor Unresponsiveness

The shortage of qualified search and content processing engineers is increasing. Google, Microsoft, and Yahoo! are in part to blame. The vendors themselves require a continual influx of new engineers. Organizations with businesses dependent on content processing need engineers. The result is that a shortage of technical expertise exists.

When your vendor does not get back to you quickly or notifies you that an engineering team will be available in 12 weeks, the reason for the delay is a shortage of qualified personnel. Most vendors try to respond to licensee requests quickly. The problem surfaces when the first-level support person must pass the request to more senior engineers for resolution. Senior engineers may be faced with multiple jobs varying from the trivial to the complex. When there are too few senior engineers, licensees see the non-responsiveness of the vendor.

There are some actions you can take to resolve the problem. First, look for qualified consultants familiar with the vendor's systems. Each of the Big Three and the superplatforms have relationships with firms able to resolve problems. Second, you can query the local university to find out if a professor on staff has expertise in the particular problem you face. Third, you can hire a person, train the individual, and delegate the repair to that individual.

None of these solutions is ideal, but because of the demand for top-quality engineering expertise, the staffing problem will not be resolved in the near future.

3 Automatic Classification Misclassifies Documents

Automatic systems can deliver accuracy comparable to a human subject matter expert doing manual classification. There are some caveats attached to this "comparable accuracy" statement.

First, the automatic classification system can be fooled when new concepts and categories appear frequently in modest document flows. Automatic systems typically use statistical, knowledge-based, or hybrid systems to work their magic. Statistical systems typically perform better when a certain volume of content flows through the system. These statistical systems can be fooled by highly volatile terminology.

Second, knowledge-based systems require updating. Some systems recognize new categories and automatically add these to the classification system knowledge base. You may have to perform some manual editing. Errors in classification must be identified, and then modified term mapping can be made via the system's administrative interface. You may find that you will need to formalize procedure to monitor misclassifications. Some of the systems profiled elsewhere in this study allow individual users to perform mappings. Be forewarned. Some well-meaning users can enter erroneous mappings.

Finally, the content processing system itself must be configured, resourced, and maintained correctly. We've covered this point in our discussion of costs.

4 Updating the Index Corrupts the Index

In the days of the Excite (Architext) system, adding the 17 millionth record could corrupt the index. Thankfully, those days are mostly behind content processing vendors. The challenge of index corruption is a tough one. There are numerous causes of index update problems. Some are as simple as trying to index too many documents. The indexing subsystem cannot process the flow and slows down. Other causes are improperly configured settings which create "hangs" or abruptly terminated processes. In some situations, one index is designated a production index, and the index being

updated is the staging index. After updating takes place on the staging index, which is then promoted to the production index, the system then redirects queries to the updated index on the newly promoted production index. A glitch in this process can lead to unexpected indexing results. Some vendors slipstream code updates. An installation bug or a flaw in the updated code can cause index updating problems.

As you can imagine, each of these problems must be considered. Different problems require different resolution procedures.

How do you deal with indexing problems? Let's look at the options:

- You make sure you maintain a hot spare of the last known good index. When a problem occurs, you send queries against that index. System architecture can prevent a total loss of query processing so users are not inconvenienced by a dead system.
- You reduce the number of documents you index. If you slow throughput and update the last known good index, you can recover without having to reindex the documents. This quick fix buys time until you determine if you need additional hardware or other remediating actions.
- You turn off certain rich text processing functions. Some advanced systems' performance characteristics cannot be appreciated until the system is in actual operation. Until and unless you resolve the engineering or resource issues, you may not be able to use some of the advanced text processing features described in the profiles that make up the bulk of this study.

5 Duplicates

A duplicate to a human and a duplicate to a computer system are not necessarily the same. Humans can often look at a page of a document and know that the data are not the most recent. A computer may look at the same document and determine that the document is identical to another document, not necessarily considering the specific data on the page the human "knew" to examine.

Duplicate detection techniques vary from vendor to vendor. Some use file date and time, and the file name to determine which one is the most current version. A number of duplicate detection systems create a number based on cyclic redundancy check (CRC) algorithms. When two documents yield an identical CRC value, the documents are identical. A human, however, may "know" that the most recent document contains the incorrect data because the person assembling the document picked up an older, inaccurate data set, not the correct information.

Content management systems often contain multiple versions of a Web page or document. Other systems maintain pieces of a document, and each component is updated as required. When a "final" version of a document is needed, the CMS system assembles the pieces flagged as "final." The system outputs a final version.

There is no easy, automatic way to ensure that search systems contain only the "best and final" version of a document. Manual systems are expensive and slow. What's the fix?

The answer is “Rules.” You will need to program your content processing system to identify and process only those documents that match specific characteristics. Remember that rules are unforgiving, so expect to spend time identifying what makes a document “best and final.”

Once you explore rules-based duplicate detection and de-duplication processes on existing collections, you will probably resign yourself to living with duplicates or variants.

Content processing systems can be fed only “best and final” documents by individuals and by enterprise publishing systems. The hurdles you will have to jump over include:

- Colleagues may disagree on what constitutes a duplicate or a “best and final” version of a document
- Documents assembled from XML repositories may exist only as entities or components. A “final” document is built from these objects. In this situation, you will need to narrow the definition of “duplicate” to these constituent elements. Data may reside in a database. In this case, there may be no “best and final” version of an object, because an object may be changed at any point in time. Therefore, the “best and final” version may be created only if certain data files are restored from a back up medium. A distinction as fine as a version created from a data restoration process may make no sense to most workers. But in a legal matter, the distinction becomes more important.
- Management does not understand the cost benefit of implementing a duplicate-free content processing system. When faced with minimal management support, you may have to invest time in justifying the costs of the de-duplication project.
- Rules must be maintained. The definition of a duplicate document can change, often unexpectedly. Cost control becomes an issue.
- Automatic systems behave in an unpredictable manner. You may have to live with these problems or abandon de-duplication efforts.

6 Structured Data

Most of the systems profiled elsewhere in this study can handle structured data. *Structured data* means information that resides in or can be correctly represented in a database table. Here, database table means a structure supported by an industry-standard relational database management system such as IBM’s DB2, Microsoft’s SQLServer, Oracle’s database, or any other Codd RDBMS.

Vendors have different strategies for manipulating structured data. Some attempt to process the data automatically; others require that specific fields be identified for the content processing systems. Purely numeric data in database tables can pose problems for content processing subsystems. Depending on the CPS, non-textual information may be excluded from content processing.

One approach that may warrant testing is the following procedure:

- Identify the data you want to have displayed in a report

- Create a script to generate one or more reports in file formats that the content processing system can identify and process without manual intervention or an excessive number of documents that cannot be processed
- Place the generated reports in a folder that the content processing system indexes.

Some systems perform a roughly similar process as part of their content transformation procedures. Most systems provide an administrative interface to configure the transformation component to handle each database containing content you want to index.

In our work, the effort invested to output a report from each database table we want to index has proved useful; however, the variation in database table data definitions makes testing different approaches a worthwhile use of your time.

7 Proprietary Information Indexed

Here's the situation: an authorized user gets access to information that particular user was not supposed to see. The problem becomes more difficult when one employee circulates the proprietary data to other employees.

Some organizations are generally unaware of what resides on their behind-the-firewall servers. In my experience, I have had to deal with soccer club information, minutes of local government meetings, and similar effluvia produced by employees using an organization's computer system with little thought to what's appropriate and what's not.

This problem is a consequence of flawed processes related to employee awareness of information policies, security procedures, content guidelines, and non-enforcement of access control lists. Most search systems use the security system already in place at an organization. In this situation, the "problem" may not be resolvable by the search administrator. The firm's security policy becomes the key to resolving the problem.

But for the immediate problem, the search administrator may have to purge the index, identify the source of the problematic documents, and either remove those documents from the content processing system's index or purge the existing index and reindex. Neither approach resolves the problem because new unauthorized or inappropriate documents will creep in.

Content processing systems ingest content copied to the processing queue by the content acquisition system. In some systems, scripts residing on servers in the organization automatically transfer new or changed documents to the content processing system. In this case, the cause of the problem is an individual who places a document in a folder "watched" by the content processing system.

Whether the security policy or an individual employee is the cause of the breakdown, the search system itself is not technically operating incorrectly. Here are steps you may want to follow:

Determine what document was retrieved, its security classification, and its "owner."

- Verify that the access controls for this document are correctly configured and that the search system is accurately exchanging data with the security system. In some organizations, different staffs may be responsible for these systems.
- Remove the offending documents from the system and re-index or adjust the access control flags.
- Perform checks to verify that the security flags are operating properly.

Note that if an individual inadvertently copied a document without observing security procedures, you will need to deal with that problem following your organization's policies. In some cases, the only way to ensure that security flags are properly recognized by the content processing system is to go back to ground zero with the content processing system. If there are habitual breakdowns in security enforcement, you have a management issue that requires escalation.

8 Timing Out

Some users report that the CPS does not respond. Others complain that their reports are not found and the system does not display error messages.

These problems are intermittent, and are, therefore, very difficult to troubleshoot. You will need to enlist users to help you recreate the situation. In my experience, many users can provide general information, saying "I click the icon for the report and I don't get the report" or "I will run the query for you. Oh, look at that. Now the system works."

There are three areas to investigate before you can begin troubleshooting the specifics.

First, intermittent behavior may be caused by bottlenecks or abnormal network or system demands. Look in the log files for information about resource utilization. Search systems are often sensitive to having the storage and computational resources available when processes run. If there is a bottleneck in a content processing subsystem, you must either gate the number of users of that feature or add computational resources in some way. If you cut back on users, you will increase the risk of user push back because the system forces a behavior change on the users. If you have adequate resources, you may be faced with a problem related to the interaction of the content processing system and the software or hardware responsible for resource allocation.

Second, you may have an overloaded network. The content processing system does its job and sends the data to the user. But network contention delays or fails to deliver the needed content. Some reports make extensive use of client-side functions and may push both data and instructions to the user's software client. Functions may not operate or pages may partially render. The fix for this problem is to find a way to reduce network overload. In most organizations, expanding bandwidth is difficult and time consuming because infrastructure must be modified, devices configured, and other work must be completed. Try to identify the processes that choke the network and explore ways to minimize the problem of trying to push your infrastructure to its limits.

Third, you may be able to use the administrative tools that are included with the search system to change the priorities or functions of the content processing system. You may

be disabling some advanced processes, but unless you cut back on what you ask the system to do, you will not be able to implement a quick fix for some problems.

I won't repeat the specifics of balancing content flow, text processing functions, and infrastructure. When these are out of whack, users may be confronted with intermittent or erratic system behaviors.

9 Specialized Data Not Handled by Incumbent's System

Audio and video files are becoming more common in organizations. These files can be indexed, but special content processing procedures may be necessary. Additional software, either from your content processing vendor or from a third part, may be required.

Before you jump into audio and video content processing, find out how many of these content objects you have to process. Remember to determine if there are duplicate versions of the audio and video files. Also, what is the rate at which audio or video files are added or changed? You will need these data to calculate storage devices and figure out how you will deliver audio or video to a user requesting these files.

A number of vendors offer subsystems that "listen" to audio and video, create a transcript, and then index the words and concepts in the transcript for each video object. Another approach is to have human indexers create bibliographic records, index, and classify each audio and video file. The audio and video files are not indexed by the system. Your content processing system indexes the bibliographic records associated with each audio and video file. Several firms offer software that "understands" audio and video. In general, these systems are likely to be too expensive or unreliable for most enterprise applications.

I have recently fallen back on human indexing of audio and video files. I then instruct the content processing system to index these bibliographic records with links to the audio and video files. For low-cost storage, I use third-party content delivery systems such as Amazon's new and very price competitive S3 (simple storage system).

10 Vendor Forces an Upgrade/Update You Don't Want

You have happy users. Your search and content processing systems are stable. Your vendor notifies you that support for your present version will be terminated, and you must upgrade to the current version.

You don't want to upgrade. In this situation, you have a choice. You can refuse to update your system. You will take responsibility for maintenance, customization, and any other support the system requires. As long as the present system meets your needs, you may want to consider this approach. The vendor may grouse, and you may be able to defer the upgrade until you need the new features or you know that the upgrade will be relatively bug-free.

If you upgrade, there is no guarantee that the new version will be fully compatible with your particular environment. The hassles of troubleshooting may be worth the time and money if the new features are ones your users require.

Think about this issue before you sign the license agreement. You may be able to insert language that gives you the option to defer an upgrade without penalty such as losing access to the vendor's technical support team.

Vendors want licensees to install the most recent versions of their systems. Upgrades may generate revenue for the vendor. Consider your particular situation and act accordingly. It is also helpful to seek comments from other customers about their experiences with a vendor's upgrade/update processes.

11 Management Mandates a Vendor Change

A large pharmaceutical company's Board of Directors okayed the firm's standardizing on the SAP enterprise software. The search and content processing system was orphaned. The SAP system included a search system called TREX (pronounced tee-rex), but it was on its way to refurbishment.

What can you do when this situation arises? The answer is simple: "Adapt." Changes imposed by mandate can be sidestepped if you are clever. If you are not as clever as the boss, you will be in a tough spot.

With mergers and acquisitions taking place, the likelihood of a mandated change continues to creep upward. In my experience, mandated changes are part of the business process today. Also in my experience, the search and content management issues are sufficiently troublesome that sometimes a housecleaning is the only solution that a Board of Directors or president can make to resolve an existing problem. As the truism says, "You are part of the solution, or you are part of the problem."

12 Vendor Is Acquired or Goes Out of Business

The best way to deal with a vendor that is acquired or a vendor who goes "belly up" is to anticipate the problem before the license agreement is signed. You want to have the vendor place a copy of the source code in escrow with a bonded third party. If the vendor is acquired or goes out of business, you can obtain a copy of the source code for the search or content processing system. In theory, you will be able to create and possibly modify the software. Without the source code in escrow, you will be forced to live with the system until you can license a replacement.

One reason why some organizations consciously operate multiple search and content processing systems is redundancy. If one of the vendors goes out of business, the service, in some form, can continue without having to go through the expense of dealing with source code or procuring an alternative system.

13 Spaghetti Code (Bonus Tip 1)

No vendor will admit it, but the code in your seven figure behind-the-firewall content processing system may be like a yarn box in a tangled mass. You discover that if you change a setting for one function, another unrelated event occurs. You roll back the change and learn that the unexpected behavior remains.

Unfortunately you may discover this tangled-yarn problem after you have signed a contract and made a payment to the vendor. The reason for the situation often varies

from vendor to vendor. Some license bits and pieces from individual contractors or specialty vendors. Some companies grow via acquisition. Their management team insists the pieces have been integrated. Well, what would you expect them to say? Others just take short cuts. Regardless of the cause, when you encounter the problem, accept that there may be no easy fix and or even a slow fix that your budget can afford.

Let me give my rules of thumb this situation:

- Do nothing until you have a backup of the system. Test the restore function to make sure it works. In my experience, one-third of backups do not restore 100 percent. Once you have a known-good backup, then you make your change.
- Document where you found the information about the change, what you did, and when you did it. If something goes amiss, you will need these pieces of information to get guidance about remedial actions.
- Set up a procedure so that no one person – full-time, part-time, or contractor – can make a change without some checks and balances. A single engineer who enjoys learning via trial and error can trash the system.
- Hold off making any changes until you have a three-tier set up for the search system. This means that you have a development server where the change is created and tested in a preliminary way; you have a staging server where the change is put in a environment identical to the one that is live and working on your Intranet; and when the change works as you expect, then you move the change to the production server.

Dissuade yourself that you are able to make *ad hoc* changes, solve vendor scripting errors, or write a patch that addresses an annoying problem. You may be able to do this, but if you muck around with the system, you may find yourself violating the license terms. Formal procedures, documentation, and prudence are the watch words when spaghetti code is the main course.

14 Specialized Content (Bonus Tip 2)

Most organizations will not have to search, process, or make searchable certain types of content. The content in question is complex because it contains text and also other high-value features or elements that are almost impossible for most content processing systems to handle. Among the problematic document types are:

- Engineering drawings and their associated tables of components, costs, and other data.
- Mathematical equations, particularly those rendered in traditional mathematical notation or in computer code. Added complexities are either the graphic outputs of these equations or an associated program that intakes the equation output and generates another object; for example, a technical rendering of some type.
- Patent applications, patents, and their associated drawings, data sets, and supporting materials. Processes for legal discovery are different from those used for competitive intelligence.

- Compound documents that include multiple content objects; for example, filings for drug approval. These filings include formulae, data, and text.
- Chemical structure information that may include graphics of the chemical structures, specialized codes and data representing the structures, text, and referenced data such as substantial collections of research data.
- Documents collected for a legal matter. The process of discovery can yield a wide range of content, ranging from photocopies of varying quality, electronic mail, digital data and information of many different types, photographs, and sometimes video such as surveillance video files.

When you are asked to create a search, text mining, data mining, or some other type of retrieval mechanism for these types of content, you need to determine the number of documents, the frequency of change in the content set, and legal requirements or guidelines, if any. Armed with these data, you can then begin the process of determining if your existing search vendor's system can handle these materials. If you find that the vendor cannot commit to a yes or no answer and offers a "Well, it depends" response, you will have to procure and deploy a separate search system.

In my experience, a system to handle special content requires as much work as an organization-wide system. Specialized content is difficult for a person unfamiliar with the information in the source material. Almost any employee can read a marketing memo and have an idea about the subject discussed. Contrast that with a chemical structure.

Do not assume that your existing systems can handle these specialized data types. You may want to get a consultant specializing in the particular type of content you want to make available to your users. You may want to test your incumbent systems to generate a benchmark for the collection. You will be in a better position to evaluate the specialized systems from which you select a vendor.

Getting to More than Key Words

If you are reading this study, you want a system that delivers more than a search box. In Key word searching a user types 2.5 words on average in a search box, hits the Enter key, and views a list of results.

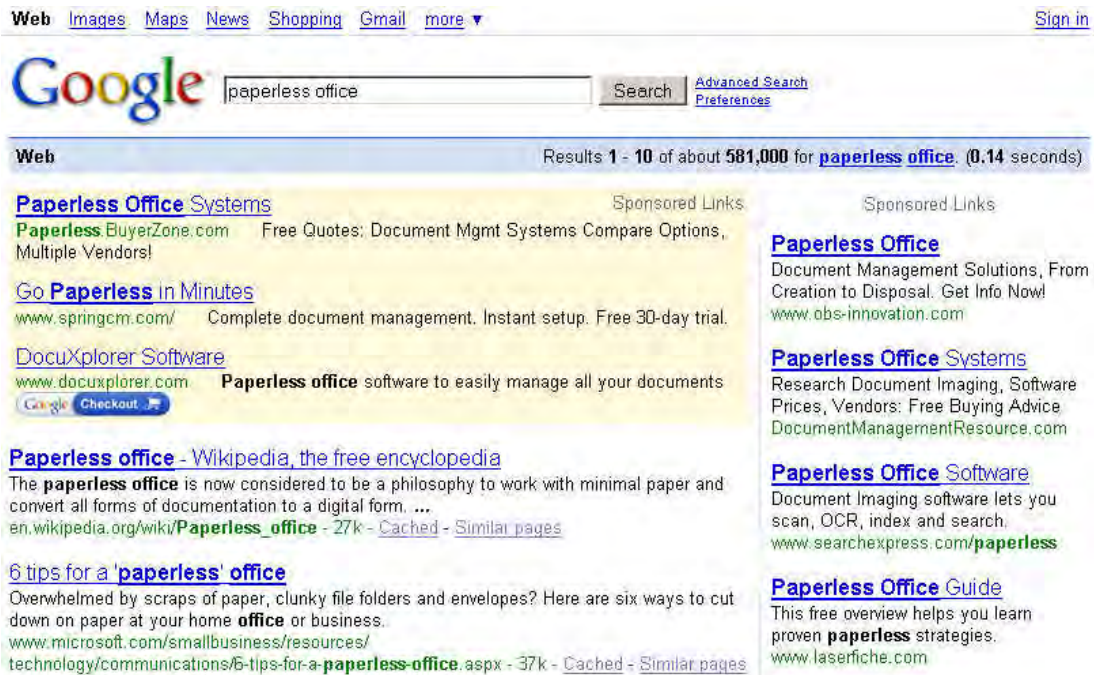


Figure 1: Google's Ubiquitous Search Box

The ubiquitous search box as it appears on Google. Google generates ads, and for many users, these messages lead the user to pertinent information. Users of behind-the-firewall systems expect a Google-style search box and Google-style results, including suggestions of other useful links

Once, key word searching was a marvel, particularly when compared to looking for information the old fashioned way with printed reference books and microfilm.

Google and other modern systems recognize that the two words paperless office go together in English as a bound phrase. The notion behind this type of display, which is ubiquitous in search and retrieval, is that the user will click on the result at the top of the list. The implicit assumption is that a result on the first page of results will be more relevant to the user's query than a result on page 12 of results. In order help a user decide whether to click on a result and see the source document or Web page, systems provide a snippet of text or a summary. But the user has to sift through the results, clicking and scanning, hunting for a document that provides the needed information. Once the novelty wears off, clicking and scanning is work. Users want a better way to locate needed information. Key word searching, therefore, doesn't meet some users' needs. Setting aside the academic fights over the best way to search, users want to break out of the search box. Users, including your author, want ways to go beyond search.

A number of public Web search services are shifting from laundry lists of results to interfaces that offer users an alternative to the search box. Endeca is one of the companies that recognized the tyranny of the search box. The company's success is in part due to its *guided navigation*. I will use the phrase *point-and-click interface* to describe the combination of a search box with hot links to other content that the search system displays in response to a user action. An exemplary implementation of the Endeca approach is the *Guardian* newspaper's [Web site](#).³

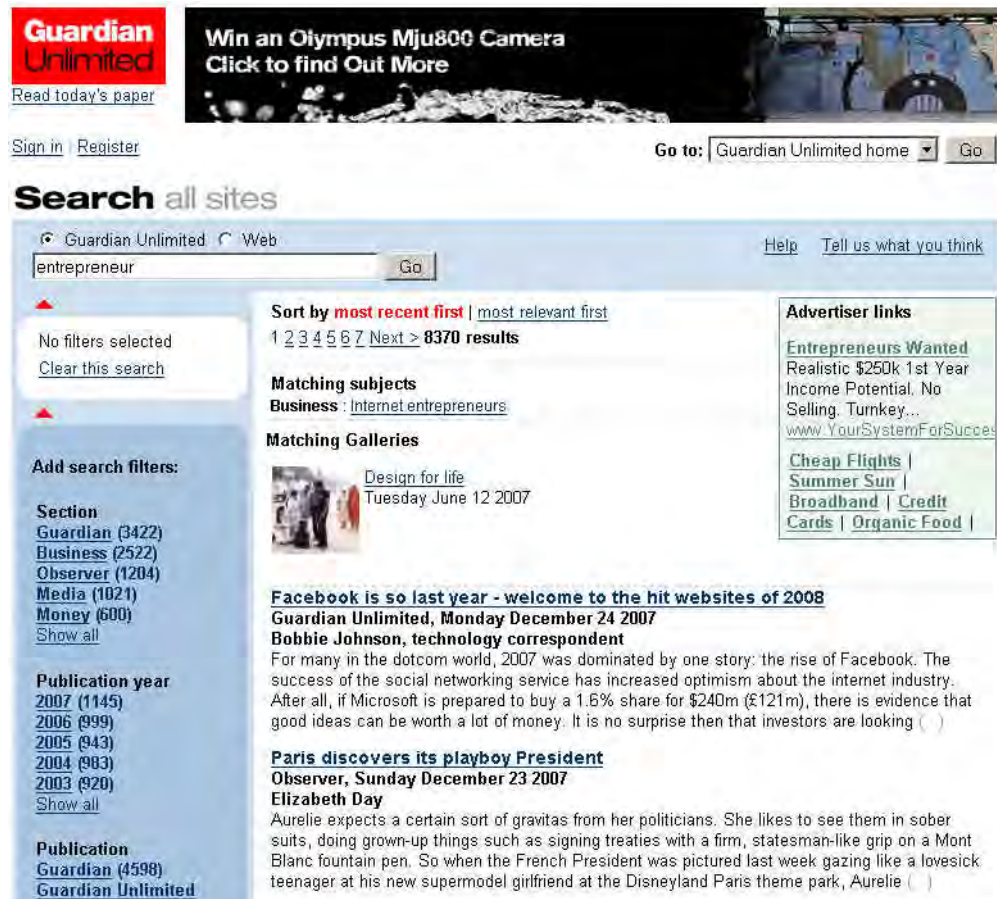


Figure 2: Endeca's Guided Navigation

The links, outlined by the bounding box, allow the user to explore other relevant information automatically identified by the Endeca system. Endeca's editorial controls allow a splash page to be generated containing the most recent or important information available. The licensee usually gives the Endeca system rules to follow that will refine its built-in, automatic functions, a practice followed by Endeca's competitors. © Endeca 2007

The Symbiosis of an Interface and Search Technology

Users of behind-the-firewall search or content processing system see the search system as the interface. In my experience, most users explain their search wants in terms of the

³ Navigate to <http://www.guardian.co.uk> or Endeca's list of live demonstrations accessible here: <http://endeca.com/technology/index.html>

interface. For example, in one focus group in September 2007, users said (and I paraphrase), “I want to see suggestions for other relevant content.”

The problem is that the interface can display only what the system supports or makes available. If the underlying system generates “See Also” and “Use For” references, displays with news content germane to the user’s business area, and presents a report of facts instead of a laundry list, the perception is that the interface is the search system.

The problem is that the interface does not supply the information. The underlying content processing system must be able to pipe categories, near real-time news, personalized information features, and ready-for-distribution reports. Therefore, both the interface and its interaction with the underlying system are important. In order to meet user needs, both components must be given their due.

When procuring a search or content processing system, the “what you see is what you get” approach can be misleading. Many vendors offer carefully orchestrated demonstrations of their systems. How can you determine if you are seeing a “real system” or a “demo system?” You can’t, so you have to ask.

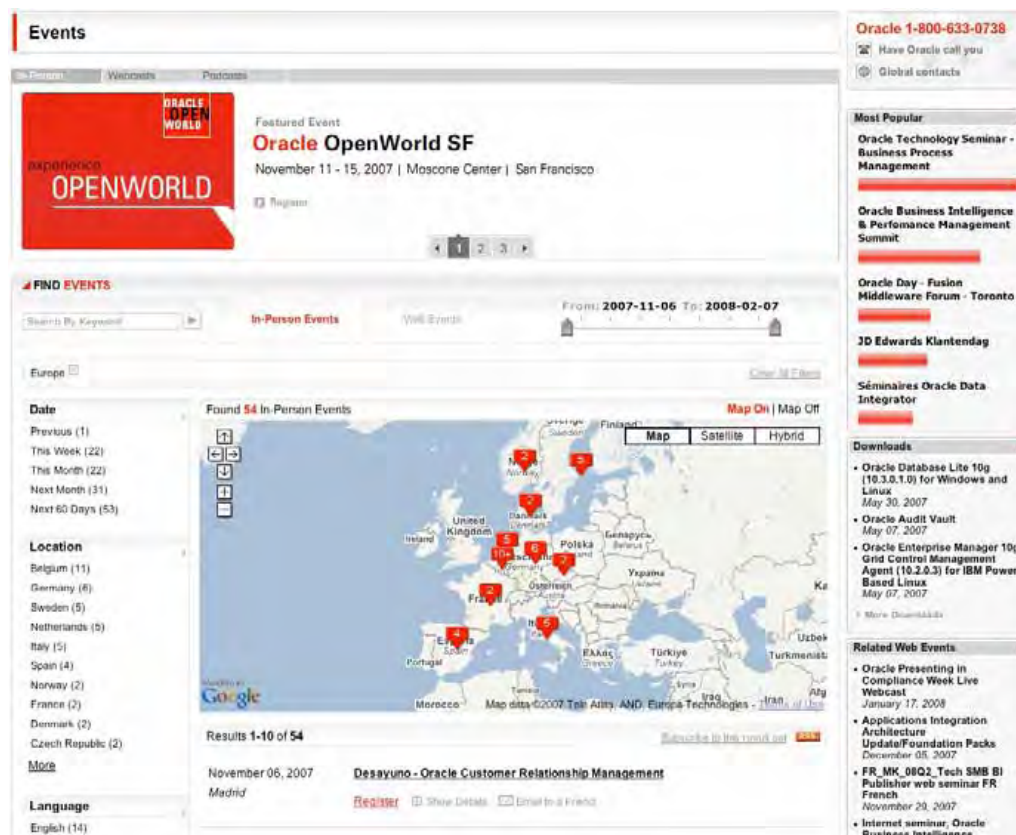


Figure 3: Siderean "Snapped Into" Oracle SES 11g

The Siderean Software system allows a dashboard, point-and-click, or an assisted navigation interface to be used. For this Oracle implementation, the Siderean system “snaps into” the Oracle SES 11g system, thus adding semantic functions that complement Oracle’s database and search technology.

Let's set aside fine points of interface design and assume that your content processing system displays a combination of:

- A search box
- Hot links to other content discovered automatically or displayed when the system follows rules that you specified when the system was installed
- A list of results presented when a user enters a query or clicks on a hot link
- Default content such as boosted content, news, or boilerplate text.

The key point is that the system automatically outputs this information. No human intervention is required once the system has been set up. Systems that require continual human editing, tweaking, and tuning are too cumbersome and costly to operate. There are some systems in intelligence, law enforcement, health care, and government agencies where the requirements mandate a human-intermediated system. But even in these specialized implementations, automation is needed to reduce bottlenecks and certain operational costs.

To move beyond key word search for behind-the-firewall information access, systems that require constant baby sitting are less and less desirable. Financial and technical resources are under increasing pressure. Manual intervention translates to higher operational costs.

The interface, while very important, cannot be separated from the system's ability to convert content into a form that permits automatic generation of a point-and-click interface. Accordingly, I want to focus on characteristics of a "beyond search" interface, not the merits or flaws of a particular design for rendering the elements.

Beyond key word search has these characteristics which may or may not be implemented on the interface the licensee presents to users:

Making Suggestions to the User

My work reveals that users like interfaces that make it easy to find useful, pertinent information. In a behind-the-firewall search implementation, a user can spot information more quickly than they can figure out a query.

The use of categories and classifications that contain relevant material is a form of suggestion that is gaining in popularity. The content processing that makes suggestions possible is, however, more complicated than the old method of building an inverted index of words.

Illustrated is an example from a British government Web site. The Autonomy technology is used to process content. Notice that the suggestions appear in tabs. The use of a tabbed folder metaphor helps reduce visual clutter.

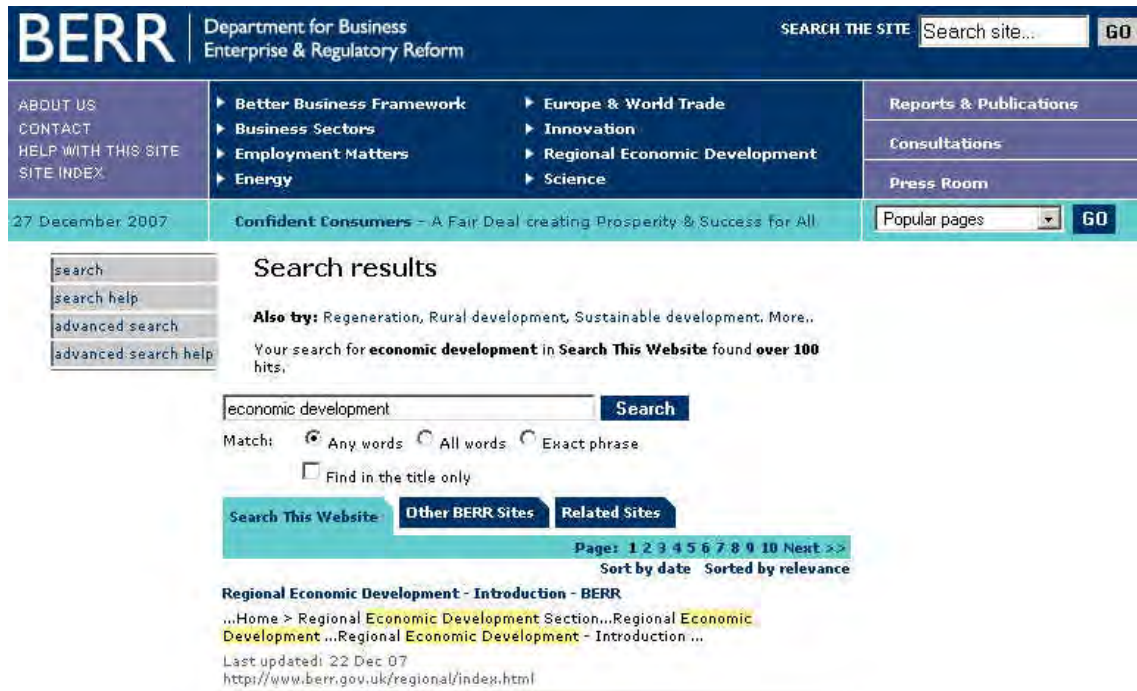


Figure 4: Autonomy IDOL Hot Links

Tabs in the center of the display are hot links to other potentially relevant content. This is the Business, Enterprise, and Regulatory Reform Web site powered by Autonomy IDOL. © Autonomy Ltd. 2007

Names of People, Places, and Things

Entity extraction identifies and indexes the names of people, places, and things in documents and other content. Shown below is Inxight Software's entity extraction function. The system generates a list of names indexed for use by other subsystems or to present hot links to information about a person, place, or thing. Inxight, a company spun out of Xerox Palo Alto Research Center (PARC), was one of the first companies offering tools such as entity extraction.

The graphic representation is complicated, but its purpose is to identify by color the different entities the system found in the snippet of text from a processed document. Traditional key word indexing systems cannot perform this type of function. Entities, once extracted, can be classified and related to concepts, other entities, and documents.

A user can set up the content processing system to watch for a particular entity. When the system identifies and tags an entity, the system can alert the user that new information about that entity has entered the system. Some systems perform automatic summarization, generate an e-mail containing the digest, and include a hot link to the source document. Other functions can be set up once the entities have been indexed.

The proposed merger between Mega, Inc. and CNA Systems, Incorporated, has been postponed, Mega CEO Joe Smith said in an analyst call. "CNA's 1st quarter revenue dropped by 32%, and they lost 23 million dollars," Smith explained. CNA Systems sources blame weak sales in China. CNA shares (CNAI) fell 47 percent to \$9.84 on May 12, the first trading day after the announcement.

Company	Mega, Inc., CNA Systems, Incorporated
Date	May 12
Person	Joe Smith
Person Position	Mega CEO
Currency	23 million dollars, \$9.84
Measurement	32%, 47 percent
Country	China
Noun Group	proposed merger, analyst call, 1st quarter revenue weak sales, first trading day

Figure 5: Inxight Entity Extraction

The color coding makes clear the different entities that an entity extraction system can identify. Keep in mind that entity extraction uses a number of subprocesses to identify and index entities. Entity extraction is, therefore, a combination of algorithms, not a single algorithm. © Inxight Software, 2007

Classification

The human mind has an ability to perceive relationships among actions, information, and events, among other phenomena. There is no hard-and-fast rule that each person uses in order to group similar ideas, facts, or other entities. Computers follow rules. Automated classification systems use a number of different methods to group similar objects. You can recognize a beyond-text search system that uses classification to assist a user. The screen shot below is from Vivisimo in use for a U.S. government-wide index.

The screenshot shows the USA Search.gov interface. At the top, there's a search bar with 'economic assistance' entered and a 'Search' button. To the right are links for 'Advanced Search', 'Search Tips', and 'Busque en español'. Below the search bar, it says '100 results for economic assistance out of at least 8,060,000 (Details)' and 'Web results by Live Search'. On the left, there's a 'Topics' sidebar with a list of categories: All Results (100), Business Assistance (16), Community (10), Department of Social Services (6), County Department (6), Division (5), Department of Commerce (4), Lower Income (4), Responding (5), Wakegov.Com (2), Announces, House (3), and More | All. The main content area has a 'FAQs' tab selected, showing 'Federal Aid and Benefits'. Below this, there's a link to 'Economic and Humanitarian Assistance Offered to Other Countries - USA.gov'. The search results list includes: 'Food Stamps :: Economic Assistance' from the South Dakota Department of Social Services, 'Economic Assistance :: SD Dept. of Social Services', 'WakeGOV.com - Economic Assistance', and 'WakeGOV.com - Emergency Financial Assistance'. Each result includes a brief description and a URL. At the bottom left, there's a 'Font size' selector and a 'Search portal by Vivisimo' logo.

Figure 6: Vivisimo Automatic Classification

The Vivisimo system automatically classifies results. A user can browse “Topics” and explore related information by pointing and clicking on those links.

Hybrid Displays

A hybrid interface combines text, hot links, and graphics on one screen. The term hybrid interface, as I use it, is a synonym for a dashboard interface.

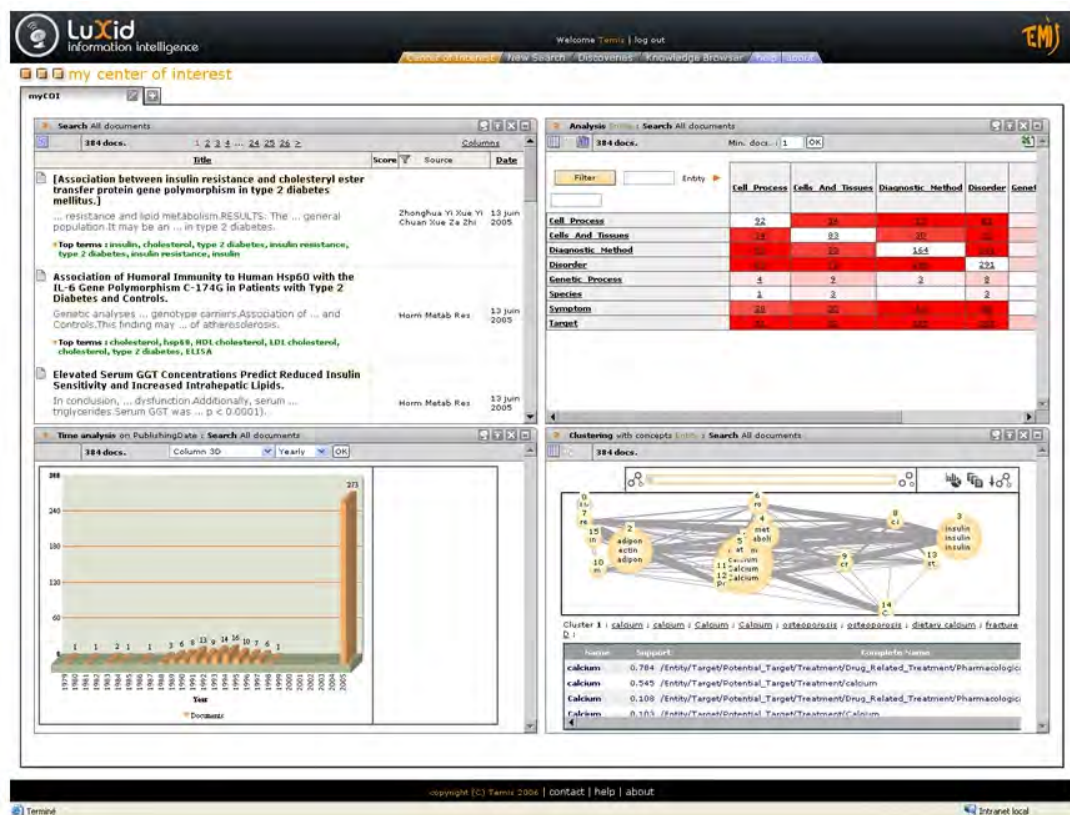


Figure 7: Temis Luxid Interface Options

The Luxid interface makes it possible to present hot links to search results via a table, a histogram, and a hyperbolic map of relationships among data in the results set. © Temis SA, 2007.

The use of hybrid displays is to move beyond key word search. No single search feature has as much sizzle as an interface that uses text and images, charts, and other graphics. Advanced content processing systems make it comparatively easy to represent relationships and other factors in a graphic. The example above comes from Temis SA, a French company that offers a wide range of content processing tools. Temis technology has strong adherents in health care, pharmaceutical, and financial institutions. The screen shot illustrates a compound interface that combines four display panels with hot links to related content

Identifiable Trends

That search is changing is the premise of this study. You and I want to have systems that go beyond key word search. We still want the search box, and we want richer interfaces. A number of other trends are discernable. It is outside the scope of this study to explore certain topics in the depth each deserves. Let's look at several of the most important. Many of these trends depend on sophisticated technologies. The "religious" wars between advocates of statistical techniques and semantic techniques are interesting. For our purposes, keep in mind that many vendors are creating systems that use *both* statistical procedures and semantic techniques. Going forward, I see more

mixing and matching of methods in order to deliver what each vendor wishes to deliver with a search or content processing system.

Leveraging Probability is the “In” Technique

The man responsible for statistical approaches to figuring out the meaning of information is a Presbyterian minister who lived in the 18th Century in Britain. Thomas Bayes is credited with codifying procedures that allowed a “rule” to be updated when new evidence becomes known Bayesian logic.

Combined with mathematical procedures from cognitive psychology and other disciplines, system vendors like Autonomy and Google built content processing “engines” that can index, classify, identify bound phrases, make suggestions, and perform other types of sophisticated functions.

The idea is that once a system begins processing content, the system can perform functions once thought to require only a trained indexing professional or subject matter specialist.

Stripping down the complicated mathematics leaves us with no easy way to illustrate these algorithms. For the purpose at hand, the statistical systems rely on probability, frequency, and similarity to make the systems “smart.” Today, when properly set up, probability-based systems work quickly and are quite useful. The downside is that if not monitored, the algorithms can stray off track. Administrative interfaces allow a human to tweak some settings to get the system back on track.

There are many different approaches to probabilistic content processing. Some vendors – like Google – use some algorithms readily available in any math book. Others create new mathematical procedures to cope with the peculiarities of human discourse. Brainware, for example, has created a patented form of statistical analysis. In either approach, the idea is to exploit mathematics to get beyond basic key word indexing.

Semantics

With ready access to low-cost, fast computers, system vendors have begun to leverage compute-power to extract the meaning of a document. I want to describe characteristics of semantic and linguistic techniques so you can appreciate the distinctions that are referenced in the profiles that accompany this study.

You have heard about semantics in phrases such as the *Semantic Web* or vendors who talk their *semantic technology*. The key idea of content processing is that a system is able to identify and index concepts and meaning in a document. Semantic systems can classify a document and sometimes generate a summary of the document.

There are a number of useful metaphors that help convey the essence of a semantic system and its semantic technology. I find it helpful to think of semantic processing as producing index tags that allow links and associations to be used to create “See Also” and “Use For” references.

In terms of what the user sees, a semantic system provides hot links to categories that may be relevant to the user's query. The best-known metaphor for this type of connection is a hyperbolic map shown in the following figure.

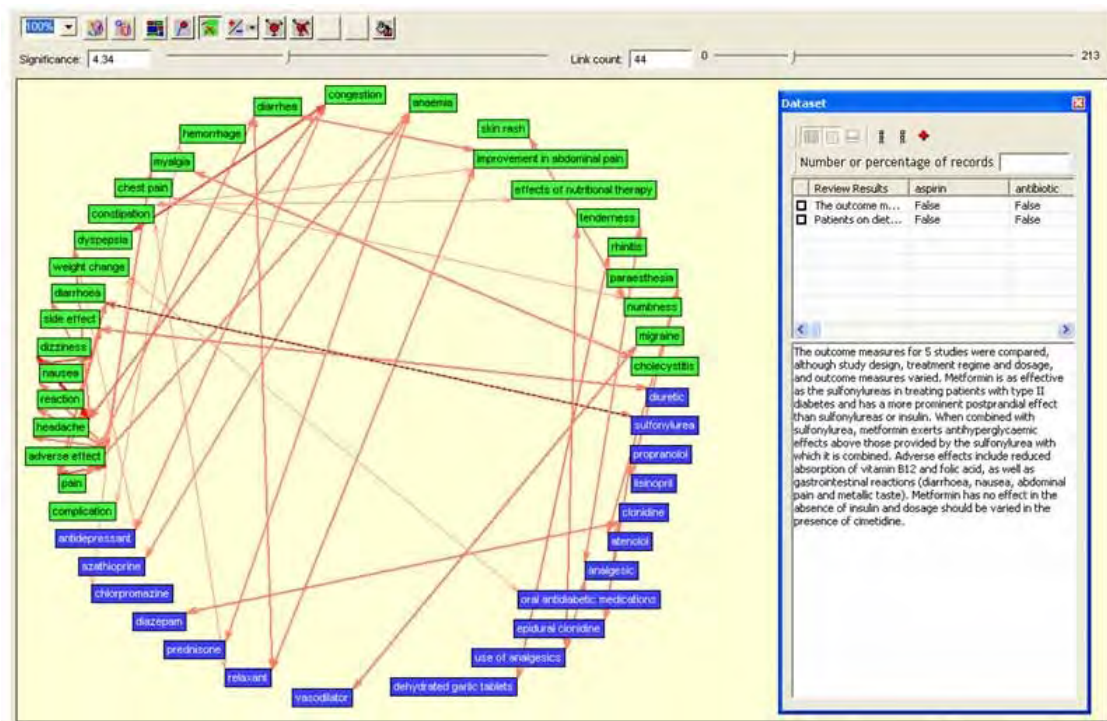


Figure 8: Hyperbolic Relationship Map

This is a wheel directory that leads to maps illustrating the relationships among items in a set of processed documents. Each link displays a source document when clicked.

Semantic content processing is experiencing a resurgence. In 2007, Google's invention of what are examples of semantic technologies appeared in a series of five patent applications published by the USPTO. Many of the companies profiled in this study offer their own approach to semantic technology. Between the approaches of giants like Google or the far smaller Siderean Software, there are many different ways to implement semantic functions.

When the phrase *semantic technology* is used to describe a system, be prepared to ask questions. Vendors will be describing some interesting but often complicated techniques. Furthermore, probing into the guts of a semantic technology may bring up some surprisingly philosophical questions as well as the pragmatic ones. Semantic systems are about meaning, knowledge, and concepts. The strong interest in taxonomies and ontologies is a reflection of the impact semantic content processing now has. For the purpose of this study, think of semantic content processing as generating information about a document that goes beyond key word indexing. Key word indexing allows words to be located. Semantic indexing allows meaning to be identified and used in assisted navigation, suggestions for other, potentially relevant content, and supporting different interface techniques for discovering information.

Computational Linguistics

A number of vendors have embraced content processing techniques sometimes referenced as linguistic text processing, computational linguistics, or just linguistic systems. One vendor – Linguamatics – uses the root *lingua* in its company name to make certain that its technology’s family tree is clear. A vendor of a content processing identifying the system as performing Natural Language Processing (NLP) is signaling you about one of the system’s key architectural features.

The idea is that linguistic systems use discoveries about the knowledge a human needs to use and understand language. Linguistic systems use discoveries about the syntax of language to create models or procedures that can understand a document in the way a human does; that is, grasping concepts and ferreting out meaning from a page of text.

A content processing system will contain one or more representations or models of the engineers’ view of how natural language works. The systems will process a document using multiple techniques, ideally executed in parallel, to speed up the many calculations required for linguistic processing. A system implemented with computational linguistics will chop content into parts of speech, identify phrases by discovering them or looking them up in a word list, use the model to help the system identify concepts, and perform other functions that result in generating index tags for concepts, categories, relationships, and meaning.

There are fierce debates about how to make a computer program understand information in a human way. Some techniques rely more on statistical processing within linguistic text processing than algorithms that rely on dictionaries. Most of the systems profiled in this study use a combination of techniques. Commercial vendors have to be pragmatic, so the more esoteric techniques may be shelved for a brute-force procedure or a clever mathematical shortcut. There is no one best way to get a computer to understand in the way a person does.

One catchphrase that comes up in discussion of NLP is latent semantic indexing (LSI), a technique for extracting and representing the similarity of words and phrases. The techniques for LSI vary, of course, as with any sophisticated approach to content processing. The core of an LSI system is mathematics. Content is represented as values in a matrix. I’ve been reprimanded by some for describing these procedures to creating a big box filled with colored ping pong balls in which the query is converted to a value I visualize as red or blue or green in the box. The system “looks for” ping pong balls the same color; that is, having a similar score.

Regardless of the aptness of this metaphor, algorithms create and populate a dimensional space with values. The values represent the information in the document. The similarities between vectors for words and contexts allow the system to find matches or relevant items.

In addition to indexing the key words a document contains, LSI systems examine the document collection as a whole to see which other documents contain some of those same words. When documents contain words or concepts in common, the documents

are considered semantically close and, therefore, more likely to be relevant. It is easy to see how LSI generates suggestions for documents that may be useful to the user.

One “gotcha” in LSI is that queries must be processed in the same way documents are processed. From a practical point of view, you want to make certain you have the computational horsepower necessary to perform these calculations. Because LSI is mathematical, its techniques can be applied to a number of information processing functions, from translation to a more mundane search for a purchase order.

Transformation

The word transformation means the process of changing a document in one format into a different format. Some vendors use these terms as synonyms for transformation: normalization, conversion, or data integration.

With regards to content processing, transformation is a little-understood aspect of a behind-the-firewall system. It is an important function, and understanding your specific transformation needs is a task best undertaken when you first decide to enhance your search system.

The different types of files in most organizations are like a fruit salad. The files can be quite different. Therefore, the transformation or file conversion process can be more difficult than you expect. For example, an Adobe PDF (Portable Document Format) can convert if it is not protected. If it does require a password to open, you will have to insert a work step to handle the file. Unprotected PDF files can be deceiving. Some are image files wrapped with information to display the content. There may be no text in these files. Alternatively, you may encounter a PDF with a text, image, and machine instruction components.

The costs of file conversion and transformation can swell under these circumstances:

- You encounter variants of a standard format that your content processing system does not support.
- A vendor introduces a new file type such as Microsoft's DOCX and your content processing filter does not have a filter or code widget to handle this type of file. The fix may be to convert DOCX to a DOC or RTF file and then process those files.
- A proprietary system may not permit an import filter to extract by field. The solution may be to write a query, generate reports in XML containing the needed data, and convert those reports using your transformation system.

Structured Data

You will have data and information residing in database tables or possibly XML. These are usually described as structured files. In my experience, I've encountered structured files that require special handling. Not long ago, a Chamber of Commerce was unable to open files created in two database systems. One was Reflex and the other was Alpha. The Reflex data had to be retyped from a paper version of the information. The Alpha

files required a little hunting to locate a version of the program and export the data in a form that the search system could understand. An insurance company had tapes containing flat files generated from IBM's CICS system – no problem other than the fact the insurance company no longer had the IBM mainframe.

Structured data usually comprises about 20 percent of an organization's total information. The cost of transforming these data into a form understandable to a search or content processing system is tedious but, in general, not a horrific problem and much of the transformation can be handled programmatically.

A tip is to inventory your structured data file types so you can estimate the time and cost of the transformation project. You will probably be able to do this work without shipping the information to a third-party data conversion shop.

Unstructured Information

Some issues may arise when it comes to unstructured information like e-mail, files produced by desktop applications, or specialized applications such as InDesign, Versions 1 through CS3. In the organizations whose content I have audited, about 70 percent of the total information in digital form is unstructured. (If you wonder where the missing 10 percent is, that information is specialized file types like audio and video files that cannot be processed by most of the systems described in this study.)

The volume of unstructured information is significant, and it seems to be growing at double digits each month. So a typical organization that starts a calendar year with 100 gigabytes of unstructured information will finish the year with 250 gigabytes or more data. However, much of this may become structured data, which is now growing rapidly due to the surge of interest in XML.

Within this large collection of unstructured information are a number of “flavors” of files. If your organization uses Microsoft Office for its word processing, spreadsheet, and presentation work, you will encounter different versions of Word, Excel, and PowerPoint files. But when you encounter files generated by different versions of Adobe PageMaker, InDesign, or Framemaker, you will encounter different format variations. Some of these new variations cannot be opened in older versions of these programs. Adobe PDF (Portable Document Format) files present additional challenges.

Keep in mind that when a content processing system operates on files, it must recognize the file type and then have the appropriate file transformation procedure available. If a file is not recognized, it is kicked out as an exception, noting the file name in the system log. Also, when a file is password protected, it may trigger an exception error.

A routine check of exception files should take place on a scheduled basis. The exception file is your finger on the pulse of the content processing function.

Dashboards

One of the better-known pioneers in LSI is Ramana Rao, formerly at Xerox PARC and later at Inxight Software (now a unit of Business Objects), who told me several years ago:

No one has to teach a human to recognize the glint from a tooth in a dark hedge. Our ancestors gave us the ability to spot important information very, very quickly. Reading is orders of magnitude slower than this innate ability to spot what's significant.⁴

The notion of heads-up displays, and visual cues, takes advantage of a human's ability to process visual information quickly. The phrase *dashboards*, at a glance, applies to these interfaces. Remember that the snazzy graphics work only if the underlying information is in a form that can be crunched, massaged, and manipulated by scripts and algorithms.

The trend to supplement the search box with hot links, recommendations, and suggestions is becoming more pervasive. But users can tire of overly busy or too clever interfaces. The functionality is what matters once the first visual impression has been absorbed. If you don't have a point-and-click interface as an option for your behind-the-firewall search system, you may want to anticipate user demand and explore how one of the companies profiled in this report can give you that functionality. As more content is tagged and indexed for concepts, entities, relationships, and other value-added elements, the more practical dashboard interfaces become. Slapping a flashy interface on a key word indexing system won't work. The effect is similar to putting a \$1,000 paint job on a jalopy. The underlying vehicle has to match the exterior.

⁴ The statement was made at dinner in London, England. Dr. Rao and I were speaking at the International Online Show there. I think of him as the inventor of the interactive hyperbolic map.



Figure 9: Conoco's Dashboard Display

This is a dashboard display implemented at Conoco. © Luxid, 2007.

Payoffs and Liabilities of Rich Text Processing

The search box won't disappear anytime soon. For some types of queries, a key word search works remarkably well. For example, if you know there is a consulting services firm that your company has on retainer called Troshkova & Associates, you can enter Troshkova in a search box and generate a list of documents quickly. However, if you don't remember the name of the firm, then you will want other ways to identify this company.

Furthermore, when the volume of content processed by the system grows, key word queries can return too many results for the user to examine. Most users of search systems have difficulty formulating complex queries using Boolean AND, OR, and NOT syntax. The special switches or commands also give most system users headaches. When the person looking for information has limited time or is under pressure, a search box can be particularly frustrating. Users report that entering key words is a guessing game where the user tries to figure out the magic words to use to get the system to provide the answer.

Search systems, therefore, must adapt to these user behaviors. Consequently, point-and-click interfaces, with many different ways to access, find, discover, explore, or find information, are here to stay.

The Upside

The benefits of a next-generation, beyond search, system include:

- Giving users more ways to locate information. Instead of a “naked” search box and laundry lists of results, users can browse categories, explore suggestions for related information, or even look at results in a graphic display. In short, the metadata makes it possible to expose information in useful, different ways. The payoff can be measured directly by system usage or indirectly by measuring time saved.
- Existing content indexed only by key words lacks dimension. Rich text processing, in contrast, adds additional handholds for users and other computer processes. Documents that share information about a particular person can be easily related to others by or about that person. Documents that have been indexed as belonging to a particular category can be sliced and diced by time, geographic location, and other factors. Software is not yet able to perform like a human, but it can uncover nuances, easily overlooked details, and make some connections among items of information a busy professional may overlook.
- The beyond-search systems can put different types and sources of information in a single display, sometimes called a dashboard. The idea is that a user may need information from different enterprise systems, plus the Internet, and from servers running behind the organization’s firewall. Rich text processing systems have functions that make it possible to pull information from many different places from a single interface. One of the more interesting dashboards, Figure #, is used at Conoco. Note that the user can click on categories and drag controls to obtain information from the system. Too futuristic? Perhaps. But the search box may go the way of the buggy whip in some organizations.

Some Cautionary Considerations

Rich text processing is not without its downsides. Keep in mind that key word search is a complicated, resource-intensive function. Rich text processing is additive; that is, key word indexing still exists. Words are stemmed or reduced to their root by discarding inflections like -ing and -ed. Rich text processing, therefore, is also a complicated suite of systems and subprocesses that inter-operate with key word indexing operations. Never forget that rich text processing is complicated and requires adequate hardware, storage, and bandwidth.

Other issues to consider include:

- Staff. Rich text processing systems that take key word systems into new frontiers require care and feeding by trained professionals. You can get the technical help you need in a variety of different ways. You can obtain professional support from the vendor. You can hire additional staff. You can train existing staff to manage the system. You can hire independent consultants. Regardless of how you add staff, you will need more trained hands to configure, maintain, and operate the rich text processing system.

- **Hardware and infrastructure.** Rich text processing is additive. If you have existing servers, you will need to make certain that you have the computational and storage capacity to handle the outputs of the rich text processing system. Some of the newer systems build a metadata repository and maintain source documents in a dedicated storage subsystem. When different chunks of content are required from different sources, pulling the source document over the in-house network can create bottlenecks that interfere with other processes. Some rich text processing systems scale gracefully. That's a plus. Others don't. In either case, you cannot quickly get new hardware and other pieces of infrastructure and have them up and running in a day or two. Advance planning is essential, and most organizations lack a core competency in infrastructure related to search and rich text processing. Your information technology team must be confident that the system issues are well in hand and be competent to manage them. In reality, inadequate or flawed infrastructure engineering is one of the two leading sources of headaches with search and rich text processing.
- **Dealing with problem content** requires additional workflow processes and often additional staff. Content transformation can chew up a significant portion of a search budget. Information can be structured; that is, organized in a traditional database or tagged as well-formed Extensible Markup Language using standardized document type definitions. Information can be unstructured; that is, lacking structure, having an inconsistent structure, or partially structured. Rich text processing systems can be particular about content file types. Customized filters often are needed or the default filters may require customization. If a content transformation bottleneck occurs, the system will not be comprehensive or contain current data.
- **Costs.** Rich text processing systems can be costly. Consider that key word retrieval has become a commodity. In fact, you can do basic search with the open source Lucene engine and derive 90 percent of the functionality of a commercial product for almost no direct cost. *Beyond search* is a premium function, and some of the vendors are reluctant to provide a retail price for a one-year license. Your cost analysis must cope with many unknowns. These range from the exact hardware you need to start up and then handle content growth over the next 12 months to estimating the cost of system customization. When you venture beyond search, you will be entering a territory with "cost unknown" signs at key decision points.

The Problem of Language, Any Language

Figuring out what a statement "means" is not easy. Humans have a difficult time understanding some people who are close to them. Steven Pinker, author of "The Stuff of Thought," goes into great detail about the problems of language. Academics logic-chop about the "semantics" of language, the "ambiguity" of an author's work, and the "meaning" of a particular sentence. President William Jefferson Clinton confounded me with his discussion of *what is-is*.

If humans can't figure out some meanings, how well do you think a series of computer instructions will do? Computer scientists and other specialists have made remarkable

progress in letting numerical recipes analyze textual content. Some of the techniques are centuries old including some in Autonomy's IDOL system. The newer systems often rely on algorithms and techniques taught in universities for more than a quarter-century.

The progress made by companies profiled in this study is exciting. It is now possible to let a computer "read" content and identify the names of the people, companies, events, and numerical data in documents. These systems can take an e-mail or a PowerPoint presentation and determine that it is about "marketing" or "contracts." Some vendors describe their systems as being able to determine the "aboutness" of not just a document, but a collection of thousands or millions of documents.

Users can look at different representations of the processed content and search it using key words. Anyone who has looked for information about a specific event in a deal, a fact about a business partner, or the e-mail that added an all-important clause to a contract knows one thing – today's systems often leave the user to figure out what he or she needs.

To get around these known problems of language, vendors have taken full advantage of the blazing performance of today's central processing units (CPUs), low cost storage, and an enormous wealth of research into indexing, advanced mathematics, statistical processes, and the digital reference materials about word roots, term occurrence, and grammar.

Dig into a modern search system, and you will find that it includes dozens of different approaches to figuring out what a document is about. In a sense, today's content processing systems are similar. That similarity gives analysts and procurement teams headaches. Because until you have installed a system and processed content you know well, you cannot pinpoint the exact differences among search and text analytics/text mining systems. A search box is a search box. A list of hot links that you can click to "explore" content look similar from vendor to vendor.

You also don't know if the system will "run" on your infrastructure. What works perfectly on a test system may display surprising behaviors when made available to several thousand employees. Surprises range from results that have little or no relevance to the query to results that have the "answer," but it's buried at the bottom of a long list of items. Other systems update slowly or erratically, so users can't find documents they know are in the system. Other systems seem to be working, but corrupt the index, causing the search administrator to have to re-index or restore the previous index. The inventory of "gotchas" can be extended indefinitely.

As you read through this high-level overview of next-generation content processing systems, pay particular attention to the profiles of more than 20 companies who are at the forefront of addressing the difficult problems in search, information access, and content processing. You are learning about the future of search. Key word queries will continue to play an important part in information retrieval. However, users need and demand other ways to get at needed information.

Progress is rapid, but language itself guarantees that a great deal of work remains. Buzzwords zip around sales presentations like angry hornets. The scientific-sounding terms usually mask the weaknesses in systems available today. In the next decade, finding information will be improved. For now, newer systems offer organizations a compelling reason to embrace newer techniques or shift to a different search and content processing system.

But if I still misunderstand my wife after 38 years of marriage, you understand why computer scientists have a long, difficult journey ahead of them.

Market Context

You may have some immediate questions about advanced content processing solutions already incorporated into behind-the-firewall vendors with a high profile and hundreds, even thousands of customers. I don't want to retrace the information in the *Enterprise Search Report*. I played almost no part in the 2007 fourth edition, yet the information about IBM, Microsoft, and 16 other vendors is complete, so from my vantage point, use *ESR* for comprehensive discussions of these vendors' systems.

I will touch upon a small number of vendors for the purpose of illustrating how more sophisticated content processing is finding its way into general purpose behind-the-firewall systems. Some of the interest is due to the use of Web services to tap additional functionality and services. Web services is an umbrella term including techniques for hooking different functions together from different systems. There are many approaches to the use of Web technology available to the *Big Dog*, superplatform vendors, and each has embraced Web services in order to get the benefits of the technologies. For example, IBM has embraced Web services and introduced a standard called UIMA, an acronym for Unstructured Information Management Architecture. Microsoft, on the other hand, has enhanced its Dot Net framework and literally most of the key touch points of its server products to exploit Web services. As you know, both of these giants talk about "easy integration" and "standards." You also know that once you embrace one company's architecture, you discover that some technical sticky pads exist to keep you firmly in each vendor's camp.

A number of other interesting developments are taking place as I write this section of the study. Accordingly, I have provided some broad perspective on a number of different firms. Some of these companies' technologies are mentioned only briefly. Other companies are profiled in the in-depth discussions that appear elsewhere in this study. Rather than succumbing to the appeal of creating an encyclopedia with profiles on large numbers of companies I want to give you a sense of the options available for replacing or enhancing a behind-the-firewall search.

The plan for this section is to:

- Describe three superplatforms and their approaches. The companies discussed briefly are IBM, Microsoft (Fast Search & Transfer), and Oracle.
- Talk briefly about the Autonomy and Endeca systems with some references to Fast Search & Transfer, which I will refer to as the "Big Three." Despite Microsoft's purchase of Fast Search & Transfer, it will be business as usual for most of 2008. I will comment on the options Microsoft must consider as it embraces the Fast Enterprise Search Platform.
- Discuss briefly four up-and-coming vendors. These are Coveo, Exalead, ISYS Search Software, and Siderean Software. Note that this list could have been extended easily, but I exercised editorial judgment based on geography: Coveo is Canadian, Exalead is French, ISYS Search Software is Australian, and Siderean

Software is American. My goal is to make clear the internationalization of content processing technology.

- I want to offer some thoughts about what comes next in behind-the-firewall search and content processing.

The Superplatforms

Search and content processing are secondary functions to these companies. In their overall revenue mix, none breaks out sales of search-centric applications. When one looks closely at each company's offerings, each has a large number of products and services. A Fortune 1000 company can license search and content processing from any of these companies, and directly or as part of another enterprise product, deploy a usable, reliable search solution. The brief snapshots of these companies is intended to provide stage dressing for the new technology presented in the profiles elsewhere in this study. More detailed discussions of the search and content processing technology of each of these firms appears in the first, second, and third editions of the *Enterprise Search Report*. As of January 2008, I have not seen the fourth edition of my search encyclopedia. I do know that the three editions I wrote provide extensive detail on the inner workings and functionality of information retrieval services provided by IBM, Microsoft, and Oracle.

IBM

IBM is somewhat more advanced in search and content processing than either Microsoft or Oracle. You may express surprise at this statement. IBM is not perceived as a company offering a product comparable to those available from Autonomy, Endeca, or Fast Search. The error is easy to make. IBM resells solutions that can incorporate the search technology of Autonomy, Endeca, or Fast Search. IBM also has its own Lucene-based Omnifind solution. IBM owns the iPhrase content processing system. IBM has deals with companies as small as X1 to as well-known as Google. To cap off IBM's search and content processing array, the company introduced a software layer that any search vendor can use to plug into an IBM platform. UIMA, or unstructured information management architecture, is a specification and a framework. Any search or content processing vendor can use UIMA to become compatible with IBM solutions. IBM's laboratories remain hot houses for search innovation. Google, Microsoft, and Yahoo! have search and content processing experts who have worked at one of IBM's research facilities. Google's programmable search engine is arguably an invention that benefited from IBM's commitment to information retrieval research and development. IBM's business model is to make it easy for IBM solution experts to assemble an information solution that meshes seamlessly with IBM's hardware, software, services, and database. Remember, IBM is a service business that sells hardware. Software to IBM is a catalyst for consulting revenue and for sales of IBM's servers, storage devices, and other products. At this time, IBM is content to cooperate with the Big Three and other search vendors. However, at any time, IBM can exert tremendous pressure by bundling search with other applications much as it does with Lotus Notes and IBM servers. Alternatively, IBM could exclude companies from the IBM family, thus reducing revenues. IBM has a new partnership with Google, and it is

difficult to anticipate how this deal will impact the search and content processing sector in 2008, if it survives at all. IBM is a consulting firm more than an information technology firm. If you embrace Big Blue, then you can buy almost any function from IBM-certified vendors. I personally rely on branded IBM servers, but the company is shifting more manufacturing to Lenovo. IBM can surprise its customers and competitors. The company could shake up the behind-the-firewall search market. So far, its tie up with Yahoo! has not created any tsunamis. Companies don't buy IBM for its content processing technology. Customers select IBM because "nobody ever got fired for buying IBM." The statement was true in 1960, and it is true today.

Microsoft (Fast Search & Transfer)

Microsoft is a bit of a mystery. The company operates search in a number of different markets, and these different search initiatives are not yet tightly integrated. For enterprise customers, Microsoft offers its Microsoft Office SharePoint Server (MOSS) search solution. The system requires Windows Server, SQLServer, and a number of other Microsoft components. A 100-percent Microsoft organization will have the expertise needed to configure, customize, and tune the MOSS solution. When properly resourced with hardware and expertise, the system delivers key word searching plus some metadata-based operations such as sorting documents by file type, creator, and time. Like IBM, Microsoft has an active research and development program for text retrieval. Microsoft also has a number of Microsoft Certified Gold partners who have developed search and content processing solutions that are compatible with Microsoft's framework, its VisualStudio.Net programming tool, and the most recent innovations such as Silverlight, an interface design tool. SharePoint is a content management and collaboration platform, so search is a utility function for it. Microsoft talks about search, but it has been supportive of companies such as Coveo, dtSearch, Mondosoft, and others for key word search technology. Microsoft has also encouraged companies like Interse in Copenhagen, Denmark. The Interse technology provides advanced content processing for SharePoint installations. Microsoft's enterprise play appears to be focused on high-value applications built on the Dot Net technology and Microsoft's quasi-proprietary Web services. Microsoft has been spotted talking with each of the Big Three and tracking innovations from smaller, more entrepreneurial content processing companies. With its billions in cash, Microsoft can acquire a company like Autonomy or Endeca, gobble up specialist vendors, and roll out its own search systems with SQLServer, its forthcoming customer support solution, or any other of its products. Taking all actions simultaneously would not make much of a dent in its pile of cash.

With the acquisition of Fast Search & Transfer SA for \$1.2 billion, Microsoft has great expectations for behind-the-firewall search. The principal assets of Fast Search include:

- **Customers.** The company has more than 2,000 organizations using its ESP (Enterprise Search Platform). Customers range from small Web sites to Yahoo. With the acquisition, Microsoft brings its surging server market, its desktop hegemony, and its go-through-barriers approach to sales.
- **Engineers.** Fast Search has suffered some financial bruises, but it does have as many as 200 engineers, maybe more. These wizards can give Microsoft some

useful content processing know how and provide additional hands to handle opportunities in content processing.

- **Technology.** Fast Search is rooted in Linux. So for some period of time, blending the two companies' core innovations will take time. Based on my experience with both Fast Search and Microsoft, deciding what to do with the each company's specific innovations will be more difficult than hooking the systems together.

The outlook for Fast Search customers is the status quo. The outlook for Microsoft customers is more options. It is too soon to begin thinking about whether this deal will materially impact an organization's approach to content processing. My advice is "Wait until there is more specific information. Go on about your business because consummating the marriage of people and technology is going to take time, maybe years."

Oracle

Oracle is a very interesting player in search and content processing. The Oracle database and Oracle applications are prevalent in more than 75 percent of the Fortune 1000 and leading organizations across most vertical markets from banking to pharmaceuticals. The company's flagship database – now a robust data management environment – comes with a built-in search function. The company has an advanced search and text processing system that consists of technology obtained via acquisition (Applied Linguistics and Triple Hop, for instance) supplemented with its home-grown code. What's interesting about Oracle is that coincident with the company's announcement of its Secure Enterprise Search system, the Oracle Applications unit announced a deal to resell the Google Search Appliance. Oracle's management decided that betting on two horses was better than betting on one in the search system derby. Like IBM and Microsoft, Oracle has a flotilla of partners offering search and content processing solutions. Its PeopleSoft and Siebel units include search with their systems, and each of these acquired companies has deals with other search system vendors who have plug-and-play content processing tools for these enterprise applications. It's not clear how Oracle will move forward in search. Its search management team is publicly quite confident about the company's ability to compete successfully against IBM and Microsoft. To hedge its bets in middle market opportunities, Oracle has the lower-cost Google option available as well as solution from its partners. I've heard rumblings of increasing agitation with Oracle's share of the behind-the-firewall content processing business. I'm keeping my eyes open for indicators of an acquisition or some staff shake ups in the search unit at Oracle.

Business Implications

To step back, several observations are warranted.

First, the superplatforms seem to be developing in-house solutions, working with partners who develop "certified" search and content processing add-ins to the base systems, and either buying or looking to acquire promising search technology vendors. Only Microsoft has not partnered with Google. Otherwise, these three companies seem

to be following generally similar strategies. As these firms look for new markets to conquer, the customer bases of the Big Three (Autonomy, Endeca, and Fast Search) seem particularly attractive. The superplatforms are not losing sleep worrying about The Big Three forcing a superplatform out of a Fortune 1000 account. The superplatforms are more likely to eye the customers of the Big Three. The technology may have some use within the IBM, Microsoft, or Oracle product mix. The value of one or more of the Big Three is their established customer accounts. A superplatform can take this initial relationship and make an attempt to sell more, thus eliminating the time and cost of a traditional sales cycle.

The benefits of working with a superplatform are well known. In fact, most organizations with more than \$250 million in revenues probably have a relationship with these superplatforms. The reason is that the superplatforms can definitely make a system work. Superplatforms are known commodities to investors and shareholders. Finally, the superplatforms are not going to go out of business, orphaning an exotic technology that no one but the entrepreneur who coded the system knows how to make work.

The downside of working with a superplatform is that like a supertanker, the relationship has momentum. Management cannot easily stop the superplatform, change its direction, or refurbish the system. Everything is a process, and the customer is buying into the implicitly understanding that “IBM, Microsoft, or Oracle” knows best. The technology on offer is not cutting edge, but it either works or can be made to work. A company standardizing on IBM servers and the Oracle database accepts a certain cost base in exchange for the benefits delivered by affiliating with the superplatforms. The truism applies today as it did in decades ago. No one gets fired for buying IBM (or Microsoft or Oracle, for that matter).

With staff moving from one superplatform to another, there is not significant technical difference in what the companies deliver. The difference is cultural. IBM is stodgy. Microsoft is combative. Oracle is aggressive. Put representatives of each firm in a room, and you would have a difficult time figuring out who worked for whom. These organizations pose a significant threat to the Big Three because the “big three” are tiny in comparison with the billions in revenue, technical resources, and market clout IBM, Microsoft, and Oracle have. When search vendors talk about a platform, the superplatforms are essentially indifferent to these protestations. A platform is a platform. A superplatform is a multibillion-dollar-a year operation with enormous influence, prestige, and power.

There is no such thing as “enterprise search.” The phrase is one of those marketing buzzwords that became widely used and rarely considered. Autonomy, Endeca, and Fast Search & Transfer have a number of similarities. Let’s run through the major ones.

Autonomy, Endeca, and Fast Search (The Big Three)

Each of these companies offers key word search and retrieval. Their software is not designed to be installed or maintained by the licensee. Each of the companies derives revenue from maintenance, customization, and technical support. Each of the

companies reports generally positive financial news. Autonomy and Fast Search are publicly traded on stock markets that operate outside the United States. Endeca remains a privately held company, but chatter about the company going public or selling to a larger firm fuels gossip-mongers. More significantly, each company talks about its approach in terms of a platform or a framework. The idea is that these organizations offer more than search of structured and unstructured information. Their respective technologies allow a licensee to build information-centric applications.

Furthermore, these companies share a number of customers. A good example is IBM, the former arbiter of computing, that inks deals with numerous search and content processing companies. IBM is now a consulting and services firm, having a wide range of choices for its customers helps drive consulting business. The Big Three also share numerous U.S. government customers. These range from wild and woolly General Services Administration to the technical enthusiasts in America's intelligence, law enforcement, and defense entities. Information, after all, is the key to 21st Century war fighting. Fortune 1000 firms often have all three Big Three search systems in operation. Acquisitions explain some of the overlap. But other enterprise software vendors like BEA Systems include a version of Autonomy in their software. A large company, therefore, has a Big Three engine, but may not think of it as a separate installation. Finally, because of the general dissatisfaction with search, many organizations look for greener pastures. The incumbent search engine remains in operation, and the most recent search system is positioned as the solution to search challenges. Not surprisingly, the sales presentations made by each of the Big Three often echo one another. Each Big Three system requires dedicated servers, storage, and personnel. Each Big Three system can be extended by a knowledgeable programmer almost infinitely. Each Big Three system can integrate, replicate, and emulate any other information service available.

With these three companies generating collectively about \$600 million a year in gross revenue, it's clear that none of the Big Three is performing like a Cisco, Google, or Microsoft. In fact, the collective revenues generated by the Big Three underscores the revenue "glass ceiling" that holds down the search sector. Wall Street mavens ask, "If information is such a hot sector, why are Autonomy, Endeca, and Fast Search & Transfer not growing faster and spinning off more profits? Why does Autonomy restate its finances? Why does Endeca keep pulling back from an initial public offering? Why does Fast Search & Transfer continue to get tangled in financial tar pits?" More problematic, a procurement team often finishes a series of presentations by each of the Big Three asking, "What exactly is the difference between and among these vendors?"

What Are the Differences?

As you might expect, on the surface, these companies have much in common. Get some hands on experience, and you find that the systems are indeed quite different. Unfortunately, the significant differences are deep in the engineering "guts" of each system. Let's look at some major distinctions and relegate finer points to the table below:

Vendor	Distinguishing Features
Autonomy	Sales oriented. Aggressive, Bayesian technology, growth via acquisitions. Asserts that its various technologies are integrated. Has thousands of licensees. Complex. Interesting CEO who is a knight and a fish lover. Wants to be Number One in search.
Endeca	MBA-centric with great positioning: “faceted navigation.” Sells to top management. Conducts studies. Proven track record with structured data for e-commerce and its work flow approach behind the firewall. Gentlemanly and very bright.
Fast Search	Technology-centric. Some Google-like aspects in its technical approach. System consists of original code, open-source software, licensed technology, and acquired companies’ technology. Low-key engineering style.

Table 3: The Most Significant Differences

Technical Foundations

First, the companies have different technical foundations. Autonomy is based on Bayesian statistics. The core idea is that the system processes content, generates a wide range of outputs, and uses these outputs to determine what the information is “about.” The core Autonomy technology is owned by Cambridge Neurodynamics, and Autonomy licenses a “black box” of algorithms from them. At the very heart of Autonomy are proprietary algorithms that allow the system to operate automatically, hence the name of the company. These algorithms are extremely flexible. For example, language does not make a difference to the algorithms. Autonomy’s content processing works as well on English as it does any other language. In addition, Autonomy’s algorithms are speedy. A licensee who runs the engine in “automatic” mode, feeding it sample documents to “teach” the algorithm thresholds and value, can crunch a large volume of content quickly. Autonomy has used its “black box” of mathematics to provide such services as fraud detection, indexing of video content, and identifying problems in call center traffic. Over the years, Autonomy has wrapped its “black box” with a wide range of software functions. Today, the system makes it possible for a licensee to use knowledge bases, process structured data in Oracle databases, and integrate Autonomy’s “integrated data operating layer” into other enterprise applications. The foundation of Autonomy, therefore, is mathematical algorithms, an approach that is very similar to that taken by Google.

Endeca relies on its IAP, or information access platform, to deliver its guided navigation services. Endeca handles unstructured information, but it has a strong structured data capability. Endeca also has tools to link information to specific work activities and embed information retrieval into employee workflow. To create a compelling value proposition, Endeca’s founder Steve Papa realized that most people using search systems were not comfortable creating queries. Endeca, therefore, used the metadata/index to create what the company labeled “faceted navigation.” To emphasize Endeca’s business-like approach to search and retrieval, the company hired professionals with technical savvy and degrees in business administration. The mixture of guided navigation and a solid business case that speaks to prospects about efficiency,

cost control, increased productivity, and ROI (return on investment) has allowed Endeca to generate upwards of \$90 million in revenue in calendar year 2007. More importantly, Endeca talks business benefits first to the prospect's non-technical managers. From its inception, Endeca worked to tap into an organization's information residing in other applications and databases such as DB2, Oracle, and SQLServer. The foundation of Endeca is handling structured data and hooking those data into actual business processes.

Fast Search & Transfer has its roots in indexing Web content. John Lervik's decision to sell its Web indexing and advertising businesses to Overture, which was then acquired by Yahoo! in 2003, was a turning point for the company. The decision marked Dr. Lervik's vision that Fast Search could generate more revenue in the enterprise sector than in Web indexing and advertising. In the period between 2003 and 2007, Google grew to reach \$15 billion in revenues from a base of \$100 million. Fast Search has reached \$200 million, a difference that underscores the challenges facing enterprise search vendors from Wall Street's point of view. In the last four years, Fast Search has wrapped its Web indexing engine with ESP, an enterprise search platform layer of software. The company has acquired, licensed, and created functions that allow a licensee to handle almost any type of enterprise content processing job. Fast Search has an impressive customer list, a remarkable annual trade show for its licensees and partners, and a range of specialized versions of the Fast Search system that covers most vertical markets, horizontal applications, and advanced processing functions such as clustering, automatic classification, and presentation-ready report generation. The foundation of Fast Search is Web indexing with other functionality layered on or wrapped around the core engine.

What do these root differences mean to an organization wanting to process behind-the-firewall content? Obviously, any one of these systems can do basic indexing. Each can handle structured and unstructured content, Internet content, and third-party content from providers such as Factiva. At a more discriminating level, my work leads me to make these observations about the technical differences among the Big Three:

Autonomy is well-suited for content processing where the domain is narrowly defined such as medical, pharmaceutical, and health data. When the content covers numerous and diverse topics, the licensee will not be able to take full advantage of the automatic operations at the core of IDOL. Autonomy does well when one of its ready-to-install components is exactly what the customer needs; for example, its fraud detection component.

Endeca is perfect for applications related to retail, scripted sales or customer support applications and for processing specific collections of content. Structured data is a core competency of Endeca. Guided navigation makes it possible for Endeca to deploy a solution quickly, so a licensee's employees can start finding information quickly. Endeca can consume significant computational resources when processing large flows of content.

Fast Search & Transfer is ideal for indexing Web-centric content. The system scales well and when a licensee wants a hosted or managed service, Fast Search's enterprise search

platform can handle almost any indexing chores without a hitch. When a licensee wants to customize the Fast Search system, the blend of open source, third-party, and custom code can be difficult and expensive to tailor. At peak periods, some licensees find that Fast Search must delay some for-fee work due to a lack of qualified engineers.

In summary, the technical differences are largely neutralized because each company has added functions, adaptors, and vertical builds of their search-and-retrieval system. At some point, one or more of the Big Three will be bought. Autonomy acquired search vendor Verity in December 2006, and on the strength of Autonomy's diversification outside of the narrow confines of search, has emerged as the leader of the Big Three. Although Fast is being acquired by Microsoft at this writing, its position among the Big Three will remain unchanged for at least twelve months.

Scaling and Extending the System

Scaling and extending the Big Three systems is a critical concern. All three organizations have embraced Web services, and there are some points about each that you will want to keep in mind. Because each vendor continuously makes changes to their search system, you will have to update the information provided in this abbreviated discussion.

Autonomy's core design dates from the late 1990s, and it was purpose-built around the Bayesian black box of algorithms. Expanding the core IDOL platform, therefore, depends on the hardware you use to run the system. When you try to boost the functionality of secondary systems, you will find that the engineering requires careful planning. The disadvantage of the Autonomy system is that it has grown in complexity through acquisition of Verity and by Autonomy's efforts to "hook" in video, audio, and other advanced functions. Hot spots in processing and throughput bottlenecks are fixable but the solutions are not as simple as adding a server or throwing more storage into a subsystem.

Endeca's system for delivering guided navigation requires appropriate resources. Licensees report that when content and transaction flows are stable – that is, do not vary widely from hour to hour – Endeca is a model of decorum. When a performance bottleneck surfaces, hardware alone may not resolve the problem. Endeca's internal processing itself may require fine-tuning. In general, Endeca was not engineered to process Google-scale content processing chores. Building a system to handle terabytes per day requires planning, design, and engineering. Time equals costs before knowing what hardware or infrastructure to change.

Fast Search & Transfer's Linux platform scales by snapping in more servers, storage devices, and random access memory. However, Fast Search consists of many different components, software modules, configuration files, and features. Many Fast Search systems are similar to custom installations even though the licensee perceives the system as an off-the-shelf version of the search system. Hot spots, performance bottlenecks, or unexpected system behavior are difficult to diagnose and repair. Fast Search's best engineers are those who have a strong grasp of the underlying code and the interaction of other functions with that underlying code. An engineer trying to learn

by experimenting will find it extremely difficult to scale and extend a Fast Search system.

To sum up, none of these systems are easy to scale, extend, and tune for optimum performance. The situation is due in part to the complexity of the search functions themselves, and in part to the engineering decisions made when the companies were founded almost a decade ago. Keep in mind, the Big Three have been around for almost a decade, a long time in our fast-paced world of ever more powerful hardware.

The downside of these systems is that each is complex. The only way these three systems scale gracefully and can be extended easily is if the licensee matches the right system to each's sweet spot: Autonomy for automatic processing of content in a tightly-constrained domain, Endeca for collections that contain reasonable amounts of unstructured data and gigabytes of structured data, and Fast Search for indexing Web-centric content. Get outside of these core strengths, and you will find escalating costs and interesting technical challenges.

Pricing

After chit-chat about technology and features, the question becomes price. Each of the Big Three takes different approaches to pricing. None of the quote prices are likely to be the final price you or another customer pays. Each of the companies negotiates with a licensee for:

- The annual license fee
- Maintenance
- Support
- Professional services.

If you ask a Big Three customer what their total annual costs are, you will get a number. In my experience, unless the person you ask is the licensee's chief financial officer, you won't get an accurate cost estimate.

Autonomy's first-year licensing fee can hit six figures. Think in the \$300,000 range as a working number. There are some indications that Autonomy's professional services can match or exceed the licensing fee. Some of the high-profile systems can cost a \$1 million or more by the time work has been completed on a new system. Complicating cost estimates is the acquisition of content processing firms by other companies. For example, what will Microsoft charge for a Fast Search & Transfer license? No one knows, and it is not clear what the fees will be for existing Fast Search licensees going forward. The same situation existed for Verity licensees when Autonomy acquired that company several years ago. Prices can vary widely. Grandfathering also takes place when a customer is guaranteed a certain price with immunity for price increases or specific limits on percentage increases in an annual fee. Pricing is and will remain somewhat fluid.

Endeca's pricing is somewhat more difficult to nail down because the company does front-end consulting, customization, and licensing. The company's estimate almost

always includes a factor for the software license, maintenance, support, and professional services. But based on information provided to me by Endeca licensees, Endeca takes a high-end consulting firm approach to its consulting business. One licensee said,

Endeca is similar to a Booz, Allen & Hamilton or a McKinsey. Its focus is on the strategic use of information. We didn't think we needed this type of support. What we found out was that our second-year costs for customizing our installation and addressing performance issues jumped significantly.

Endeca does not provide a price list, but comments from licensees suggest that the first-year fees start at \$500,000 and go up. The recent investments by Intel (\$10 million) and SAP (\$5 million) may be precursors to one of these companies buying Endeca. The smallest of the Big Three, Endeca's annual revenues are in the \$80 to \$90 million range, and the company has postponed its initial public offering because of market conditions. I expect the company to be acquired possibly by SAP or a similar enterprise application vendor in the next nine to 24 months.

Fast Search & Transfer has several different pricing models. For an enterprise client, the company offers a multi-year license that works out to about \$80,000 per year for three years. However, special versions of Fast Search's ESP platform for newspapers, for example, hit \$100,000 or more in first-year license fees. When professional services, training, and maintenance are included, Fast Search's costs are similar to Autonomy's and Endeca's. Fast Search offers a hosted solution. One of Fast Search's top engineers told me in 2006, "We offer a hosted and managed option. The costs for this depend on the amount of data we process in the data center and the customization the client requires." Based on anecdotal information, hosted annual costs begin in \$36,000 range, but can go up depending on the specific needs of the licensee.

The money paid to any of the Big Three, then, depends on your requirements. The less support and customization you need, the less you will pay. The more hand-holding required, the more costs go up.

Business Outlook

You know that users are increasingly disenchanted with key word search and for several thousand organizations, the disenchantment, or some of it, may be caused by one of these three companies' systems. The Big Three offer functions and interfaces that support personalization, discovery, alerts, and many, many other advanced features. You also know that when trying to select a search system for your organization, you are faced with a difficult task of figuring out which system is best for your needs.

The Big Three find themselves in a strategic box. As you will learn by reading the profiles of the 24 companies discussed in this study, there are some very interesting competitors making sales and gaining attention of sophisticated corporate customers. Any one of the Big Three can partner, license, or buy technology from these newcomers and upstarts. But the entrepreneurs and inventions are likely to keep coming for the foreseeable future. Search is too important and potentially lucrative to ignore. The Big Three, therefore, face pressure from smaller companies or larger companies willing to

attack a market via a smaller unit or subsidiary. A single insect bite won't kill most people. A hundred thousand insects swarming can do the job. The Big Three have to escalate their marketing sizzle, offer options that thwart the newcomers, and keep their business models intact. Not easy.

Autonomy, Endeca, and Fast Search & Transfer also find themselves under pressure from the superplatforms. These are companies whose business is not search or even content processing. These organizations offer 360-degree enterprise solutions. Search, therefore, is just an add-in, an option, an extra. Superplatforms apply top-down pressure. The Big Three assert that each is a platform. The engineers at superplatforms are hyper sensitive to smaller vendors making a sale in one of the super platforms' key accounts. The reality is that the search vendors are no significant threat to IBM, Microsoft, and Oracle, but superplatforms see a potential threat and are responding aggressively to protect their under belly. The diagram below depicts in an overly simplistic way the plight of the Big Three who are "caught between a rock and a hard place":

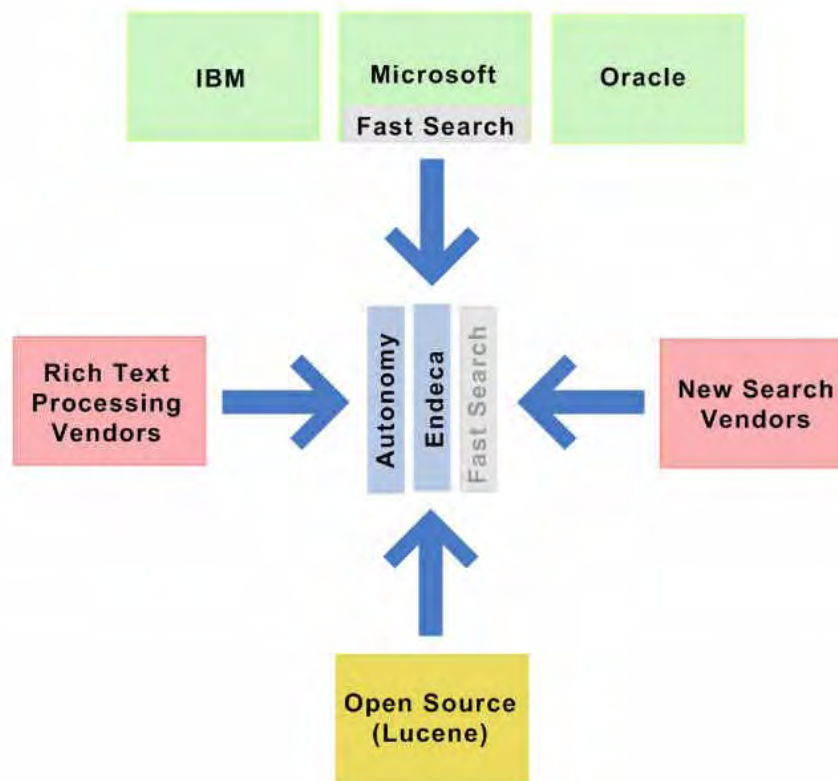


Figure 10: The Squeeze on Autonomy, Endeca and Fast

The middle is proving to be a very competitive place for Autonomy, Endeca, and Fast Search & Transfer with pressure coming from multiple directions at once.

Going forward, these companies' management and engineers face an increasingly problematic market. Customers can look for lower cost approaches or wait until an IBM, Microsoft, or Oracle includes a search function as part of another enterprise application. Microsoft's recent SharePoint includes search and some modest rich text

processing functions but the acquisition of Fast indicates more search options can be anticipated. Some organizations who would have licensed Autonomy, Endeca, or Fast Search solutions may choose to use what IBM, Microsoft, or Oracle provides as part of another, possibly higher value application.

Up-and-Coming Vendors

Before looking briefly at the rich text processing functions of the Big Three and the superplatforms, I want to call attention to a small group of search vendors who are winning customers from the Big Three and crafting deals with large enterprise software companies.

These are up-and-coming search vendors. This grouping could be expanded to include more than 18 companies. Four vendors stand out, and each warrants a brief summary. Again, keep in mind that in the third edition of the *Enterprise Search Report*, considerably longer and more detailed descriptions of these companies are available.

The four companies are:

- Coveo Solutions, Inc., “People with Knowledge Drive Business”
- Exalead, “The Other Search Engine”
- ISYS Search Software, “Enterprise Search Solutions for Real People Doing Business in the Real World”
- Siderean Software, “Navigation for the Digital Universe”

These companies approach search and content processing in different ways. The tag lines tell quite a bit about each company’s approach. For example, Coveo is nudging forward the idea that smart people help a business. Exalead has positioned itself as a Number Two, but it’s not clear whether the company is second in Web search, behind-the-firewall search, or both markets. ISYS Search Software is communicating its pragmatism as opposed to less practical search techniques. Siderean suggests that its system makes it possible to explore vast quantities of digital information.

Each of these companies makes available some type of point-and-click interface or what can be described as, to use a phrase I heard Siderean’s president, Michael Schmitt use, assisted navigation. Each of the companies’ technology classifies and identifies entities. Each of the companies makes an effort to price its system as much as \$150,000 less than a comparable system from larger, better known vendors.

Finally, each of the systems has embraced Web services, provides a customizable interface, and includes numerous content transformation tools; for example, direct support of content in Lotus Notes databases, content management systems such as EMC Documentum, and Windows SharePoint systems, among others.

In my view, as the Autonomy, Endeca, and Fast Search drama unfolds, these four companies may be the principal beneficiaries of the squeeze on their larger, better-known rivals. If consolidation continues, they may well be tomorrow’s dominant vendors of behind-the-enterprise search and content processing systems.

Coveo Solutions

The Coveo system has been a reliable, easy-to-deploy alternative to the search system included in Microsoft Office SharePoint Server and its SharePoint predecessors. In the last year, the company has recoded portions of the Coveo system to make it easy to use Coveo in non-Microsoft environments. The company's system can ingest a range of content, perform on-the-fly classification, generate metadata for documents, and deliver features once the exclusive domain of systems costing four to five times as much in annual license fees.

Coveo is stepping up its marketing efforts and expanding its presence in the United States. Based in Québec, Coveo has been growing and has just announced a major change at the top with an infusion of new capital. Like Mondosoft (a company now owned by SurfRay in Denmark), its search solution provides fast, cost-effective relief from the pains of SharePoint's native search function. Coveo's president is keenly aware of the dissatisfaction with most traditional enterprise search solutions. He told *Beyond Search* in November 2007:

We think there is a significant opportunity in the SharePoint market and in the broader enterprise search market for our solution. It's fast, provides assisted navigation, and performs some advanced functions without requiring a massive computer infrastructure."

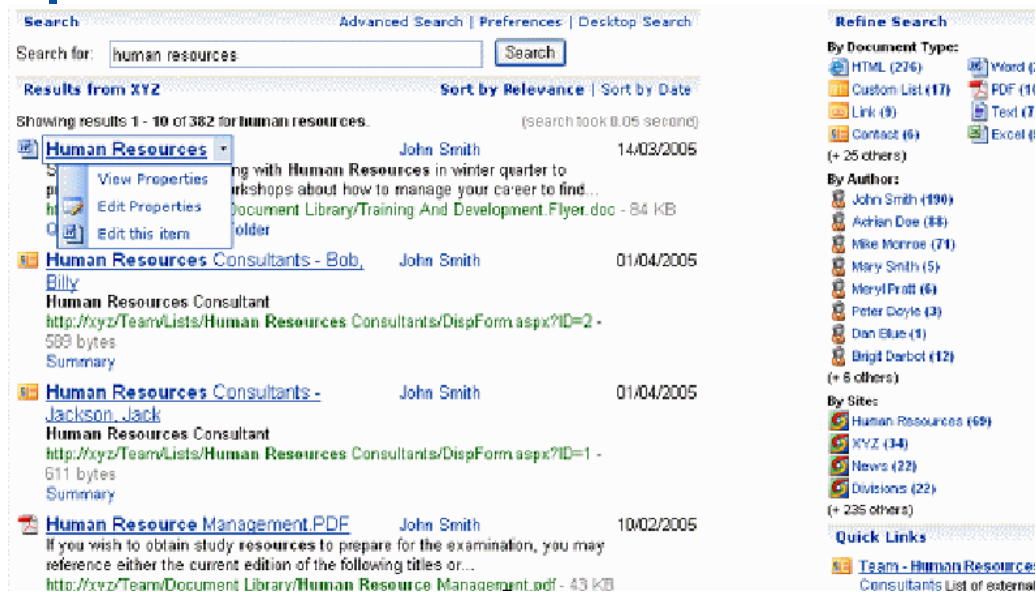


Figure 11: Coveo and SharePoint

Coveo processes SharePoint content and offers easier installation and administration than Microsoft's current search solution. Once the Fast ESP technology is integrated into SharePoint, Coveo may face increased market resistance.

Exalead

Exalead has been a solid performer for numerous customers in France, ranging from financial institutions to government organizations. At one time, America Online in France used Exalead instead of Google for its French language Web search. What few

know is that Exalead shares some DNA with Google. Like Google, Exalead's technical foundation benefited from Digital Equipment's AltaVista.com search. Founder François Bourdoncle set up his own company while some of his AltaVista.com colleagues signed on with Google.

From a performance and scaling viewpoint, Exalead is among the most sophisticated engineering organizations from the group of companies discussed in this study. In terms of the company's enterprise search solution, Exalead offers key word search, manipulates metadata, and provides a robust application programming interface. Enterprise licensees can integrate Exalead into most third-party applications. As content processing volume grows, Exalead's Linux-based architecture scales without headaches. Exalead offers a number of interesting features. These range from assisted navigation to the ability to display a graphic thumbnail of the source document. Exalead offers a number of interface options. The one illustrated below includes a number of the system's rich text processing features.

Exalead, like Fast Search & Transfer, offers a hosted and managed service. Keep in mind that the company's roots are in France, so be prepared to experience a bit of Paris with each Exalead interaction. The company has been growing rapidly and has a growing presence in Europe and the United States.

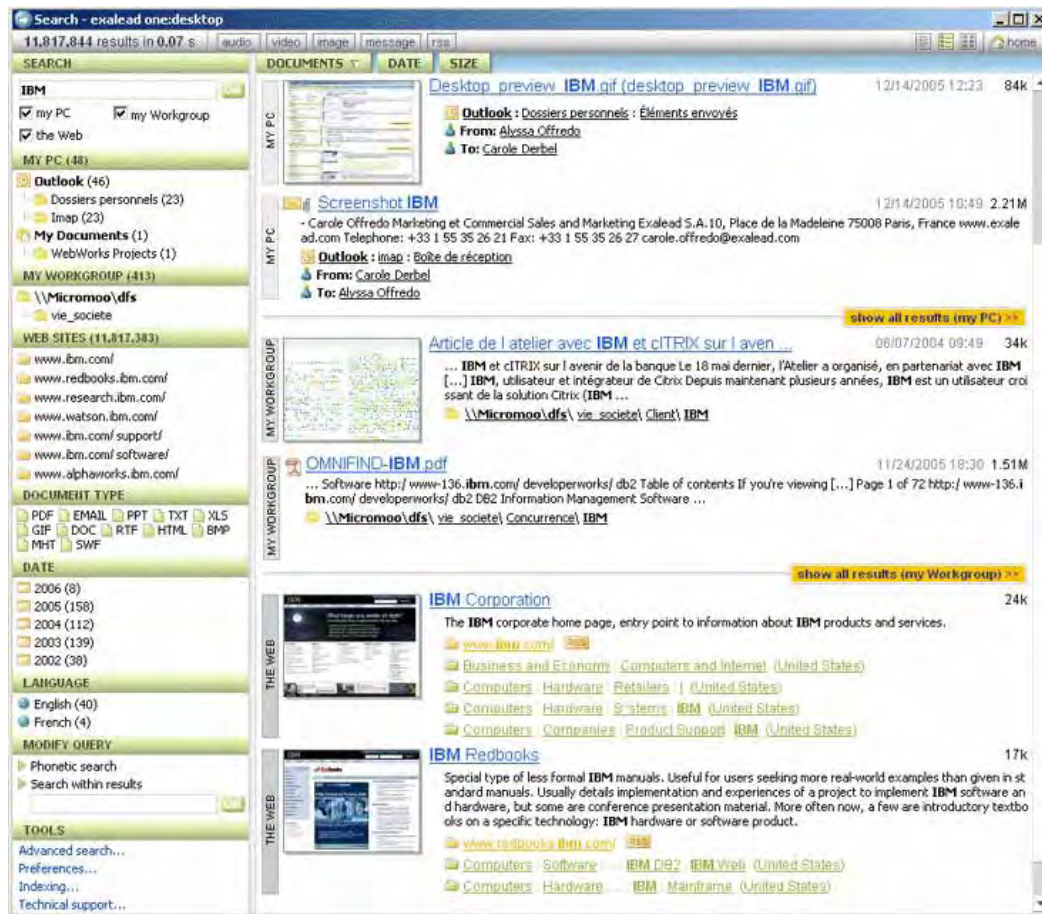


Figure 12: Exalead's Interface

Exalead incorporates point-and-click functions to look at results by file type, data, and other attributes. In addition, the result list features thumbnails of each document.

ISYS Search Software

This Australian company has been growing rapidly. Despite its low profile, ISYS has enjoyed success in law enforcement, legal, and organizations of many types. Version 8 includes a number of enhancements to an already solid system. Document processing is among the speediest I have tested.

The system extracts entities, classifies documents, and supports a basic search box as well as a comprehensive set of search operators available for power users. The current version delivers the assisted navigation similar to other systems that cost orders of magnitude more.

The ISYS system can be integrated into other third-party applications, and the interface can be easily customized. The system includes a software development, adaptors for structured and unstructured data, and support for SharePoint and a number of other third-party applications. The company makes an effort to provide speedy technical support. Development takes place in a suburb of Sydney, Australia. The company has a sales office in the United States.

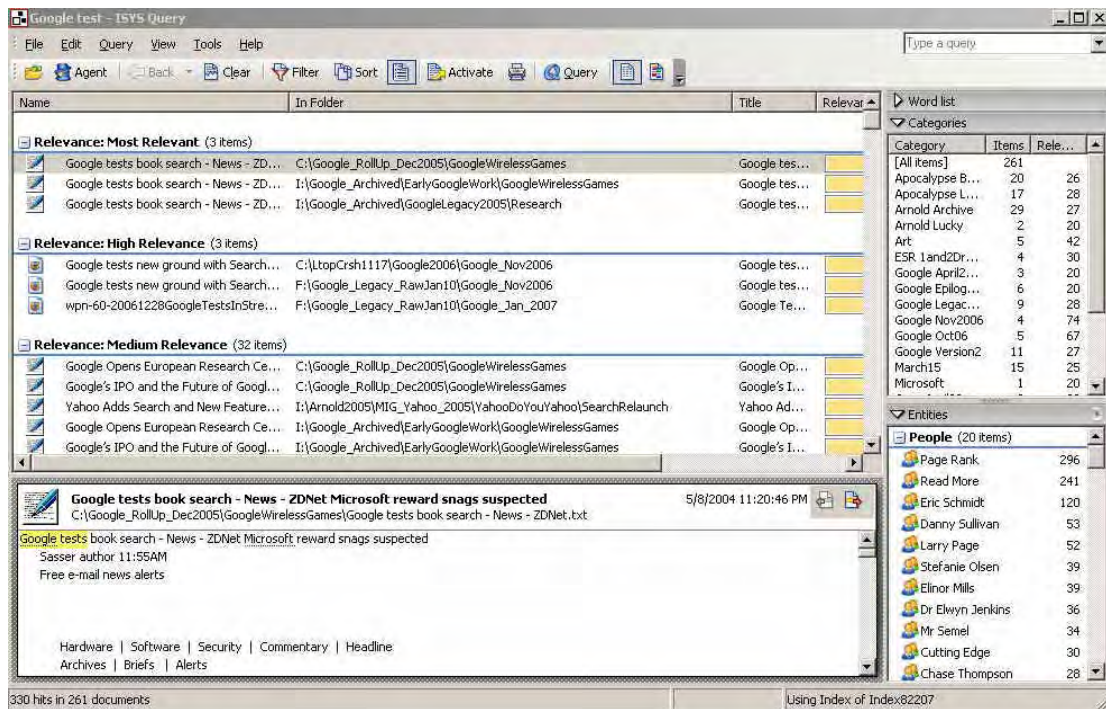


Figure 13: The ISYS Customizable Interface

ISYS panels display search results, point-and-click navigation hot links, and a preview of the documents matching the user's query. The interface is easily customized.

Siderean

Siderean embraces the tenets of the semantic Web in a pragmatic way. According to Bradley Allen, founder and chief technical officer:

We take information and make it easy for a user to see what's available, explore, search, reverse field, then dive into a particular document. We do this without asking the user to know how to formulate a query, worry about the format of the information, or have specific expertise in a subject.

Siderean has landed some major accounts, including Oracle. One of Oracle's implementations of the Siderean technology appears below. Perhaps more notable was Siderean's winning the *Financial Times's* contract. Rumor has it that the high-profile incumbent could not match Siderean's fast-cycle deployment of semantic content processing. After almost a year of frustration with a solution provided by a larger, better-known vendor, Siderean implemented its technology in a few days.

Siderean makes it possible to display information in various graphic formats with suggested content available as assisted navigation links. A search box can be displayed on each page of the interface with sliders or controls that the user can drag with a mouse to alter the information display. Real time news can be filtered to display with results from processed content in structured or unstructured form. Like Exalead and Fast Search & Transfer, Siderean offers a hosted or managed solution, an application programming interface, and a range of engineering support services.

The screenshot displays the Oracle OpenWorld SF event page. At the top, there's a navigation bar with 'Events', 'In-Person', 'Webcasts', and 'Podcasts'. The main header features the 'Oracle OpenWorld SF' logo and the event dates: November 11 - 15, 2007, at the Moscone Center in San Francisco. Below this, there's a 'Find Events' section with a search bar and filters for 'In-Person Events' and 'Web Events'. A date range is set from 2007-11-06 to 2008-02-07. The 'Location' filter is set to 'Europe'. A map of Europe is shown with red pins indicating event locations, each with a number. The map is powered by Google. Below the map, there's a list of results, with the first result being 'Desayuno - Oracle Customer Relationship Management' in Madrid on November 06, 2007. The page also includes a sidebar with 'Most Popular' events, 'Downloads', and 'Related Web Events'.

Figure 14: Siderean Graphic Interface

Siderean's technology extracts entities, classifies, and discovers relationships in a way that is somewhat analogous to Endeca's procedure. Siderean's outputs can be displayed on maps (shown above) or a variety of textual or graphic representations.

More Choices, More Functionality

From among these four companies, you may find a solution that costs less than other systems, delivers more functionality, and demands less babysitting than many other systems.

Organizations have more choices than at any other time in the 40-year history of online search and retrieval. The sophistication of the systems continues to go up. In fact, savvy procurement teams can assemble a search and content processing system to meet the needs of even the most skeptical and demanding user. The reality is that technology is not the problem. The vendor is not the problem. The user is not the problem. The problem in search is the implementation and resourcing of the search system.

Those responsible for procuring a search or content processing system often find that the need to make a decision interferes with thorough fact finding. In our work in the last 15 years, we have yet to encounter a procurement that includes a benchmarked, head-to-head test of the top candidates' search systems. Little wonder that users are frustrated with behind-the-firewall search. Their organization makes a decision without

solid, reliable facts derived from tests of candidate systems on the organization's content. When slowdowns occur, the licensees are quick to blame the search system vendor. There are times when the vendor is responsible for the issue. But in most of the cases with which I am familiar, the search system requires appropriate hardware, infrastructure, and technical support. These are engineering and technical needs. A short cut when the system is installed may not make its presence felt until the volume of content processed crosses a threshold. When the system slows to a crawl or crashes, a quick fix is not likely to work. I am not willing to lay blame on problems with search and content processing at the feet of the vendor. The licensee and the vendor share responsibility. When the search system runs on inadequate infrastructure and is maintained by a person with inadequate technical expertise in the system, the crises in search that many organizations face are created by the organization's budget and management processes.

Lucene

I want to mention Lucene in the context of these up-and-coming vendors. Lucene will continue to put pressure on vendors with inexpensive solutions. If you know how to code, you can install Lucene and customize it. Siderean has used Lucene as a key word search system in some of its installations. A commercial version is available from Tesuji, a vendor with offices in Italy and a technical center in Budapest. Lucene also plays a role in IBM's search solutions. It works, and it is fine for an organization with savvy technical resources. Because it is "free," Lucene exerts bottom-up pressure on the commercial vendors. It's worth a look, and, who knows, it may meet your needs. Keep in mind that you may have to do some "fast dancing" to handle indexes with more than 2 million documents.

Google

I'm not sure if you have noticed, but I have not referenced Google in this battle for the enterprise. Google is an environmental force, and it is operating on a broader time horizon than the companies mentioned. Google has more than 8,000 users of its Google Search Appliance (GSA). Google buys companies, and it continues to act like a gravitational force on search and content processing system vendors. I have written extensively about Google's enterprise "pull" tactic, the Google Search Appliance, and the significant OneBox API. I want to reiterate one point: The OneBox API makes it possible to use the GSA to generate virtually any of the functionality described in this study. For more information, please read my analyses of Google in *The Google Legacy* (Infonortics, 2005), *Google Version 2.0* (Infonortics 2007), and the *Enterprise Search Report* (CMSWatch, 2003-2006).⁵ I also provide some Google information in my *KMWorld* column, which runs each month and in my *Beyond Search* Web log. I have written a separate chapter for this study that talks about Google's leap-frog strategy. Although not released for commercial use, Google is far from abandoning or ignoring

⁵ The two Google studies may be ordered from Infonortics at <http://www.infonortics.com> and the *Enterprise Search Report* may be ordered at <http://www.cmswatch.com/Search/Report>

the content processing needs that exist. Google is coming at the problem from an interesting direction, and you can read about it in the separate Google chapter in this study. Google, based on my research, is a truly significant player, but it is now involved in a game not fully appreciated by most of its competitors. That will change in 2008, and by 2009, Google's gravitational effect will be significant.

What's Next?

Each year I am less and less comfortable predicting what will happen in information retrieval. Nevertheless, a study of this type cannot ignore the future. Those involved in behind-the-firewall search want to know my view of the future, if only to disagree with me. Accordingly, here are the major trends in behind-the-firewall search that my research seems to support.

Consolidation

With pressure being exerted on the middle of the market, my view is that Endeca may be an acquisition target. Autonomy may find itself in the position of an army trapped on top of a hill, surrounded by enemies on all points of the compass. I am fascinated with the drama now being played out among the superplatforms, the up-and-coming systems, Lucene, and, of course, the dozens of content processing organizations profiled in this study. A remarkable number of new companies enter the competitive fray, sometimes two or three a week. I have a tough time following the established firms. I'm often asked about a company selling content processing, and I have to tell the caller, "I have never heard of them." There is not enough oxygen in the content processing space to allow the current contestants to survive. Staying ahead of the innovations is expensive. Supporting existing customers is expensive. Making sales is expensive. In short, search is hot, but it is not the easiest system to build, maintain, enhance, and support.

Search Becomes a Commodity

Key word search is now either included in modern operating systems, downloadable from key players such as Google and Microsoft, or free from programmers who want to provide a better search solution. Organizations can use Lucene, sidestepping the hefty license fees associated with some search-and-retrieval solutions. Third-party applications include key word search or more advanced systems in their applications. One license fee covers the enterprise applications and nice-to-have functions like search. For organizations strapped for resources, there are also cloud-based, hosted, or managed search and content processing solutions. Some of these solutions can cost a few hundred dollars a month. Others can hit six figures. In short, search is everywhere.

The outlook then is for more sophisticated information processes to become key differentiators. Most users find the search box acceptable, but not without headaches. Assisted navigation, automated alerts, point-and-click interfaces for mobile access or small screen access, or search that operates in the background and displaying information only when the user takes an action – each of these techniques will morph in the "new" search or the "next-generation" in search. The short term outlook for

established vendors is more competition. Over the longer term, search will become the next application interface and shift from key word queries to other types of interaction between the user and the processed data and information.

Market Drift

Each of the major market sectors are moving. For example, superplatforms like IBM, Microsoft, and Oracle want to capture more sales from small- and mid-sized businesses. These organizations can give away or bundle their content processing solutions, thus squeezing other vendors.

Organizations with a steady influx of youthful technologists may be more open to cloud-based content processing solutions. Lucene may be familiar to some of these engineers and provide an acceptable solution without the cost and complexities of commercial software. Google may be pulled into organizations by recent graduates.

Companies with search solutions from well-established vendors may find that these companies are transforming and upselling from search to a more expensive, more complex information platform.

Start-ups eye the enterprise customers of IBM, Microsoft, and Oracle as well as the licenses of Autonomy, Endeca, and Fast Search & Transfer solutions. Start-ups may offer some combination of features, service, functions, and cost.

Specialist vendors of content processing tools and subsystems may expand their offerings to provide *business intelligence (BI)* or some other high-value solution.

What's going on is a movement of the key vendors into and across different market sectors. Search vendors want to become platforms. Platforms want to provide full spectrum solutions to organizations of any size. Tool and utility companies are shifting from specialist roles to broader solutions.

When multiple clusters of vendors chart a course to another market sector, the effect on potential buyers is significant. Confusion is now endemic. These shifts are only now becoming evident, so the market forces guarantee instability and potentially rapid change for user, licensees, vendors, and analysts.

Marketing

Marketing and selling content processing solutions is in a state of flux. The traditional approach of trade shows, magazine advertisements, direct mail, and face-to-face sales calls is almost prohibitively expensive. When the cash runs out, the marketing and sales grind to a halt. What's happening now is that vendors are trying to find ways to make sales without the long decision cycles and the punishing costs of traditional marketing. One of the consequences of increased competition is an escalation of claims, offers, and bargaining. You as a buyer will have an increasingly difficult time figuring what system can actually do a specific function. You will have to dig through the license agreement to find out what's included and what's a "gotcha." Hint: consider customization and engineering support. You will have to expend considerable effort estimating the total

cost of ownership for a system. Hint: consider content transformation, infrastructure, and customization costs. In the 1980s, I thought marketing hype had reached a peak. I was wrong again in the 1990s. Now, in 2008, I'm not making that mistake. Marketing "volume" and "noise" will definitely rise.

Implications

The implications of these three large-scale trends vary by one's point of view.

- Licensees. For the next 12 to 18 months, there will be more options from which to select a content processing solution. Prices, at this time, are negotiable. The competitive arena makes it possible to craft an agreement that can be on the buyer's terms. However, the volatility in the sector increases the risk of a vendor being acquired or falling into financial trouble. Many potential buyers will find high-profile, well-known solutions less risky than an unknown vendor's product.
- Vendors. The marketing imperative means that confusion is likely to become an almost permanent part of the landscape. To get a message out, more marketing resources will be required. In some vendor organizations, marketing may suck resources from engineering or another technical unit. Vendors will have to balance their need to make sales against the potential starvation of essential technical resources.
- Investors. The potential for a big payday from content processing, semantic technology or some other content processing play exists. In the next 12 to 18 months, the present saturation of the market and the repositioning moves of the vendors increase risks for some new ventures. Content processing is a definite home run, but juggling the unknowns increases the odds for each player.
- Users. Users are likely to see little significant change in information access in the next year or so. The new solutions described in this study have yet to achieve ubiquity. Over time, assisted navigation and other next-generation content services will become more widely available. Unfortunately, even when an organization embraces a "beyond search" solution, it will take months to deploy, transition, and educate users about the new services.

The stakes for the dominant vendors mentioned in this study are rising. The Big Three (Autonomy, Endeca, and Fast Search & Transfer) and the superplatforms (IBM, Microsoft, and Oracle) have uncertain seas to navigate. Not only must these organizations compete within their respective grouping, but each must compete in the other's core customer base. Managers responsible for these companies' content processing solutions run the risk of focusing on one competitor and missing a move by another. The more these six companies focus on one another, the easier it becomes for each to overlook Google or another upstart.

To sum up, 2008 will be pivotal for some of the best-known names in search and content processing. Upstarts and newcomers, if each can deal with its own technical and financial challenges, may have an opportunity to break out and generate substantial revenues because larger organizations lack the resources or foresight to avoid a strategic blunder that benefits smaller, more nimble organizations.

Google and Dataspaces

Google is deeply involved in content processing. The company crunches with a single process – one called MapReduce – about 20 petabytes per day with the volume creeping up. As of January 2008, Google is the leader in content processing. Based on my study of Google’s public engineering documents, its YouTube lectures, and its public documents, Google is working intensely to improve its processes. Some technologies that other companies rely upon are insufficient for Google’s needs.

Content processing is a complex suite of processes. Most of the technologies described in this study simply cannot operate at what I call “Google scale.” In fact, the problem is not with identifying entities, classifying documents, and performing other types of metatagging functions. The problem is larger, and Google has embarked on a research track that aims to leapfrog current technologies and thus gain a competitive advantage. Google wants to have a solution that can operate at Google scale with the speed and reliability its users and customers expect.

In this section, you will learn about one research initiative at Google which aims to move beyond databases into the realm of dataspace. A *dataspace* is a representation of information that sidesteps the bottlenecks with current technologies and permits new types of queries and applications at Google scale. It’s not clear if Google can commercialize this research, but it underscores Google’s willingness to tackle fundamental problems in content processing in an effort to increase its market share, its opportunities, and its revenue.

In January 2008, Google is not an active participant in the rich content processing market in which the vendors profiled in this report compete. However, the company’s Google Search Appliance or GSA can be made to perform most of the functions described in this study. But, and this is a major exception, Google leaves it to its licensees, resellers, and partners to use the OneBox API as an integrating doorway.

Google, true to its approach to my inquiries, ignored my requests for comments. What I want to discuss I have had to assemble from engineering papers, presentations, patent applications and patents, and information available on the Web. I want to make clear that the information in this section may be spot on, generally correct, or just wrong. Nevertheless, I feel it is important to provide some information about Google’s context processing technology.

Semantic Technology at Google

In my Google Version 2.0 and the 2007 Bear Stearns’ report titled “Google’s Semantic Web: The Radical Change Coming to Search and the Profound Implications to Yahoo! &

Microsoft”, I dig into the five Ramanathan Guha inventions pertaining to a “programmable search engine”⁶ (PSE).

The PSE is a system that imparts semantic metatags to content. Its inventor worked on the World Wide Web Consortium’s semantic Web standard. The five patent applications that comprise this invention reveal a system that captures meaning of documents and associated objects. The tags don’t exist in a vacuum. A context is generated for each object. Together with usage data and other information in the Googleplex, the PSE has some interesting and potentially far-reaching implications.

For example, the invention can generate a master ontology of the information processed by Google. Don’t confuse a taxonomy with what the PSE can generate. The Google knowledge base will contain categories and relationships. It also can contain programmatic instructions that permit operations on the metadata and the objects themselves. Without succumbing to the temptation to repeat what’s in my other Google analyses, let me suggest that Dr. Guha’s invention makes it possible for Google to eliminate the roadblock that keeps the 3WC’s vision for a Semantic Web from becoming a reality. Through it Google becomes the Semantic Web and becomes the way to use unstructured and structured data in a programmatic way.

Some of the companies profiled in this report are on the same track. Siderean, for example, has anticipated if not invented a different solution that delivers similar functionality to its customers.

An even more remarkable invention appears in two patent applications mostly ignored in the technical literature. US2006 0230350, “Nonstandard Locality-Based Text Entry,” makes possible an automatic, behind-the-scenes query, described in paragraph [0096] of the patent application. This invention dovetails with an invention that can invisibly but actively monitor a user’s actions via a mobile phone or other continuous network connection with a user’s computing device. US2006 0224448, “Obtaining Content from an Electronic Device,” provides a glimpse of the company’s recognition that key word search is not appropriate for certain types of queries.

From a content processing viewpoint, Google’s engineers are observing behaviors, analyzing a user’s usage data, and tracking real-time information such as the user’s geospatial location information.

The result is that when the user looks at the device display, the information, which Google’s algorithms have predicted the user will need or want, will be there or at least in a cache so there’s little or no latency in delivering this information. Endeca’s integration of stored queries (saved searches) into an Endeca licensee’s work flow is a somewhat similar innovation. Google, however, is operating on a much larger scale if I

⁶ The Bear Stearns’ report is not a public document. You may request a copy from a Bear Stearns’ office or locate it in the Investext service. It appeared on May 16, 2007, as the firm’s equity research analysts. Attribution to me appears on page 2 of the report.

read these two patent applications correctly. Google is moving into what it calls “I’m feeling doubly lucky” or what I describe as “search without search.”

Transformic: A Meta-Meta Approach to Content

As remarkable as these two “beyond search” examples are to me, both are less ambitious than Google’s dataspace research. What I want to describe has not been revealed in a patent application as of January 2008. However, if you navigate to Google.com and run a query for “Alon Halevy” +dataspace, you will be able to review the information that is publicly available.

History

In 2006, Google purchased a little-known company named Transformic, Inc. The most significant asset in this acquisition was Dr. Alon Halevy (née Levy). In January 2008, as I write this section of *Beyond Search*, outside of a select circle of content processing experts, Dr. Halevy’s technology is unknown.

Google’s Search Appliance, sometimes abbreviated to GSA, takes some hard knocks. For example, the fourth edition of the *Enterprise Search Report* finds fault with the Google Search Appliance because it lacks features. Whether this judgment is warranted or not underscores the lack of knowledge pundits have about Google. The GSA does what it is designed to do. It allows basic key word search to be deployed quickly and without most of the hassle, fiddling, and hair-pulling associated with better-known search solutions.

What I want to do in this chapter is cover three topics briefly:

- Describe Dr. Halevy’s dataspace technology and its relationship to content processing
- Highlight how dataspace may mesh with other Google enterprise functionality
- Comment on the potential disruption of an already-tumultuous group of markets that make up behind-the-firewall information access.

Dataspace

Most people are familiar with databases. These are the tables containing structured information in software systems such as Oracle’s relational database, Microsoft’s SQL Server, IBM’s DB2, and dozens of other products, some free, some expensive.

Dataspace, however, is a broader notion.⁷ A dataspace can contain the data residing in databases. It can also contain any other information accessible to and processed by the

⁷ The information in this section comes from Dr. Halevy’s papers and presentations available as open source information. I have drawn upon more than 24 of his writings, and since this is not an academic study, you can locate the materials by running the following query at Google.com “alon halevy” +dataspace. Errors in interpretation are mine, not those of Dr. Levy or his colleagues such as Dr. Jennifer Widom.

dataspace system. The diagram in Figure # shows the additional metadata about information in a dataspace. Traditional content processing does not automatically capture the social metadata associated with a document, its sources, and the individuals involved in the content process.⁸

The formal definition of a dataspace is a representation of many types of data; for example, the aforementioned databases whether relational or XML, files of different applications, code collections, Web services, software, data from sensors, real time newsfeeds, and other sources.

A dataspace also contains information about these data. Some of the data are obvious, such as the file name, the file type, and the file's date and time stamp. But dataspaces permit relationships to be identified and manipulated. For example:

- Full schema mappings; that is, views of other schema and replicas
- Information that content object A was created from content object B and content object C
- Information that content A is a snapshot of content object B on a certain date
- Content object A and content object B reflect the same underlying entity, but are different
- Content object A was sent to me at the same time as content object B.

Dataspaces, therefore, federate structured and unstructured data, permit classification and other advanced text operations, and contain relationship information.

The dataspace, then, can be queried. Traditional search-and-retrieval and point-and-click discovery are supported. But dataspaces enable different types of information access. Examples are:

- Search and query on data, schema, or what Dr. Halevy calls “meta-anything”⁹
- Query the lineage of a content object; that is, “Where did X come from?” and ask hypothetical questions
- Perform text and data mining
- Set up work flows, triggers, etc.
- Monitor for special events, new content, and changes
- Impose soft or fuzzy constraints on sets.

The dataspace takes content processing into a new dimension. With federated, normalized, and dataspace-processed information, the dataspace makes it possible to model uncertainty. A dataspace makes it possible to relate the lineage of data to certainty. Dataspaces can manipulate programmatically different types of uncertainty;

⁸ This diagram is based presentations about dataspaces by Dr. Halevy. See, for example, Alon Halevy, “Principles of Dataspace Systems,” PODS Keynote June 26, 2006.

⁹ See Dataspaces: A New Abstraction for Data Management, undated by Dr. Alon Halevy, Mike Franklin, David Maier, and Jennifer Widom.

for example, is the uncertainty in the data? Or is the uncertainty a result of approximate integration and translations?

Keep in mind that Attensity and Siderean, for example, have technology that performs similar operations. The difference with Google's approach is scale. Attensity and Siderean have solutions designed for a single organization. Dr. Halevy's dataspace notion operates on a heterogeneous index of everything. The Google version of a dataspace resolves multiple references to objects in the dataspace.

Dr. Halevy acknowledges the manual tasks required to implement a dataspace in an organization. These include finding sources and classifying them. The organization must manually map between and among sources or train an automated system to perform these mappings. The system requires rules or training to find related sources or must be smart enough to determine which sources are related and create associations between and among the data items.

Dr. Halevy wrote in "From Databases to Dataspaces: A New Abstraction for Information Management" in 2005:¹⁰

Dataspaces can be seen as an umbrella for much of the research that is already being actively pursued in the database community; In fact this was one of our original goals. We have also, however, tried to outline several new research opportunities that arise from making a more holistic view of emerging "dataeverywhere" challenges. These are challenges that the database research community is uniquely qualified to address, and we look forward to continued progress extending the applicability of data management technology.

¹⁰ Alon Halevy, Michael Franklin, and David Maier, "From Databases to Dataspaces: A New Abstraction for Information Management," ACM SIGMOD Record, December 2005.

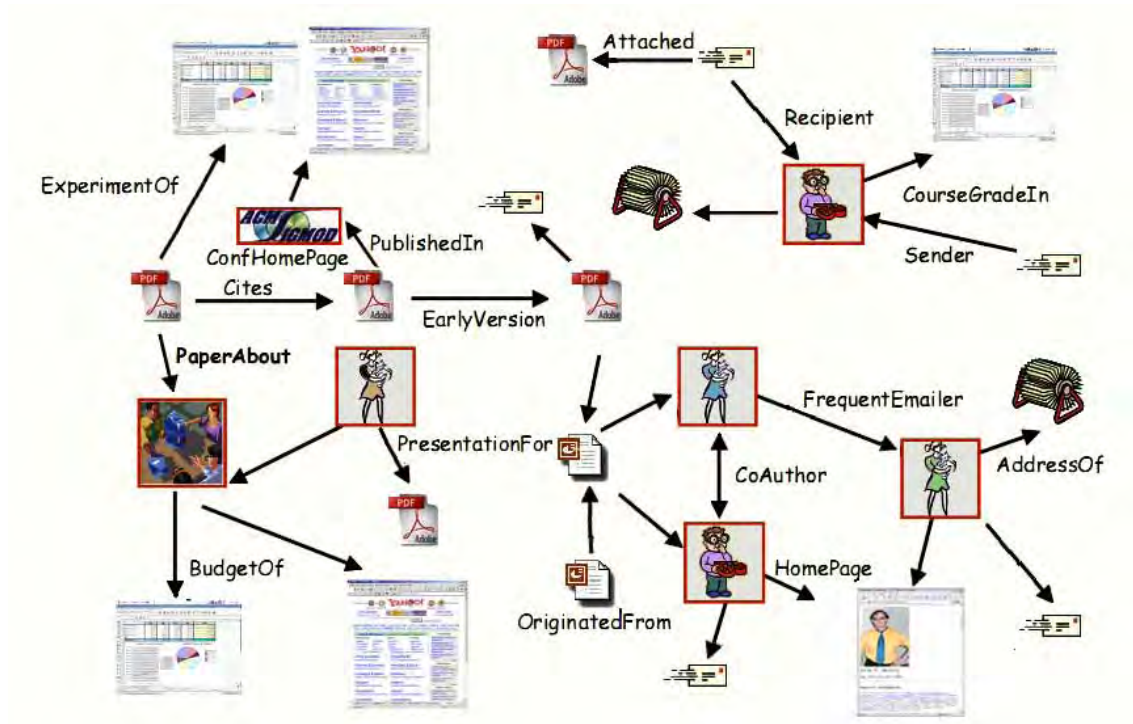


Figure 15: Interactions not Captured in Most Content Processing Systems

This is a diagram that shows the types of interactions not captured in most content processing systems. These are social interactions, system operations, and time-based activities. Source: Dr. Alon Halevy, "Principles of Database Systems," June 26, 2006.

Dataspace as an Enabler

Dataspaces are a way to manage information and make it possible to run certain types of queries impractical in traditional indexes and databases. After reading the dataspace information authored by Dr. Halevy and his associates, including the prescient Dr. Janet Widom at Stanford University, dataspace is a construct that integrates many separate indexes, their metatags, and data. In a dataspace, the user no longer worries about a specific collection or even what type of information has been processed. Whatever the dataspace system has processed is available. Google's "universal search" is one of the first glimpses of the dataspace technology in the Google system you and I use for Web searching.

In a sense, a dataspace sits on top of the Google PageRank engine and the PSE. A dataspace "knows" about actions that most systems cannot "see." For example, in a dataspace configured as Dr. Halevy describes it in words and equations, e-mail sent with an attachment to a reviewer will yield useful metadata. The fragments of human activity and system actions that are neither captured nor mapped to specific objects will be processed, tagged, and used to deliver information needed by a user.

I want to steer clear of privacy, security, and any legal implications of Dr. Halevy's work. From a content processing point of view, dataspace performs federation on a large scale and makes possible very search-useful enhancements.

If we consider Google processing its existing information in a dataspace, a user could run one query and get results from Web logs, Web sites, news, books, and other sources. Google's universal search is a step in the right direction, but dataspaces would take federated searching much, much further.

With the proliferation of smart mobile devices, Google has an "environment" in which the PSE and "search without search" inventions can make personalization a more refined tool than it is today. Buying Transformic and getting Dr. Halevy on staff marks the first steps in what I call "Google's envelope strategy." The metaphor suggests that Google puts everything – data, information, tags, metadata, and process information – in a structure in order to eliminate:

- Expensive content transformations
- Time-consuming reconciliation of indexing inconsistencies such as "IBM Corporation" and "I.B.M. Corp."
- The now-outmoded technology of traditional database systems such as those available from Microsoft, Oracle, and IBM, among others
- The latency incurred when users wait as information from different sources is gathered, normalized, and rendered for a user.

In short, Google smooths out the differences that we users and systems must reconcile with old-fashioned procedures.

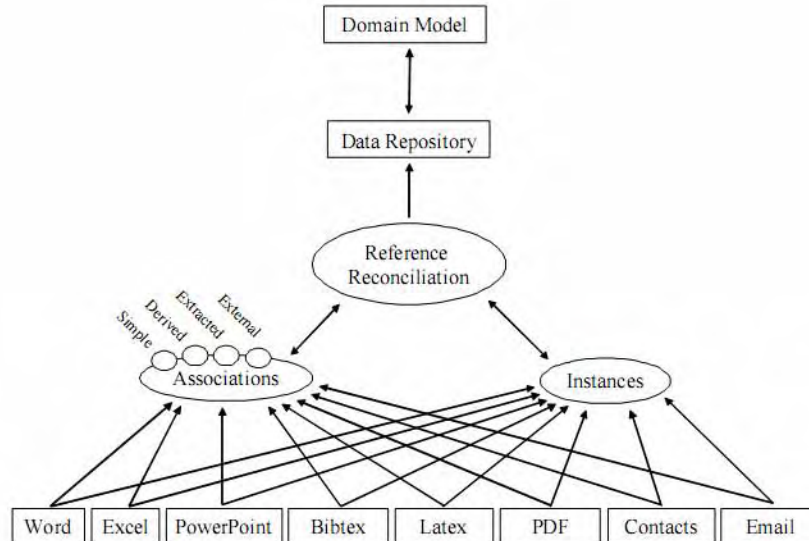


Figure 16: Semex: Mining for Personal Information Integration

This diagram is from "SEMEX: Mining for Personal Information Integration," a precursor to Transformic's technology. The paper was written by Alon Halevy, Xin Dong, Ema Nemes, Stephan Sigurdsson, and Pedro Domingos. (The paper appears to have been a workshop presentation at Knowledge Discovery and Datamining (KDD 2004) Conference, Seattle, WA, August, 2004)

New Types of Search Functionality

The more exciting implications of Dr. Halevy's work are the new types of queries that a dataspace permits. After reviewing more than 50 companies' content processing technology, I can assert with reasonable confidence the following:

No other company offers dataspace-type functionality in their currently available systems. Some researchers at the University of Illinois and the University of Wisconsin-Madison are making solid dataspace progress, but at this time, Google is the commercial firm with a significant commitment to this information retrieval and data management approach.

One example will illustrate the potential of the Transformic dataspace technology. Assume the dataspace system is operating. The system processes and tags conflicting data. A simple representation of this situation appears below, and it comes from a journal article written by Dr. Halevy.¹¹

(Karina Powers, { 345-9934 | 345-9935})
(George Flowers, 674-9912) ?

The question is: Which of the two phone numbers for Karina Powers is more likely to be correct? Dataspace technology as described by Dr. Halevy can support probability processes that can provide the user with a score for each phone number. The user can use the score as an indication of which number is most likely to be correct. Another application of this probability is an "uncertainty score" that can provide to a user only the fact that is most likely to be accurate, such as when multiple addresses exist for an individual or multiple employers and many other situations in which data are contradictory.

At this time, users, not content processing systems, have to reconcile conflicting information. Dr. Halevy's work suggests another interesting functionality that dataspace permit. In order to calculate "uncertainty" scores, the system must have the ability to look across the processed data in order to determine its lineage and evaluate the reliability of each content object in that lineage. Users and systems cannot currently provide insightful information about the probability a source is reliable across a content objects and time.¹²

Dataspace Management

The diagram below appears in a number of Dr. Halevy's talks. This version comes from a presentation by Dr. Halevy called "Principles of Dataspace Systems," dated 2005.

¹¹ Alon Halevy, Michael Franklin, and David Maier, [Principles of Dataspace Systems](#). This paper, which seems to date from mid-2006, contains an extensive bibliography of dataspace-related references.

¹² There are rumors that Kroll, a unit of Marsh & McLennan, offers Engenium's reputation management technology to certain clients in their Ontrack product. See <http://www.kroll.com/news/releases/index.aspx?id=16510>

Google has the computing and storage resources to handle a dataspace layer for the content it processes. Dr. Halevy's presentations in 2005 and 2006 suggest that his team is building a dataspace management system. Its principal components appear in his figure below.

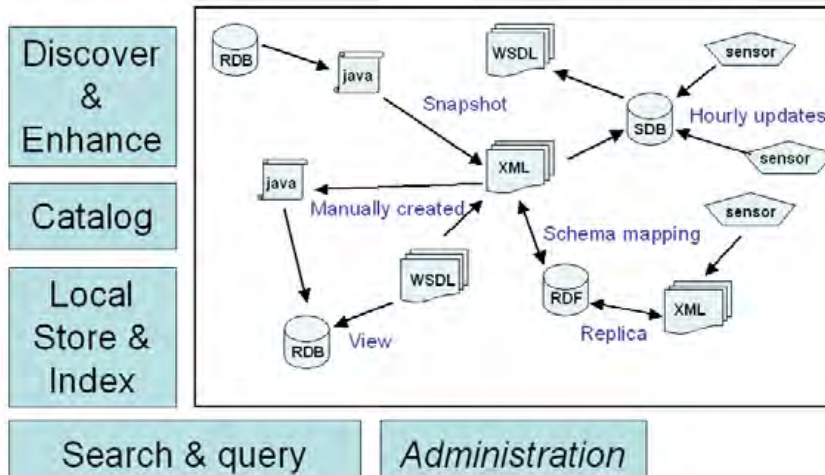


Figure 17: Managing Dataspaces

This diagram identifies the administrative and system functions needed to create, operate, and manage dataspaces.

At this time, Google provides no information about when these administrative and management functions will be available to users, developers, and partners. These tools are needed to “cross the structure chasm,” a phrase that seems to suggest that dataspaces become possible when content is in well-formed XML, traditional database systems, and can be converted when necessary into a dataspace-compatible form.

Possible Impact of Google's Dataspace Technology

Most readers will have little knowledge of dataspace technology. When you read more about dataspaces, you will discover that I have simplified a hugely complex innovation. The engineering and mathematics are more sophisticated than my summary reveals. Without veering too far from the basics presented in this brief discussion, I want to offer several observations.

First, the idea that Google is indifferent to the needs of users and customers with regard to sophisticated content processing is wrong. The firm's purchase of Transformic and the flow of papers and presentations about dataspaces are signals that the company is exploring this information retrieval avenue.

Second, the notion of dataspace is big, far larger, in fact, than the content domains that the systems profiled in this study tackle. Even the most robust content processing systems have not been engineered to handle Google-scale content flows. The implication of scale means that Google is operating in a research area largely without competition from the companies profiled in this study and from other information giants like IBM, Microsoft, and Oracle.

Third, the modest hints of dataspace technology that I have been able to identify – the “universal search” visible in the screen shot below, for example – may be viewed as harbingers of similar functionality in the Google Search Appliance.

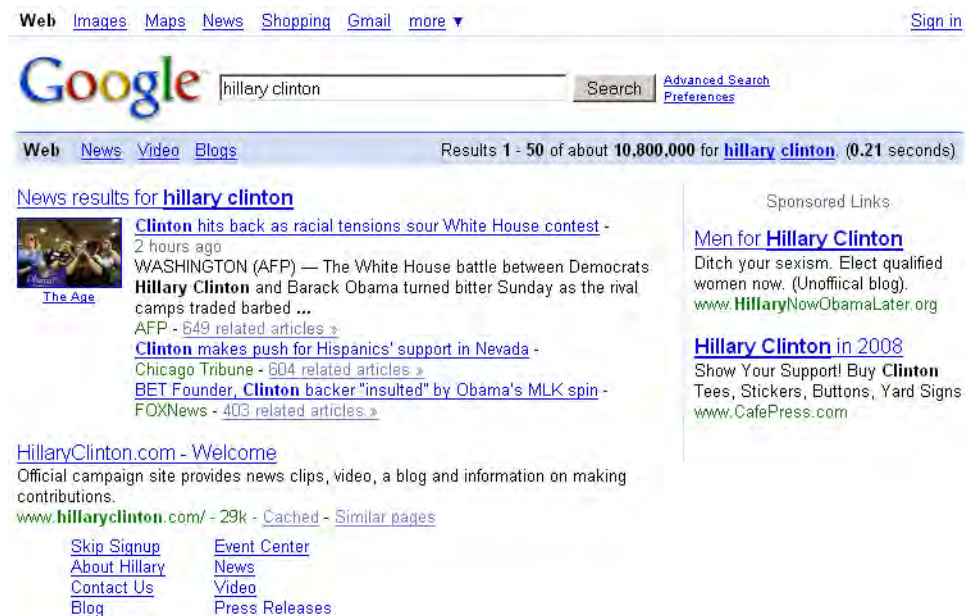


Figure 18: Google Universal Search - Multiple Content Objects

This result for the query Hillary Clinton presents different content objects extracted from a single Web site.

These add up to Google’s off-the-radar, pressure-generating tactic. Companies offering advanced content processing may find themselves out-flanked in dataspace. However, you should not defer your exploration and use of the systems described in this study.

My take on the dataspaces’ initiative is that Google is willing to invest significant resources to find a solution to the bottlenecks, transformation costs, and scaling costs that accompany most of the commercial content processing systems available today. Prudent actions include:

- Monitoring Google’s innovations
- Tracking changes to the OneBox API, which seems to be the mechanism for adding additional functions to the Google Search Appliance
- Watching for other vendors to introduce similar technologies; for example, IBM’s acquisition of APTSoft in January 2008 may be a signal that these hyper-functions are gaining traction. APTSoft has technology that performs complex event processing, or CEP, a function related to portions of the dataspace research.

II. Tracking the Players

A Snapshot of the Beyond-Search Market

A change is taking place in the search-and-retrieval “space.” As you know, Google has an effective monopoly on Web search. In other market sectors, there are many different leaders.

Let’s take a quick run through of the companies that claim leadership in these more-or-less consistent sectors described below. As you read these very brief market “thumbnails,” please, keep in mind that I am providing this information in order to set the stage for the market map for the companies profiled in *Beyond Search*. You can, when equipped with the technical savvy, make almost any search system work in a number of different settings.

eCommerce Search

This sector consists of vendors who provide a database system and various utilities to facilitate eCommerce. Contenders in this search segment include Mercado, Endeca, Saqqara, EasyAsk (now a unit of Progress Software), and Amazon. You may be surprised at my inclusion of Amazon, but the company makes available its S3, EC2, and SimpleDB systems as fee-based services. Taken as a group, you can create a reasonable back end for a Web-centric system. Who is the market leader? Based on information available to me, Endeca and Mercado are neck and neck in this race. Endeca has been more successful in making sales for behind-the-firewall content processing. Nevertheless, the Endeca system is quite useful for eCommerce. In the first three editions of the *Enterprise Search Report*, I included Endeca as a behind-the-firewall system because of its trademarked faceted navigation. I have not included Endeca in *Beyond Search* because I believe other companies have caught up with it. These upstarts offer somewhat easier scaling and more flexibility in situations where a licensee must work around a processing bottleneck.

Social Search

Social search is an invented category. Google is, at its core, a social search system, yet few recall that voting and links are human-centric features of information. A number of vendors will tout their ability to deliver collaborative, group-centric, and social functions. For example, Autonomy and Fast Search & Transfer (prior to its acquisition by Microsoft) assert that social search features exist in their systems. I don’t disagree. Neither of these companies’ more exotic functions are included in this set of profiles. In general, be certain you test certain types of search before you release them to your users, and in particular, the social search functions of start ups. Social search can have unexpected consequences. Companies in this sector include Tacit Software Inc., Eurekster, and the previously mentioned Autonomy, and Fast Search & Transfer. The leader in this segment is Tacit Software.

Database Search

In the market map developed for *Beyond Search*, you will find two companies identified as database-centric. Many companies included in these profiles use XML data structures in their system. But in the broader market, there are five database management systems that account for about 80 percent of the market in commercial organizations and government entities. These companies include a search function in their database system; therefore, these companies have a larger footprint in database search. You may be familiar with these four companies and the strengths and limitations of their build-in search-and-retrieval systems based on the structured query language (SQL). The companies are:

- IBM DB2
- Microsoft SQL Server
- Oracle
- MySQL and its variants (Sun Microsystems now “owns” MySQL).

The commercial market is split evenly among IBM, Microsoft, and Oracle. You can find various studies that show one company with a larger market share. In my experience, Microsoft has been nibbling into the customer base of IBM and Oracle. You will learn about vendors with next-generation systems that can do “regular” database work plus more sophisticated content-centric operations. The discussion of Google’s dataspace is a look into the future. Databases are likely to be supplanted by dataspace. But since dataspace systems are not yet widely available, I urge you to look at the database-centric vendors and the companies profiled in this report who offer XML databases. My bet is that Google will make a run at the dataspace business, leapfrogging the incumbents and the companies profiled in *Beyond Search*.

Hosted or Managed Search

Cloud-based services received a set back with Amazon’s S3 crash in February 2008. A cloud-based service delivers an application via the Internet. A hosted solution to me means the vendor offers the licensee a subscription and service level agreement to use a search or content processing system running on the vendor’s servers. You can customize hosted services and exercise control over what happens to your information. A managed solution means that you lease the machines or own them. You could have the servers on your premises, a third-party data center, or in the vendor’s data center. You specify in the license agreement and the service-level agreement exactly what the vendor will do to manage the search and content processing system. For example, you may have the servers at your place of business and specify that the vendor will hire, train, and manage the staff to maintain, customize, and tune the system. Alternatively you can specify that you own the machines, leaving the choice of data center up to the vendor. The vendor handles certain management tasks related to the system, and your systems engineer provides oversight and project management. A full discussion of the upside and downside of hosted versus managed solutions is outside the scope of *Beyond Search*. In general, hosted solutions offer you less control over how the specifics of the deal with the search vendor are handled. A managed approach lets you

spell out exactly what you want the vendor to do. Based on my experience with hosted and managed search, security is not an issue if appropriate security procedures are already in place at the licensee's organization.

A number of companies offer this service, including Blossom Software, Fast Search & Transfer, Sideran Software, and others. The benefits of this approach distill to three:

First, you can deploy a hosted search application more quickly than you can deploy most enterprise search systems, with the exception of Google Search Appliance, which can be deployed almost as quickly as a hosted or managed search solution. This means you can turn off your incumbent search system and provide search and assisted navigation without the cost and time required to remove the incumbent solution, install and debug the replacement solution, and plug the new solution into your existing interfaces. The hosted solution does the content indexing and query processing. You just handle the interface, editing files provided by your vendor.

Second, your internal information technology team does not have to learn a new system. Assuming that your team has customized your interface, the amount of their time and effort required to deploy a new solution is reduced by roughly up to 80 percent. In my experience with Blossom, for example, to reindex and customize an application using Blossom, the Threat Open Source Information Gateway, took us less than two hours. Your overhead may vary, but the burden on your technical team should be reduced.

Third, updating and tuning the system along with other maintenance is handled by your search and content processing vendor. Instead of waiting for your engineer to figure out the problem or getting direct support from the vendor, the technical load is assumed by the vendor. You can tailor your license and its service level agreement to meet your specific requirements.

Why has it taken longer for hosted or managed search and content processing to capture market share? In the customer relationship management (CRM) sector, NetSuite and Salesforce.com, among others, are reporting strong growth. The reason is due to vendors' marketing and positioning. Going forward, hosted and managed search solutions will attract more customers. Some of the search systems profiled in this study are too complicated and sensitive for the licensee's technical team to manage. I think cost, time, and complexity will make hosted and managed search solutions more desirable in the future.

Search "Toasters"

The idea of receiving a ready-to-run search appliance is still a novelty in many organizations. There are five companies struggling to be the leader in delivering a search appliance. These are:

- Google and its Search Appliance or GSA
- Index Engines. This is an appliance designed to index and make searchable information stored in backup devices

- EPI Thunderstone. The Thunderstone appliance preceded Google in the search toaster market. It is designed for customization and easy scaling.
- Planet Technologies. This company offers a search toaster designed for Microsoft Windows environments.

I include one advanced content processing engine in the profiles in this edition of *Beyond Search*. The market leader in search toasters with more than 8,500 installations as of January 15, 2008, the last date on which I received second-hand information about Google's GSA market share.

The Beyond Search Market Map

In my files I have information on more than 150 organizations offering search-and-retrieval, content processing, entity extraction, linguistic processing, and other technologies for indexing textual information. I have five or six vendors who index images. I track several companies performing audio-to-indexed text from spoken data streams. To keep this study manageable in terms of size and to provide a representative overview of a very active business sector, I had to exercise editorial judgment. As you read about the companies whose technology I have examined and described, you will wonder why I did not include other vendors. The selection criteria I used for these profiles were:

The company's technology goes beyond key word indexing. For that reason I excluded Lucene and other high-profile systems such as Microsoft's SharePoint search. I don't think key word systems are too useful in behind-the-firewall search. Key word searching is a commodity and not interesting to me.

The company's system implements some interesting advanced content processing functions. I was not overwhelmed with some of the systems' overall performance. There were some disappointments on some systems' ability to handle my test queries. But, in general, the companies profiled are pushing the boundaries of search.

The technologies are in use and demonstrate that the software, with varying degrees of success, can "understand" textual information and make that information available to users in a variety of ways. You will find that the vendors offering enhanced search or database-centric search perform as well or better than Autonomy, Endeca, Fast Search & Transfer, and many other widely used systems. But each of the vendors I have included has increased the innovation factor by a significant amount.

I excluded some companies for a variety of reasons. For example, I did not want to recycle the profiles of companies I analyzed in depth for the first three editions of the *Enterprise Search Report* or my two Google studies. You can buy a copy of these books, scan the profiles, and review the market map information in *Beyond Search*. Armed with these pieces of information, you can quickly determine if your search requirements warrant Oracle's SES11g system or one of the systems I have identified as a next-generation system. Oracle SES11g is a secure system, but its core functionality is getting a bit long in the tooth. Similar analyses can be articulated for the Teratext, IBM Omnifind, and Funnelback systems.

I also dropped from the list of the final 24 some vendors because their financial stability, their turnover, and their technology raised doubts in my mind about the viability of the company over the longer term. I have tried to include companies that I have determined are near break even or profitable. I also excluded some vendors because I was not sure what business the company was in. When I can't figure out what someone is selling, it's one tiny signal that something may be amiss. However, if your favorite vendor is not profiled in this study, write and tell me why I should include them. Maybe I will do so in an update or another edition of *Beyond Search*.

I am not endorsing any one company over another. In fact, there are several companies on this list with which I found it very, very difficult to develop comfort. You will have to select a company using your judgment. You will have to make the system work. If you run aground, that is your responsibility, not mine. My inclusion of a company in my list of 24 does not mean that you will agree with my assessment. I know there are some companies on my list that have outstanding products. I also know there are several that would give me a stroke if I had to work with their managers and their technology.

A Patchwork of Niches

Here's the "map" of the principal sectors of the beyond-search market place. I want to review its principal features and offer several observations about how to use this breakdown.

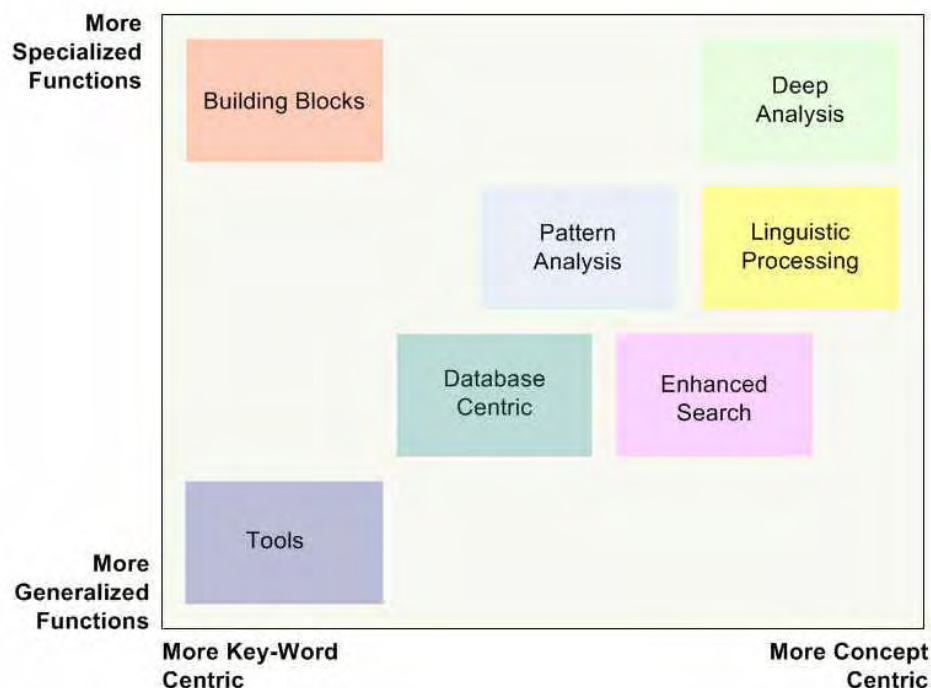


Figure 19: "Beyond Search" Market Sector Functionality

The principal features of this map are the rectangles that represent specific types of beyond-search functionality. These rectangles have fuzzy edges. In fact, most of the vendors profiled in this report assert that their systems include both statistical and

linguistic/semantic processes. Most of the vendors support XML, handle structured and unstructured data, and classify content. Nevertheless, I have made distinctions between vendors with database-centric systems, enhanced search, pattern analysis, and linguistic processing. The idea is that a system placed in a particular category **emphasizes** a particular technique or approach. My intent in making these fine distinctions is to assist you in asking more pointed questions about specific systems. It is not useful to view all 24 vendors as identical. The vendors are quite different in their approach, technology, and implementation of enhanced content processing.

The second feature of the map is the two axes. The x-axis represents a spectrum from “More Key-Word Centric” approaches to “More Concept-Centric Approaches”. Note that the “Enhanced Search” vendors sit somewhere near the middle. These vendors support key word searching, but incorporate more sophisticated content processing systems. The market segments placed farther toward the concept-centric end of the spectrum in my judgment rely upon more complicated sequences to process content. The systems are not necessarily better; these systems are approaching the problem in a different way from the vendors in the Enhanced Search sector. You can use my market map to help you focus on vendors who offer you a number of building blocks that you can combine to build a system that meets your needs.

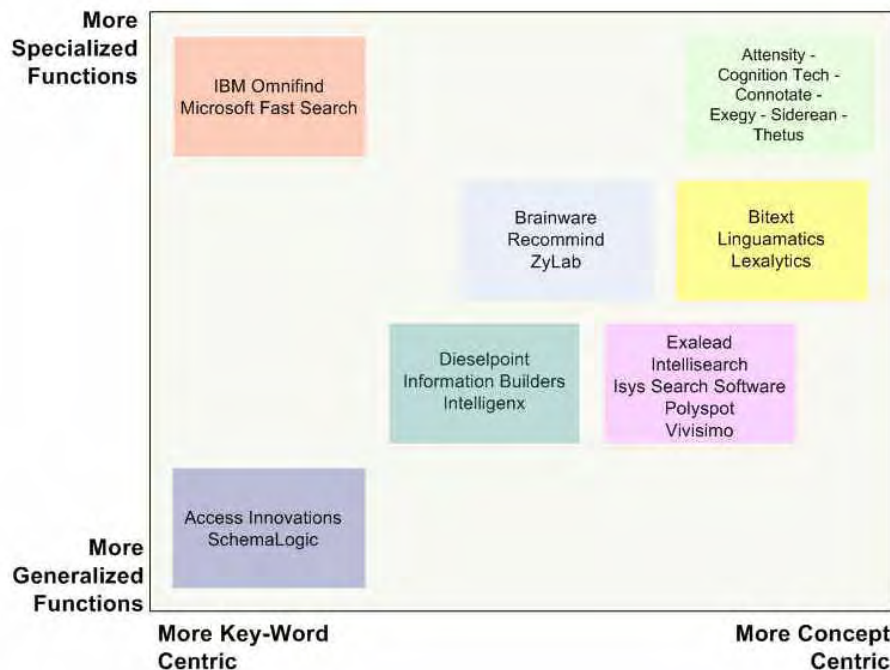


Figure 20: "Beyond Search" Market Sector Vendors

The y-axis is also a spectrum. Closer to the x-axis are vendors whose systems offer “More Generalized Functions”. A vendor at the other end of the spectrum is identified as “More Specialized Functions.” The vendors of more generalized systems give you a tool that you use mostly “as is.” Each of the components can, of course, be customized, but these vendors assemble custom systems by mixing and matching modules, not

developing original procedures to handle your content. The other end of the spectrum is “More Specialized Functions.” You will be able to customize these systems extensively.

Let me characterize briefly the meaning of the seven niches or sectors in the market map. I will discuss these in alphabetic order. Vendors profiled in this study are in bold face.

Enriching the Map

Now let’s look at how the 24 companies slot into these categories.

Building Blocks

These are systems that offer you a list of functions or modules. You select the functions and modules you need for your search system. Examples of vendors offering building block search and content processing solutions include Autonomy, Endeca, Exalead, Microsoft-Fast Search & Transfer, Oracle, and SAP. In *Beyond Search*, I narrow my focus to IBM and Microsoft–Fast Search. In my opinion, these other systems are well-known and covered in my *Enterprise Search Report*, first, second, and third editions. If you are reading this study, you probably have a “building block” system and want to remediate it or replace it.

Database Centric

These are systems which have as their core technology a database system. Because of this, these systems are adept at handling data management, content repurposing, and generating reports from the content that reside in the system’s database. Keep in mind that these systems can perform statistical and linguistic processing. The key feature, however, is their proprietary database system. Examples of vendors with this type of system include Teratext, Dieselpoint, Intelligenx, and Information Builders.

Deep Analysis

The vendors in this niche are pushing search and content processing in new directions. The vendors have very different technical approaches, but a unifying thread ties them together. These vendors use combinations of techniques. The use of multiple processes in iterative cascades point to the direction search and content processing is moving. Simple key word indexing is a Model-T Ford to these vendors’ finely tuned machines. Vendors in this sector include Attensity, Cognition Technologies, Connotate, Exegy, Siderean Software, and Thetus.

Enhanced Search

The vendors in this category offer key word search plus many of the features that users want. Classification, entity extraction, and point-and-click access to related content are quickly becoming “must have” features. However, many organizations find themselves unable to afford the seven figure price tags of some of the higher profile systems. Other organizations have one, maybe two, of the high-profile systems and considerable dissatisfaction from users and management. These companies want a better solution at a more competitive price. Most of the vendors in this sector offer a more attractive price to features ratio. I think one or two of these companies can become increasingly

competitive with Autonomy, Endeca, and Microsoft-Fast. Vendors in this sector are Exalead, Intellisearch, ISYS, Polyspot, and Vivisimo.

Linguistic Processing

As I began research for this study, I assumed it would be a straight-forward process. I would select a dozen companies using linguistic and semantic technology, profile them, and be 90 percent done with my analysis. I was wrong. Almost every vendor's system incorporates linguistic technology. Even almost "pure" statistical methodologies accommodate external knowledge bases to make their systems "smarter" when it comes to figuring out content. This sector is represented by three firms in this study: Bitext, Linguamatics, and Lexalytics.

Pattern Analysis

I use the phrase "pattern analysis" to describe systems that at their core are statistical. The archetypal statistical system is Autonomy. I profile Brainware, Recommind, and ZyLab. Each of these systems incorporates other functions, but each has a statistical foundation that arguably performs as well or perhaps better than Autonomy's system now elaborated and extended with functions obtained through its various acquisitions.

Tools

Tools are software that perform specialized search-related tasks. Most licensees of search systems don't know what they don't know. Once you have some experience with behind-the-firewall search, you have a better understanding of the importance of controlling and managing metadata. I profile two vendors with useful, professional taxonomy and controlled vocabulary tools; namely, Access Innovations and SchemaLogic.

In the first three editions of the *Enterprise Search Report*, I prepared a table that provided an "at a glance" reference to the vendors whose systems I discussed. I have prepared a similar table, but, please, keep in mind, you will be able to use my summary and the profiles in this study as a way to get oriented to search and content processing options. You can use my information to determine with which vendors you should speak, what questions to ask, and formulate a method to compare two or more systems. I cannot presume to tell you which system is better for your particular needs. You will find information that complements the detail in this study on the *Beyond Search* Web log at <http://www.arnoldit.com/wordpress>.

III. Vendor Profiles

In the pages that follow, you will learn about 24 companies and their systems which go “beyond search”. The companies come from innovators and visionaries from Australia to Spain, from Canada to Germany, as well as the U.S.

Each profile is designed to be an informative, “quick read”—almost like class notes in a university course.

The profiles follow a general format, but I have modified the general structure in order to highlight what my work indicates is the most important feature in a system. For example, you will learn about the Semantic Web in the Siderean analysis and about sophisticated statistical methods in the Brainware analysis. In this way, I make an attempt to provide you with a case example of an important idea or technique in search and retrieval without academic impedimenta.

Each section contains information about the genesis of the company’s approach to search, a description of its “beyond search” functionality, an example, sometimes two, of the company’s system in action; that is, a client implementation or a screen shot of the system’s administrative tools.

The most important part of each profile is my view of the strengths and weaknesses of each system. Unlike other search analysts, I discuss my views with the vendors. If a vendor disagrees with my assessment, the vendor has an opportunity to “push back”. In general, I listen, but I make up my own mind based on my use of the system, my tests, and my experience. Some of the vendors profiled in this study, therefore, would have liked me to reword my assessment of their system. I have tried to be fair, but the views are mine, not the vendors’ nor the publisher’s.

I have also included a “net net” comment. Years ago, I worked for a large newspaper, and we used this buzzword as the title for a for-fee newsletter. Then and now, “net net” means the real bottom-line about a company. The “net net” section of each profile, therefore, is my opinion of the system. My “net net” can and does change. Vendors add features and improve. Like a grade in college, a vendor can improve over time. My “net net” observations can evolve.

To assist you in keeping the vendors’ systems differentiated, I’ve prepared a table that summarizes the key points about each of these vendors’ systems. In a pinch, the table will be sufficient guidance for your to perform your own due diligence about each of the profiled systems.

You can also use the overview in the Executive Summary to this study to aid you in determining which vendors you want to invite to make a presentation to your search procurement team.

1. Access Innovations

www.accessinnovations.com and www.dataharmony.com

The founders of Access Innovations in Albuquerque, New Mexico, have a long history in the information industry. The company has a stellar reputation among commercial database producers and highly-regarded publishers. Marjorie Hlava and Jay Ven Eman know how to organize large-scale content processing operations.

We're not talking about indexing text, which the company can do, but complex technical content: medical studies, chemical research, and analyses of nuclear phenomena. Handling this type of information requires specialized knowledge and software tools. For years, Access Innovations built content processing utilities for the company's own use. Several years ago, Access Innovations launched a commercial version of its software and marketed that product under a new unit called Data Harmony.

Item	Quick Facts
Product	MAIstro Suite
Price	~\$65,000. Custom price quote is recommended
Key Feature	Automated machine indexing with manual and semi-automatic options
Purpose	Build and apply ANSI-standard controlled terms, taxonomies, and knowledge bases
Clients	Reader's Digest, National Geographic Society, Consumers Union, Discovery Communications, Special Library Association, Lockheed-Martin and US government agencies
Company	Privately held
Contact	Telephone: 505-265-3591

Table 4. Quick Look at Access Innovations

Today, Access Innovations is one of a very small number of companies with a software system that incorporates a methodology for building, maintaining, and managing controlled term lists, taxonomies, and what are commonly known as knowledgebases.

Dr. Jay Ven Eman told Beyond Search, "We find that our existing customers tell other people about our approach to managing metadata. In fact, when a major search engine vendor experiences a meltdown, we get a telephone call asking us if we can get the indexing back on track. Automatic systems are generally okay for most business documents. But if there's a glitch, the entire system can go off track. Our software prevents and remediates those problems."

The system incorporates decades of experience creating these complicated beasts. You can sign up for a taxonomy boot camp elsewhere to learn everything you need to know about indexing in eight hours. These "boot camps" work about as well as a day at a

Beyond Search: Access Innovations

weight reduction spa. You get a taste of the real deal, but you come away essentially unchanged.

With Access Innovations' software, you get a system that delivers results. The Access Innovations system can operate automatically, in a semi-automated way, or in manual mode. You can shift between modes in order to deal with new terms or shifts in how concepts can be organized.

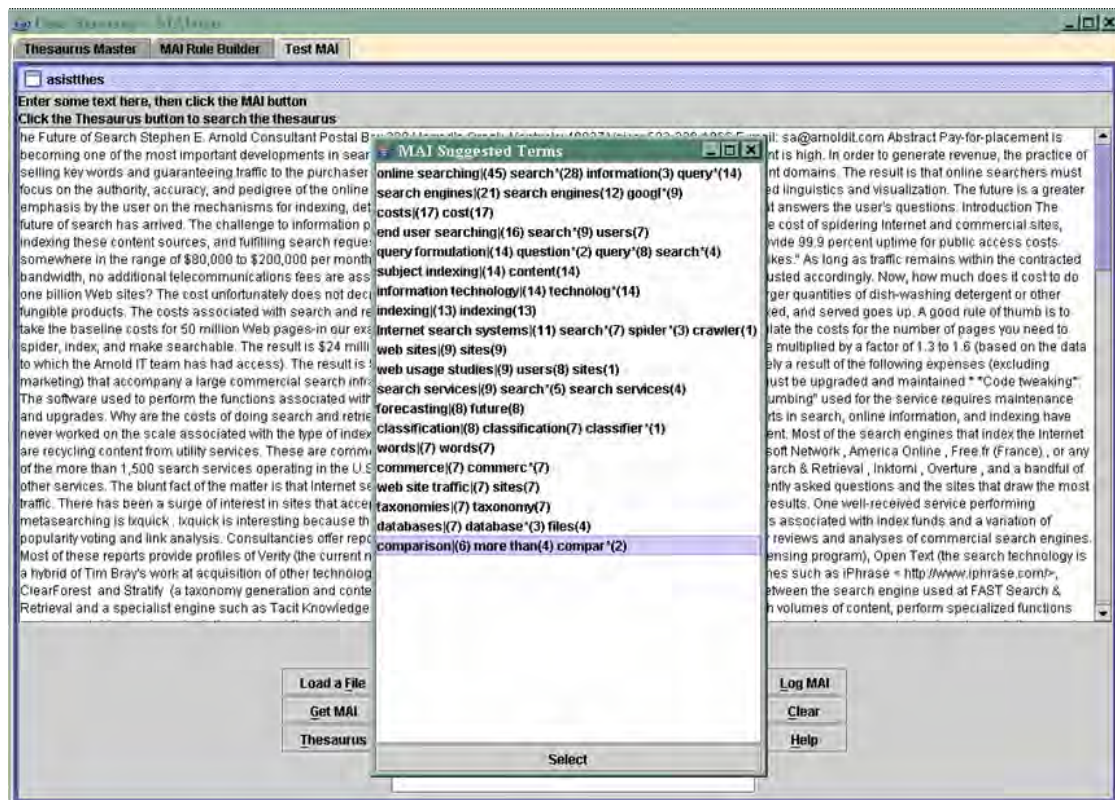


Figure 21: Access Innovations' MAIstro Interface

The MAIstro interface displays the source document and the terms automatically assigned to that document. In semi-automatic mode, an administrator can accept or reject the automatically-assigned terms. In automatic mode, these terms are used to index the document.

MAIstro Suite

The firm's flagship product is a bundle of integrated components called the MAIstro suite. The "MAI" stands for machine-assisted indexing. Founder and President Marjorie Hlava says, "Systems need automation. The volume of digital information is large, and in most organizations, it's doubling every nine to 12 months. In 1980, I believed in manual indexing. Now, it is machine-assisted indexing or you won't be able to cope with the data flows."

Function

MAIstro provides a knowledgebase for indexing and text management using the Access Innovations' proprietary *Rule Builder* module. The module allows an editor to create, edit, and review rules for the use of indexing terms.

Other components of MAIstro suite are:

- Concept Extractor - tool that reads the data against the knowledgebase to present suggested index terms that the editor may accept or reject.
- Statistics Collector - a software module that gathers and stores the index experience of the system. The "counts" guide the indexer in term distribution and use. An indexer can use these data to refine the knowledgebase and its associated knowledge domains.
- The Thesaurus Master TM (available in both W3C and ISO-compliant version) - implements the ANSI/NISO-compliant creation and maintenance of a full taxonomy. Thesaurus Master allows the licensee to develop a custom vocabulary or as a tool to import taxonomies from external sources for editing and maintenance. Access Innovations also offers an ISO-compliant version that incorporates multiple broader term functionality.

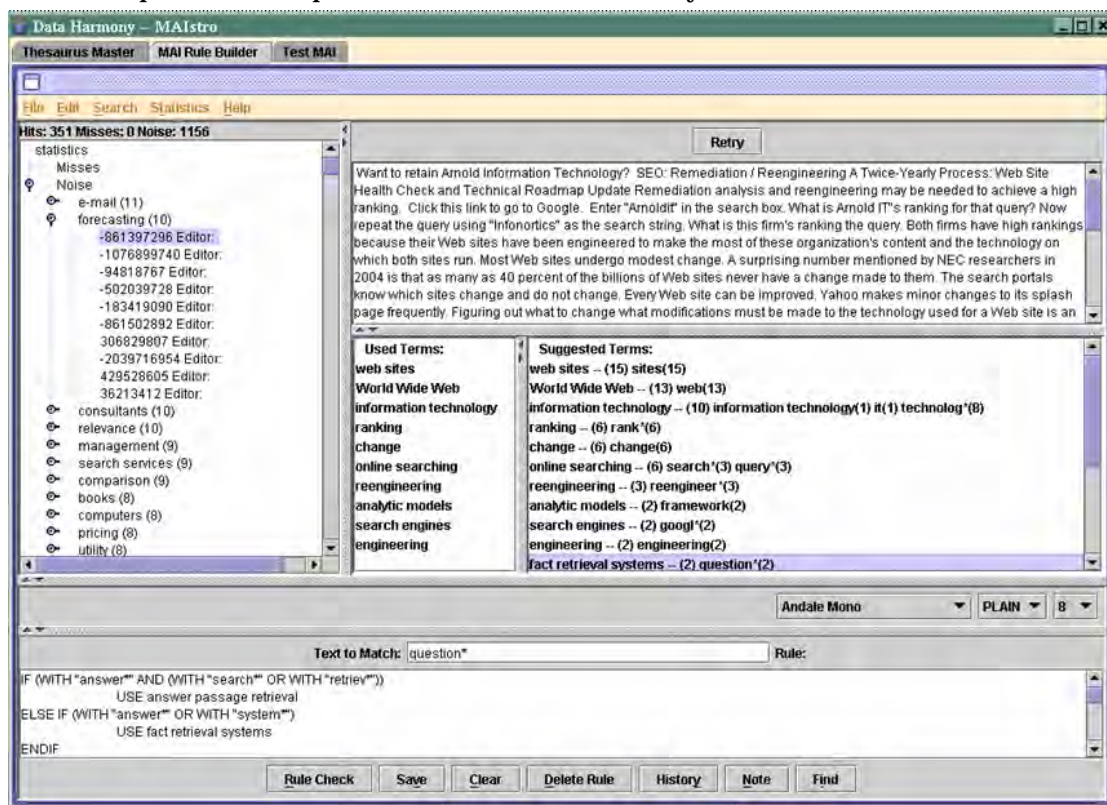


Figure 22: MAIstro Rule Building Interface

The rule-editing mode allows a licensee to configure the conditions under which terms are applied. Unlike some systems which use a "black box" for indexing, the Access Innovations' system allows fine-grained control of indexing and classification operations.

Extensions

The company has created what it calls “extensions.” These optional software components extend the functionality of the core components. Extensions available from the company include:

- MAIChem™. This module allows a user to identify chemical names in long and complicated documents such as research papers and patent documents. MAIChem identifies new chemical names in processed text.
- MAILib™. The software adds function to associate Library of Congress Subject Headings (LCSH) with items in a knowledge collection or knowledgebase.
- MAI STAR. This software runs in the Cuadra’s STAR information management system in order to add MAI services to the STAR system.
- XIS™, the XML Intranet System, is an object-oriented, platform independent CMS. Content becomes a database record in XML.

Access Innovations can support licensees with custom-built knowledgebases, taxonomies, and controlled vocabularies for a number of different vertical markets. Now available are vocabularies for medicine, geospatial, and education content domains.

Access Innovations knowledgebases are expressed in hierarchical form with cross-references to related concepts or aspects of a topic. A well-conceived hierarchy of concepts is needed in order to represent information in a way that makes it easy for a user to locate what’s needed. Each of the concept terms in the taxonomy has scope notes and *Use For* information. Access Innovations also includes guidance on how and where a term fits into the hierarchy. Each term is annotated with “rule” (associated information) about how to use and apply a specific term.

Technology

Access Innovations software is written in Java with published Java application programming interfaces (API’s), an approach taken by a number of vendors in this study. The use of Java allows the software to run on most platforms. The API’s make it possible to integrate Access Innovations’ technology with third-party applications.

The company uses its own proprietary data structure to house the objects manipulated by the system. The Access Innovations’ database engine generates XML objects for search, content processing, and enterprise publishing system use. The design of the Access Innovations’ system eliminates the need for creating custom scripts to move or transform indexes and knowledgebases.

When the system automatically processes content, the parsing process matches terms in the document to the index terms and concepts in the XML database file. When a match is identified, those tags are generated, attached to the source item, and the metadata is written to the database table. The data in the table can be processed by a third-party application such as an enterprise search system or output in a format for use by an analyst.

When the system is used in an interactive mode, the matching process is the same with an additional step added. Assigned terms or candidate terms are displayed to the indexer who can accept or reject terms. If an alternate term is needed, the Access Innovations system displays candidate terms and provides the indexer with a browse function if access to the controlled vocabularies is required.

Machine-Aided Indexer

The Machine Aided Indexer makes it possible for human indexers to increase their indexing efficiency and consistency. M.A.I.™ facilitates selection of terms from controlled vocabularies, authority files, or full thesauri. It presents a list of suggested terms to the editor for selection, which saves time looking up terms manually and speeds processing time.

M.A.I. allows flexible term entry because the editor can add or reject terms as needed. All editorial actions are gathered by the Statistics Collector, which then submits hit, miss, and noise lists to the Rule Builder module for continued improvement of the rule base.

Indexing using M.A.I. mines the entire depth of the vocabulary applied, improving document retrieval, relevance and precision for the end user.

Thesaurus Master

Thesaurus Master allows a licensee to:

- Create and/or import the terms you use
- Define the scope of use in the rule base
- Fine-tune and refresh your definitions easily as terms evolve
- Maintain hierarchical consistency while adding terms at any level
- Link directly to automatic indexing and your documents

The system operates independently in managing your structured taxonomies. It accommodates the formal thesaurus structure of Broader Term, Narrower Term, Use and Used For refs, History, Related Terms, and Scope Notes. Restricted access allows maintenance controls to ensure database integrity. The system can be used in conjunction with M.A.I. (Machine Aided Indexer) and XIS, the XML Intranet System. Thesaurus Master adheres to thesaurus standards ISO 2788 (monolingual), ISO 5964 (multilingual), and NISO Z39.19-2005.

The Notation Module represents a new approach to structuring material in a thesaurus. Notations are prepended onto thesaurus terms, which notations then form the basis for the ordering of the terms within the thesaurus. Thus, the hierarchy may reflect other than an alphabetical structure, in addition to (and parallel to) traditional thesaurus structure. The “concepts” applied by some automated systems become more useful when the prepended notation data are used as points of entry to tagged information. In addition, the notations increase a licensee’s options when building a thesaurus. The

notations allow a more accurate mechanism for including user-defined weighting and prioritization of terms. Notation allows a thesaurus to reflect:

- Process structure - Thesaurus structure may now match the steps of a process or workflow, in the order that they are followed within the business or industry.
- System structure - An annotated thesaurus can prioritize as well as accurately map the structure of a system, from top level to component to sub-component.
- Organizational structure - Departments of an organizational structure may be placed according to priority, funding, etc.
- Filtering levels - Assigning a notation to a term simplifies filtering of levels of information or data against which the thesaurus is applied. Data discovery tools can be set to recognize terms with the same levels of notation (a “3.2” term and a “6.2” term could be mined similarly, based upon the “.2” element) to select which areas of data are examined first.
- Security levels - Thesaurus structure can reflect levels of security or access within a system or organization.
- Report or manual structure - Terms in the thesaurus can be ordered to structurally reflect a report or manual used within a business or organization; the thesaurus itself forms the index structure of a document.
- Frequency weighting - Terms can be arranged within the thesaurus to reflect frequency of term appearances within databases or other material.
- Chronological order - Terms may reflect the order in which they were added to the thesaurus, or may reflect a fiscal cycle.
- Multiple sources or editors - Prepend notations may reflect the source of a term, where several thesauri are being combined, so as to provide ready visual cues for those combining the terms within a new structure. A notation prefix may also be assigned to each editor involved in creating or maintaining a thesaurus, to easily “brand” work as it is done.

Customers

Access Innovations’ customers are interesting. First, one of the firm’s clients is the American National Standards Institute, which uses the product to develop its own taxonomies and vocabularies. Not surprisingly, the MAIstro suite is certified by ANSI to generate taxonomies, word lists, and knowledge bases that comply with the institute’s own rigorous engineering standards. Marjorie Hlava says: “Standards are your friends in text mining. Standards-compliant taxonomies and controlled term lists lead to richer, more informative information product. Furthermore, standards promote interoperability and consistency. When systems support standards, information can be more easily repurposed.”

Other customers include:

- American Chemical Society
- Elsevier Science Publishers
- SLA (Special Libraries Association)

- Weather Channel
- U.S. Department of Transportation, National Transportation Library.

Benefits

Access Innovation's system is one of a very few designed to minimize "editorial drift." Vocabularies and classification can shift away from core concepts as new information is processed by an automated system. The manual controls and the semi-automatic mode make it easy to correct or nudge the system away from incorrect assignments. MAIstro is one of small number of tools that allows an organization to develop an ANSI standard taxonomy and controlled term list for use with third-party text mining systems. In addition, the technology improves the accuracy of automatic indexing of other systems. If your search system is not indexing accurately, the Access Innovations' system can remediate your existing system without down time.

Other benefits of Access Innovations technology include:

- Ability to construct and maintain controlled vocabularies and term lists for individuals of interest and their "use for" aliases
- Advanced indexing modules, which can be integrated into an editorial or intelligence analysis workflow, amplifying the accuracy of routing, alert, and filtering systems
- Technology to handle complex scientific, technical, and medical technology as well as more colloquial language used to describe the concepts.

Feature	<i>Beyond Search</i> Comment
Knowledge Base Support	Yes. Controlled term lists, taxonomies, and knowledge bases with extensions for chemical structures
Query Types	Supports key word, Boolean, and assisted-navigation interfaces
Visualization	Third-party tools may be integrated via an API
Entity Extraction	Can extract entities via term lists, rules, and automatic processing
Platforms Supported	Java-based system supports most platforms and operating systems
Export	XML, MARC, OWL, Comma, or tag delimited
Third-Party Support	Scriptless integration with most enterprise applications
Vertical Support	Medicine, education, and others available from the company
Analytic Functions	Third-party tools may be integrated via an API

Table 5: Technical Highlights for MAIstro

Downside

There are some drawbacks associated with licensing Access Innovations' technology. These include:

- Does not include semantic or linguistic techniques when processing text
- The system does not identify relationships among entities, although a third-party tool can be used to process the generated metadata
- A “by the book” approach, which is often at odds with the “run-and-gun” indexing implemented by some vendors.
- Access Innovations' system assumes that staff with subject matter expertise will be involved in the development of the rules and the knowledge bases.

Net-Net

Access Innovations' tools are designed for organizations where advanced systems require standards-compliant taxonomies or text mining systems where humans interact with the system.

A rule-based approach to determine the “aboutness” of a digital object has both strength and weaknesses. It is the most accurate approach. It scales. It is consistent. It deploys more rapidly and has a lower life-cycle cost than alternative approaches.

The current version does not have automatic novelty detection built in. However, existing approaches to concepts from third-party sources can be integrated with Access Innovations products. A rule base can be generated more rapidly and at less cost than alternative approaches for certain content domains.

The consistency of indexing is a key characteristic of Access Innovations' approach. Although Access Innovations was not designed as a text mining tool, it can provide tremendous value to text mining initiatives, particularly when accuracy is needed. The persistent nature of the Access Innovations metatagging resides with the content after an initial text mining analysis. Properly assigned subject terms significantly enhances content value.

Although Access Innovations supports fully-automated text mining systems, Access Innovations' bias is toward text mining systems that involved human analysts at key steps in the process. The company starts with a focus on controlled terms with technology a hand maiden to the larger intellectual challenge of getting the knowledgebases right.

2. Attensity Corporation

www.attensity.com

Attensity is one of the leaders in squeezing facts and relationships from unstructured text. Most of Attensity's customers focus on getting the obvious bits and pieces of data from customer feedback found in emails, service notes and surveys and found on the Internet in web forums, Blogs and reviews. Attensity takes a more interesting and computationally complex approach. Attensity's technology performs what the company calls "exhaustive extraction," which identifies Facts and Relationships and can get to the level of cause, if expressed in the text.

Attensity is less about generating a list of people and more focused on producing information about thematic roles and discourse processing. Yes, it's text mining, but it's text mining on steroids.

Item	Quick Facts
Product	Attensity 4.0
Price	License fees begin at \$150,000. A custom price quote is required.
Key Feature	System can rapidly and exhaustively extract the facts from unstructured text without the need for time-consuming predefinitions
Purpose	Convert unstructured text into structured data, thus making the transformed data available for analysis
Clients	Whirlpool, The Hartford, Travelocity, Bose, Dept. of Homeland Security, Law Enforcement Agencies
Company	Attensity Corporation
Contact	(800) 721-0560

Table 6: Quick Look at Attensity Corporation

Getting software to figure out what's behind or implied by a comment is a spin that Attensity has leveraged into venture money and some juicy contracts with the intelligence community.

Like other text mining systems, Attensity provides a search mechanism, a suite of query and visualization tools for exploring the data, and an SDK (software development kit) to allow Attensity technology to be customized or used in third-party applications.

Attensity has harnessed a number of advanced linguistic processes to its technology platform. Attensity analyzes text to discern roles and themes. Like other text mining companies, Attensity can ingest available dictionaries and taxonomies. If these are lacking, Attensity can generate lists of words, people, places, and things.

What's interesting about the Attensity system is that the company has approximately 17 patents pending (with 6 already granted) that focus on taking discovered entities, applying metatags to items, and converting unstructured text content into database

entries. Attensity's technology then goes a step further. Its processes take the data from other structured databases and fuses it with the newly-structured data from its text mining processes.

Voilà. An organization has its structured and unstructured data in a consistent set of database tables. These tables can be analyzed, sliced, diced, and processed by Attensity's tools or tools from other third parties. Attensity's "exhaustive extraction" and its database creation tools put Attensity at or near the top of the text mining front-runners.

Rather than depending on key word triggers or word frequency to "interpret" unstructured text, Attensity takes a new approach that uses sequences of processes to find information needed to answer who, what, why, when, and where without making an analyst manually process documents. Attensity's system boils down an analyst's work to asking a question and getting an answer or clicking on a relationship diagram and getting access to the chunk of information behind that discovered fact.

To sum up, the Attensity Server and the company's other analytic applications enable more sophisticated analytical processing. Attensity automates the transformation of written language into structured, relational data. Attensity asserts that its approach reduces the need for manual fiddling with separate collections of data. By putting everything in one fused database, the analysts have more time to explore relationships and plan responses.

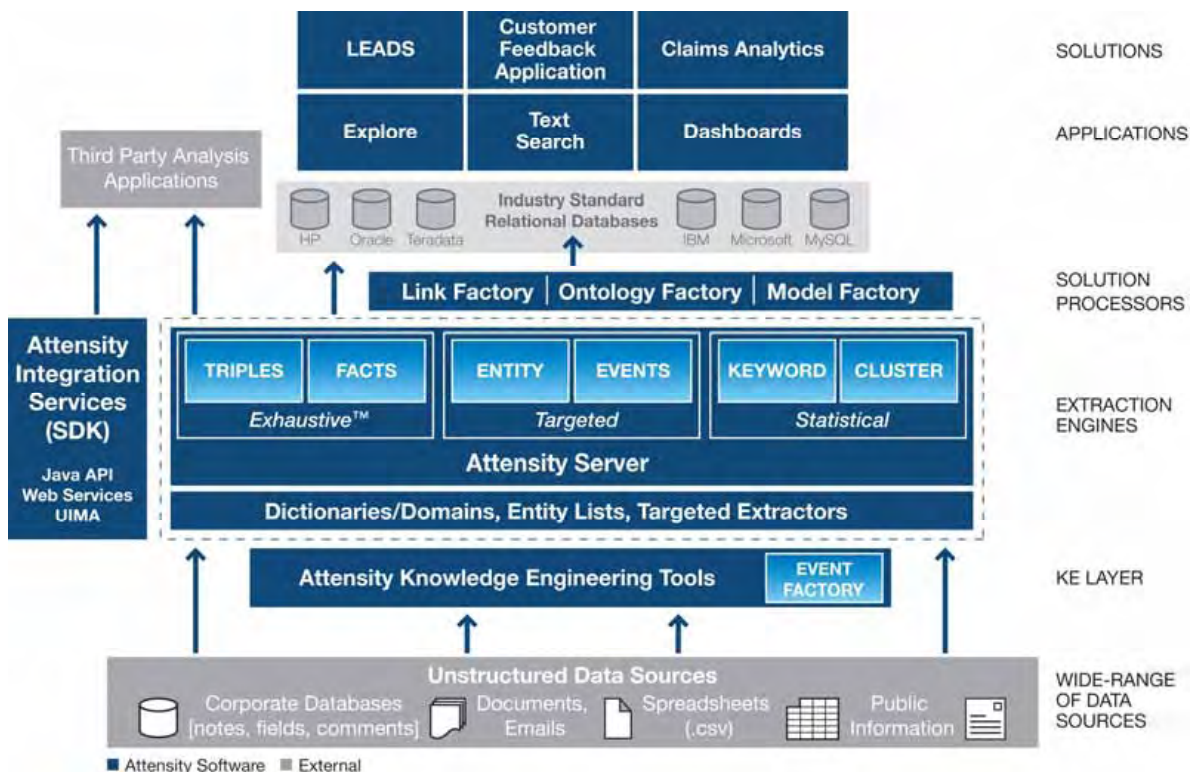


Figure 23: Attensity product stack

The System in Action

Attensity remains a somewhat secretive organization. In its presentations and the videocasts available to qualified individuals, Attensity provides snapshots of how its “unnamed customers” use the text analytics technology. Three examples are summarized below:

Manufacturing Company

A major U.S. appliance manufacturer collects tens of thousands of service records and warranty claims annually, representing direct warranty costs of two to three percent of revenues (total warranty expenses are typically five to ten percent of revenues). Effective analysis of service logs and warranty claims requires an understanding not only of dates, times and part numbers, but also of unstructured information captured as notes and comments which make up the majority of information in the service report, such as: What failed? What were the circumstances? How is this failure related to other incidents that have been reported? What did the customer experience? Using the Attensity Text Analytics solution to uncover this information, this manufacturer is able to substantially reduce warranty costs, improve overall product quality and boost customer loyalty.

Insurance Company

A top ten U.S. insurance company stores and manages millions of claims and applications forms, and related documentation (police, medical, engineering reports - depending on the type of policy). It is estimated that over 80% of the content of these documents is unstructured and this unstructured data usually contains information that can indicate patterns reflecting loss exposure, reserve calculations and other valuable, detailed information related to claim root cause and resolution. With Attensity's Text Analytics solution, rapid, on-going analysis of this information gives this insurance company an edge, allowing the company to identify emerging trends that have an impact on policy holder satisfaction and profitability.

Government Agency

Government agencies must process extremely large amounts of unstructured information in various forms - field reports, email, electronic content, cables, transcripts, etc. This unstructured data contains information regarding suspects and crimes. The principal tools for addressing critical unstructured content have typically categorized, searched and retrieved documents. Unlike Attensity, these methods are unable to extract or use the “relational facts” from text that describe not only the activities or entities referenced, but the associated what, when, where, how and why. Today, leveraging the Attensity-supplied ability to extract relational facts, a federal investigative agency has been able to successfully identify, locate and prosecute a variety of suspects.

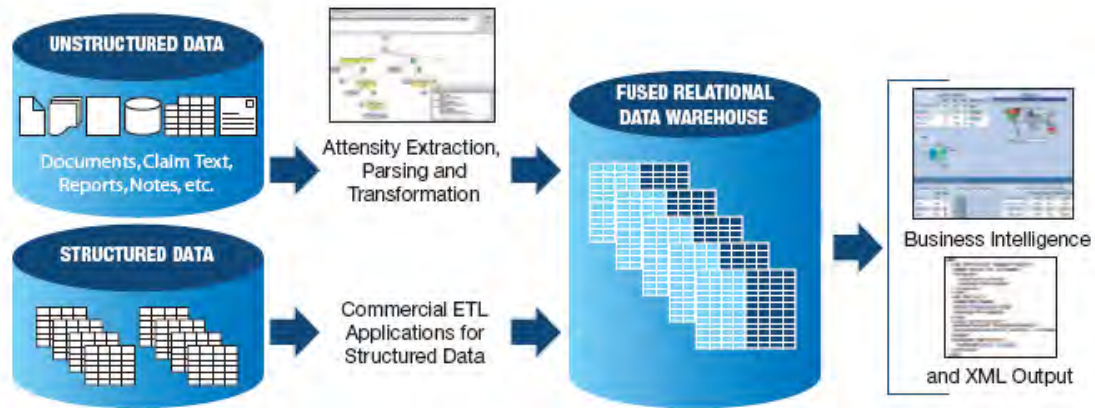


Figure 24: Attensity unstructured data transformation process

Technology

Attensity's architecture pivots on its extraction engines. These are processes that discover the facts residing within unstructured text, tag those items, and perform various sub-processes to ensure that what's discovered is accurate.

The term engine implies that Attensity uses a number of moving parts to produce its XML and relational outputs. The job of the extraction engines is to pull out and transform various forms of unstructured information in email, customer service and technician notes, claim forms, public records, research reports, Web forums, and pages, faxes, newsfeeds, and other sources. The extraction engines convert a "bunch of stuff" into a structured form. Once in a structured form, other processes can be used to explore these newly structured items.

Attensity provides generic engines intended for general business intelligence and more specialized engines tuned to meet the needs of intelligence agencies, for example. The targeted extraction engines make use of word lists, event definitions, knowledge dictionaries, controlled vocabularies, and taxonomies to pull out pre-identified facts and concepts.

Both targeted and generic engines can be operated in what Attensity calls "exhaustive mode". The system processes content until each content object has been mined for maximum information value. Two examples of exhaustive extraction from Attensity are AAO and FRN or what is more commonly known as "doubles" and "triples."

First, the AAO (actor action object) engine offers an exhaustive record of the actors, actions, and objects extracted from each processed document. It contains a generated key value for the record itself and for each actor, action and object. It also contains a file ID that links back to the FILEINFO file, and a sentence ID that links back to the SENTINFO file. It records the byte offsets of each actor, action, and object. These byte offsets record both the full phrase and the head noun or verb of the extraction. For example, if *Seattle-based Microsoft* were extracted as an actor, beginning and ending byte offsets for both *Seattle-based Microsoft* and *Microsoft* are recorded. Finally, the

file contains both the head noun or verb and their morphological root forms, for example, *buying* will be stored as the head verb, but *buy* will be stored as its root form.

Second, FRN (fact relationship network) deflects the extensive knowledge-engineering-intensive process in which content also includes modifiers. You would need to determine which facts to extract, and then determine how to extract those particular kinds of facts. With FRN Attensity has a better idea; it extracts all facts, not just some, and dumps them in a “fact relationship network” (FRN). The FRN is two relational tables, one for facts and one for relationships, suitable for copying to a Teradata or other commercially available data warehouse for business intelligence.

Extraction Engine: Advanced Text Processing

Attensity’s Extraction Engine application, known as Attensity Server, breaks the full text down into individual words and then applies numerous elements of computational linguistics to arrive at a result that has a accuracy usually in the 90 percent plus range.

Keep in mind that at its core, Attensity Server is a linguistic engine with some statistical and basic keyword engines. Attensity’s approach is to understand via algorithms what many systems approach by having a Subject Matter Expert (SME) and a Linguist working together. They would map the grammatical patterns and elements detected in text into columnar data elements of a relational database. Many vendors transform textual data into relational data. But Attensity’s software approach eliminates human intervention.

Attensity’s new approach was to bypass the human factor altogether and have the text analytics exhaustively capture as many entity/relationships as possible and then store that information in just a few relational tables. Stated simply, these tables contain an entity and relationship table. Once the tables are formed, standard business intelligence tools reveal facts.

For example, data revealed that a weak weld joint was responsible for a failure in a mounting bracket in a car. The simplicity of the Attensity approach is one reason the company avoids detailed explanations of the “inner workings” of the system.

With the discovered entity-relationship information, Attensity generates relational database tables. The Attensity insight was that RDBMS tables can easily handle millions of rows of data. These existing query and analysis tools can perform better because the data in the table are more accurate. Attensity can extract a fact out of text like most search and text mining systems, but it extends that to the point of exhausting all roles and relationships, events, causes and locations. But getting a fact is only part of the problem. The user wants to join the text data with other data stored in relational database table. Attensity, therefore, provides a function to merge these different tables. The approach essentially performs a join on a person’s name with a credit history, for example. Multiple tables can be joined so a holistic view of the person emerges from the Attensity system.

Attensity’s approach is to extrapolate representations of meaning from word content and proximity. The company says that its technology can understand English and other

languages via computational linguistics which parse sentences into fundamental linguistic elements. The resulting elements are analyzed using the firm's proprietary algorithms. Although Attensity will not reveal the workings of its Discover engine, the system appears to use techniques such as:

- Syntactic processes; that is, lemmatization, linguistic parsing, and parts of speech identification
- Role identification routines; that is, algorithms that deal with the assignment of event-specific roles to the entities mentioned in a piece of text
- Theme identification; that is, what is the document about and how does that theme impact an object in the document
- Domain extractions; that is, language used within a topic area such as medicine, finance, or manufacturing

Attensity uses regular expression pattern matching, part of speech identification, and semantic grammar rules. When its engine identifies the word *purchase* as a verb, the subject is identified as a possible customer. If *plastic explosive* is used as an object, the subject is tagged as a potential enemy.

Poking under the hood of the Attensity engine reveals a number of esoteric concepts about language. To provide a flavor of Attensity's approach, we'll look at the problem of figuring out what words in a collection of unstructured text are bound phrases like *white house* or *stock market*. To do that, the company has added what it calls "Directed Learning" technology. The user interacts, via the Attensity Workstation, which supports the company's example-based model. The user supplies sample content that teaches the engine how to find and extract the types of events needed from a corpus of documents. The user needs only to be conversant in the language of the document set and then responds to examples derived by Attensity from the document set and displayed in its training interface. Based on the feedback received from the user, Attensity derives grammar and extraction rules, to be applied to the data set. It will output structured data in a form that is loaded into the user's preferred repository, application, or business intelligence solution.

The result, is Attensity's approach incorporating an "automatic" mode with a "training" mode, blending the two most commonly used processes in one text mining system.

Analytics

Attensity's currently available Explore 4.0 module allows the licensee to investigate the facts and relationships generated by the extraction engine and server application.

This version includes a "fact co-occurrence". An authorized user can find facts that identify an issue with a group of customers and then analyze other facts occurring within that group. The function allows an analyst to probe business issues to understand root causes typically hidden in customer service, agent field reports, and repair notes, emails, feedback surveys, and other text-based sources.

Beyond Search: Attensity Corporation

The system includes an interface that permits drilling down and visualizing unstructured data. One feature of the Explore module is that it can display newly-discovered facts from unstructured text and any other structured data that organizations already store in spreadsheets, databases, and business intelligence applications.

“Rather than spending weeks or months attempting to articulate what they're looking for, customers can use Attensity Explore 4.0 to focus on whatever facts jump out from the exhaustive extraction of text automatically compiled for them,” said Craig D. Norris, Attensity’s chief executive officer.

Other features of Explore 4.0 are:

- Sharing function that allows authorized users to create and share queries or reports from the module
- Support for frequency counts of specific issues revealed in the text
- One click access to a function that sorts items or counts into categories for analysis.

The application is available as a hosted or installed application, and now includes support for the Teradata platform in addition to other platforms and databases already released.

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	Can use existing word lists and taxonomies. A taxonomy can be discovered by the system.
Query Types	Keyword, concept, natural language, and point-and-click words and concepts
Visualization	Included in visualization module. Outputs from Server can also be processed in third-party analysis and visualization tools.
Entity Extraction	Yes – people, places, things, relationships
Platforms Supported	Windows, Unix, Linux
Export	Yes – export function to CSV and other popular formats
Third-Party Support	Teradata, Oracle and support for other storage/data warehousing systems
Vertical Support	
Analytic Functions	Multiple analytic functions provided from simple tabular reports to link analysis

Table 7. Technical Highlights for Attensity

The SDK - Attensity Integration

Attensity's Software Developers Kit (SDK) is available to the licensees of the Attensity system. The SDK opens the system technology architecture to developers via a Java API. The SDK allows a developer to perform text analytics in real time from any application such as customer relationship management, Web and search applications.

The Attensity SDK provides business users with the text analytics extraction capabilities of Attensity Server, enabling them to leverage the valuable, sometimes hidden, information residing in their unstructured data.

The SDK, Attensity's Web Services interfaces and Attensity's support for IBM's Unstructured Information Management Architecture (UIMA) allows licensees considerable flexibility in exploiting the Attensity text mining technology.

New Features

The company has integrated its extraction and discovery products, and added search and auto-classification to simplify the text analysis process. The test of these new features will be in how well they pay off by bringing in new partnerships.

Other recently added features enable users to view and export raw text when an insight is found using the "Actors, Actions and Objects" tabular view of the text extending that analysis to events and entities. Attensity also optimized wildcard queries and counts by records to better help customers find facts, track trends, tag threats, and analyze co-occurring issues and relationships.

Attensity 4.0 delivers improved performance for the company's linguistic technologies to extract the facts from unstructured text and organize those facts into a relational database. Each row in the table represents an event, and each column an attribute of that event, such as location, time, action, or actor.

Upside

Attensity's payoff to licensees is significant. The system delivers greater than 90 percent accuracy on most document collections. Furthermore, the system categorizes actual events, not the documents containing an event. Throughput on text mining systems will vary; however, Attensity's document processing hits in the range of five megabytes per minute on basic systems.

The benefits of the Attensity approach include:

- Fused data sets
- Identification of cause and effect event relationships
- Extraction of nuances from text and use of those nuances to enrich the information associated with extracted entities and their relationships
- A good alternative to the manual, expensive, error-prone task of document analysis and tagging

- Generates tabular, analysis-ready data structures from text sources up to 20,000 times faster than a human analyst
- Provides robust handling of “noisy” input, such as poor grammar, misspelled words, shorthand and unknown terms
- Integrates easily with existing data sources and targets.

Downside

Attensity’s system is computationally intensive. The company says that the system will run on a single computer; however, most licensees will want to deploy the core system across a series of servers with a data warehouse backend. Like other advanced text processing systems, temporary files can consume significant disk space. Analytic and visualization processes can consume significant computing resources; therefore, a robust infrastructure is required to derive maximum benefit from the system.

Other drawbacks include:

- The company says that Attensity can operate without manual set up and training. However, the system benefits when training, word lists, and knowledgebases are made available to the document processing subsystem. Like Autonomy, “automatic” does not mean “hands off” from set up through operation.
- The licensee will want to have appropriate staff or resources available to configure, customize, and tune the system.
- For mission-critical applications, setup and testing can easily consume three to four weeks.

Net-Net

Attensity continues to have an impact on the text mining market. ClearForest and Inxight, among others, have found that Attensity’s approach has strong appeal for situations where features, accuracy, and integration are vital to the information mission.

Attensity has raised the bar for finding nuances in unstructured text. Attensity offers an alternative to traditional statistical text mining systems, training-based systems, and text mining approaches that generate less than stellar accuracy.

The patented extraction engine generates some of the richer metatags we have examined. The system’s ability to add structure to emails, customer reports, web forums, faxes and other sources is impressive.

One function that we found of particular value was the system’s ability to generate structured data and then fuse the newly-structured data with other data in various database tables. For organizations with the need for industrial strength text mining, Attensity’s single data warehouse approach is a logical solution. In 2007, Attensity upgraded and expanded the software modules for analysts and visualization. The end user can interact with the Attensity outputs in a standard browser, eliminating the need

for the end user to wrestle with desktop applications and fat clients such as Excel. The drill down approach makes underlying data easy to explore. In addition, the Java API makes it comparatively easy to integrate Attensity functions into other third party applications. Financial institutions' and intelligence agencies' analysts may have preferred tools for certain tasks and need only Attensity functionality in a familiar software environment.

The system provides a security backbone, essential for mission-critical applications. Attensity provides tools to allow the licensee to tweak word lists and customize certain functions associated with entities and relationship discovery. The single-user version provides an organization with modest text mining needs to derive value from the Attensity technology. However, we believe the workstation has more value in an organization with a full Attensity system and individual analysts with specific requirements that may best be met on a single, secure workstation.

Keep in mind that Attensity's techniques are complex, and they are not intuitive. A sale may boil down to Attensity's ability to explain what their system does and how it differs from the competitors.

At this time, we believe Attensity offers one of the more sophisticated text mining systems on offer today. Attensity is one of the leaders in next-generation text mining.

3. Bitext SA

www.bitext.com

Madrid, Spain, is a hot spot for search and advanced text processing. Most North Americans think of Madrid and recall an Ernest Hemmingway novel. Bitext wants the association to hook into natural language processing from “the bits and text company.”

Bitext’s founder and CEO is the lean, handsome Antonio S. Valderrábanos, a graduate of Universidad Autonoma de Madrid. With a PhD in Linguistics, Mr. Valderrábanos founded Bitext in 2001 launched the company to carry out consulting services on language technologies after his long experience in this field in IBM, with Wordperfect and Novell. By 2004, the company’s principal focus was NLP or natural language processing technology.

Item	Quick Facts
Product	NaturalFinder
Price	Begins at 23,000 Euros
Technology	Natural language search system
Key Feature	Performs automatic synonym and query expansion; supports NLP queries
Purpose	Allow a user to interact using sentences in English or Spanish and other languages such as Basque and Catalan
Clients	RENEF, Ministry of Defense
Company	Bitext SL, Madrid, Spain
Contact	info@bitext.com

Table 8: Quick Look at Bitext SA

Bitext’s linguistic technology can also be applied to computer-assisted translation environments. One of the leading companies in this area, Atril, developers of Déjà Vu X, uses Bitext’s linguistic technologies to improve results for searching in translation memory databases. In addition, Bitext has participated in computer-assisted translation projects funded by the European Commission.

Bitext is a privately-held firm. The firm’s customers include:

- RENFE (the Spanish Railroad Company)
- Public Administration National Institute, Spanish Government
- Ministry for the Presidency, Spanish Government
- Ministry of Defense, Spanish Government
- TYPSA
- Sitesa, Grupo EP, distributor of Google Search Appliance in Spain.

In addition, Bitext has developed adaptors to link its NLP technology with dtSearch (a Microsoft-centric search system) and the Google Search Appliance.

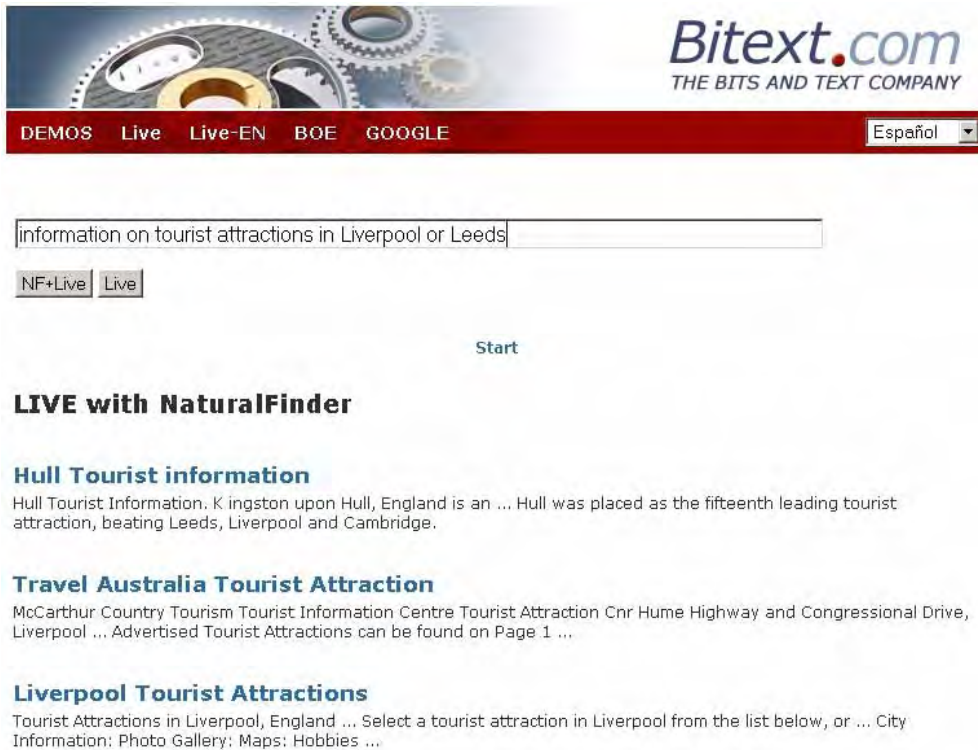


Figure 25: Bitext NLP Adds Functionality to Microsoft's Live Search

The Bitext NLP system adds functionality to Microsoft's Live.com search. Note that Bitext integrates with SharePoint as well. The user can enter a complex query without worrying about Boolean operators and query syntax. The system delivers results that are more specific to the query. Bitext, when executing a Boolean OR, sees term occurrence as significant. The first result in this list points to a site with information about both Leeds and Liverpool.

The Bitext Data Suite

The core functionality of NaturalFinder resides in what the company calls its DataSuite. These are subsystems that perform the heavy lifting required to process content and queries in the NLP system.

DataGrammar

DataGrammar is the subsystem that interprets natural language. It is built into the Bitext system. One feature of the DataGrammar subsystem is that it can “learn” as it processes content; for example, when an unknown phrase appears in a document, DataGrammar recognizes this phrase and adds it to the knowledgebase in the system. Learning is automatic and for most general content does not require the intervention of a subject matter expert.

DataLexica

This is a built-in lexical database. It uses linguistic stemming; that is, the subsystem removes inflections from words. The Bitext stemmer makes context-based decisions. Bitext asserts that its approach is “more than an intelligent stemmer.” For lemmatization and conjugation: DataLexica returns the lemma or root of words along with morphological information. For example, given the Spanish word *casa* DataLexica returns the following morphological information to the system:

- The root *casa* as a feminine singular noun
- The form is the third person of the present tense in indicative mood of the verb *casar* and *casarse*)
- Other forms of the verb *casar* such as *casando*, *casado*, *casada*, *casados*, *casadas*, *caso*, *casas*, etc.

Feature	Beyond Search Comment
Knowledgebase Support	Includes a lexicon, thesauri and knowledge base
Query Types	Natural language
Visualization	None. Third-party tools may be integrated with the API
Entity Extraction	Built in via proprietary algorithms and a knowledgebase
Platforms Supported	Linux and Windows
Export	The API allows expert functions to be defined
Third-Party Support	Can be integrated with third-party systems
Vertical Support	Builds for English, Spanish, and other languages available
Analytic Functions	None

Table 9: Technical Highlights for Bitext

DataSpell

DataSpell is the built-in spelling correction mechanism. It includes more than three million correctly codified words, not including proper names in Spanish in this count. DataSpell determines whether or not a word is correct in a specific language and, if it is not correct, it suggests alternatives. For example, for the Spanish word *inmobiliario*, DataSpell offers *inmobiliario*. DataSpell is configurable, and it can be integrated into a wide range of third-party applications, ranging from search engines to enterprise resource planning systems.

DataNet

The DataNet subsystem houses the rules regarding semantic relationships. The subsystem discovers relationships automatically and performs synonym expansion. DataNet makes use of existing thesauri, taxonomies, and ontologies. For example, a user can specify that the words *auto* and *coche* are related because of their similar

meaning; or we can specify that Italy is part of *Europe*. An administrative interface is provided to allow the system administrator to customize DataNet's relationship tables.

System Developer's Kit (SDK)

The SDK makes it possible for a licensee to integrate NaturalFinder components into another application or search engine. The SDK includes libraries for Windows and Linux, sample code, and documentation.

Technology

Bitext's system is available both for Linux and Windows. The company also offers a version of the system suitable for use as a hosted service.

NaturalFinder supports a wide range of documents and file types. These include documents, Web pages, and structured data. In addition, the system can make use of existing metadata, thesauri, and ontologies.

The system can process hundreds of thousands of words per second at the lexical level. Grammatical processing handles thousands of sentences per second.

The minimum recommended memory for the server running NaturalFinder is 256MB of RAM. The SDK makes it possible for licensees to add support for other languages.

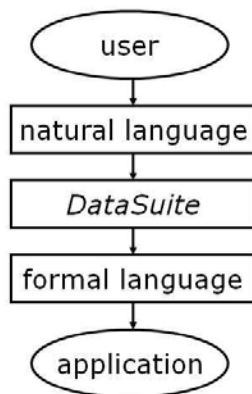


Figure 26: The Bitext data flow is straightforward.

The System in Use

Bitext has customers in Spain, Canada, and Germany. A representative installation is RENFE's use of dtSearch and Bitext. dtSearch is a Windows-centric key word searching system developed by a firm in the Washington, D.C. area. Bitext engineers integrated its NLP system with the dtSearch system.

"Our users at RENFE have been very pleased with the resulting application. Users particularly liked the speed, reliability and precision of searches, and the overall ease-of-use of the application," said Mr. Valderrábanos. "For Bitext, this agreement proves

that its linguistic technology makes a difference in the content management systems of large corporations.”

Bitext's DataSuite for RENFE includes DataLexica, which consists of a large and complete lexical database containing more than three million words classified according to their linguistic features.

At INAP, a digital library in Spain, Bitext installed NaturalFinder. The system was able to index content from different servers in different file formats. The user was able to search one or more of the collections from a single natural language interface. Bitext technology supported federating the content and providing users with the NLP interface. Other features of the INAP installation were filtering by document type, and support for a stringent security system.

Upside

The upside for Bitext's NaturalFinder includes:

- Built in knowledgebase, lexicon, and semantic mappings. These are supplemented with a knowledgebase administrative interface.
- Runs on Linux or Windows
- Supports NLP
- Supports voice or spoken queries when integrated with speech-to-text applications

Downside

The downside for the Bitext system includes:

- The system requires careful set up and configuration. Adequate bandwidth, computational resources, storage, and random access memory are essential for system performance.
- The API for Windows Live lacks some features; for example, the API only shows the first 250 results that Microsoft returns and will not process queries longer than 20 words. Check with Bitext for the functions available in the API.
- Hits with multiple occurrences of terms can be ranked above hits that are directly about a query, for example, in a search for city information.
- The system's relevancy improves with longer queries. Some users enter two to three word queries or prefer clicking on suggested links or categories to discover information without having to formulate a query.

Net-Net

Bitext illustrates the interest in NLP and linguistic search in Europe. Along with Exalead, PolySpot, and Sine Qua Non, entrepreneurial activity in rich text processing is increasing. The Bitext system can add NLP to almost any search system. If you have a search system and want to add automatic entity extract, NLP, and other functions to your existing system, Bitext is definitely worth a look. For companies in Spain, Bitext may well be the NLP system of choice.

4. Brainware, Inc.

www.brainware.com

For years, Brainware was a property of SER Solutions, a company providing call center systems. In February 2006, Brainware was acquired by company executives in a buy out. Today's Brainware owns the search technology and is focusing on content processing.

According to a company official, "more than \$100 million has been invested in Brainware's technology.

"We grew at a rate of 50 percent in 2007, and if our estimates are accurate, we will double by the end of 2008," James Zubok, the company's CFO, told Beyond Search. Not surprisingly, Brainware is chock full of smart people. Werner Voegeli, Director of Research and Development, and his flock of PhDs from the University of Oldenburg in Germany continue to support the "associative memory" technology. The company's approach is patented and discussed briefly in another section of this profile.

Item	Quick Facts
Product	Brainware
Price	~\$500,000 (enterprise edition)
Technology	Natural language search based on the patented "associative memory" technology
Key Feature	System identifies relevant documents using the digital DNA identified by the associative memory procedure
Purpose	Allow rapid identification of relevant information from documents and information that may not be processed by traditional content processing techniques
Clients	Law firms, SirsiDynix, Reynolds & Reynolds
Company	Brainware, Inc. Privately-held
Contact	Yegor.kuznetsov@brainware.com

Table 10: Quick Look at Brainware, Inc.

Brainware has set up shop in suburban Washington, D.C., not far from Dulles Airport. In the U.K., Brainware has an office near Oxford and one in Nottingham. There's an office in Neuchâtel, Switzerland, a small city in easy reach of Zurich and Geneva.

"It is encouraging for us to be able to expand our offices so rapidly," Carl Mergele, Brainware's CEO, told Beyond Search. "We are riding a wave of demand for tools that help companies automate the process of extracting information from the high volume of content that flows through organizations today."

The company's products are powered by Brainware, which, asserts the company, is "the world's only engine that does not rely on exact definitions." Word definitions means word recognition."



The advanced search interface for the Brainware system provides a list of named entities and a graphic display of the relevancy of the document set. The system includes a built-in document viewer.

The system includes “an intelligent” data capture component. If you have scanned pages, the system can transform a PDF into ASCII. The Brainware system then processes the text. The data capture component is versatile, able to convert faxes or email. As the system processes textual information, the system categorizes the information and extracts information from these documents. For example, the Brainware content processing system can process and make searchable invoices, loan applications, and similar problematic documents. The Brainware system includes workflow tools. You can scan paper documents or process Adobe PDF files and send that data to another enterprise application. Brainware can handle authorization and other procedural functions.

According to the company, some of its customers require a tightly-integrated document capture, conversion, and retrieval system, which Brainware offers. The technology can

be implemented with only the specific features a licensee requires. Brainware can be deployed as a desktop search solution or an enterprise-wide content processing system.

Brainware's technology allows it to recognize and find data through inexact definitions, patterns and context, mimicking the way the human brain processes and sorts information. The company processes data and information in databases, emails, document archives, and images.

The search function provides structure-free access to processed information in databases and unstructured content as well as images. Brainware shares some similarities with ZyLab, a company profiled in this study.

Associative Memory: The Brainware Innovation

Brainware's technology is one of the most interesting approaches discussed in this study. Most of the companies providing key word search and sophisticated content processing create inverted indexes of terms.

Brainware takes a different approach and then uses a more traditional inverted index for certain, special situations. The foundation of Brainware's approach is the rough equivalent of creating a digital fingerprint that represents the document. Like a fingerprint, software can process the digital representation quickly, while performing necessary analyses.

The company describes its patented pattern matching technique in this way in its January 2006 US patent:

[The procedure requires] coding each document or a part of it through a corresponding feature vector consisting of a series of bits which respectively code for the presence or absence of certain features in [the]... document; arranging the feature vectors in a matrix; generating a query feature vector based on the query document and according to the rules used for generating the feature vectors corresponding to the stored documents...; for those columns of the matrix where the query vector indicates the presence of a feature, bitwise performing one or more... logical operations between the columns of the matrix to obtain one or more additional result columns coding for a similarity measure between the query and part or the whole of the stored documents; and said method further comprising... retrieval.¹³

The idea is to create a numerical lattice of ones and zeros. Through matching and other mathematical techniques, Brainware discerns patterns, identifies entities, and performs automatic concept tagging and classification. The speed of the system results from a combination of optimized code and use of hardware memory registers so one and zeros

¹³ USP6983345, "Associative Memory", January 3, 2006. See also US6976207, "Classification Method and Apparatus", December 13, 2005, and WIPO WO/2003/044691, "Method and Apparatus for Retrieving Relevant Information," May 2003

can be flipped at processor clock speeds. Exegy, profiled elsewhere in this report, uses hardware to achieve higher performance, as well.

In short, Brainware creates a digital DNA identification for documents and its constituent elements.

Features

Brainware is an intelligent system, able to learn as it processes content. The system can, for example, be set up to request assistance from a human if it encounters an unknown object. Despite its ability to learn, the system can make use of available knowledge bases, taxonomies, and controlled term lists.

The search system comes in enterprise and personal editions. Both permit high-speed retrieval of information from centralized corporate repositories. The system can access Lotus Domino, mail servers, file servers, databases, the Internet, and more. It can also retrieve information from over 250 file formats including Adobe Acrobat PDFs, word processing documents, scanned document images, and spreadsheets.

The system supports a wide range of third-party applications and their content outputs, for example, Oracle, SAP, IBM Lotus Notes, and enterprise resource planning environments.

The system includes APIs that support Microsoft Component Object Model and its variants, Java, and SOAP (Simple Object Access Protocol). The content processing system can be integrated into third-party applications and used to power Intranet portals.

Feature	<i>Beyond Search</i> Comment
Knowledge Base Support	System can operate automatically or use local or remote knowledge sources, including Wikipedia, the CIA Factbook, and similar sources
Query Types	Key word, phrase, segments of documents, entire documents
Visualization	None. Third-party rendering systems can be integrated with Brainware via the API
Entity Extraction	Entity extraction and other object manipulations are supported
Platforms Supported	Microsoft Windows
Export	XML and other formats supported
Third-Party Support	Can be integrated with third-party systems including content management systems, databases, and ERP systems from JD Edwards, Lawson, etc.
Vertical Support	The scripting language and the API permit integration with almost any vertical applications
Analytic Functions	Log files and usage reports

Table 11: Technical Highlights for Brainware Inc.

Querying

You can interact with the system in different ways. For example, you can use a “fuzzy query”, which delivers correct results without requiring exact matches, in spite of OCR errors, misspellings, and other inconsistencies. You can enter a phrase, a paragraph, or an entire document. The system permits exact matching queries. My tests reveal that the system delivers its most useful results by copying a paragraph from a relevant document and letting Brainware use that extended chunk of text as a query.

Content Processing Functions

The technical result of Brainware’s implementation of associative memory includes these rich text processing features:

- Fuzzy, phrase, sentence, paragraph, keyword and exact search capabilities. A user can enter a word, phrase, or an entire document as a search statement. An “assisted navigation” interface allows the user to point-and-click on topics, entities, and concepts.
- Fuzzy logic: Search logic that helps information seekers avoid limitations of search results due to simple misspellings of queries, plus helps searchers discover additional “did you mean?” results.
- Categorization engine: Brainware can discover categories, and it can process content such as the body of information in Wikipedia. The system can then automatically build relationships among other data such as a formal thesaurus or specialized word lists and taxonomies such as those available from the National Institutes of Health or any other source.
- Ability to search range of data sources and formats. The system can process information in more than 250 formats. Among the file types supported are community/social networking data (for example, user tags and reviews), tables of contents, book reviews, digital collections, crawled Web content, and specialized, third-party content from Factiva or LexisNexis.
- Stateful, URL-based searching: Enables information seekers to build “saved searches” and supports creation of external links
- Full-text document searching: Ability to search full texts of documents, serving as search engines for libraries’ growing digital collections or for content obtained in the legal discovery process.
- Look and feel flexibility: CSS templates offer the look and feel of SirsiDynix public library interface products (HIP 3.x, iBistro/iLink and Enterprise Portal Solution), making it easy for current customers to add new search solutions to existing library interfaces with minimal disruption to library users; an administrative interface will make it possible for sites to design their own customized templates
- Single- and multi-byte character support: For libraries with records in a range of languages and whose users search in a range of languages
- Support on local servers or via SaaS / ASP hosted solution: New search solutions available on sites’ local servers or via SirsiDynix’s world-class

SaaS/ASP hosted operations, through which SirsiDynix manages all hardware and software support, maintenance, and upgrades — leaving libraries to focus on providing outstanding user experiences for information seekers

The System in Use

Brainware lacks the profile of better-known systems offered by Autonomy and Endeca, for example. Nevertheless, Brainware has landed a number of high-profile companies. These include commercial and government entities.

One licensee is the U.K.'s Her Majesty's Prison Service (HMPS). The agency selected Brainware's Accounts Payable solution—A/P distiller—to automate HMPS invoice processing operations. Brainware has allowed the HMPS to reduce the costs of its accounts payable function. Brainware told *Beyond Search* that HMPS has become “a business-value generator.” After shifting to Brainware, HMPS “is saving taxpayer funding that can be reallocated to further its mission.” Brainware permits high-volume line-item extraction and verification and the automation of matching of invoice data against purchase orders.” Savings in the six figure range are anticipated.

The Reynolds and Reynolds Company adopted the solution to rapidly retrieve critical data across all information repositories. R&R is one of the leading providers of software and services to automotive retailers. Brainware allows R&R to make a highly specialized, source-independent knowledgebase available to employees and authorized users.

Upside

The upsides of the Brainware content processing solution include:

- Quick deployment. For content processing, the system can be installed in less than one day. For modest content collections, the “personal edition” might be sufficient. For fast-cycle jobs larger jobs on a tight deadline, Brainware offers a hosted service.
- The company provides technical support. With offices in Europe and the U.S., the company is able to respond quickly to on-site calls. As of January 2008, the company was adding technical staff and reported no delays in responding to client requests for engineering support.
- The company provides documentation, training, and technical support to licensees.

Downside

Brainware has marketed by focusing on specific companies in legal, finance, and services sectors. The company is expanding its sales and marketing effort in 2008, but the low profile of the company may be a factor in some procurements. Other downsides include:

- An approach to content processing that is interesting and promising. However, it is different from the key word-based systems with which you may have the most familiarity.
- The company is growing rapidly. Despite a hiring push, at times the company's executives can be difficult to reach. Like other content processing firms, there may be times when the engineer with a particular expertise may not be reachable quickly.
- The bundling of scanning, work flow, and functions that hook into such enterprise applications as accounts payable are of great value. However, you may find that some extra support is needed to decide what components of the Brainware offering are appropriate for your needs.

Net-Net

Brainware is an effective discovery tool. The system makes finding documents relevant to a particular subject quite easy to locate. Lawyers and competitive intelligence professionals will find much to like in the Brainware system.

Unlike some content processing systems, Brainware wants its licensees to learn how to integrate the system into enterprise environments. There is a modest learning curve, but for a complex system, Brainware's product is straightforward from the system administration point of view.

In my tests of the system, I was delighted with the system's ability to process multiple email repositories, index both the content of the message, generate metadata, and process attachments. Email "stores" for individual users pose few problems to content processing systems. However, when a system must process email from dozens, even thousands of employees, the process in many systems is thrown off track by compressed files, obscure file types, or links to external content. Brainware's content processing system handled these problems with aplomb.

The company's growth has come via word of mouth and from organizations frustrated with better-known systems. These customers have had to find Brainware. The company prefers to grow organically. I would suggest that law firms, analysts, and organizations dealing with problem content test the Brainware system.

5. Cognition Technologies, Inc.

www.cognition.com

Cognition Technologies, Inc. is a privately-held search technology company, based in Los Angeles, California. Like Powerset and several other next-generation text processing companies, Cognition asserts that its technology is able to deliver much higher relevancy and recall within search results than is possible with traditional search technologies. Cognition, unlike Powerset and others, has a robust proprietary dictionary, ontology and thesaurus which includes virtually all of the words and phrases within the common English language, and it has customers, including the highly-respected LexisNexis Concordance™.

Cognition's executive says that their engineers have crafted a technology that is "the next evolution" in search. That remains to be proven, but Cognition, like a number of rich text processing companies have jumped into advanced search with verve. However, compared to other search newcomers, Cognition has uniquely solved one of the biggest hurdles toward increased precision and recall by understanding both the meaning of user queries and the searched content. Through this understanding it is able to resolve both the ambiguity and synonymy of the English language.

Item	Quick Facts
Product	CognitionSearch
Price	Between \$15,000 and \$500,000, based on the amount of data indexed. Custom quote required.
Key Feature	Proprietary linguistic knowledge base that enables the system to "understand" word, phrase and concept meaning within a query and document set
Purpose	Improved precision and recall
Clients	LexisNexis, litigation support companies, life science companies
Company	Cognition Technologies, Inc.
Contact	learnmore@cognition.com

Table 12: Quick Look at Cognition Technologies, Inc.

Through its patented linguistic meaning-based search architecture, known as CognitionSearch, the system delivers significantly greater numbers of relevant search results than is possible with currently used search technologies.

Linguistic Approach

Cognition's linguistic meaning-based search technology, CognitionSearch, employs a unique mix of linguistics and mathematical algorithms which has, in effect, "taught" the computer system the meanings (or associated concepts) of nearly all the words and the frequent phrases within the common English language. More remarkable is that the

system incorporates algorithms that replicate a bit of the psychology a human uses to understand words and content.

CognitionSearch--unlike Autonomy, Endeca, and Fast Search & Transfer--does not rely on mathematically-based pattern-matching technology. These systems still perform string matching to locate a particular word pattern. Cognition Search “understands” the meaning of words in context; in both the query and in the document base. As a result, Cognition Search delivers results that are more precise and relevant. Cognition’s executives assert that it offers the only commercially available linguistic search engine on the market, a claim with which Inxight and Teragram.

CognitionSearch employs true natural language query capability, which means that a user can simply enter a statement or question in the query box without the need for complex Boolean expression. As a matter of fact, in order for a user of CognitionSearch to retrieve the best results, users are asked to enter their queries in complete sentences with appropriate capitalization. Regardless of how the user frames his query and how the answer was written in a source document, CognitionSearch finds the desired material. Instead of a laundry list of semi-relevant or irrelevant results, CognitionSearch delivers a more complete and precise answer.

Technology Background

CognitionSearch is an incarnation of technology with roots in research germinated at IBM’s artificial intelligence project, and additional work done for the Department of the Army. Dr. Kathleen Dahlgren and her colleague Professor Edward P. Stabler, Jr. received a US patent for the technology used in Cognition’s system. The invention “Natural Language Understanding System” was filed in October 1997, and granted in August 1998. This patent includes 30 patent claims making Cognition an early player in NLP.

Over the past 15 years, the underlying semantic technology has been adopted for several applications, most recently as applied to content search within enterprises, applications and on the Web.

Management

Dr. Kathleen Dahlgren, the founder of the company, has been involved in a number of search- and NLP-based systems. Prior to becoming an entrepreneur, she worked at IBM’s Los Angeles research facility (focused on artificial intelligence engines) after receiving her Ph.D., and then she helped start ITP (Intelligent Text Processing). The Dot Com collapse in the late 90s shuttered that firm. Two years later, Dahlgren and a team of computer scientists and linguists formed Cognition Technologies.

Investors in the company include the Tech Coast Angeles, Scott Jarus (the company’s CEO), Tim Draper and other notable angels and VCs.

In January 2006, Scott Jarus, former chief executive of j2 Global Communications, the billion dollar public company that developed and markets the eFax service (fax to email), joined the company as both an investor and CEO. Mr. Jarus hopes to supply the

business and strategic knowledge needed to leverage the technology into the next big thing in the search arena. Mr. Jarus, who left his position as president of j2 Global Communications, Inc. to join Cognition, was in 2005 the Ernst & Young Entrepreneur of the Year for Media/Entertainment/Communications.

CognitionSearch™ Beta
Data • Information • Knowledge • Understanding

Home Government ? Help

fatal fumes in the workplace

For best Search results, use a plain English phrase or sentence with appropriate word capitalization.

Send Feedback Search

The following word meanings were selected.
Use dropdown menus to change meanings.

fatal	1) adj: causing death: he has a fatal illness
fume	1) n: a vapor: The fumes from the auto shop dulled her senses.
workplace	1) n: the place where someone works

Submit Changed Meanings Advanced Query Edit

fatal plus fume plus workplace: 7 files — displaying 7 results

1. **UNITED PILOTS ASSN. v. HALECKI, 358 U.S. 613 (1959)**
The administratrix of the estate of Walter J. Halecki brought this action against the owners of the pilot boat New Jersey to recover damages for Halecki's death, [...]
<http://caselaw.lp.findlaw.com/cgi-bin/getcase.pl?court=US&vol=358&invol=613>
► The following stems were matched exactly: fume
2. **WILMINGTON STAR MINING CO. v. FULTON, 205 U.S. 60 (1907)**
On January 27, 1901, Samuel Fulton, while working as a trackman and mine laborer in a mine operated by the Wilmington Star Mining Company, in Grundy county, Illinois, [...]
<http://caselaw.lp.findlaw.com/cgi-bin/getcase.pl?court=US&vol=205&invol=60>

Figure 28: Cognition's Search Interface

Cognition's search interface uses tabs and horizontal "panes" to allow the licensee to interact with the content processing subsystem.

Examples of the System in Use

Cognition has become the advanced search engine for the LexisNexis Concordance application, a litigation case-management software service used on more than 65,000 desktops. In the life science area, the first enterprise target is a large university medical school with whom the company plans to further augment its life science-specific terminology, including those driven by the Human Genome project. Within the coming weeks, Cognition will be launching a series of specialty Web Search portals intended to bring its linguistic meaning-based search technology to professionals and consumers on the Web. The first two Websites will be <http://MEDLINE.cognition.com> and <http://WIKIPEDIA.cognition.com>. Future Web content will include the USPTO full text data base, news feeds and other deep content best served by a deep search engine, such as CognitionSearch. The company may brand these websites: "SemanticPATENTS, SemanticMEDLINE, and SemanticWIKIPEDIA" to communicate the meaning based access to each of the datasets.

The Knowledgebases

The unique aspect of Cognition's technology reaches back to finding a way to reduce the computational burden usually imposed by natural language processing (NLP). Even with the plummeting cost of hardware and storage, deep pockets are needed to tackle large volumes of content. Not surprisingly, NLP vendors have faced Sisyphean tasks to make sales.

Dr. Dahlgren asserts that when CognitionSearch is compared to conventional pattern-matching search tools, CognitionSearch's technology can significantly increase both precision (fewer but more relevant hits unrelated to what you want) and recall (more results on target to your query). CognitionSearch better understands the meaning of the query. For example, in a query related to energy legislation, CognitionSearch knows that energy bill does not mean an invoice for electricity services and that senate bill, SB 47, and senate initiative may also be relevant.



MR. JUSTICE STEWART delivered the opinion of the Court.

The administratrix of the estate of Walter J. Halecki brought this action against the owners of the pilot boat New Jersey to recover damages for Halecki's death, allegedly caused by inhalation of carbon tetrachloride fumes while working aboard that vessel. The action, based upon the New Jersey Wrongful Death Act, N. J. Stat. Ann. 2A:31-1, was brought in the federal court by reason of diversity of citizenship. Under instructions that either unseaworthiness of the vessel or negligence would render the defendants liable and that contributory negligence on the part of the decedent would serve only to mitigate damages, a jury returned a verdict for the administratrix, upon which judgment was entered. The Court of Appeals affirmed, holding that the New Jersey Wrongful Death Act incorporates liability for unseaworthiness, as developed by federal law, and adopts the admiralty rule of comparative negligence when death occurs as a [358 U.S. 613, 615] result of tortious conduct upon the navigable waters of that State. 251 F.2d 708.

For the reasons stated in *The Tungus v. Skovgaard*, decided today, ante, p. 588, we hold that the Court of Appeals was correct in viewing its basic task as one of interpreting the law of New Jersey. For reasons also stated in *Tungus*, we accept in this case the Court of Appeals' determination of the effect which New Jersey law would accord to the decedent's contributory negligence. But even if the Wrongful Death Act of New Jersey be interpreted as importing the federal maritime law of unseaworthiness, the court was in error in holding that the circumstances of this case were such as to impose liability under that doctrine.

Figure 29: Cognition Highlights Multiple Search Concepts

Cognition displays the selected result with the key passage highlighted. Each of the search concepts appears in a separate color to permit fast scanning.

The company's meaning-based linguistic search technology is an ambitious undertaking. Cognition's approach is based on computational linguistics, and is an attempt to mimic, as best as is technically possible, a human's understanding of language.

CognitionSearch introduces a model of natural language semantics with a complex relation between entities, linguistic expressions and meanings. CognitionSearch's lexicon adds commonsense or "naive" knowledge, wherein word meanings (concepts) are naive beliefs, as in "lemons are typically yellow, but some lemons are brown". CognitionSearch indexes very large textual databases and scales to an indefinite size document base.

Cognition's approach is purposefully built to reduce the computational resources required for NLP and scales to accommodate increasing size of data sets such as those contained on the Web.

The system uses modules that deliver specific functionality to the system. For example, CognitionSearch content processing relies on:

- A naive semantic lexicon that permits the system to reason about the meaning of words and phrases
- The system classifies concepts by considering how words and terms in a document relate to the knowledge bases included in the system.

One of the interesting features of the approach is that the lexicon includes "psychologically-motivated representations of human concepts." and extensible common sense knowledge. A licensee can make the system smarter.¹⁴

Most Search engines don't understand that the same word can have multiple meanings (ambiguity). Therefore, they return many false positives and miss relevant information.

Cognition's patented technology combines formal linguistic algorithms with semantic representations to create a "naïve" semantics that speeds up the computational parsing.

Key Features

Several search systems contain thesauri and taxonomies. Convera and Oracle provide vertical term lists to provide licensees with a way to begin processing text without a drawn out editorial exercise.

CognitionSearch includes a bundle of knowledgebases--what the company calls computational dictionaries. These data sets eliminate the time and cost of building word lists, taxonomies, and ontologies.

Cognition takes this approach one step further. The Cognition system includes:

- 506,000 word stems. Stemming or truncation facilitates clustering
- 536,000 concepts or what Cognition calls "word senses"
- 17,000 ambiguous words selected because each is frequently used in English
- 7,000 nodes for the tree structure of the taxonomy
- 191,000 multi-word phrases
- Over four million semantic contexts useful for disambiguation.

Licensees may modify or amplify these knowledgebases. Recall that CognitionSearch has been designed to work efficiently by eliminating the CPU-hogging recursive processes that other systems require for rich text processing.

¹⁴ See "Natural Language Understanding System," August 11, 1998, US5,794,050

Enterprises employing the software are provided with tools to add their own specialized terminology, e.g., product name lists. In the case of very large term expansions, Cognition will supply a professional service to assist the customer with implementation of CognitionSearch.

It has several linguistic components to analyze text at many levels from tokenization to sense disambiguation.

The Advanced Search mode for CognitionSearch offers five basic search approaches: plain English search, linguistic Boolean search, quoted (or phrase) search, pattern search, and fuzzy search (a variation of the pattern search). The Advanced Search mode will seem familiar to professional searchers who have a lot of experience dealing with database services. The complexity and field searching approaches may seem new and somewhat difficult to end users, however.

Over the company's 20 year history, Dr. Dahlgren has lead a team of 12 Ph.D. linguists, 19 specialists with advanced degrees, and numerous computer scientists to build out the largest computational dictionary known to exist.

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	Ships with several knowledge bases designed for natural language processing
Query Types	NLP, keyword, Boolean, and concepts
Visualization	No, but third-party tools may be integrated with the system
Entity Extraction	Yes
Platforms Supported	Linux, Unix, Windows
Export	Text
Third-Party Support	Third-party tools may be integrated. No native support for third-party applications
Vertical Support	Currently legal, life sciences, financial, insurance, energy, computing, military, entertainment, and travel, however, the company is launching many general consumer Web Search portals in the coming months
Analytic Functions	No. Third party tools may be integrated.

Table 13: Technical Highlights for CognitionSearch

Upside

One possible upside for users of CognitionSearch is that it allows an organization with a need for intelligence-agency grade text processing to implement an advanced, linguistics-based search system.

Life sciences, pharmaceutical, financial services firms will be able to process large volumes of textual data, obtain on-target search results, and gain an information advantage over firms using more traditional search-and-retrieval systems.

Downside

At this point, one clear difficulty lies in the lack of an option to display results in a reverse chronological order on all document bases for those interested in the most recent work in a field. Relevance ranking is the default display mode offered, though sophisticated users could conduct a series of Advanced Searches in some files using the date field. The system does not support non-text content.

Net-Net

Beyond Search believes an intelligence-centric organization will want to test CognitionSearch to determine its effectiveness in providing deep search within complex deep content.

6. Connotate Technologies

www.connotate.com

A Content Processing Riff

Connotate has its roots at Rutgers University and Dr. Tomasz Imielinski's interest in music – hard rock, to be more precise. A computer scientist, entrepreneur, and lead vocalist for The Professors, Dr. Imielinski recognized almost a decade ago that key word search was not appropriate for some enterprise information tasks. Boundaries had to be pushed; new ideas implemented. Dr. Imielinski told *Beyond Search*:

Search is more difficult to master than music – even rock. Our approach blends some unusual elements. We had early support from the U.S. government's Defense Advanced Research Projects Agency (DARPA) and the private equity firm, Trautman Wasserman & Company, Inc.

Item	Quick Facts
Product	Agent Community GEN2
Price	Starting at \$50k - \$125k
Technology	Built on the Microsoft .NET framework and runs on the Windows operating system
Key Feature	Connotate's machine intelligent software Agents interact with the Web and internal information sources to discern high value information, provide analysis and alerts. Agents deliver to an array of output devices and create a Web 2.0 ecosystem for information access. Agents can be created without programming.
Purpose	To enable end-users to quickly access, share and deliver derivative rich data, new content, and on-demand applications by providing them with a solution for reaching deeper and more accurately into information sources, without the need for programming or IT involvement.
Clients	Goldman Sachs, Dow Jones, Top-tier Hedge Funds, Government Agencies, Leading Global Publishing/Media Firms such as The Associated Press, Reuters, Interactive Data (FTID)
Company	Connotate Technologies, Inc.; privately-held
Contact	+1 732 296 8844

Table 14: Quick Look at Connotate Technologies

Today, this privately-held, Goldman Sachs funded company is on a growth track. It delivers solutions that perform content monitoring, harvesting, acquisition, transformation, integration and content processing. Wall Street and investment banks have demonstrated an appetite for a solution that puts information in an interface that a harried trader can use without special training, without IT involvement or programming expertise. The solution provides a Web 2.0 ecosystem that can ingest

information from on-premises servers and Internet-accessible sources, giving licensees the flexibility to design interfaces suited to particular users in an organization.

Connotate positions itself as a software company that empowers the end-user to quickly create and share on-demand applications. The solution — Agent Community GEN2 — can be used to mashup, monitor, mine and extract user-defined content from enterprise applications and Web sources. The indexed and normalized information is accessed via a search box, a point-and-click assisted navigation interface, or via alerts. These “pushed messages” ensure that a user gets current information regardless of time, device, or location.

Founded in 1999, the company now has a seasoned management team and a streamlined marketing and business development focus. Compared to some content processing companies, Connotate’s value propositions have a Madison Avenue flavor unusual among its competitors. For example, Connotate sums up its system’s benefits by stressing for business intelligence professionals “Harness the value of information from the Web and enterprise” and for financial analysts, “Delivering More Alpha Using the New Research Platform.”

Connotate’s solution can be installed within the enterprise, or is available as a hosted solution. Providing the enterprise with a complete information access ecosystem, the solution enables rapid configuration of on-demand applications that deliver valuable reports and data. Bruce Molloy, the company’s senior executive, told *Beyond Search*:

Connotate’s system obtains information, normalizes it, standardizes it, and makes it available for search or as inputs into third-party applications such as business intelligence systems. What sets us apart from other content processing vendors is the use of machine-intelligent Agents, dramatic scalability, simplicity of use, and the empowerment of the end-user. We have worked hard to minimize the time and effort required by our customers searching, monitoring, retrieving, and analyzing information. Other approaches typically require time-consuming, expensive custom scripting and programming.

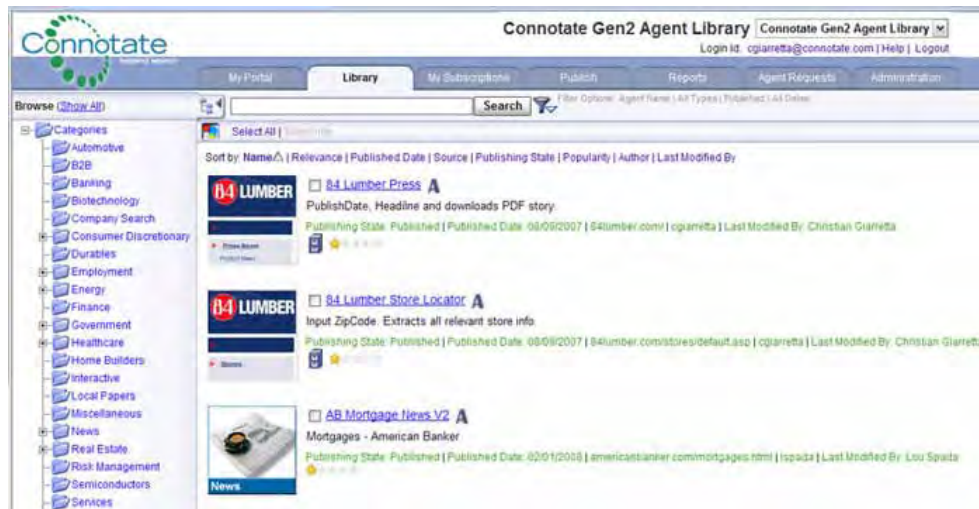


Figure 30: Connotate's Agent Library

The Agent library makes it easy for all community members to share and personalize Agents.

Connotate's Agent Approach

The company's flagship solution – Agent Community GEN2 – makes it possible for the end user to configure machine-intelligent software Agents to perform the functions of acquiring, normalizing, indexing, and supporting search and retrieval. Patented machine-learning algorithms make Connotate's system "smart". The Agents can be trained, often in a few minutes, to do anything a human can do to interact with Web or internal application sources. Configuration and customization is essentially a point-and-click process. Built on Microsoft's .NET framework, there is strong support of Web Services and SOA capabilities. Neither custom scripting nor programming is required for most installations.

"Connotate's patented Agent technology automates content interaction," said Mr. Molloy. "A Connotate Agent can act on the behalf of a user, type in information, search terms, can click on links, can know his/her password and keep it protected. The agents can go get the information needed, bring it back, format it to the user's specifications, and deliver results to a variety of output devices in a number of formats. That's why we say, 'Connotate goes beyond search.'"

Connotate supports semantic and Web 2.0 functionality. The system can accept RSS (really simple syndication feeds). Connotate's APIs are XML centric, making it easy to accept information in almost any format, transform it, index it, and make it available without the manual drudgery other systems impose on their users.



Figure 31: Connotate's Flow

Connotate's solution can generate outputs in a variety of formats. These range from an email containing the needed information, to Excel files that can be manipulated using the analytic functions in the spreadsheet, to mobile alerts.

Agent Community GEN2 includes the following components: Agent Studio, Agent Library, and Agent Server.

Agent Studio

Agent Studio (AS) is a point-and-click Agent creation interface designed for non-technical users such as analysts or researchers. AS makes it possible to create and deploy customized Agents, in essence creating on-demand applications. In organizations where information needs to be created or changes quickly, AS gives the users the ability to create new Agents or tune existing Agents without waiting for an engineer from the information technology department to code the adjustment.

Capabilities include:

- Surveillance and Monitoring of sources for business, market or competitive intelligence, providing a situational awareness snapshot of key companies, topics, products or markets
- Quantitative Analysis and Time Series data collection and correlation
- Data extraction & Aggregation from multiple sources
- Integration among sources

Agent Library

The Agent Library (AL) is, what the company calls, a browser-based hub for collaborative intelligence. AL makes it possible for licensees to share and personalize Connotate Agents. Features and functions can be assigned to a software agent via a graphical interface. AL ensures that different agents can share configurations, thus facilitating collaboration among users and agents themselves. AL provides a single administrative interface for managing agents, functions, and customization.

A licensee can create different agents, which can be subscribed to with a single click. Agents can be further customized with execution schedules, filters, and alerts. AL permits **mashups** – essentially customized reports or information displays – without programming. Mashups and agents can be created for a single user or for a group of users. Extensive set of AL supports such delivery options as email, portals, iFrames, databases, mobile devices and desktop alerts.

Agent Server

An enterprise-level engine for running the system agents, Agent Server handles the execution of agents and the delivery of resulting content. The server includes an administrative interface for managing, analyzing and adjusting Agent executions.

Scalability

The platform includes a scalable server. In addition to hosting tens of thousands of Agents, the server permits monitoring and mining millions of pages on a daily basis. The server provides value-added processing such as advanced filtering, classification, and change detection. The server can be deployed on a single server, a server farm, or as a managed service. The server includes an administrative interface for managing, analyzing and adjusting Agent executions.

Professional Services

One-month QuickStart services that will streamline your evaluation and adoption of Agent Community GEN2 include:

- Connotate Analyst – utilizing a structured approach for efficiently gathering information, Connotate will collaborate with your subject matter experts to identify initial automation and innovation goals for Web and enterprise information.
- Hosting – employing a secure, hosted computing environment, Connotate will create a private community for you, as well as manage the environment, to ensure performance and availability.
- Agent Production – leveraging best practices learned from building armies of Agents, Connotate will produce the initial agents in your community.
- Training – advancing from basic Agent creation to complex approaches for penetrating the deep Web and enterprise applications, and best practices for managing communities, Connotate prepares you to successfully capitalize on the use of Agent Community GEN2.

The System in Action

Connotate's customers consist of large multi-strategy hedge funds, global publishers and media companies, trading firms, Internet-related firms, and Federal and State government agencies. Some of the firm's most recent customers are in the pharmaceutical sector.

Feature	Beyond Search Comment
Knowledgebase Support	The system does not require knowledge bases or controlled term lists. If available, the agents can be configured to use these in the information processes.
Query Types	User-defined or administrator tailored “reports” or outputs to Excel and other applications
Visualization	Third-party tools may be integrated via API
Entity Extraction	Supported
Platforms Supported	Microsoft Windows
Export	Data may be output to XML, databases, files or formats specified by the licensee
Third-Party Support	The system can be integrated into almost any enterprise application or environment via Web services or the Connotate API
Vertical Support	Finance, publishing, government and healthcare
Analytic Functions	Excel and third-party applications may be used for data analysis

Table 15: Technical Highlights for Agent Community Gen2

Goldman Sachs uses the Connotate solution to supply its professionals with content mining technologies to create information tailored to Goldman Sachs’ specific needs. What’s interesting in the Goldman Sachs’ implementation is that the system has been configured to process proprietary internal research, third-party commercial information, and publicly-accessible Web content in one federated system. Goldman Sachs then makes these data available to certain Goldman Sachs’ clients, thus supplementing the traditional analyst notes and reports with a near real-time, live online system.

Upside

An important payoff from Connotate’s agent-based approach to information retrieval is reducing the time a knowledge worker spends hunting for and monitoring certain types of information. If your organization routinely monitors certain companies or a large numbers of sources, you will want to use an automated system such as Connotate’s for this type of job. Humans are good at knowing what they need to make a decision. But humans can be incredibly inefficient performing certain types of routine information-centric tasks. Agent-based systems like Connotate’s are, therefore, one way to make on-point information available at lower cost.

Other upsides to Connotate’s system include:

- Putting end-users in control of the agents via a point-and-click interface to build machine-intelligent Agents without any programming knowledge

- Automatically, transparently combining data from multiple, disparate sources whether that information resides on servers behind your firewall or on the public Internet
- Displaying an instant preview of what the content result will look like
- Supporting individual or group personalization features for agents themselves and for content outputs
- Permitting content sharing among individuals or groups in the organization using Connotate
- Permitting content sharing with individuals outside of the organization via unique URL sharing of portal data

Downside

The Connotate system is a next-generation information processing platform. Key word search has given way to software machines – what Connotate calls *Agents* – that acquire, transform, filter, and present the information a user requires. Depending upon the needs of your organization, you will find the Connotate system a welcome change from the key word indexing and complex content processing procedures used by other vendors mentioned in this study. However, if you want to replace one key word search system with another, you may want to give Connotate a quick look, but concentrate on vendors focused on the search box and assisted navigation techniques.

Other considerations include:

- Making certain that your users will make use of the information and data that the Connotate system obtains, transforms, and delivers. One of the long-standing complaints of systems that do most of the heavy lifting for a user is that the user must invest time in adopting then using the outputs.
- Integrating the Connotate agents into existing enterprise applications is not technically difficult. The challenge will be getting users to change their information habits. Next-generation systems often meet with cultural resistance, particularly in organizations where key word searching is the standard way to find information.
- Managing large numbers of agents is greatly simplified with Connotate's administrative tools. But a system administrator must ride herd on the agents. Without continual, appropriate oversight of an agent-based system, machine and bandwidth resources can be stretched to the breaking point. Connotate's system can operate in the background, but you have the responsibility to allocate adequate staff time for routine housekeeping tasks. However, Connotate's solution provides both system monitoring and reporting so that processing can be appropriately managed and balanced.

Net-Net

For almost a decade, Connotate has been a leader in developing agent-based content processing technologies and systems. The firm's software is sleek, well-designed, and very good at delivering end-user ready information. As the stress fractures become

more evident in traditional key word search and retrieval, Connotate's approach is a pragmatic and innovative departure from what most organizations perceive as behind-the-firewall search.

Beyond Search believes that the Connotate system warrants a test drive. The firm's hosted solution and quick start program makes it relatively painless to deploy a Connotate system. Once you have a grasp of the basics, you will be in a better position to determine how to use the company's agent-based platform.

Beyond Search believes that agent-based systems will become more widely available in the future. Disenchantment with key word based systems and the cost / complexity trade off for content processing subsystems will be one driver. The other reason agent-based systems will be attractive is the need to reduce the amount of time a professional spends looking for basic information. Key word search may never disappear, but far-sighted organizations will want to make information access more efficient and, therefore, less expensive.

7. Dieselpoint Inc.

www.dieselpoint.com

Chris Cleveland, a political science grad with an MBA, launched Dieselpoint in January 2000 from software he had developed at Genesee Development, the Lincoln Park, Illinois-based industrial business technology firm he founded in 1990. Mr. Cleveland sold Genesee's consulting arm in order to run Dieselpoint.

He told *Beyond Search*:

Providing high performance faceted search and navigation for terabyte sized datasets with millions of items is what Diesel-point was designed to do. However, we are often selected, because we are 100% Java with elegant and open APIs.

Rumor has it that Google took a long, hard look at Dieselpoint's technology, then in a *Googley* way was distracted.

The name Dieselpoint refers to the strength and power of what the company was trying to create. One engineering textbook describes the 'diesel point' as the point at which the combination of heat and pressure in a diesel engine causes ignition. Dieselpoint intends for its software to ignite possibilities for its customers too.

Item	Quick Facts
Product	Dieselpoint 5.x
Price	Begins at \$100,000
Key Feature	XML and parametric search with navigation
Purpose	Search and faceted navigation for structured and unstructured data
Clients	Waterstone's, Federal Reserve Bank of New York, and Northrop Grumman
Company	Dieselpoint Inc.
Contact	sales@dieselpoint.com

Table 16: Quick Look at Dieselpoint

Dieselpoint has some high-powered clients with plenty of search experience under their belt. Instead of believing the marketing brochures, these customers have licensed the Dieselpoint technology and dropped their licenses for other, better known advanced text processing vendors. Dieselpoint allows you to manipulate document attributes as well as document text. It can be put to best use as a searching tool for unstructured documents, semi-structured XML or fully structured SQL databases. It can work like a search engine, perform a full-text search and also SQL-like queries for parametric searches. This search offers a powerful full-text search syntax including linguistic tools.

The company focuses on scalable search and faceted navigation, which enables search results to be ordered and classified in multiple ways based on like attributes. Result sets

of any size can then be navigated using dynamically-generated menus. Menus are generated from the underlying document attributes or metadata. These are designed to give users context-dependent browse capability, allowing them to see what options are available to them at each step. Dieselpoint's software is used in a variety of applications including e-commerce, document search, site search, PLM, ECM and OEM. Dieselpoint is a text processing and search system, not an XML database.

Dieselpoint provides text processing and search for enterprise-class applications. System licensees are currently using Dieselpoint for XML search, PDF search, catalog search, and Intranet search, and OEM search applications.

Key Features

The ability to tap into the power of metadata is a key strength of Dieselpoint. All search software provides full-text search capabilities. Dieselpoint Search fully supports full-text search, but the software differentiates itself from other search software products by allowing metadata facets to be exposed in the search interface, enabling guided browsing via dynamically-generated hyperlinks.

One licensee HMV, a U.K. based retail operation, uses faceted navigation to allow customers to do things like drill-down through categories (for example, song genre) to find results, rather than just doing a text search.

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	Dieselpoint uses available metadata, including knowledge bases and taxonomies if available
Query Types	Supports key word, Boolean, and faceted navigation (point-and-click interfaces)
Visualization	None
Entity Extraction	Yes, the system can identify entities
Platforms Supported	Any system running Java can run Dieselpoint, including mainframes and IBM systems running OS/400
Export	Dieselpoint indexes can be exported in XML
Third-Party Support	Native support for Documentum, Lotus Notes, and Vignette, and a half dozen other enterprise systems
Vertical Support	Sample applications are available for ecommerce, document and parts search
Analytic Functions	Yes, a range of reports on system performance are included

Table 17: Technical Highlights for Dieselpoint

HMV users can navigate large information spaces without feeling lost. The HMV interface (illustrated below) guides the user toward potentially-interesting choices. The end result is that HMV customers find information they are interested in quickly and efficiently. Because Dieselpoint is written in Java, the system is accessible from HMV's mainframe-based point-of-sale terminals, a feat few other systems can duplicate.

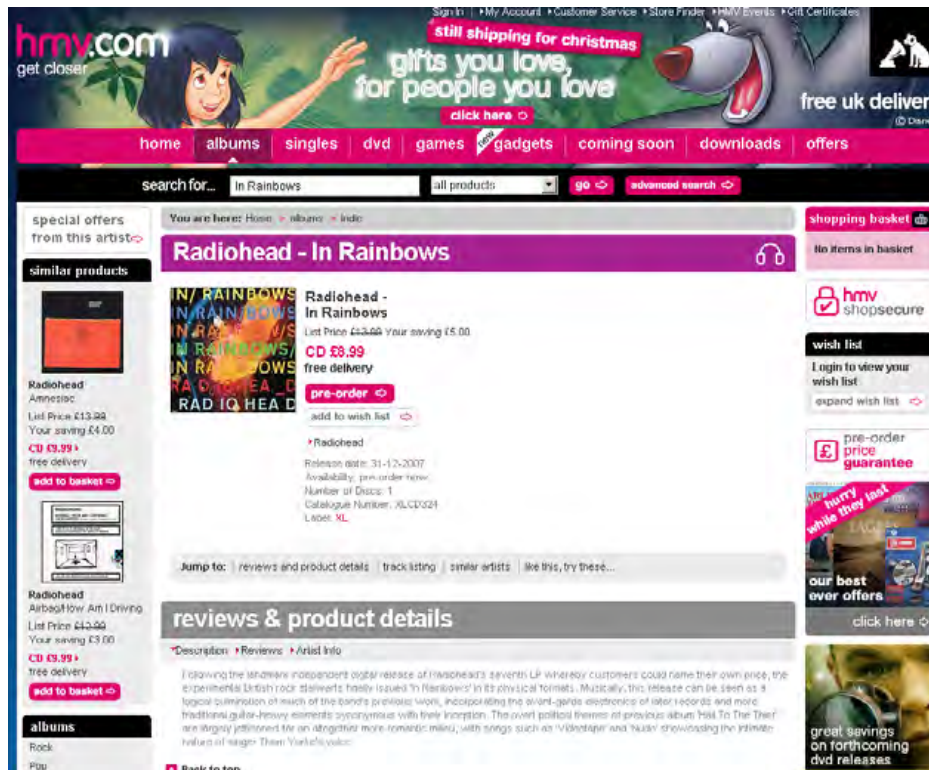


Figure 32: HMV's Dieselpoint Interface

A large UK retailer uses Dieselpoint to handle queries. The implementation allows fast response time and gives the licensee the ability to manipulate text, images, and other content objects.

With this current release, Dieselpoint has enhanced support for taxonomy-driven searches. In search software terminology a taxonomy is a knowledge model organized into a hierarchy of major and minor concepts. In the HMV implementation, song lyrics might be categorized by genre (for example, rock) and each genre may include sub-categories (e.g. *soft rock*, *classic rock*). Dieselpoint uses a taxonomy data type that allows a developer to exploit interrelationships between and among content attributes expressed as metadata. The system includes redesigned internal indices for taxonomy attributes that are automatically generated.

Open Pipeline

Dieselpoint's new Open Pipeline architecture is similar to Autonomy's IDOL and Fast Search & Transfer's ESP. The idea is that a layer of software allows different content sources to be "plugged into" the search system. Some vendors like TIBCO call this an information bus; others refer to it as a framework. Regardless of the terminology, the search system using this technique can crawl data from a variety of sources, process it, and route it quickly and easily. Systems that make it possible to obtain content from different sources are sometimes described as federators or federating search engines. The idea is that the search engine's interface allows access to content from multiple sources through a single interface. This compares favorably with search engines that index only a single server's content.

Open Pipeline implements a publish-and-subscribe model for data feeds. Subject to security and access rules, the system allows an authorized user to subscribe to feeds that support such standards as HTTP, Atom, and RSS.

Throughput

By opening up the process of analyzing, representing, and routing data, Dieselpoint functions as “middleware for search”. One twist in Dieselpoint’s implementation of federation is that the company has engineered parallelization into the content consolidation function. This change increases system throughput which can reach content processing in 500+ megabytes per an hour range when properly resourced and configured. The Dieselpoint approach also includes replication so that queries do not choke the system when indexing and query processing hit peak loads. Keep in mind that these types of features work only if you are running appropriate hardware with adequate bandwidth.

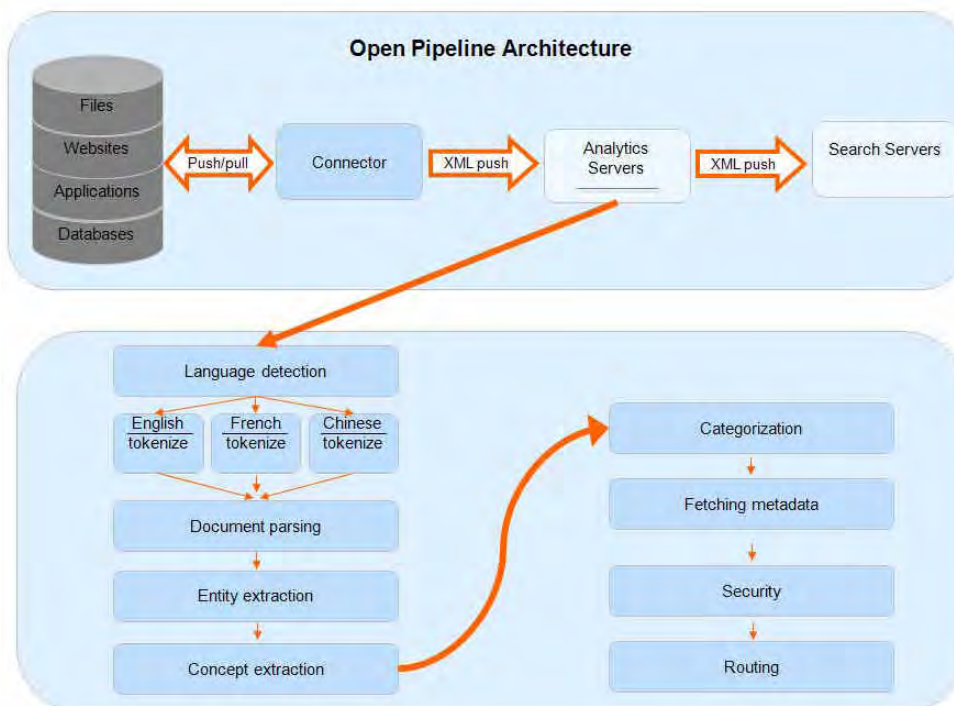


Figure 33: Dieselpoint's "Open Pipeline Architecture"

Dieselpoint's "open pipeline architecture" delivers advanced functions such as entity extraction and categorization.

Connectors

Dieselpoint arrives with a number of software connectors or filters. The current version supports content in repositories or index from:

- ANSI standard databases or proprietary databases that can export comma separated value and other common file types
- Documentum

- Linux, OS/400, UNIX, and other common file systems
- Vignette

JCR (JSR 170) support for:

- Documentum
- FileNet
- Lotus Notes
- Interwoven
- Sharepoint
- OpenText
- Vignette

Dieselpoint's SDK makes it possible for a licensee to create other adaptors, filters, and connectors as required.

Basic Features

A significant number of Dieselpoint installations support electronic commerce and parametric search. However, a growing number of licensees are using Dieselpoint as a federating system to make content from CMS systems, databases, the Web, and various servers in an Intranet available from a single search interface.

Indexing

Dieselpoint indexes documents and data specified by the user and then executes queries against those indexes.

Dieselpoint indexes documents and data retrieved by a crawler from Web sites, directories, and databases. It can index documents (XML, HTML, PDF, Microsoft Office), databases (via JDBC), and flat files (comma-separated, tab-separated, and so on). Data in other formats can be indexed via calls to a user-implemented API. The indexer extracts data in the form of attributes, such as document metadata, XML elements and attributes, and database columns. A preprocessor allows user-written code to modify, categorize, or reject items before they are indexed.

Dieselpoint uses a proprietary query language, which supports full-text and parametric searching. Search clauses can be joined in any way by AND, OR, NOT, and parentheses, and can include comparisons (=, >, >=, <, <=, <>), wildcards, and regular expressions. Full-text features include stemming, thesauri, stop words, misspellings, relevance, hit highlighting, and support for 40 languages and 140 dialects. Search results can be returned as a JDBC result set or an XML document and can be sorted by relevance or attribute value.

XML-specific features include searching by element or attribute and by XML path. (The system indexer preserves the XML hierarchy.) The query engine can return complete documents or fragments, and can also treat fragments of a document (headed by a particular element name) as separate documents. Dieselpoint understands both

ECCMA (an XML language for catalogs) and Dublin Core and provides special processing for both. In addition, it can handle XMP metadata (RDF documents) embedded in PDF documents.

Analytics

The system includes a range of analytic functions. Standard reports include system usage and query data. Other analytic functions can be integrated via the SDK, or third-party analytics tools from Visual Sciences (formerly Web Side Story) or other vendors.

Administrative Interface

Dieselpoint includes an administrator's interface for performing such tasks as managing indexes, defining data sources, and scheduling the crawler. It also contains a Web server and servlet container.

Technology

Dieselpoint's APIs and search platform are written entirely in Java to simplify implementation and enable interoperability within a range of environments.

The system delivers near-real-time incremental index updating. In addition to supporting forty languages it provides stemming for European languages. Additional features include spell checking, synonyms, special weighting of search results, and extensive search administrator reports and controls.

Dieselpoint is one of the first "pure Java" text processing systems. It indexes metadata and field attributes as well as text and documents.

API

The core Dieselpoint Search engine is implemented as a Java library. It has a simple, intuitive, but extremely powerful API. Index and configuration files reside in a single directory structure, making it easy to move and take backups of indexes.

Indexes are stored in a Dieselpoint-proprietary file format. No external database is required. The file format and index structure is fault-tolerant, requiring no rollback or recovery procedures after a server crash. After an unexpected server reboot, a Dieselpoint Search application will simply pick up where it left off, giving you peace of mind that your application is reliable.

The product ships with three ancillary modules: an administrative interface, sample applications, and a bundled JSP/servlet container/ app server suitable for common uses.

The administrative interface makes it easy to create, update, and search indexes. Wizard-like screens capture requirements and write them out to configuration files formatted as XML. The search engine uses these XML files to index data and define system behavior.

Unlike other resource-heavy solutions, Dieselpoint Search integrates easily into Java apps. By default, no configuration is required at all, although several hundred configuration and indexing options are available.

Other

Dieselpoint is written in Java will run in any J2EE-compliant application server. It is designed to be called from a user-written application and its API is designed with such applications in mind. For example, it returns metadata about search results so applications can dynamically create user interfaces relevant to those results. Applications can call Dieselpoint through a Java API, a JSP front end, JDBC, or XML. For users who do not want to write their own application, Dieselpoint ships with a number of sample applications (including a product catalog application) and a generic, JSP-based user interface that is “suitable for common uses”.

Third-Party Tools

Dieselpoint licenses XML parsing technology from the finish company Davoris, the leading developer of Java-based conversion tools.

Dieselpoint in Use

Customers are currently using Dieselpoint for XML search, PDF search, catalog search, and intranet search, and OEM search applications.

AC Nielsen (a unit of the Dutch publishing giant VNU) shifted from Autonomy IDOL to Dieselpoint in 2007. According to Daniel Morse, DPM Technologies, Inc., the integrator working on this “rip and replace” project said at Enterprise Search West in November 2007, “Dieselpoint exceeded out expectations. We implemented facets, clustering, integration, and reporting on time and at a lower cost than the Autonomy solution.” They currently search three million documents in a Vignette repository. It was implemented quickly after a two-month Endeca implementation failed.

A leading defense contractor deployed Dieselpoint’s easy-to-implement solution for a large parts database. More than 1,000 Northrop authorized users use Dieselpoint to find information required for projects.

Upside

The upside for Dieselpoint’s system includes:

- The system delivers a good balance of flexibility, support for enterprise content sources, and performance
- The Dieselpoint Open Pipeline engine makes it comparatively easy to access content from many disparate sources. Unlike federating tools from such vendors as Vivisimo, Dieselpoint minimizes “script fiddling”
- Dieselpoint offers a system that makes the “rip and replace” approach to fixing a rich text processing and search problem feasible

Downside

The downside associated with Dieselpoint includes:

- The company's lack of visibility may make convincing a procurement team to license the technology a difficult job
- The Java-centric approach increases flexibility, but to get the most out of the Dieselpoint system, appropriate resources are necessary
- The low-profile of Dieselpoint may make it difficult for a licensee to find an integrator to handle tuning and customizing tasks. Dieselpoint can perform this work, but the company has a small staff, so you may have to wait to get access to technical specialists. Dieselpoint does assert, however, that they have integrators readily available in both the United States and the United Kingdom, including, but not limited to DPM Technologies, Raritan and Bluetab.

Net-Net

Dieselpoint's functionality is similar to that delivered by Endeca and Siderean, moving from rich text processing into the mainstream of enterprise content access. Users are able to explore search results by pointing and clicking on categories. Because Dieselpoint is parametric, the system understands numbers. Therefore, a licensee finds it trivial to allow a click on a price range to generate subsets within a specific range and perform additional calculations such as generating a monthly payment estimate. Definitely worth looking at for faceted metadata at a lower price than most others.

Dieselpoint is one of those text processing solutions that delight customers savvy enough to find the vendor. Dieselpoint has a low profile, which inhibits its market penetration. The company's technology delivers exceptional value when compared to better known competitors such as Autonomy, Endeca, and Fast Search & Transfer. For ecommerce, parametric, and federating applications, Dieselpoint can deliver comparable functionality, scalable performance, and extensibility and save you \$250,000 or more in license fees.

8. Exalead

www.exalead.com

France has been a hot bed of search innovation, but many of the efforts seem impractical; for example, Kartoo. There are some polished systems – the image search and retrieval from LTU Technologies to the interesting Pertimm system to the little-known Lingway system. French engineers are generating search solutions in as many varieties as French cheese. Some are surprisingly good; others require cultivating one's sense of taste.

Exalead is one of the more intriguing French content processing solutions available. The system boasts a configurable interface and a number of metatagging functions engineered for high-speed content processing. The company's catch phrase is "search by serendipity", which is a clever way of suggesting discovery plus key word search.

Item	Quick Facts
Product	Exalead one:enterprise
Price	Begins at \$50,000
Key Feature	High-speed content processing with automatic classification based on statistical linguistics
Purpose	Index text, database, and rich media content in an organization
Clients	Infonxx, Rightmove, YPG, Alstom, Sanofi Aventis, American Greetings, BNP Paribas, Air Liquide
Company	Privately-held

Table 18: Quick Look at Exalead

The search by serendipity notion, as founder François Bourdoncle explained it to me, is built on the assumption that, "Most people don't know what they are looking for, though they can recognize what they need when they see it. Our system lets a user set out on their 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."

What Exalead does not say is that the system shares some DNA with AltaVista.com. For those of you unfamiliar with the machinations of Hewlett-Packard, AltaVista.com was the search system orphaned by the senior managers of HP when it tried to digest the Compaq Computer acquisition. Compaq had previously bought Digital Equipment Corp, and its management brain trust didn't know what to do with a Web search engine, AltaVista, tied to hardware running the sophisticated and expensive DEC Alpha chip.

Beyond Search: Exalead

Well, Google's founders knew what to do with AltaVista.com's disenfranchised engineers. Mr. Bourdoncle could have joined Google along with Jeffrey Dean and many other high-profile AltaVista.com wizards. Instead, Bourdoncle took the knowledge gained from his AltaVista.com days and founded Exalead. Today, Exalead not surprisingly shares some of Google's performance characteristics - plus innovations crafted by Exalead's Paris-based engineers – and you can see the influence of the founders pioneering clustering work from AltaVista in the company's “search by serendipity” approach.

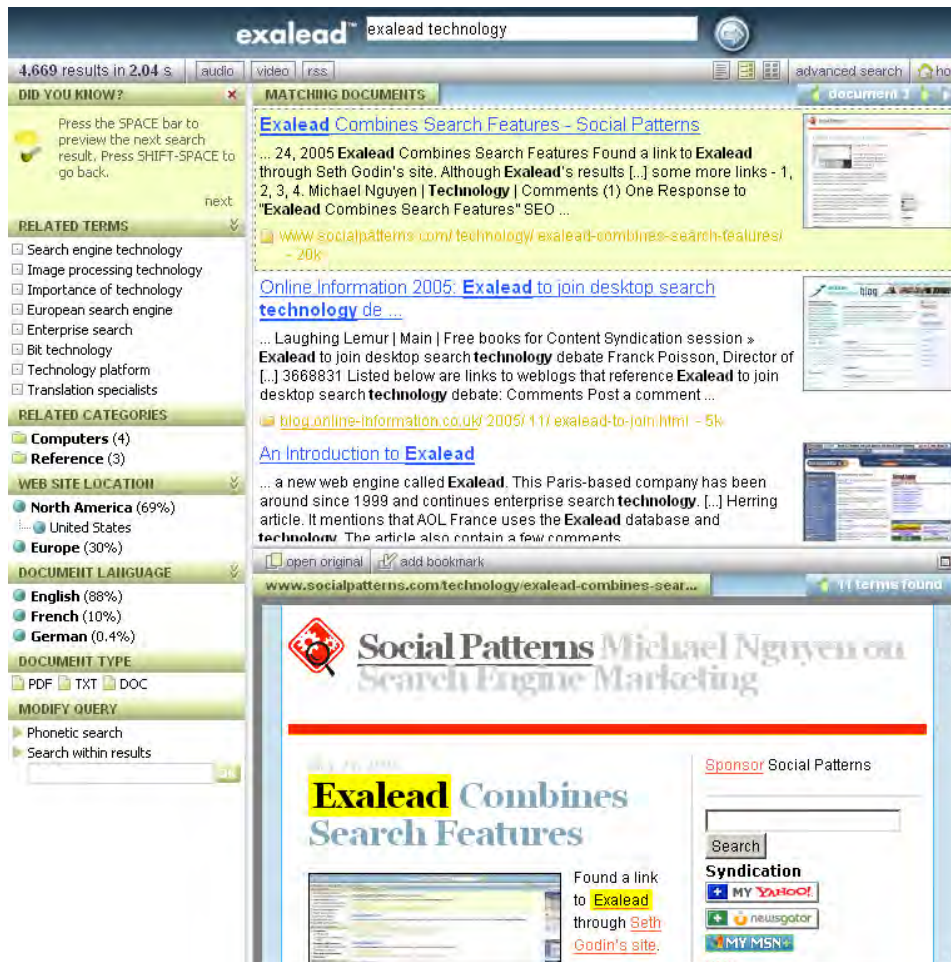


Figure 34: Exalead's Panels

The Exalead interface makes use of panels. The left-hand column shows related terms and other hot linked information germane to the query. The main display shows thumbnails. A click renders the source document below the results.

For behind-the-firewall content processing, Exalead is an interesting solution. Now that their sales and support network is growing in the U.S., the U.K. and elsewhere, it's easier to test that hypothesis. To date, the challenge for any non-French customer will be navigating the cultural maze that the French government and French companies find as logical as a math puzzle. Prospects without sensitivity to these cultural nuances were likely to have found the experience invigorating.

The Company

Exalead was founded in 2000, but work on the system began in 1996. Francois Bourdoncle earned his Ph.D. from École Polytechnique, France's top engineering school, in 1992. Bourdoncle was a colleague of Louis Monier, the founder of Alta Vista.com, an eBay technologist, and now a senior engineer at Google. Bourdoncle left AltaVista.com and returned to teach at his undergraduate alma mater, l'École des Mines de Paris, another top science and technology training ground in France. In addition to guiding Exalead, Bourdoncle continues to teach at l'École des Mines, as well as at the Centre de Mathématiques Appliquées.

Exalead is a privately-held company with about 110 full-time employees. Most are in Paris with a growing contingent in the United States. *Beyond Search* previously estimated that the company would generate about \$10 million in revenue in 2008 as it continues its strong growth. However, we just learned they already surpassed that mark in 2007, hitting \$12 million with a 100% increase over 2006. The company expects similarly strong growth in 2008.

Bourdoncle put up the initial money to create Exalead with two partners, Eric Jeux and Patrice Bertin, and in 2001 accepted an injection of cash from SCA Qualis, a French investment fund, rumored to be in the tens of millions of euros. Exalead will also likely receive cash from l'Agence de l'innovation industrielle (the French government's Agency for Industrial Innovation) when and if the EU approves proposed funding for the Quaero project.

Quaero is, a search technology R&D project dubbed by some a "Google killer". Spearheaded by the French and German governments, and comprised of both public and private research organizations, the initiative has suffered from some momentum swings, with multimedia-centric research currently on the up cycle again. But Exalead has remained unaffected by any bumps in the road caused by the whims of the politicians in Berlin, Brussels, and Paris.

The 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.

Clustering and Math

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

The infrastructure engineering, based on information available to the *Beyond Search* team, shares some similarities with the AltaVista.com approach. 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 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. If you want to have a local installation of Exalead, you can obtain an on-premises license. Mr. Bourdoncle told *Beyond Search* that his engineers have focused on reducing the number of servers typically required to process content in high-end applications.

Real Time Processing

The company uses its computational efficiencies to generate what it calls “real-time dictionaries”; that is, word stemming, identification of word groups (bound phrases like White House), and thesauri that are “fully automatic and incremental.”

One interesting feature of Exalead’s architecture is that when new content is processed by the system, it becomes available to the users in a “few seconds”, as Bourdoncle notes, providing essentially real-time processing. The system also automatically recognizes languages.

Modularity

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 Linux boxes. The system comes with code widgets that can import most common file types.

The core system is Java-centric and exploits XML. Although Java has been characterized by some coding gurus as an end-of-life technology, Exalead has used Java’s philosophy to develop its own scripting language, the Java-inspired configuration language ExaScript. A licensee or Exalead’s engineers can tune the system’s indexing and querying modules using this language and what Exalead calls “standard Java APIs”.

Metatagging

The Exalead system uses both named entities automatically extracted from indexed documents and hierarchical metadata, or categories, as illustrated at the upper right.

One of the firm’s richer interfaces exposes this metadata in an “assisted navigation” presentation. This assisted navigation system has been patented by Exalead in both

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 rely on contributed dictionaries or require human intervention. 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 clustering system.

You can see the Exalead content processing outputs at <http://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 (blogs, forums), language, or file type (e.g., Adobe PDF).



Figure 35: Exalead Assisted Navigation

To summarize the content processing functions, the system provides:

- Natural language processing; that is, lemmatization (stemming)
- Categorization of documents by the available tags
- Dictionaries or word lists aimed at discovery for the corpus; for example, the related terms or 'See Also' suggestions.

Exalead points out to *Beyond Search* that Exalead's public search engines has an index of 8 billion web pages, which they claim is the Web's third largest.

Product Line Up

Exalead offers several different versions of its content processing technology, all of them based on a unified platform called *Exalead one:search*. These versions include:

- one:workgroup
- one:enterprise
- one:datacenter

one:workgroup extends *one:desktop* (a search systems that runs on a user's PC and indexes local files and information for which the user has access rights, such as Microsoft Exchange or Lotus Notes). A test drive of one: desktop provides a good insight into how the workgroup system operates.

one:enterprise

Exalead's enterprise content processing system includes support for structured data. The system supports fielded search, numeric searches, and sorting. Industry-standard database content can be processed by the system. Upon installation, *one:enterprise* can access content immediately from IBM, Microsoft, and Oracle databases. The system includes a crawler so that content from Internet servers can be processed and made available to users. The system includes easy-to-customize scripts for accessing sites requiring a user name and password.

The system includes adaptors to access content from HTTP or HTTPS Web servers (HTML, XHTML, XML, etc.) as well as file systems, LDAP and Active Directory, IMAP, Lotus Notes, Microsoft Exchange, and NNTP message stores, ODBC databases, Microsoft Office Sharepoint 2007, EMC Documentum and eRoom.

The enterprise version includes an administrative interface. An authorized user can maintain taxonomies and word lists used by the system. The automatic functions can be tuned as authorized users add, delete, or cross-reference terms to create 'Use For' and 'See Also' relationships.

The system includes wizards to help reduce the time required for routine tasks such as identifying a collection to process, adjusting hit boosting functions, or modifying the schedule for alerts based on standing queries for specific users.

one:datacenter

The *one:datacenter* "product", optionally available as a managed service, is Exalead's high end offering for clients who need an industrial-grade platform for processing a nearly unlimited number of documents. The system features a redundant, clustered architecture that can index billions of documents in real-time, even on high traffic Web sites. Deployment and administration are centralized.

Feature	Beyond Search Comment
Knowledgebase Support	System can use word lists and taxonomies if available
Query Types	Supports Boolean, free text, phrase search as well as point-and-click access via hotlinked metadata
Visualization	Renders thumbnails and previews of documents. Third-party visualization tools may be integrated via the API
Entity Extraction	Common file attributes; identifies related terms and categories
Platforms Supported	Hewlett-Packard Tru64, Sun Solaris, Microsoft Windows, 32 and 64 bit
Export	System outputs may be controlled via filters created in Exalead's scripting language
Third-Party Support	Lotus Notes and Microsoft Exchange. Third-party applications may be integrated via the API
Vertical Support	None
Analytic Functions	Third-party applications may be integrated via the API

Table 19: Technical Highlights for Exalead

Customers

The company has some high-profile international companies, along 3 types of deployment: B2B2C Market (classified with Rightmove in the UK, online directories like INFONXX and YPG, and e-commerce with American Greeting), Pure B2P (enterprise search with Alstom, Sanofi Aventis, ARF, CapGemini, BNP Paribas, and the French equivalent of the U.S. Department of Homeland Security), and OEM agreements (Exanet, Messaging Architects and H&S).

Upside

Exalead processes content quickly. Exalead offers licensees a quick installation, a platform neutral technology, and customization options via style sheets and an API. The enterprise version of the system processes both structured and unstructured data. Filters and connectors to acquire common file types are included with the system.

Other upsides include:

- A balance of basic search and retrieval and more sophisticated content processing
- Automatic summarization
- Ability to integrate structured and unstructured data, as well as indexing Lotus Notes content

Downside

Exalead is increasing its profile in the U.S., and it has been involved in a number of head-to-head competitions with Autonomy and Endeca. Clearly Exalead's technology puts it in a league where it can compete effectively against blue-chip vendors of enterprise information systems. Once again, the firm's low profile in the U.S. is a drawback. Technically, there's no issue. Over time, the firm's visibility will improve, and as it becomes better known, the company will capture more accounts. Other downsides include:

- The taxonomy and knowledgebase tools are useful, but they are not as mature as systems offered by specialists such as SchemaLogic and Data Harmony, among others.
- The firm's content processing is good, but it is not yet comparable to the deep extraction and tagging functionality delivered by other companies profiled in this study.
- The company has a tendency to discuss opportunities in detail. Although this ensures technical thoroughness (a delightful characteristic of French engineers), decision making can become more Gallic than Silicon Valley.

Net-Net

The Exalead system is a good solution when traditional search is paramount, and users are asking for more search options. The system doesn't deliver a full-blown "assisted navigation" interface like that offered by Endeca or Siderean Software. The company's approach provides useful suggestions within the context of key word searching and traditional lists of relevant results.

Although Exalead's system scales economically, some organizations are less concerned about hardware costs and more focused on getting a solution that meets their particular needs. Exalead permits a wide range of customization and supports integration via its Java-centric technology.

In short, for many organizations frustrated with the cost, complexity, and sluggishness of their existing systems, Exalead merits a road test.

9. Exegy

www.exegy.com

Hardware+Software

Exegy — echoing *exegesis*, a Greek word with its root in *to lead* or *to seek* — offers a hardware-accelerated appliance that extracts previously unknown information from data. Technically, Exegy's products are text mining and data mining devices that leverage stream processing to output information for an analyst or to ingest into another enterprise application.

An appliance is a pre-configured server that can be unpacked and deployed without the delays associated with traditional hardware procurement, installation, and deployment. Exegy, like other appliance vendors, reduces the headaches associated with getting a near-real time, high-throughput content processing installation operational.

Item	Quick Facts
Product	Text Miner
Price	Begins at \$100,000 per year with lease options available. Custom quote required.
Technology	Hardware-accelerated coprocessors, proprietary software and dedicated servers
Key Feature	Entity extraction and metatagging
Purpose	High-speed content processing and high-volume unstructured data searching
Clients	U.S. Federal government, and Financial Services and Large Enterprise integrators
Company	Exegy Inc.
Contact	info@exegy.com or call 314-218-3600

Table 20: Quick Look at Exegy

The Company

Exegy is a privately-held company. When founded in 2004, the company operated as Data Search Systems Inc. The company renamed itself Exegy in January 2005. The company introduced its first commercial product, Text Miner, in April 2006. One of its investors is Washington University. The firm has over 50 full-time employees. The company has offices in Saint Louis, MO, Washington, D.C. and in London, England. University spin outs have an uneven record of success. Exegy's shift from start up to operational company will be interesting to watch. Its technical approach is refreshing because the company's founders have tackled a problem that most content processing organizations ignore.

The company has its roots in Washington University in St. Louis, Missouri. The company's founders have significant experience in developing high-performance systems that use customized hardware and software that comes alive in the special-purpose, dedicated appliances. Exegy's founder and CTO, Ron Indeck told Beyond Search:

Exegy enables the integration and deep analysis of massive data never before considered practically available by processing information hundreds of times faster than conventional systems.

Speed is the distinguishing characteristic of the Exegy content processing solution. Think in terms of processing hundreds of gigabytes per minute without bottlenecks or downtime.

Dr. Indeck's value proposition is that when decision-makers are able to use outputs from near-real-time deep analysis of the data available to their organization, better and more timely decisions are possible. But a plug-and-play piece of hardware is irrelevant unless it is secure and sufficiently flexible to be used by an analyst or another enterprise application¹⁵. Exegy's server has been designed to deliver high-performance content processing that complies with stringent government security and regulatory requirements.

Exegy's approach is to combine content processing and hardware that can transform structured and unstructured information, analyze the data, filter the content, and also provide search-and-retrieval services. Exegy reports that its hardware-software combination processes content at a sustained rate greater than 50 gigabytes per minute, fast enough to chew through the proliferating digital information in most organizations.

The Exegy appliance is "intelligent". When the content volume changes, the Exegy appliance can automatically add processing resources. It reacts dynamically and without intervention by a system administrator. Recall that many enterprise systems choke when the flow of content exceeds the indexing subsystems' capacity. Exegy's system adjusts, thus eliminating the complaints that some enterprise systems cause when bottlenecks bring the system to its knees.

Revenue

Beyond Search estimates that the company's revenues are in the \$1.5 to 2.0 million range. Exegy is a privately held company and doesn't publicly report its revenue. According to Dr. Indeck, Exegy is "pre-revenue, but on the way up." With its appliances starting at \$100,000 per month and some installations requiring three devices, content

¹⁵ Dr. Indeck is director of the Center for Security Technologies at Washington University, where he is the Das Family Distinguished Professor.

processing can cost upwards of a \$1 million per year, excluding professional services and system customization.

Customers

The company's client list includes numerous high-stakes financial firms that Exegy declined to name because of the competitive nature of the business. Its first customer was the U.S. Federal government. After a test, the agency shifted from a supercomputer using commodity hardware to Exegy's servers and experienced a performance boost of about 20 times.

In fact, Exegy works for customers who shun the limelight. However, some information has become available about Exegy's "super secret" installations. For example, scientists at Lawrence Livermore National Laboratory, where nuclear weapon research is on the agenda, are working in collaboration with Exegy to develop high-speed search applications to improve national security. Livermore scientists use the system for processing multi-language scientific and general textual data, fast database insertion, and deep analytics for various research initiatives.

Several financial institutions use the Exegy system to process real-time financial information to support programmed trading and special-purpose quantitative analyses. In automated trading, delays of small fractions of a second can translate to millions of dollars gained or lost. Bottlenecks are not acceptable when the financial stakes or human risks are high.

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	The licensee may reference external libraries of terms, taxonomies, or entities
Query Types	SQL-enabled Boolean or non-Boolean, including regular expressions. This includes a rule-based system
Visualization	Third-party tools may be integrated via API
Entity Extraction	Supported. mathematical operations may be performed on historical data or real-time data
Platforms Supported	UNIX, Linux, Windows. Other platforms can be supported at the request of the licensee
Export	Data may be output in XML or formats specified by the licensee
Third-Party Support	Major financial institution applications are natively supported. API permits integration of the system into almost any enterprise applications
Vertical Support	Finance and law enforcement
Analytic Functions	Mathematical and statistical functions included

Table 21: Technical Highlights for Exegy

Technology

Exegy's approach makes use of hardware and software innovations. The company delivers an appliance or group of appliances with the software pre-installed. Some configuration is necessary, but the system can be up and running in a day or two in most commercial organizations.

Dr. Indeek uses an analogy of a dictionary to explain what is different about the technology. If, for example, someone wanted to look up the phrase *golden retriever*, the person would start by going to the letter g then progressing to the two letter pair *go* then to *gol* and so on. Conventional systems use this sequential approach.

Exegy's technique is to grab the entire word or phrase as a whole instead of parceling it out into individual symbols, thus speeding up the process dramatically. Thousands of such phrases can be used concurrently.

Keep in mind that you will need to edit or create the rules that fuel the system. In many cases, these rules will be Boolean statements, non-Boolean expressions, and may include a multitude of regular expressions such as email addresses, phone numbers, social security numbers, dates, etc. In other cases, you will need to create word lists and code the filters needed to extract the information you require.

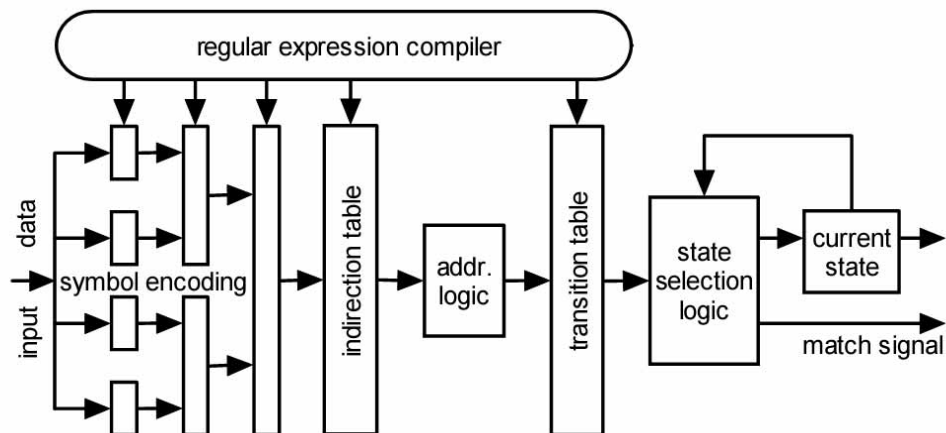


Figure 36: The Exegy System's Data Flow

This diagram from Exegy's 2006 patent USUS7139743 lays out the basic flow of data through the system. Note that the regular expression compiler operates as a meta function, guiding the specific or programmable functions of the appliance as it processes data.

Patent US 7139743 B2: Fusion of Hardware and Software

"Associative Database Scanning and Information Retrieval Using FPGA Devices", filed in May 2002 and granted in November 2006 provides a window into the Exegy approach to content processing.

Be forewarned, the patent consists of 28 figures and 10 pages of explanation and claims. Assigned to Washington University, the invention set forth in this patent is a

fusion of hardware and software. The embedding of certain functions in hardware with other functions provided via software and configuration files sets Exegy apart from most of the organizations profiled in this study.

Exegy's insight is that the bottlenecks that plague many content processing systems boil down to hardware limitations. For example, most vendors allow licensees to run their systems on available hardware or to acquire dedicated hardware for the content processing system.

These systems, unless carefully designed for high-performance, choke when the flow of content exceeds the capacity of the system. Quick fixes are possible, but these often succumb to bottlenecks because the additional hardware suffers from the same technical limitations as the original servers.

Most information technology departments trust their preferred hardware vendor to provide the machines needed to run standard business applications. While this approach is suitable for most enterprise software applications, standard servers lack the engineering required to deal with real-time newsfeeds and the escalating amount of digital information flowing through an organization's network and residing on its departmental and Web servers.

The result is a non-functional content processing system. Vendors of search and rich text processing are often unable to resolve the bottlenecks because the problem is hardware. Software vendors deal with code, so addressing the hardware problem is outside of the search vendors' span of control. The unhappy consequence of this bifurcation of plumbing from content processing software is often dissatisfaction with the search software. In reality, the root of the problem is the licensee's hardware infrastructure. [Fixing a Search System.]

The Exegy patent makes it clear that unless the hardware and software are tightly integrated, and able to reconfigure certain operations on the fly, performance will remain lackluster. Exegy approached the problem by designing a content processing system that binds hardware and content processing software into one homogeneous "appliance".

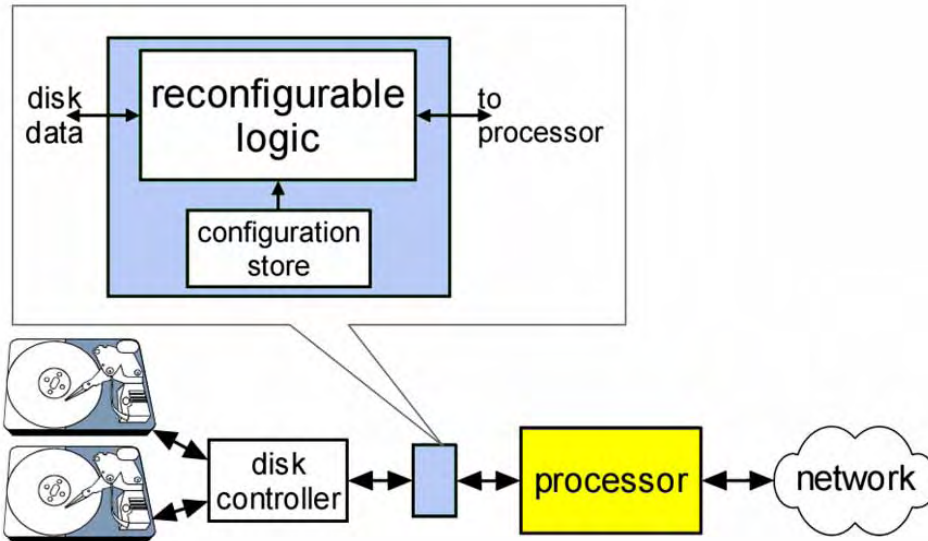


Figure 37: The Exegy Appliance

The appliance contains firmware that makes it possible to reconfigure the system resources on the fly. As content processing demands change, the reconfigurable logic module automatically adjusts the cache to optimize system throughput. This eliminates most bottlenecks associated with rich text processing.

System Gestalt

Exegy's servers use a custom board that offloads certain text and stream processing work from the CPU. The board integrates with other server components to enable high-speed operations for incoming data and for outflowing content outputs. The system includes a library of pre-defined modules that perform indexing, entity extraction, and mathematical operations. Due to this design, Exegy is able to add additional functionality without requiring the licensee to remove the appliance from service.

The target throughput is about 500,000 documents per hour. The server is a three form factor, measuring about 5.5 inches in height. These can be stacked in a standard rack mount.

In an example demonstrated at SuperComputing '07, a year's worth (over 800,000) Reuters news articles were processed in under 4 seconds.

Each server contains random access memory, storage, and central processing units. A persistent cache of more than three terabytes is available.

One key to Exegy's performance is the server-side high-speed persistent cache. The cache may be used for creating in-memory, unstructured operational datastores or special purpose datamarts. If the system is set up to handle real time feeds, the cache may be used to buffer the streaming data. Consequently, there is reduced latency since the content processing subsystem is not choked by comparatively slow disc accesses.

Another tweak is the use of devices that permit seamless scalability. Exegy uses devices that incorporate non-commodity hardware such as high-speed Field Programmable

Gate Arrays (FPGAs) combined with high-performance off-the-shelf hardware. Exegy's engineers have designed the servers to minimize traditional bottlenecks encountered when processing large volumes of data on standard servers.

Exegy's engineers have inserted reconfigurable logic into its server. The firmware intermediates among the disc controller, the CPUs, and the configuration data for the installation. As loads or processing demands change in real time, the logic core is able to reconfigure certain server components. For example, the cache may be allocated to accommodate a particular processing demand. When the peak has been passed, the logic reconfigures the caching system as warranted.

Exegy makes uses of massive parallelism across the FPGAs and across the appliance's CPUs.

The sustained throughput is greater than 50 gigabytes per minute. The appliance can be configured with a variety of interfaces including a gigabit Ethernet and 10 GigE connectors, fiber channel connectors, and InfiniBand connectivity. The server includes an API accessible via C, C++, Java, or Perl.

The Exegy hardware is compatible with IBM's high-speed storage solutions, among others. The system "snaps in" to most storage area networks. Each appliance provides 4.4 terabytes of local high speed RAID storage.

The Exegy hardware has been designed to facilitate the search and other text processing functions of the system. Keep in mind that each Exegy appliance weighs about 100 pounds and requires appropriate power, cooling, and bandwidth.

Product Line Up

Text Miner Software

Text Miner makes it possible to query terabytes or petabytes of content with three built-in search systems. Engines are automatically called depending on the specific content and query passed to the system.

The features of the Text Miner software include:

- Up to 10,000 concurrent term exact matching, wild carding, approximate matching, proximity searching, and pattern matching operations
- Full support for Unicode
- Data encryption and decryption supported
- Support for Boolean queries
- Local and remote data sources are supported

The Text Miner also supports queries with up to 50 regular expressions to identify text that fits a specific pattern. Telephone numbers and credit card numbers can be processed.

Structured Data Miner Software

This is a hardware-software co-design built and optimized for processing data that are structured. The system can handle real-time market feed data such as stock and bond price/volume data, data for ingest into a transactional system, data outputs from a transaction system, and data derived from a database extraction.

The features of Data Miner include:

- Statistical computations can be applied directly to the data flowing through the appliance to calculate volume weighted average price, volatility, variance for stocks and other financial instruments.
- Real-time anomaly detection such as distance of a purchase from a cardholder's home or risk analysis
- Calculations may be simultaneously historical and real-time
- API to permit the Exegy system to be integrated into real-time, algorithmic trading systems or proprietary quantitative subsystems developed for a financial institution
- Outputs may be piped directly into a real-time message system such as those provided by IBM or Tibco.

Outputs from the system may be configured via scripts or style sheets to a look-and-feel appropriate to your organization.

The New Ticker Plant 2.0

Exegy has introduced a version of its system for the financial services industry. Ticker Plant is able to process one million messages per second (MPS) and has been designed to meet regulatory guidelines for brokerages and related institutions, specifically MiFID or the Markets in Financial Instruments Directive and others. A high-performance device is needed because of the rapid increase in data flowing through financial trading systems.

Ticker Plant operates with latency of about 80 microseconds at a throughput rate of two million exchange messages per second. As of January 2008, Exegy combines fast throughput with calculations, exchange authorization, time stamping, and related functions without latency due to bottlenecks in the processing subsystems.

The appliance is the first hardware-accelerated market depth appliance design for traders. The device supports the specific order and market procedures in use at the New York Stock Exchange, NASDAQ, and in major European markets.

The product makes it easy to incorporate an "add on framework". A licensee can extend the built-in statistical and mathematical functions with customized analytics, including an index arbitrage calculator.

The Ticker Plant 2.0 API is available for Windows, Linux, and Solaris. Exegy can customize the API if you are running a different operating system.

The recommended system configuration is a pair of Exegy Ticker Plant appliances side by side for automatic fail over. A third appliance is used as a pre-production staging and test device.

The Ticker Plant offers system backup and real-time support. If a system fault occurs, the Exegy engineers will rectify the situation so throughput is not compromised.

Upside

The upside for Exegy's system includes:

- Application software chooses appropriate hardware-based content processing engine
- On-the-fly configuration makes it possible for each processing engine to be tailored to problem at hand
- Large volumes of structured or unstructured data can be processed with a latency measured in thousandths of a second.

Downside

The downside for Exegy's system includes:

- Exegy, HyperFeed Technologies, Inc., and PICO Holdings, Inc., HyperFeed's parent company, are engaged in litigation arising from Exegy's termination of a Contribution Agreement which provided for Exegy's acquisition of HyperFeed. HyperFeed has filed for a Chapter 7 bankruptcy, and the litigation is pending in Delaware bankruptcy court.
- A query across structured and unstructured content requires multiple Exegy appliances or multiple data passes
- Some overhead is associated with the configuration processes, so switching configurations can slow throughput in certain circumstances.

Net-Net

For real-time, high-volume content processing, the Exegy system warrants a demonstration. Be aware that Exegy is often pre-occupied with its work and, therefore, can be somewhat "interesting". The company seemed to take some extra vacation days in January 2008, making communication "interesting" and hit or miss. The Exegy hardware – software approach is clever. Just allow time to deal with the wizards as you determine if Exegy's system is right for you. Organizations wanting to process a small number of documents will find that other systems provide a less costly solution. However, when the volume of data and the need for near-real time performance are key requirements, Exegy is one of a handful of vendors offering a plug-and-play solution.

10. IBM Corporation

www.ibm.com

Run a Google search for “IBM text mining” and you get a link to IBM Research’s Computational Linguistics and Text Mining Group.¹⁶ Run a query on Google for “IBM enterprise search” and you get links to OmniFind in various versions, including Analytics Edition, Discovery Edition, Enterprise Edition, Starter Edition, and the Yahoo! Edition.¹⁷ Now run a query for “IBM Web Fountain” and you get different hits. Poke around for unstructured information analysis and search and you end up with UIMA, shorthand for **Unstructured Information Management Architecture**.

UIMA is, according to IBM, “an open, industrial-strength, scalable, and extensible platform for creating, integrating, and deploying unstructured information management solutions from combinations of semantic analysis and search components.”

Item	Quick Facts
Product	OmniFind, DB2, Cognos, SearchManager/370
Price	Begins at \$20,000. Custom price quote will be provided by IBM
Key Feature	Extensible, scalable content processing with IBM’s own software or certified partners’ third-party systems
Purpose	Handle industrial-strength content processing
Clients	Virtually all Fortune 1000 companies, U.S. government agencies, leading financial services firms
Company	Publicly traded
Contact	Call the IBM office in your city

Table 22: Quick Look at IBM

UIMA is a framework and software development kit (SDK) for developing text analytics and related applications. For example, a third-party vendor can “hook” into OmniFind and perform additional content processing. Let’s imagine you want to identify entities such as persons, places, or organizations. UIMA makes it possible for the third-party to decompose the application into “components.” One component would handle sentence boundary detection. Another would detect entities. With each component’s interface defined by the framework, the third-party component generates XML that identifies the component. The UIMA framework manages these components and their data flow. UIMA makes it possible for a third party to make components available as network services, thus allowing the components’ functions to scale. The firm’s many deals with

¹⁶ IBM Research is located at <http://www.research.ibm.com/dssgrp/>

¹⁷ IBM OmniFind is located at <http://www-306.ibm.com/software/data/enterprise-search/>

vendors have matured into the UIMA standard. A UIMA-compliant vendor can integrate easily into an IBM environment, such as, WebSphere, DB2, OmniFind, and other components in IBM's massive software arsenal.

Figuring out the sweep of IBM's activities in behind-the-firewall search and content processing is time-consuming. In the last two years, iPhrase's advanced content processing system and text mining system has been integrated into the OmniFind Discovery product.

IBM's influence on other search and content processing companies is significant in a less visible, less publicized way. IBM alums have migrated to Google, Microsoft, and Yahoo!. Some have moved from IBM to academia and back again, influencing future engineers. A few ex-IBM professionals have joined start-ups or started new search and content processing companies. The intriguing InfoDesk is just one example.¹⁸

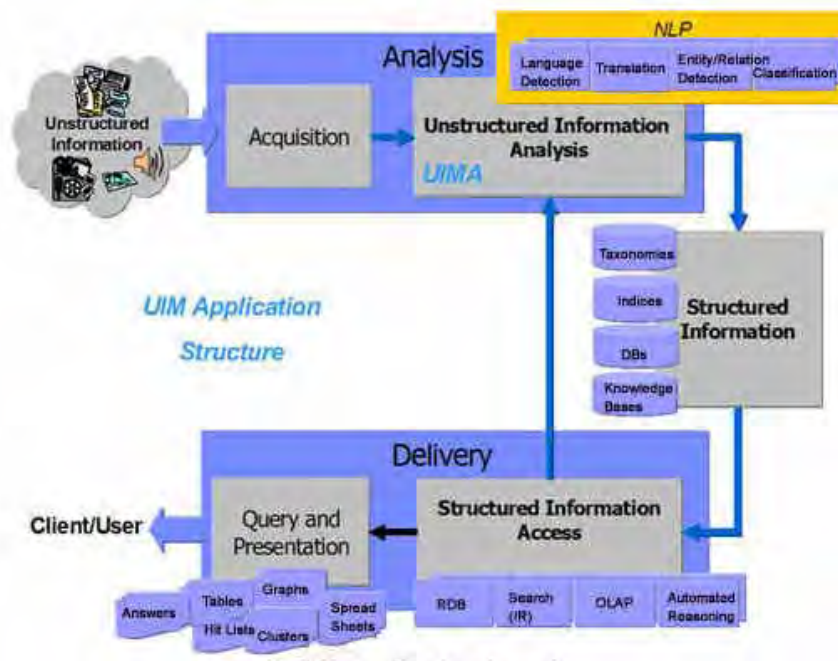


Figure 38: The IBM OmniFind Architecture

The OmniFind architecture uses subsystems for major functions. OmniFind can run in a massively parallel distributed environment, so there are no upper limits on the number of documents or content objects the system can index.

The challenge facing an IBM customer or a person trying to understand IBM's search and content processing offerings is getting a coherent picture of what this \$98 billion behemoth offers.

¹⁸ Sterling Stites managed the development of IBM's "World Avenue", one of the first electronic transaction services available on the Internet.

IBM Content Processing Products

In my analyses of IBM's search and content processing products, until recently, the products were fragmented. At this time, you can find most of the entry-level pricing and product information on a series of graphically consistent Web pages. However, IBM's own Web search systems are often sluggish. IBM has licensed search and content processing technology from such companies as Endeca and Fast Search & Transfer. It is difficult for me to determine whose search engine is used on certain IBM Web sites. When I queried IBM about the search and content processing systems in use, my request was ignored.

You, if you want to deploy an IBM-branded search system, can choose from a number of different versions of OmniFind. Two are comparable to the other systems discussed in this study. Let's look at each briefly.

Start with OminiFind

The easiest point of entry is the OmniFind product lineup. OmniFind is the umbrella "brand" for different builds of the search and content processing system. The basic key word engine seems to be based on Lucene, but I have heard conflicting information. Note that getting IBM engineers to sing from the same song book is difficult. There is a sales-and-marketing layer running interference for product engineers. IBM researchers are just as difficult to track down because of the meetings, the tight scheduling, the travel, and the difficulty of keeping track of the fluid nomenclature, phone numbers, and e-mail addresses.

The basic OmniFind system can handle millions of documents and thousands of users. The system includes what IBM calls "pre-built integrations" for indexing data and content from file shares, databases, collaboration tools, content management systems, Web logs, wikis, and fora. Keep in mind that OmniFind is unabashedly optimized for the Lotus Domino and WebSphere portal environments. This means you will enjoy the smoothest operation when OmniFind is loaded on a true-blue platform with IBM servers, IBM-certified peripherals, IBM software, and IBM management utilities.

Pricing for the enterprise edition of OmniFind begins in the \$20,000 range, but you will find that other charges may be assessed; for example, a CPU charge, maintenance, and services.

The question is, "Can you perform sophisticated content processing with the enterprise edition?" The answer is, "It depends." If you have the requisite technical expertise, you can make the enterprise edition deliver most, if not all, of the functionality of other high-end systems.

The trick is UIMA.¹⁹ OmniFind is UIMA-compliant. This means that any third-party vendor who supports the UIMA standard can interact with OmniFind as well as other

¹⁹ Information about the UIMA Java Framework is located at <http://uima-framework.sourceforge.net/> and <http://incubator.apache.org/uima/>

parts of the IBM WebSphere system, including Lotus Notes. OmniFind incorporates a range of semantic technology.²⁰ You can integrate additional semantic, linguistic, and metatagging functionality by using UIMA-compliant third-party systems and subsystems from these selected vendors:

- Attensity
- ClearForest (acquired by Reuters and now a unit of Thomson Corporation)
- Endeca
- Inxight Software (acquired by Business Objects, and now a unit of SAP)
- Nstein Technologies
- Siderean Software
- TEMIS

OmniFind Discovery

The Discovery edition requires WebSphere. You use this edition to get functionality that:

- Identifies and tags the context of a query
- Supports assisted navigation and point-and-click discovery interfaces
- Includes tools that allow system tuning and monitoring

Includes vocabularies for finance, pharmaceuticals, and other vertical markets.

Once you have licensed the basic engine that begins at \$12,000, you can add a wide range of additional modules; for example, classification, adaptors to link to a Siebel Systems' application, and Web self-help subsystems. OmniFind Discovery is available with templates, scripts, and adaptors for use in commerce, self-service, and case resolution applications.

²⁰ For details about OmniFind's semantic functions, start here: <http://www.alphaworks.ibm.com/tech/wssem>, and here: <http://www.ibm.com/developerworks/db2/library/techarticle/dm-0508lang/>

Feature	Beyond Search Comment
Knowledgebase Support	OmniFind can support controlled term lists whereas their search and content processing products can be configured to use word lists and knowledge bases
Query Types	Boolean. Other query types can be supported with IBM extenders or third-party systems certified by IBM or supporting UIMA
Visualization	Supported in Cognos. Other IBM systems can use the Cognos system or integrate with third-party visualization tools
Entity Extraction	OmniFind supports entity extraction. Third-party tools can be used to process content in other IBM systems
Platforms Supported	S/390, AIX, Windows, Linux, Unix
Export	Tab delimited, comma delimited, XML, and other formats supported by adaptors or filters coded using the API
Third-Party Support	The UIMA standard allows third parties to integrate into a WebSphere environment
Vertical Support	Vertical builds of products are available. To see if your industry is supported with a purpose-built build, contact your local IBM office
Analytic Functions	Native analytic functions are included in OmniFind Analysis; Cognos can be integrated into an IBM environment. Third-party tools can be integrated via UIMA or directly by a certified IBM integrator

Table 23: Technical Highlights for IBM

SearchManager/370

Younger information retrieval and engineers don't know about STAIRS. The acronym is derived from SStorage And Information Retrieval System, a mainframe search-and-retrieval program dating from the late 1960s.

You can license a derivative of this program if you want to perform what IBM calls "full context retrieval" of documents in an S/390 (formerly an OS/390 virtual machine). Make no mistake – SearchManager/c is a capable and extensible search system, designed for manipulating content in a mainframe environment. SM/370 supports forward, middle, and inflection lemmatization. There's an API, support for Boolean queries, and a graphical interface so users with a Windows operating system can act as a client to the MVS or VM server. An "extender" is available to allow the SM/370 system to access DB2 data. You can use the thesaurus tool kit to implement controlled vocabulary support for system users.²¹ You can get license prices from your organization's IBM account manager.

²¹ SearchManager/370 information is here: http://www-304.ibm.com/jct01004c/systems/support/machine_warranties/warranties_licenses_maintenance.html

One question I am asked is, “Can I integrate SearchManager/370 with OmniFind Discovery or OmniFind Analytics?” The answer is, “Yes, but ...” You can integrate any IBM software and system. The “but” triggers the trade off of cost and benefits. You may want to make OmniFind Discovery, for example, be the system that uses data in a mainframe environment and makes it searchable within the OmniFind and WebSphere environment, but not the MVS or VM environment. The way to accomplish this is to deploy a dedicated subsystem that receives data exported from the mainframe system, then breaking the connection to the mainframe. The subsystem then converts the content to a form that can be processed by the OmniFind Discovery system. The subsystem can be configured to generate “reports” that become a document to a person searching for information via OmniFind. The data in a transformed format becomes the source for OmniFind Discovery value-added processing. This approach requires an intermediating system and, in my experience, direct, real-time connections between WebSphere and OmniFind often require significant hand-holding and baby-sitting. A mistake with a real-time interaction can corrupt memory in the MVS or VM environment. To make life easier, the use of an intermediating subsystem warrants careful consideration.

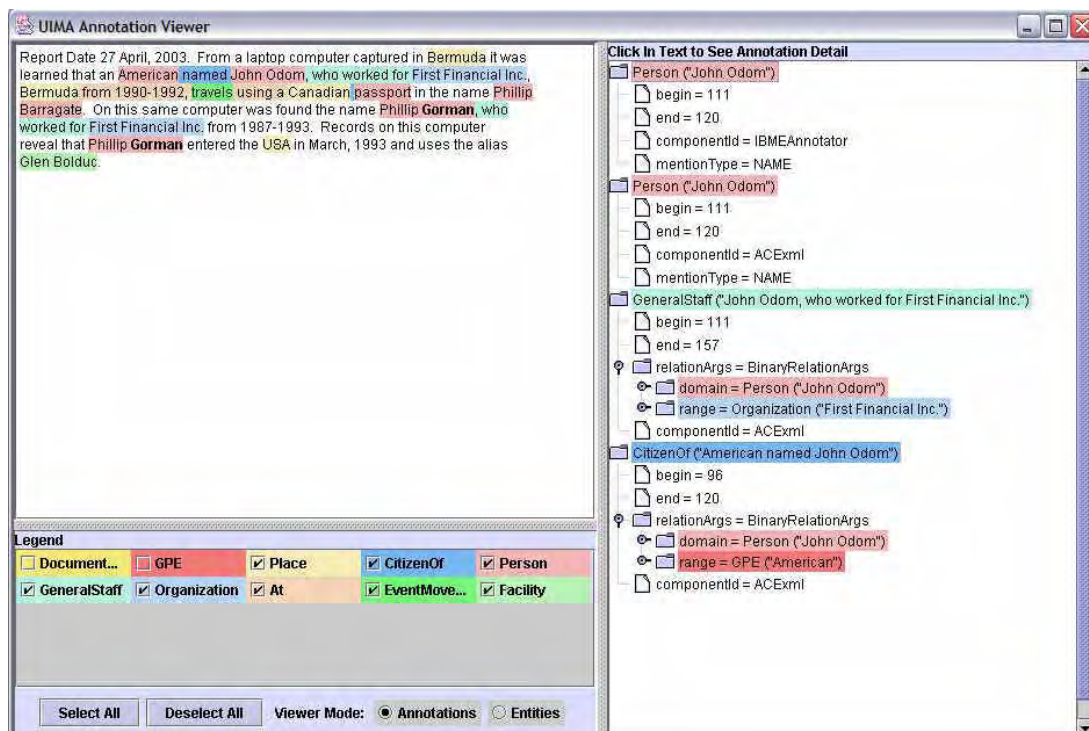


Figure 39: The IBM UIMA Annotation Viewer

The screen shot shows the UIMA annotation viewer. Notice that terms are highlighted in color. The left hand panel shows the detail of the syntactic analysis. © IBM Corp. 2006

DB2

IBM's description of OmniFind says that the search and content processing system can index structured data. Many organizations use DB2, IBM's flagship database, as a data warehouse and data management system for a wide range of applications.

How can you search a DB2 repository? Obviously, you can license OmniFind and use that system as a front end to DB2. IBM offers “search extenders” to add functionality to the DB2 SQL query functions. In fact, you can build a search-and-retrieval engine on top of DB2 to manipulate the data and XML objects in a DB2 table. You can also use third-party tools to add search functionality and speed to the DB2 SQL procedures. These components are available from vendors such as Copper Eye in the United Kingdom and SurfRay in Denmark. Discussion of these tools is outside the scope of this summary.

Keep in mind that many third-party applications use the IBM DB2 database and data management system in their enterprise software solutions. You may want to determine if you have a DB2 instance already running on your organization’s servers.

Cognos

With IBM’s purchase of Cognos, IBM now has additional analytic tools, data and text mining, work flow, and report-generation functionality. It is not clear what Cognos functions will be integrated into the Analytics’ version of OmniFind. IBM’s traditional handling of acquisitions is to, at some point, roll the new technology into an established brand. Cognos, however, has a strong market presence. IBM may use the Cognos name as a way to rationalize other “information on demand” products and services available from IBM. Information on demand, as I understand IBM’s use of the phrase, means Business Intelligence (BI). BI depends upon data and text mining (content processing). If this sounds like the snake biting its own tail, IBM will have to find a way to clarify its data and content processing services.

Comments about These Options

You are probably asking yourself, “How can I determine what IBM product I need to perform my value-added content processing?”

The answer, as anyone with a long, rewarding relationship with IBM will tell you, is, “Pay IBM to advise you.” IBM is a very large company for a reason. Large companies can rely on IBM to recommend products and services that can be made to work. Compared to IBM’s product lineup when I wrote the first edition of *Enterprise Search Report* in 2003 for publication in 2004, today’s rationalized product offerings are easy to grasp.

IBM Partners/Developers

Most organizations integrate third-party systems into IBM environments. IBM invented the business of certifying partners and resellers. The key to building the system that meets your exact requirements is to accept three facts of IBM life:

1. IBM will help you identify vendors, consultants, and resources that are known to work with IBM servers, network management tools, and the other components manufactured, coded, and supported by IBM. You must use “true blue” hardware, software, and people-ware. If you don’t, you can invalidate your warranty or be refused technical support.

2. You will have to pay IBM to assist you even if you buy your hardware and software from IBM. IBM's principal cash cows are mainframes and services. If you try to do this work without IBM's involvement, the likelihood of a problem is close to 100 percent. IBM's systems are complicated and often very difficult to get working without access to the technical information available only to IBM's own engineers and certified partners. "How to's" for a mainframe-to-OmniFind connector are not easy to find on the Web. The number of people who know how to do this work is small in comparison to the number of VisualStudio.net code jockeys.
3. IBM's products and services can and do work together in high-demand, high-availability situations. If you can't get your system working, the problem is almost always resources. The IBM system needs people, money, hardware, memory, and expertise. Take a short cut, and you will have some issues to resolve.

If you don't have a relationship for services with IBM, you can ask your preferred vendor if it is an IBM partner in good standing. If so, that vendor can integrate its system into your WebSphere environment without you having to involve IBM. The IBM partner will communicate with IBM on your behalf.

Upside

The principal benefit of working with IBM is that you know IBM will be able to make the system work. The other benefits of working with IBM include:

- IBM systems can scale to handle large volumes of content and perform at almost any performance specification you require.
- IBM provides a safety net to protect you from making flawed technical decisions. IBM's approach to third-party applications minimizes the likelihood that integration errors will bring a mainframe or WebSphere system to its knees.
- IBM will not get anyone fired. Large organizations cannot run the risk of deploying hardware, systems, and applications that are not reliable, stable, and extensible.

Downside

IBM is a mindset. If you try to understand the company's approach to enterprise applications by reading Web pages, you will have a difficult time making sense of IBM-speak and the products themselves. However, if you have experience working in an organization that is "true blue," you will appreciate IBM's approach to systems and application engineering. With IBM's strong support for Linux, the culture shock and learning curve are somewhat less steep. Nevertheless, IBM is not a Web 2.0 start-up. You must follow the IBM procedures and methodology. If you took a class in systems engineering in college, you were learning big chunks of the IBM "way." The company developed these procedures and some of them have worked for over 50 years. Other considerations include:

- The planning procedures are essential to a successful system deployment. Be prepared to spend time thinking about the basics of your content processing system. Your IT department may want to take shortcuts. In general, IBM will not. Their procedures reduce risk. Remember, whoever inked a deal with IBM wanted these management procedures applied to IT projects.
- IBM's way is your way. If you want to integrate a third-party product that is not UIMA-compliant or not certified by IBM, you will invalidate your warranty. You will be refused or required to enter into a different type of support contract. If you jeopardize the stability of an IBM infrastructure, someone at IBM will complain to the Board of Directors.
- Speed. IBM's systems can run, meet, or beat the performance of any other system in the world. To get that speed, you will have to spend money. Without proper resources, you can eat lunch as your OmniFind system fields a single query. Going faster to IBM involves an engineering solution, not a stick or two of RAM.

Net-Net

If you work in a Fortune 1000 company, a major bank, or at the Central Intelligence Agency, you will want to take a long, thoughtful look at the search and content processing solutions available from IBM. I've been using IBM servers for many years, and I speak from experience about the strengths and weaknesses of the company's policies, approach, and technology.

No matter what feature or function you want to implement, you will be able to deliver that to your users. The "gotcha" is that you will have to play by IBM's rules. With the cheerleading for Software as a Service (SaaS) and open-source solutions, it's easy to overlook the fact that when the world's most successful enterprises buy systems, a large percentage of them rely on IBM. The reason is that IBM can get systems to work and keep them online in high-demand, mission-critical implementations where a single mistake can trigger severe consequences.

11. Information Builders Inc.

www.infobuilders.com

Information Builders, located next to Madison Square Garden in New York City, has been an innovator in business intelligence, among other databased applications.

What's interesting about Information Builders is that the company has taken a pragmatic approach to rich text processing. One can argue that the Information Builders' approach gives the company flexibility and a way to take advantage of metadata, federation, and combining structured and unstructured data without distracting its engineers from the company's focus on creating near-real time reports.

Item	Quick Facts
Product	WebFOCUS 7.x
Price	\$30,000 and up
Key Feature	Business intelligence platform, search, access, and integration platform
Purpose	Implement business intelligence by processing structured and unstructured data and information
Clients	Fortune 1000 firms, Department of Defense, financial institutions
Company	Information Builders, Inc. Privately held
Contact	askinfo@informationbuilders.com

Table 24: Quick Look at Information Builders

The company prefers a low profile, well-suited to its blue-chip clientele. Information Builders serves thousands of financial institutions, service firms, and manufacturing companies with software that makes "pervasive business intelligence" a reality.

Instead of processing data and then preparing an alert or a report, Information Builders intercepts information at the message level. The idea is that when an order arrives and contains special instructions for the company, the order message is then passed to the fulfillment center for shipment. Information Builders' technology aims to process that data in real time, integrating the data into its business intelligence systems, and then taking action, if warranted, before data are placed in a traditional data warehouse.

The approach, therefore, leverages search, metatagging, indexing, and other rich text processing functions into gears within the larger Information Builders' system.

Information Builders' approach to rich text processing embeds rich text processing and other functions into the larger business intelligence framework. The approach is in sharp contrast to that taken by some of the other companies profiled in this study. Most competitors offer a framework or an add-on. Information Builders offers a comprehensive system that puts the emphasis on delivering on-point information that illuminates and facilitates business decisions.

Information Builders' system can make use of existing knowledgebases and taxonomies. In addition, there are add-on components or packaged vertical solutions for Performance Management, an insurance reporting framework (Information Builders IRF) with a data model, and an Integrated Justice or Law Enforcement framework. For competitive intelligence requirements, Information Builders offers tools to process syndicated content or the type of real-time information flowing from surveillance operations.

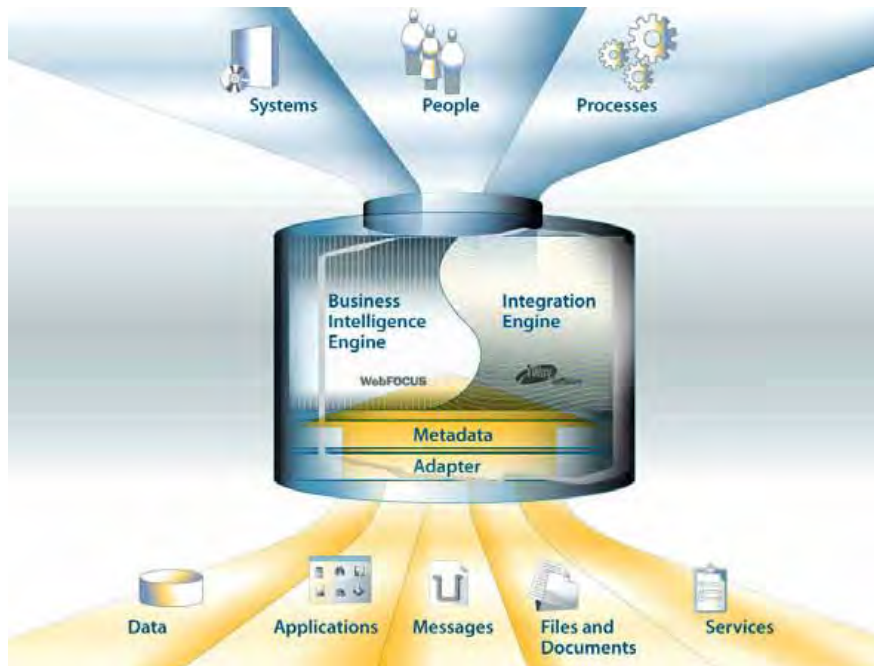


Figure 40: Information Builder's Platform

Information Builders' approach implements an intelligence platform, not just a search and content processing system.

The System in Action

The best way to get an overview of the Information Builders' approach to content processing is to look at a case example.

A large pharmaceutical company based in Montreal, Pharmascience, has been recognized for innovation and extensive investments in research and development. This innovation has extended into every facet of its business, as evidenced to its approach to data management and reporting.

When David Lavalée took over as CIO, he inherited an aging ERP system that didn't deliver the data the company needed. His strategy focused on putting data in an understandable format so that business users could access their own information. To achieve this goal Pharmascience implemented SAP R/3 and the SAP Business Warehouse in 2003.

Due to problems in reporting in the previous ERP system, Lavalée and the managers at Pharmascience decided to approach SAP reporting with a fresh set of reporting tools.

As a result they standardized on Information Builders' WebFOCUS for enterprise reporting in 2004, based on the software's ability to create complex reports that can natively access SAP R/3 and SAP BW data.

Lavallee likes WebFOCUS because it allows his team to gather information from multiple software programs and roll it up into executive scorecards. Initially his team used WebFOCUS to create sales and marketing reports. Based on their success with the software, they began creating reports for other departments as well, including finance, manufacturing, and human resources.

To achieve his goal of self-service reporting, WebFOCUS leverages SAP security by enforcing role-based authorizations, as defined in SAP, and allows users within their specific roles, to access and integrate all enterprise data (including SAP and non-SAP sources) in real time, turn it into relevant information, and share it with co-workers across the enterprise as reports.

In addition WebFOCUS leverages existing Web services built into the SAP NetWeaver Enterprise Services Architecture, which simplifies application integration tasks. As a result, WebFOCUS applications can be deployed on the SAP Web Application Server to create a cohesive user environment. Users can set up custom views that include only the metadata this is appropriate for their sphere of activities. Reports can then be created in any file format, including HTML, XML, Excel and PDF.

Information Builders technology delivers similar solutions to more than 2,000 customers worldwide.

Technology

The core of Information Builders' technology consists of three components of the company's business intelligence platform. They are Foundation Technology, Integration Engine, and Programming.

These technologies include:

- Self-optimizing autonomic servers whose workload and traffic management, and capacity planning eliminate complexity, improve system performance, and dramatically reduce TCO
- Super-linear scalability through multiple technological advantages
- A unifying integration infrastructure that accesses, reconciles, cleanses, and prepares any and all data for business intelligence use
- Service-oriented architecture support with the ability to create, consume, and publish Web services
- Simplified developer and end-user interaction, with advanced visualization and deep integration with desktop products such as Microsoft Excel and Adobe PDF.

Foundation Technology

The Information Builders Magnify Search system includes a metadata management system, content acquisition adaptors, repository access and management facilities. The approach combines a repository with a mechanism to use one set of metadata across the Information Builders' system. In effect, Information Builders provides a data transformation, metatagging, and storage service.

Adaptors allow Information Builders to look at any data source through a common metadata layer. These adaptors are more than file filters because the Information Builders' system integrates with other applications and databased content. Unlike a traditional data warehouse, the adaptors retain the functionality of the system generating the data.



Figure 41: Information Builder's Search Result Screen

The search result screen is uncluttered. In addition to the search box, the system displays the categories containing the search results for that topic.

Metadata in the Information Builders' system is more than a simple index of words or field names. The Information Builders' approach normalizes and synchronized metadata for information processed by the system. The metadata describes:

- The relationship between items
- Cross platform relationships
- Relationships among different databases
- Assisted navigation and process relationships for automating workflow actions

Integration Engine

The integration engine hooks into the foundation technology. The term integration is somewhat misleading. This subsystem monitors events that occur within the enterprise, including messages passed between and among other enterprise applications. Note that this is a layer that is typically not tapped by traditional search engines or rich text processing systems. Here monitored data are transformed automatically and transferred to another application.

The integration engine includes event monitors. The technology “watches” or “listens” to network traffic carrying messages and content. Events of interest are identified and specific actions are performed when an event occurs.

Transformation technology operates on any message type. Packets or larger objects can be manipulated by this subsystem. Outputs of the transformation subsystem are usable by other Information Builders’ subsystems or third-party systems.

The integration technology allows the consumer to run reports or call reporting tools directly from within the search results. Search terms can be parameterized as filters for the reporting data. The search results can be converted and analyzed in real time as a regular data set.

Programming

Information Builder supports Web services and programming languages associated with interactive, browser-based applications. In addition, Information Builders provides:

- iWay Software Data Migrator. This is a toolset that performs extraction, transformation, and load operations. The graphical interface makes it possible to design a data warehouse by tapping into the common metadata that the system generates.
- WebFocus Developer Studio. These tools provide a graphical environment to interact with metadata, create reports, set up parametric forms, and craft dashboards displaying different content objects on a single screen.
- iWay Software Service Manager. These tools provide access to administrative functions, ranging from process flows to text and search controls.

Other Technology

The company offers a full complement of utilities and services to licensees. These include a scheduling subsystem to simplify time-based or event-based operations. Information Builders offers a layer of software that makes it possible to create a portal; that is, a single, Web-based interface for content known to the system, business intelligence, discovery, and search-and-retrieval.

Search and Rich Text Processing

Information Builders' approach to text processing offers licensees different options. Let's look at each briefly.

WebFocus Magnify

WebFOCUS Magnify with the bundled Lucene open source search engine in the core product, is a cutting-edge tool that combines the power of state-of-the-art search, business intelligence, and integration technologies. It lets you tap into enterprise records and empower more users than ever to leverage critical corporate information. With WebFOCUS Magnify, structured and unstructured data references are compiled into the universally known and easy-to-use Google Search paradigm. By enabling users to easily locate key facts through simple keyword searches, organizations can realize significant productivity gains and enhance decision-making across the enterprise.

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	Supports knowledgebases, word lists, and taxonomies
Query Types	Keyword, natural language, assisted navigation, and automatically-generated reports with hot links
Visualization	Graphic and charting components included
Entity Extraction	Automatic identification and real-time term list matching
Platforms Supported	IBM mainframe OS/390, Linux, UNIX, Windows
Export	Multiple output formats supported
Third-Party Support	Can be integrated with third-party systems
Vertical Support	Insurance, law enforcement, financial services
Analytic Functions	Comprehensive and extensible statistical and mathematical functions

Table 25: Technical Highlights for Information Builders

Google Appliance

Like Oracle, Information Builders offers the Google Appliance as an alternative to the bundled search engine. Information Builders uses the OneBox API and other Google components to customize the Appliance to the needs of Information Builder licensees. For example, when Information Builders want a federated search across unstructured

Web content and access to information processed by Information Builders' system, the Google Appliance becomes the nerve center for this type of content aggregation. Information Builders customizes the functions of the Appliance to integrate certain live data into traditional key word queries. Information Builders' metadata are used to provide point-and-click assisted navigation of content on a Google Appliance search results page.

What's Interesting?

Information Builders' approach to search and rich text processing is that the licensee can have multiple approaches to finding information available in a single Information Builders' framework.

A person comfortable with key word queries can access processed content via key word, Boolean, or point-and-click interfaces. Users who know what information is required can receive a single document containing only new, actionable information. A user, who is not sure about what may or may not be needed, can access the system via a dashboard that includes:

- Assisted navigation for content discovery
- Automatically generated reports, documents, graphs, and charts
- A search box
- Summaries of new information identified by Information Builders in near-real time.

Rich Text Processing

The key point is that Information Builders has moved beyond search by offering multiple ways to find an answer. Search is not the *raison d'être* for the system. The system exists to make actionable information available to a user in the form that meets the specific requirements of a task. Because information needs change without warning, the Information Builders' system provides different access avenues.

New Features

Information Builders has an aggressive plan to enhance their present search and assisted navigation products. The product name for these systems and subsystems is WebFOCUS. Keep in mind that Information Builders offers a fat client solution environment as well as browser-based tools and solutions. It supports enhanced drill down functionality, that is, a user can click on an object and access the underlying or source data. The system will support one-click access to information in a database table, a document or a function or expression. Among the new features slated for release in 2008 are:

- Additional management tools for real-time indexing of message data, streaming content, and automatically-generated reports
- Enhanced dashboard technology so that a user can access data and then continue to work offline. Information Builders calls this *Active Technologies*.

- Addition of collaboration tools to allow a user to flag, route, annotate, and discuss information available within the Information Builders' system
- Support for mobile access including mobile search, mobile reporting and two-way mobile applications.

The company also will introduce a set of new Web 2.0 technologies for development of AJAX and RIA applications. Info Assist will allow ad hoc query of content and data. FLEX Enable will offer advanced visualization and interaction with content of web based applications.

Upside

The upside for the Information Builders' WebFOCUS system includes:

- A comprehensive solution to business intelligence, search, and discovery requirements
- Flexible architecture to allow licensees to integrate third-party applications such as text utilities, rich text processing subsystems, or third-party search engines
- Support for the Google Appliance with pre-coded components to allow Information Builders' metadata to be used with Google Appliance results, making assisted navigation functionality is available to users
- Strong work flow and automation tools that operate on message traffic and real-time data streams to ensure timely information in result sets or automatically generated reports

Downside

The downside of the Information Builder approach includes:

- Information Builders requires a licensee's commitment to the framework. As a result, adopting WebFOCUS is a significant enterprise software decision, not a casual decision to license a text processing tool
- Information Builders approach is an explicit decision to deploy a framework that makes business intelligence a priority.
- The system delivers good performance, supporting 1,000 or more business intelligence transactions per second. However, the system must be appropriately resourced. This means dedicated hardware, a system administrator, and appropriate training for programmers and analysts who will interact with the system.

Net-Net

Information Builders' approach points to a future in which search is a component of a larger enterprise system. The notion of search as a stand-alone enterprise application is different from the Information Builders' notion of an integrated business intelligence system.

Users who want key word searching or assisted navigation can deploy those solutions individually or in combination. If Information Builders' architects are correct in making search and rich text processing components of a business intelligence solution, vendors of stand alone products may find their market among Fortune 1000 firms shrinking. IBM, Microsoft, Oracle, and SAP offer somewhat similar solutions to their customers. These super platforms, along with Information Builders, can offer compelling reasons to acquire a business intelligence system that operates at scale.

At this time, individual text processing companies are finding a ready market for their products. The question is, "How long will it take for Information Builders and other super platform vendors to win over the world's largest organizations, if they can?"

12. Intelligenx

www.intelligenx.com

Intelligenx is one of those companies with solid technology which is off the radar. But it was Intelligenx's Discovery Engine that was the secret ingredient for the Carlyle Group when it sold Dex Media to R H. Donnelley Corporation for \$9.4 billion. Search technology from Intelligenx also substantively changed how the Office of AIDS Research manages and administers research grants at U.S. NIH. And it was their Discovery Engine that helped to transform the way in which D&B licenses data to libraries around the country.

Iqbal Talib, his son, and a cadre of skilled engineers have built technology that permits users to search and interact with incredibly complex datasets. The core product offering, Discovery Engine is unique in that it was built ground-up to enable full-text search with categorizations. The display of intuitive refinements (with counts) that are derived from the structure in data helps users to find and 'discover' information.

Item	Quick Facts
Product	Intelligenx Discovery Engine
Price	Starts at \$50,000. Custom price quote required.
Key Feature	Full-text search with categorizations.
Purpose	Provide access to structured information, so that users can interact and discover
Clients	Publicar, Axesa, MediaTel, OAR at NIH, TDS, ilocal, D&B, WebVisible
Company	Privately-held
Contact	+1-703-793-3270

Table 26: Quick Look at Intelligenx

Mr. Talib told *Beyond Search*, "Our Company was one of the first to introduce a combined full-text search coupled with navigation. What we discovered was that there are far more effective ways to let users interact with information. We also found that we could engineer systems to deliver unprecedented search features and functionalities at far lower costs and without many of the challenges and bottlenecks associated with other conventional search methods." The son, Zubair Talib, is the CTO. He attended MIT and, with some friends from school, developed the first algorithms that are still the foundation of Discovery Engine.

The Carlyle Group purchased Denver, CO-based Dex Media for \$7.05B. Over the next 26 months, Dex launched a new Internet strategy that harnessed the power of Discovery Engine. On DexOnline, users could conduct a Google-like full-text search and for the first time anywhere, they could search all the text from all of Dex Media's print directories. Users could refine the search results in order to find (or *discover*) what they were looking for. The site was responsive and users took to the interactive search

functionality. During the time Carlyle owned Dex Media, usage of DexOnline skyrocketed (10-fold increase in traffic) propelling Dex from Internet obscurity to the number 1 traffic position within its 13 state region, ahead of Google Local, Yahoo Local, and Switchboard.

Since Dex, Intelligenx has won a number of highly competitive contracts with large directory publishers around the world who use Discovery Engine to provide interactive access to yellow page information over the Internet. Mr. Talib said,

We had success with directory publishers because our technology can easily handle very large traffic volumes, large data sets, and complex business logic. Directory publishers also face challenges with how to monetize their traffic and how to scale their business models – a problem that Discovery Engine solves quite naturally.

The company's system allows you to search content from a print yellow page ad (including brands, locations and hours of operation, for instance), including the standard name, address, and category fields. A user does not have to specify which fields to query. Each result set is then presented in "buckets," or collections of on-target results, not a list of results. You can then refine or "drill down" into these buckets to find particular listings quickly and intuitively. The suggestion of results that may be related to the initial query allows you to discover information that they may not have known even existed.

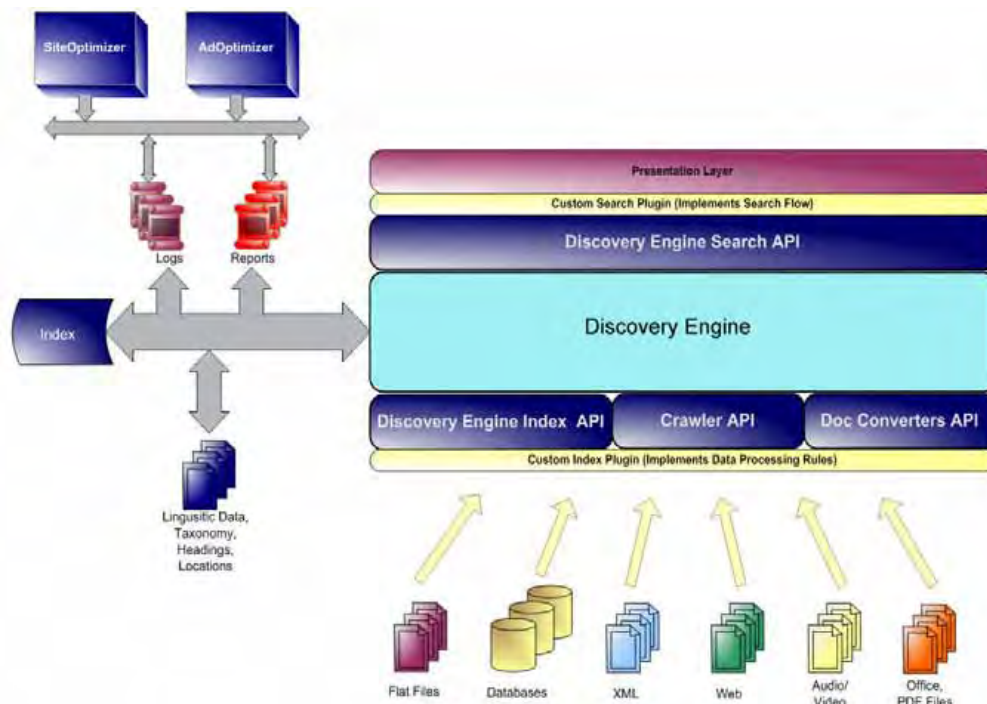


Figure 42: Intelligenx Discovery Engine

The Discovery Engine includes separate APIs: one for indexing, one for acquiring content, and one for data transformation. The system can be integrated into almost any enterprise environment.

The Technology

Discovery Engine is proprietary technology. The approach combines full-text search with fielded search. The result is that the system that provides all the benefits of and capabilities of conventional full-text search technology and all the search capabilities that exist in relational database management systems (RDBMs), combined with navigation and counts. Discovery Engine helps to exploit the underlying structure of the data for refinements and many other assisted search techniques; it also resolves failed queries.

With more than a decade of computer science and development, the Discovery Engine incorporates innovative algorithms for compressing, optimizing and searching processed content. The approach required a “ground up” rethinking of content processing, according to the company. Innovations include algorithms for data compression and storage, content processing, and distributed parallel processing. A high-level schematic of the Discovery Engine illustrates a number of incorporated components.

The system does not require a third-party database. A licensee can use commodity servers to scale the system. Like Google, the Intelligenx approach allows additional storage and servers to be added without complicated configuration and certification processes.

Intelligenx’s founder told *Beyond Search*:

Typical implementations achieve an 80 percent reduction in hardware, hosting and enterprise database costs. Our software simply bolts on to an existing enterprise infrastructure, eliminating expensive integration work. In fact, many of our customers retrofit our system into their existing data and maintenance infrastructure.



Figure 43: Paginas Amarillas' use of the Intelligenx Interface

The Intelligenx system makes it possible to display a result set with hot links to other Web pages and related categories. The two-panel display used in Paginas Amarillas displays related content in the left-hand panel of the display.

Linguistics

The system includes support for linguistic techniques to improve query understanding. The standard Discovery Engine linguistics toolkit includes spelling checkers, stemmers, stop word removers, and synonym updating functions. These tools support multiple languages including multi-byte languages like Japanese, Chinese and Arabic. The linguistics tools are used within the query transformation infrastructure that can be used to extend the capabilities of Discovery Engine. This infrastructure can also be used to perform complex query transformation tasks such as parsing complex Boolean queries, including Boolean NOTs, translating query operators from different languages, performing category matches preferentially, and constraining or loosening a query.

APIs

The architecture of the Discovery Engine includes a number of components. The application programming interfaces make it possible to integrate the Intelligenx system into other enterprise applications, Web pages, or a portal. The APIs and extensions are fully documented. The product is typically shipped with a Software Development Kit (SDK) that contains sample configuration files as well as the entire toolset required to manage a real application on a real deployment. The SDK contains a sample application

along with data, source code and display files that can be used as a starter kit for developing a customer-specific application.

The Index API provides all of the functions required to construct an Intelligenx index from a copy of the customer's data feed. The Search API provides all of the functions required to search an Intelligenx index. Particular strengths of the Search API are the very flexible and customizable ranking and sorting methods, query expansion and linguistic modifiers, inclusion of complex search logic and search trees, and failed search handling methods. The index and search plug-ins are typically application-specific code written to process the customer's raw data feed, as well as satisfy the business requirements specified by the customer. While accessible through an API or XML web service, Discovery Engine is also packaged with a presentation layer that consists of visualization pages, e.g., JSP or ASP, to accept a user's query and present the relevant results.

Other APIs available include a Crawler API for crawling the web and accumulating a web index to augment the customer's data, as well as a Reporting API for generating statistical information about the queries processed by the Search API and a Management API for administering a deployment.

In addition to the public APIs, Intelligenx provides a number of documented extension sub-systems that can be used to enhance the capabilities of the basic search engine. These extensions can be used, among other tasks, to augment the indexing process, configure the query transformation process and control the results ranking process. Intelligenx also provides a suite of pre-written implementations of these extensions that suffice to satisfy the business rules of most customers. However, customer-specific requirements can be incorporated quickly by writing fresh implementations within this infrastructure.

Intelligenx Features

The system includes a number of interesting features. For example, content processed is automatically categorized and appropriate metadata generated and linked to the content. The system can process XML, structured data, or unstructured text.

More recently, Intelligenx has packaged its internal data mining tools into rich business intelligence log analysis tools. These add-on products, Ad Optimizer and Site Optimizer, build on the Discovery Engine architecture to provide deep, interactive information about usage. AdOptimizer, tracks user behavior and generates real-time reports about those actions. One application of AdOptimizer is to permit real-time inspection of users' interaction with suggested content. These reports can be syndicated to allow advertisers, users, or licensee staff to make adjustments to certain system components; for example, content boosting or advertising fees. SiteOptimizer helps determine relationships and correlations between user behavior and how those relationships can be used to drive improvements to the search application.

Another recent add-on, Content Enhancer, crawls web pages and extracts relevant and meaningful content and entities from web pages in order to enhance the original content repository.

Feature	Beyond Search Comment
Knowledgebase Support	None needed. The system “discovers” entities and categories
Query Types	Boolean, free text, and assisted navigation
Visualization	Outputs can be displayed as tables or other representations
Entity Extraction	Not applicable
Platforms Supported	Linux, Windows
Export	Content can be generated in XML or user-defined formats
Third-Party Support	The Discovery Engine can be integrated with any third-party application
Vertical Support	Publishing
Analytic Functions	The system includes strong analytic support including various numeric functions. Additional mathematical processes may be integrated via the APIs

Table 27: Technical Highlights for Intelligenx

Other Intelligenx features include:

- Geospatial data support so results can be searched, mapped or manipulated by geo parameters
- Configurable categorization and relevance ranking thresholds
- Key word highlighting in results
- Near real-time index updating
- Multi-threaded architecture to take advantage of multicore processors
- Built in content transformation tools
- Federated search capability to search across disparate repositories

The system is language-independent and provides a configurable security model based on the operating system in use. For public access, the system supports hypertext transport protocol (HTTP) authentication. The system has no limit on the number of documents or the amount of content it can process and index.

Discovery Engine in Action

You can explore the functionality of the Intelligenx system at Publicar’s Spanish language directory portal at <http://www.paginasamarillas.com/>. Publicar is the largest directory publisher in South America. Traffic has almost doubled for Publicar since deploying Discovery Engine and the site processes millions of queries per day with high performance. Publicar will add on extensions for wireless search and SMS that will utilize the core search infrastructure built on Intelligenx technology.

Other Intelligenx current customers include:

- Axesa (Puerto Rico, formerly Verizon Information Systems Puerto Rico)
- Conselho Federal da Justiça (Justice Department Brazil)

- DeTelefoongids (Netherlands)
- Dun & Bradstreet (USA)
- iLocal (Netherlands, Belgium, Luxemburg)
- Localeze (USA)
- National Institutes of Health (US Federal Government)
- MediaTel (Czech Republic)
- WebVisible (USA)
- 411.ca (Canada)

Upside

The upsides of the Discovery Engine pivot on the system's ability to handle very large volumes of content even at extremely high loads. *Beyond Search's* tests revealed response times in the 100 millisecond range for our test queries. Other upsides include:

- Support for structured and unstructured information regardless of the source document's language or the physical location of the data.
- A scalable architecture that allows licensees to expand the system's infrastructure with commodity hardware. Note that Intelligenx also offers hosted solutions and a suite of web services for merchant-level reporting and search analytics.
- Discovery Engine has excellent failed-search handling
- A well-documented and comprehensive suite of APIs with sample code. Intelligenx makes integration and extension of its system less painful than some of the other companies profiled in this study.

Downside

The downside of Intelligenx is the low profile the company has adopted in its 10 year history. Even though the firm is projected to generate \$4 to \$6 million in profitable revenue in 2008, most information professionals are not aware of the company's high-performance, feature-rich system. And because the company has captured a number of international customers (mostly directory publishers) Discovery Engine is perceived as only a local search technology. That's not true.

In reality, Discovery Engine can bolt on to any database or content repository, including native XML files and deliver blinding performance, equal to or better than many of the features associated with Endeca's or Fast Search & Transfer's systems. If your applications require scalable full-text search with categorizations, then you ought to know about Intelligenx.

Other drawbacks include:

- The system performs best when the source content is structured; for example, content from a database or well-formed XML

- The basic system can be used in its default mode. However, tuning the system or integrating it with third party applications requires study of the API documentation and may involve writing scripts
- The company offers a range of professional services. Some of the work is performed by senior developers. If you want a large, custom project in a very short time, you may have to wait until the firm's technical highly trained staff becomes available.

Net-Net

The truth is that processing so much information so quickly is not so easy using conventional search technology. Using the wrong technology to achieve this sort of functionality has its limitations including challenges with performance and scalability. Today, Intelligenx's performance over the Internet and its high-speed indexing is closer to that delivered by Google than most other Web search systems. The software has also been battle tested under heavy loads where it has delivered the goods.

The system is adept in its manipulation of structured data. It is even possible to use the Discovery Engine as a database engine, eliminating most of the hassles and processing bottlenecks associated with traditional relational database architectures. Like Google, Intelligenx technology works on commodity class clustered computing environments so that scaling is easy and cost effective.

The product is flexible enough to support custom query transformations to enhance the user experience. As well, it can provide totally customized ranking/sorting/filtering schemes in order to accommodate the relevance and ordering of search results. A full set of APIs, interfaces and complete documentation enables rapid application development and easy, rapid deployment.

If you want to make use of assisted navigation *and* offer key word searching, you will want to take a long, hard look at the Intelligenx system. Using it as the data management foundation, Intelligenx makes it relatively easy to hook in specialized visualized, statistical, even additional content processing functionality.

13. IntelliSearch Inc.

www.intellisearch.com

Norwegians innovate in content processing. Perhaps the cause is the weather (brisk in the winter, I hear) or the herring (plentiful anytime)? The company's catchphrase is "Go beyond the search." Intellisearch understands that users want more than laundry lists of results. Harald Jellum told *Beyond Search*, "Our vision is that search will be everywhere." Mr. Jellum has founded other high-technology companies, including Cyberwatcher, which contributes some technology to his latest venture, IntelliSearch.

Mr. Jellum founded the company in 2002. He continues to serve as the firm's chief executive officer, heading the engineering effort to create specific bundles or builds of the firm's search technology for vertical markets. Today, you can license IntelliSearch for eCommerce, Web site search, and competitive intelligence, among others.

IntelliSearch is the most recent entrant in the content processing wars with Oslo, Norway, as its European and technical headquarters. The publicly-traded company now has offices in San Francisco in order to raise its profile and revenue in the North American market.²²

Item	Quick Facts
Product	IntelliSearch Enterprise Search Platform
Price	Begins at \$30,000. Custom price quote required.
Technology	Proprietary on Microsoft Dot Net
Key Feature	System can be tuned to give greater or lesser relevance boosting to specific content; controls to fine tune relevancy
Purpose	"Empowering wisdom from a single access point"; federated search
Clients	Sintef, KPMG, the Norwegian Post Office, the Red Cross, Ericsson
Company	Publicly traded on Norway's OTC market
Contact	info@intellisearch.com

Table 28: Quick Look at IntelliSearch Inc.

Like Fast Search & Transfer (another Norwegian search vendor), IntelliSearch positions its technology as a platform. The idea is that other information-centric applications may be built upon or integrated into the IntelliSearch solution.

The company's content processing solution may be used for behind-the-firewall search, Web site Search, eCommerce, and as an OEM component.

²² You can check the share price under INTS at INTS <http://otc.nfmf.no/public/otc-list.html>

One of the interesting things about IntelliSearch's approach is that the company is Microsoft-centric. The engine makes use of Microsoft's Dot Net technology, and the system ships with connectors that permit seamless access to information on Microsoft's servers and to a variety of enterprise applications; for example, SAP repositories.

In September 2007, the company inked a deal with Swedish mobile phone giant Ericsson. Ericsson has deployed the system to its 40,000 worldwide employees. Beyond Search estimates that the company has 49 full time employees and will generate between \$8.0 and \$10.0 million in revenue in calendar 2008.

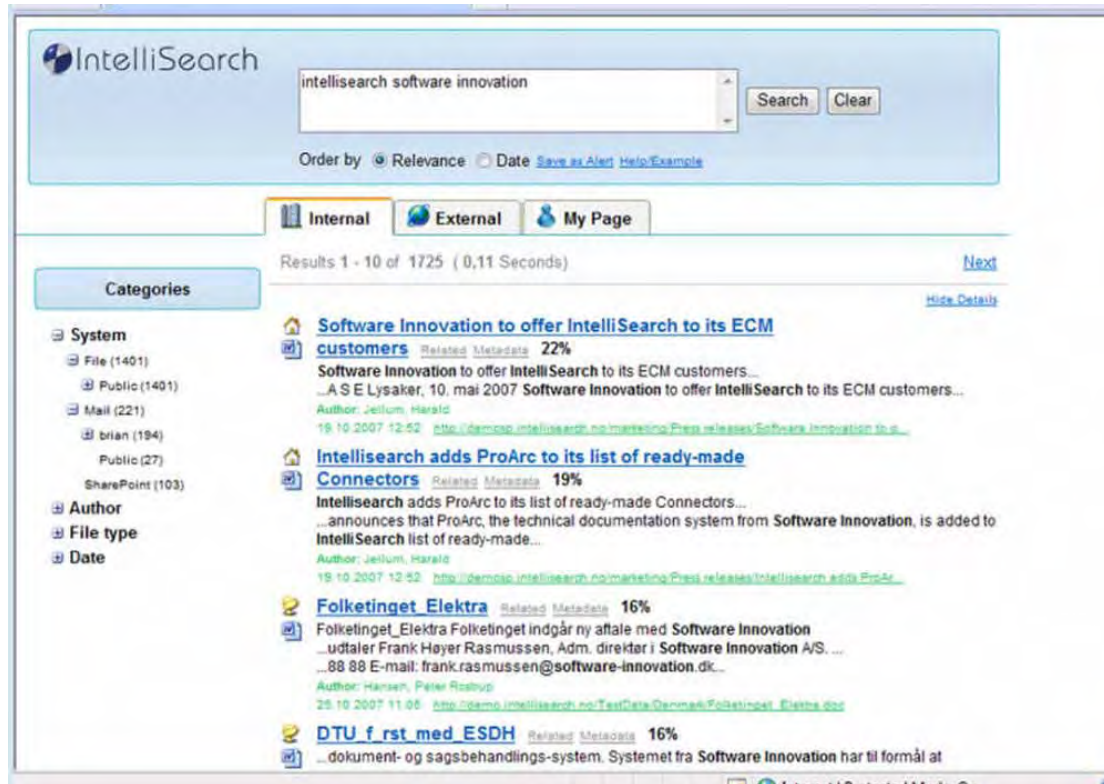


Figure 44: The IntelliSearch Interface

The user can formulate queries using either the Express or Pro interface. As with consumer systems, the user can start with a keyword search and then move on to more complex context-sensitive queries.

Technology

In the 2007 release of the product, the company expanded beyond the proprietary technology used for Microsoft's framework. IntelliSearch is now Web services "friendly." As a result, the system is easier to integrate with other enterprise applications and operating systems. Keep in mind, however, that IntelliSearch is happiest when running in a properly configured and resourced Microsoft environment. The system has been designed to scale using a distributed search architecture.

The IntelliSearch platform delivers the "bells and whistles" associated with search systems designed to serve as application platforms. If you want more detail about

Autonomy, Fast Search & Transfer, Oracle and other platforms, please, consult the *Enterprise Search Report*, where this subject is explored in greater depth.

In terms of content processing, IntelliSearch is interesting because it is a system that incorporates, as a standard feature, two rich content processing or metatagging functions: categorization of results and what the company calls prioritization.

Categorization, in terms of the IntelliSearch system, means that items in a results list are categorized according to:

- Source plus day, month, and year
- Author
- File-type
- Topic
- Geography
- Business unit (e.g. accounting)

The categories can be customized for each licensee depending on metadata available. The metatagging makes use of information in a SharePoint server, for example, or discovers these attributes based on information available to the IntelliSearch system.

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	Makes use of controlled vocabularies, dictionaries, and other knowledgebases
Query Types	Boolean and point-and-click category-based interface
Visualization	Microsoft graphing and third-party applications can be integrated into the system
Entity Extraction	Author, department, geographic location, and other items are tagged
Platforms Supported	Microsoft Windows
Export	System can export Microsoft file types; for example, comma delimited files
Third-Party Support	Lotus Notes, Documentum, ProArc, relational databases, and Microsoft SharePoint
Vertical Support	Web search, competitive intelligence, eCommerce, rich media
Analytic Functions	Standard content processing log files

Table 29: Technical Highlights for IntelliSearch Enterprise Search Platform

Prioritization is IntelliSearch's term for relevancy ranking and its tuning. For example, if the licensee wants to use the system in customer support, the licensee can weight customer support content to appear higher in a results list than information on a topic from the company's public relations department. The tuning can weight for prioritization product support documentation, customer invoices, ticket handling

status, customer related mail and emails, etc. to minimize the time-consuming process of inspecting items in the result list to locate a needed piece of information.

The current version of IntelliSearch Version 2.0 has built-in a relevance model enabling identification of similar or topically related documents. You can clip a paragraph or an entire document and paste that text into the IntelliSearch search box.

The firm's engineers have developed a language-independent spelling system based on soundex and word uniqueness algorithms. For the current release, the company has optimized certain processes so that performance in content processing and query processing has improves as it has done with each release.

Connectors

The company includes filters for most file types found in Microsoft-centric organizations. The basic system supports Microsoft Word, Excel, PowerPoint, and Visio. However, if you want to process specialized files types such as archive formats or mail servers, you will have to pay extra for these connectors. The company can create a content connector if you have a specialized file type for which no connector is available. The company offers a connector manager to simplify the acquisition and processing of content. The connector API allows developers and licensees to create custom connectors as well. IntelliSearch also uses Microsoft's iFilter technology.

Knowledge Assistant

This component is an intelligent search agent for keeping you informed about the latest news within your area of interest when working with Office 2007. The assistant will automatically search and retrieve information related to your current work. When the user writes a new document in Word, the agent will automatically search and retrieve similar and related documents in the company network and external media sources. IntelliSearch told Beyond Search, "The Knowledge Assistant solution works proactively without the user having to search."

Natural Language Search

The IntelliSearch platform supports natural language or NLP queries. The system makes use of knowledgebases; such as, controlled term lists and Use For or synonym lists. The system administrator can define a "reference profile" Model. IntelliSearch uses these models to filter or manipulate content; for example, a reference profile can specify certain terms or concepts to filter from a result set or content processing.

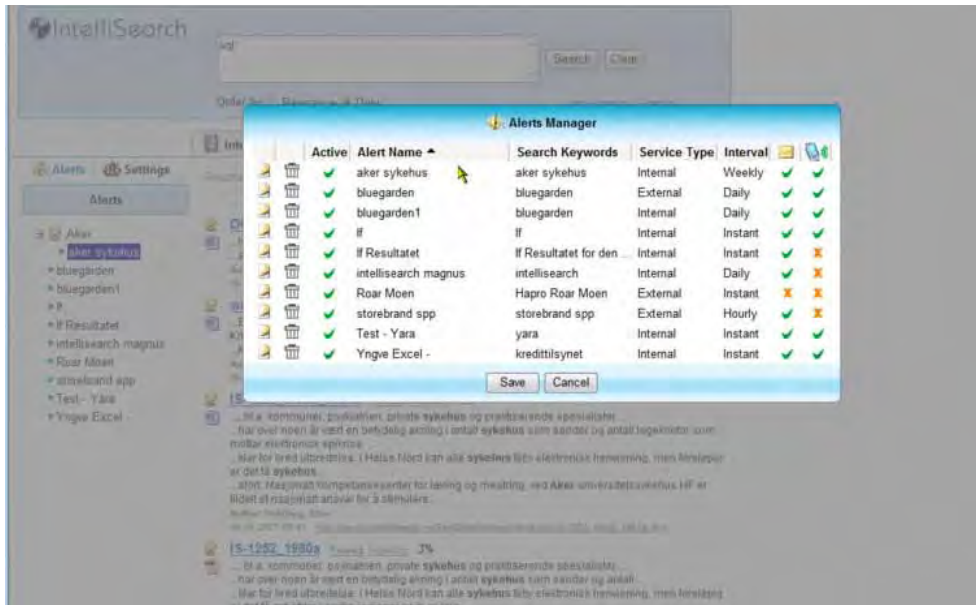


Figure 45: The IntelliSearch Alerts Manager

Alerts administrative screens features point-and-click set up.

User Profiles

IntelliSearch supports personalization via user-defined profiles. You can configure a profile for yourself or a group of users. The system displays rules, and configuring the behavior of the profile is a matter of pointing and clicking. The profile can weight certain content so it appears higher in the result list. Alternatively, high-value content can be displayed in a result list so that users have one-click access to this high priority content.

Other Features

The system also supports Boolean AND, OR, and NOT operations and phrase searching; for example, *White House*, which allows the user to locate compound word forms. The system offers a licensee's three access modes, which can be customized. These are a search box, a personalized "my page", and its alerting interface. The system can also suggest documents to users.

Upside

The upside for IntelliSearch's system includes a modest service footprint that holds down hardware costs yet delivers automatic document categorization. The system is customizable. The company also offers a subscription pricing model. Other upsides include:

- Easy integration with Microsoft Dot Net environments in general and SharePoint in particular. You will want to compare the IntelliSearch system with other Microsoft-friendly systems from Coveo, dtSearch, and ISYS Search System, among others before making a decision.

- Allows access to behind-the-firewall content as well information located on the Internet or other networks
- You can license builds of the IntelliSearch system for OEM, intelligence, Web site search, and eCommerce search, as well as for behind-the-firewall search.
- Programmers familiar with Microsoft's VisualStudio.Net coding system are widely available.

Downside

Considerations for the IntelliSearch approach include:

- The company has a low profile that may translate to some procurement teams not giving the system a close look
- Microsoft-centric search solutions can be difficult to set up so that performance remains at a high level. To resolve some bottlenecks inherent in the Microsoft Dot Net framework, additional hardware, storage, and bandwidth resources may be necessary.
- IntelliSearch integrates with Word 2007 but not older versions of Word.
- The system can be affected by the hot fixes and security issues that go hand-in-hand with Microsoft clients and servers. In some organizations, the ease of use of the Microsoft platform may outweigh security considerations and the use of Microsoft's security components and access control lists may be inappropriate.

Net-Net

IntelliSearch has a strong following in Europe, and the company hopes its deal with Groxis and New Idea Engineering, a search integration firm in Silicon Valley, will allow the company to build its customer base in the U.S. and Canada. It's too soon to tell if the behind-the-firewall search market can absorb another vendor.

Beyond Search continues to ponder Microsoft's decision to buy Fast Search & Transfer, a company with less Microsoft "gravity" than IntelliSearch. Perhaps Microsoft was buying Fast Search's customer base. The smaller, more Microsoft-centric IntelliSearch may have a technology advantage when it comes to integration with Microsoft applications, but the company lacks the more than 2,500 customers that Fast Search has.

If your content processing needs require the type of metatagging typically associated with the original 2007 version of SharePoint search or Microsoft specialists such as Interse in Copenhagen, Denmark, give IntelliSearch a test drive.

14. ISYS Search Software

www.isys-search.com

The Company

Nestled in a trendy neighborhood near Sydney, Australia, founder Ian Davies said:

In 1988, I was frustrated with search in general, so I started working to eliminate the irritants—difficult configuration, confusing or useless results, and sluggish performance. Now, we have a solution that points where search is headed and we think is the leader in delivering features and functionality. We offer search, navigation, and discovery at a very competitive price point.

Prior to founding ISYS, Mr. Davies spent 10 years writing as a consulting technical editor for one of Australia's leading computer magazines. Additionally, he spent four years with a prominent commercial software house in Australia before becoming an independent consultant in IBM mainframe fourth-generation languages.

Version 8.x of ISYS and the soon-to-be-released Version 9.0 address problems that bedevil administrators and users of search systems: complexity, cost, sluggish response, and difficult customization. Since the first commercial release of ISYS in 1993, the company has evolved to focus on Windows-based search solutions for workgroups, enterprises, and developers.

Item	Quick Facts
Product	ISYS Search Software, Version 8.x
Price	Begins at \$14,800
Technology	Key word and entity extraction
Key Feature	High-speed text processing and hot linked entities and topics
Purpose	All-in-one search solution with minimal administrative overhead
Clients	Boeing, EMC, IDG, QANTAS
Company	ISYS Search Software, (formerly Odyssey)
Contact	info-us@isys-search.com or info-au@isys-search.com

Table 30: Quick Look at ISYS Search

Still privately-held, ISYS has become one of the search systems that competes successfully with the likes of Autonomy, Endeca, and Fast Search & Transfer. Like Coveo and Exalead, ISYS has become a text processing system warranting careful consideration where features, performance, and simplified administration are important.

About two-thirds of ISYS's revenue comes from the U.S. The remaining third of the firm's revenue comes from Australia and the U.K. At the start of 2006, ISYS said, "We

have more than 12,000 customer organizations worldwide, ranging from single users through to the U.S. Federal Bureau of Prisons with over 10,000 users. Other notable ISYS customers include:

- Boeing
- EMC
- Miami (Florida) Police Department
- Perkins Coie LLP and affiliates
- QANTAS
- U.S. Internal Revenue Service
- US Department of Justice and Department of Homeland Security
- World Trade Organization.

ISYS has a strong position in the Australian government. Its customers include the Australian Crime Commission, the Federal Court, the Department of Employment and Workplace Relations, the Department of Immigration and Multicultural Affairs, Refugee Review Tribunal and the Department of Defence.

The company has about 50 full time employees and is profitable. Beyond Search estimates that ISYS's revenues in 2008 will exceed \$30 million. Company insiders report that growth is accelerating and now is about 50% per annum. ISYS is one of the success stories in the search sector where traditional key word queries are buttressed by entity extraction and other rich text processing features.

ISYS Product Line Up

The company offers two products and a software developer kit to permit integration with other enterprise applications. These products are:

- ISYS:web 8. This is the flagship system that supports search across Intranets, Web sites, portals and custom Web applications.
- ISYS:desktop 8 is designed for a single computer or LANs. It can index and search email and local documents. The desktop version can also be managed centrally to index content on other computers to which the user has access. The product is designed for networks of desktop computers searching multiple data sources while respecting domain/workgroup as well as individual security.
- ISYS:sdk 8 enables OEMs, system integrators and others to embed search technology into their custom applications.

The company also has versions of its core system tailored for email-only search, CD distribution and other applications.

Rich Text Processing

Among the most significant functions of ISYS is its advanced text processing services. In our tests, the ISYS system performed categorization and entity extraction without bogging down the document indexing process. In fact, our test corpus of 500 megabytes was processed in 23 seconds on a dual processor workstation.

ISYS:web 8 Rapid Deployment

Home :: Search Results

Your search for "Improvised Explosive" found 336 hits in 27 documents. (0.18 seconds)

Sort By: Best Match | Hit Count | Title | Date | Hit Density Results 1 to 10 of 27

1. Bomb Squad Table [Outline] [Details] [Similar]	11 May 2006	11 hits	81.5K
Procedures (RSP) Equipment Same as Type II Employ explosive tools to conduct specific or general disruption Demolition ...and level C) for Chem/Bio with associated explosives See Note 1 No PPE for Chem/Bio ...Equipment Explosive Transport Same as Type II Explosive Transport Vessel			
C:\ISYS8\WebIndexes\web7.Intelligence Data\DefenceData\bomb-squad-table.doc			
People: X Ray, Level A PPE, Level B PPE, Level C PPE			
2. EODMU 4, Det. 10 Train, Test Remote Detonation Techniques of IED [Outline] [Details] [Similar]	11 May 2006	4 hits	2.1K
Bahrain (NNS) Explosive Ordnance Disposal Mobile Unit (EODMU) ...completed the Final Evaluation Phase (FEP) of its Improvised Explosive Device (IED) training Nov. ...put our hands on the device." During improvised exercises, such as the IED FEP,			
C:\ISYS8\WebIndexes\web7.Intelligence Data\DefenceData\news_stand5-11-2006.doc			
People: Buster Standil, Elton Shaw, Tim Johns			
3. BUDGET TIGHTENING MOVES [Outline] [Details] [Similar]	11 May 2006	2 hits	24.5K
at sea), Recruiting duty & Great Lakes Instructors/RDC, Improvised Explosive Devices (IED) Defeat related			
C:\ISYS8\WebIndexes\web7.Intelligence Data\DefenceData\BUDGETTIGHTENINGMOVES.doc			
People:			
4. DEADLY THREAT OF EXPLOSIVE DEVICES IN IRAQ PROMPTS SECRECY DEBATE [Outline] [Details] [Similar]	11 May 2006	7 hits	26.5K
DEADLY THREAT OF EXPLOSIVE DEVICES IN IRAQ PROMPTS SECRECY DEBATE Author: ...THREAT OF EXPLOSIVE DEVICES IN IRAQ PROMPTS SECRECY DEBATE 1, ...Rights Reserved The Defense Department's push to counter improvised explosive devices on the battlefield has sparked			
C:\ISYS8\WebIndexes\web7.Intelligence Data\DefenceData\DEADLY THREAT OF EXPLOSIVE DEVICES IN IRAQ PROMPTS SECRECY D.doc			
People: Anna Blumenthal, Dick Cheney, Colin Powell, Scott Rasmussen			

Search For
Improvised Explosive
in Intelligence Data

Categories

2004	1
Defence Data	26
All	27

Entities

Person

President Bush	28
Clark Staten	15
Sir Robert Marks	14
Governor General	12
President Clinton	11
President George W Bush	11
Dr Abdel Rahman	10
Saddam Hussein	9
Mr Bin Laden	7
Child Wise	7

Organization

AFP Co	565
AFPA	230
ACC Police	169
Department Of State	95

Figure 46: ISYS' Default Interface

ISYS's default interface displays the search results in the left-hand box. A standard search box and navigation hot links appear on the right side of the interface. The look-and-feel of the interface can be customized via style sheets.

Ian Davies told *Beyond Search*:

We have optimized our indexing subsystem in the current release. Most of our customers want to index content rapidly and perform incremental updates without losing access to the system. We have eliminated most of the delays associated with key word indexing, on-the-fly classification of documents, and identifying the people, places, and things in a document.

Categorization

ISYS implements "on-the-fly" categorization. ISYS automatically builds the categories according to file path, while administrators can use metadata to customize this structure to address specific categorization requirements. ISYS engineers use a combination of statistical techniques and its knowledge base to assign category tags without imposing excessive system overhead during document processing. The categories allow users to "drill down" in a particular category and refine the results list to a specific topic without typing a query into the search box. The built-in administrative tools allow the search administrator to customize the categories, as required.

Entity Extraction

ISYS automatically extracts and displays entities such as people, organizations, email addresses and other patterns that can be specified in the administrative interface. ISYS uses a proprietary technique to identify, extract, and tag these items.

The API

The ISYS Search API consists of function calls arranged into logical groups, for example: basic and advanced retrieval, indexing, concept trees, and named sections, among others.

The SDK makes it easy to access these calls. The search administrator or developer includes a standard header file in the code. The header file provides the required constants, data structures, and function definitions the application needs to control the ISYS Search Engine.

One interesting feature in the ISYS API is its intelligent search agent, which lets your application provide user-level tracking of what new information has been found, and what has already been seen, thus eliminating unnecessary duplication of information in alert services, for instance.

One other function warrants comment, document indexing.

ISYS document indexing may be performed at three levels. At the highest level, a configuration file is created that contains a rule-base of how documents located on various volumes should be treated.

The configuration file may either be created through a series of API calls or, more simply, preconfigured using an ASCII editor or otherwise generated by your application. It is not necessary to use the configuration API to create the configuration file, although you may do so if you choose.

The index is automatically brought up to date by an update process, whereby new documents are indexed, altered documents re-indexed, and references to deleted documents are removed.

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	Supports ontologies, controlled vocabularies, and lists of people, places, and things. Entity references such as gene sequences are also supported
Query Types	Keyword, natural language, Boolean, proximity fielded
Visualization	None (Coming in version 9). Third-party tools may be integrated via the API
Entity Extraction	Built in via proprietary algorithms and a knowledgebase
Platforms Supported	Windows with support for Linux and UNIX
Export	The APL allow export functions to be defined
Third-Party Support	Can be integrated with third-party systems
Vertical Support	None needed
Analytic Functions	Includes a range of built-in reports

Table 31: Technical Highlights for ISYS Search

The ISYS Search Engine automatically scans the disk directories and determines which documents have been created, which have been changed and which have been deleted. Call-backs advise of indexing progress.

The second lower level indexing mechanism is known as the “low level indexing API”, and bestows complete control of the indexing and deindexing process with the application. The OEM application directs the ISYS engine to index and de-index specific files “in the active voice.” The host program provides the “file name” of the file to be indexed. The file name need not be an actual disk-based filename, but can be considered a 255-byte access key that uniquely identifies the document. The host program is returned a 32-bit handle by which the index knows the document. The application becomes completely responsible for deciding which files get indexed and when. The application also decides when files become de-indexed. A special form of de-indexing is also available which provides faster performance if the original text of the indexed document is still available, as is often the case with document management systems, for example.

The third method of indexing is “transactional indexing”, whereby the content source application constructs a transaction file containing various statements of fact, for example, “this document still exists”, “I know this document no longer exist”, and “here is a document, its identifier and its content”. The ISYS Engine reads the transaction file and updates its index according to the statements of fact. This enables applications to update an ISYS index without necessarily having a complete view of the document set in its entirety at any one time.

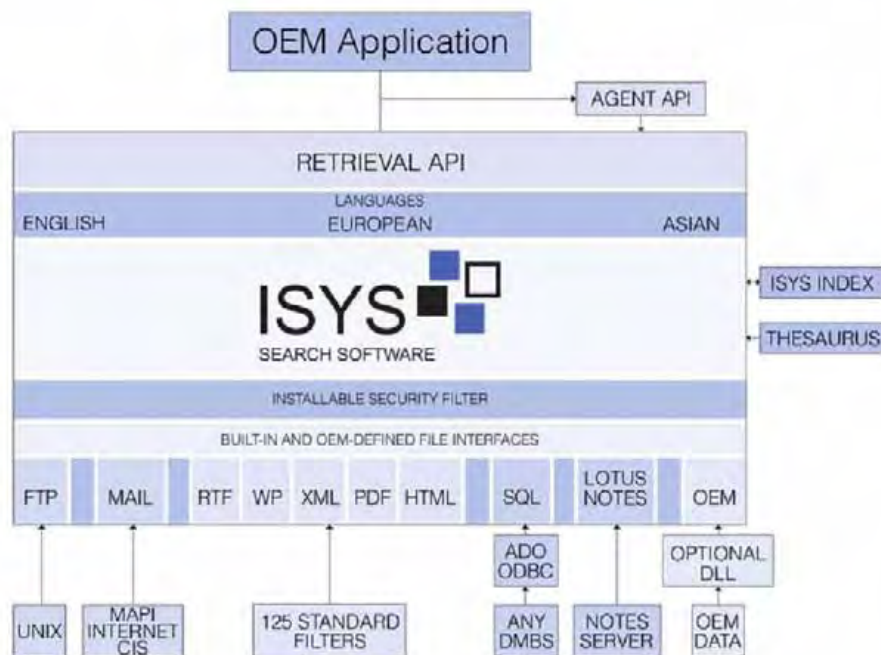


Figure 47: The ISYS API

The program you write talks to the ISYS API. The DLL reads documents; it reads and writes the ISYS indexes. Notice that the ISYS system supports UNIX and Linux system, IBM's Lotus Notes and structured data via Open Database Connectivity.

ISYS supports Windows and Linux natively, with support for indexing information residing on a UNIX box.

Other ISYS Features

The ISYS search system offers a number of features that go beyond key word search.

Multiple Query Methods

The system offers three distinct query methods. Users can select the approach that makes the most sense to them for their needs. ISYS offers a *command line query* that allows advanced users to construct Boolean, proximity and fielded searches.

ISYS also provides a *menu-assisted query*. This is a wizard-style interface that gives intermediate users a way to formulate queries without remembering the Boolean syntax.

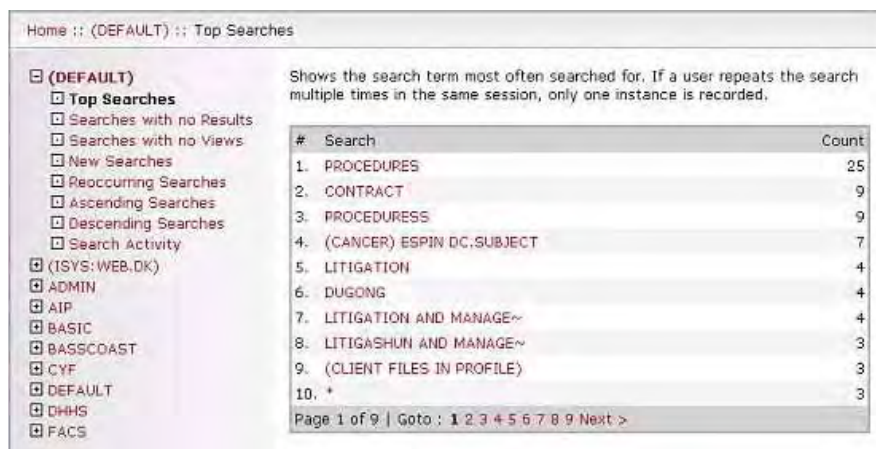
The system offers a *fielded query* or parametric interface that allows users to search structured information, such as metadata and database fields.

In addition, the system includes a "browse taxonomy" option which delivers the point-and-click or navigation via hot links that are characteristic of systems with advanced text processing functionality.

For users with a preference to a Google or Yahoo-style query. Web search systems rely on popularity and other techniques to generate useful results. ISYS users can type one or two words and get useful results. Alternatively, a user can type a question in the search box, and the ISYS system will attempt to generate an “answer”²³.

Included Extras

Other system features include a useful report function. For several years, Mondosoft was the leader in built-in search analytics. Now, ISYS has pulled ahead with its wider variety of reports to help you better analyze search activity and user behavior. Default reports include top searches, searches with no results, subsequent searches, and trends in ascending and descending order. The SDK can be used to link ISYS with a third-party tool such as those from WebTrends.



Home :: (DEFAULT) :: Top Searches

☒ (DEFAULT)

- ☒ Top Searches
- ☐ Searches with no Results
- ☐ Searches with no Views
- ☐ New Searches
- ☐ Reoccurring Searches
- ☐ Ascending Searches
- ☐ Descending Searches
- ☐ Search Activity

☒ (ISYS:WEB.DK)

- ☒ ADMIN
- ☒ AIP
- ☒ BASIC
- ☒ BASSCOAST
- ☒ CYF
- ☒ DEFAULT
- ☒ DHHS
- ☒ FACS

Shows the search term most often searched for. If a user repeats the search multiple times in the same session, only one instance is recorded.

#	Search	Count
1.	PROCEDURES	25
2.	CONTRACT	9
3.	PROCEDURESS	9
4.	(CANCER) ESPIN DC.SUBJECT	7
5.	LITIGATION	4
6.	DUGONG	4
7.	LITIGATION AND MANAGE~	4
8.	LITIGASHUN AND MANAGE~	3
9.	(CLIENT FILES IN PROFILE)	3
10.	*	3

Page 1 of 9 | Goto : 1 2 3 4 5 6 7 8 9 Next >

Figure 48: ISYS Reports

The ISYS reports provide point-and-click access to a wide range of system and support data. The API makes it possible to integrate ISYS with third-party analytics and visualization tools.

ISYS highlights hits and provides what the company calls “hit-to-hit” navigation. These allow quick spotting of terms, a function that is particularly useful in Adobe Portable Document Format files. A user can click a link and jump to the relevant portions of a document.

The ISYS engine includes file viewer technology within its system. When a user want to examine a document’s relevant portions quickly, the ISYS system can display the content without launching a third-party application.

ISYS can process more than 200 file formats in 60 different languages. The native language support includes Chinese (traditional and simplified), Japanese and Korean, and Unicode for multi-language indexes.

²³ This is natural language processing, but the technique converts the query into a Boolean string and displays results.

Technology

The current version has been engineered to permit “plug and play” scaling. Additional computational and storage resources can be added as required. The ISYS administrative interface allows the search administrator to make these available to the system. Once identified, the ISYS engine will use these resources.

Collections

The collection is a fundamental design element in ISYS. A collection is a set of documents, possibly in a department or located on a particular server. ISYS incorporates a metatag collection identifier. It allows a query to be run across multiple collections or limit the query to a specific collection such as *contracts*.

One of the important refinements in ISYS’s current release is its approach to large document collections. Each ISYS index supports 64 million documents. When more documents enter the system, a separate index is generated. Queries execute in parallel across indexes, so performance is not compromised. This “chaining” of indexes allows billions of documents to be accessible to a user. The ISYS SDK and scripting tools make it possible to extend the system to connect ISYS to uncommon formats and systems, or to inject metadata for greater structure of content.

Indexing

The ISYS system includes a thesaurus or knowledgebase for high-frequency terms. The search administrator can add additional terms to what ISYS calls a *synonym ring*. These are Use For terms. These terms are used for spelling corrections that are roughly analogous to Google’s “Did you mean?” function.

The system administrator may exclude content by drive or directory, by file extension, by file type, or other criteria. Indexes typically are between 10 and 15 percent of the size of the source documents, because of the compressed indexing algorithms employed. Our tests showed that the amount of compression varied by type of file. ASCII files reached high levels of compression when indexed. Other file types yielded indexes that were about the same size as the source documents.

Examples of the System in Use

ISYS has several thousand customers. An exemplary implementation of the system is the use of ISYS Search at Cisco Systems. The company uses the engine for its public Web search at www.cisco.com as well as for the company’s Intranet. Both structured (database) and unstructured information are processed by ISYS. The system processes terabytes of content and sustains a high query volume.

Upside

The upside for ISYS Search system includes:

- High-speed processing of structured and unstructured data

- Built in automatic document classification and entity extraction with customization options and support for pre-existing taxonomies and controlled term lists. Seamless integration of key word searching with point-and-click navigation and discovery functions.
- Robust API, which permits customization and extension of the system

Downside

The downside for the ISYS Search includes:

- The company lacks the profile of some higher-profile, more costly systems; therefore, procurement teams may overlook and underestimate ISYS
- Integration of ISYS into some third-party enterprise systems requires custom scripting. Though not difficult, some administrators may prefer native support for Interwoven, Documentum, or other enterprise systems.
- ISYS's approach to scaling may be perceived as requiring more search administrator intervention than systems that do not require collections

Net-Net

ISYS represents an excellent balance of key word search and rich text processing. In one system, the user can enter Google-style queries or point-and-click when particular items of interest appear in a list of entities or a category of related information.

More important, ISYS delivers excellent document processing performance on standard workstations and servers. The competitive price and the deep API makes ISYS a system able to meet a range of search-and-retrieval requirements without the engineering and administrative overhead required by other systems.

15. Lexalytics Inc.

www.lexalytics.com

Lexalytics is a company created by Jeff Catlin, who sold PleasantStreet Technologies to Chiliad Publishing in late 2001. Prior to forming Lexalytics, Mr. Catlin worked at LightSpeed Software, a small content management and document classification company. He was general manager for the East Coast operations, and managed sales, marketing and development efforts. In late 2002 Mr. Catlin's team created a product called the Knowledge Appliance, which was transferred to Amherst Information Group in 2003. In early 2004 Amherst Information Group changed its name to Lexalytics and expanded on the initial Knowledge Appliance. Today the company's flagship product Salience Engine leverages its sentiment analysis technology. The firm's co-founder and CTO is Mike Marshall, an alumnus of Oxford Brooks University in England.

The Catlin-Marshall team has developed products that extract metadata from unstructured content at LightSpeed, and now Lexalytics. In each incarnation of their metadata systems, the team has advanced the processes to ferret increasingly subtle nuances from unstructured text. Its customers license Lexalytics' components and install the software on their servers. Lexalytics offers consulting and customizing services to supplement its revenue from license fees.

Item	Quick Facts
Product	Salience Engine 3.2
Price	\$35,000. Custom price quote required
Key Feature	Ability to measure sentiment or tone at the document, summary and entity levels
Purpose	A suite of products that attack the problem of finding relevant information in unstructured content
Clients	Fast Search & Transfer, Cisco Systems, InMagic, IPro, and Solcara
Company	Privately held
Contact	info@lexalytics.com

Table 32: Quick Look at Lexalytics Inc.

Since 2004, Lexalytics, Inc.--formerly the Amherst Information Group-- has been committed to helping businesses extract, analyze and report on any information contained within their servers or accessible from outside data sources. The firm's tag line is:

Discover. Understand. Act.

The company's first product was the "Knowledge Appliance." By 2005, Lexalytics renamed its product line and focused on technology that can determine the sentiment or tone of a document. The system assigns a score to an information object such as a

document and then manipulates these scores to provide various types of analytic outputs.

The company offers a complete text processing suite to help companies achieve their goal of harvesting actionable intelligence from unstructured content. Lexalytics asserts that it offers one of the few “out-of-the-box solutions” for rich text processing. In fact, some customers report that Salience is installed, tuned, and processing in as little as a day. Lexalytics solutions offer access to information needed to understand the impact of a company’s brand or product messaging, as well as to reduce costs associated with filtering through non-relevant information.

Brand	Hits Near Satisfied	Hits Near Dissatisfied	Normalized Phrase Transaction
Ford	114	59	0.25
BMW	41	20	0.28
Lexus	27	4	0.70
Chrysler	N/A	8	0.47
VW	32	6	0.64
Mercedes	33	33	0.48
Jaguar	10	8	0.69
Honda	52	1	0.79

Figure 49: Lexalytics Report

A Lexalytics “report” provides a user with specific information about customer satisfaction.

Technology

The flagship product in the Lexalytics Suite is the Salience Engine. The company offers a number of complimentary components that can be mixed and matched to some degree:

- Acquisition Engine—a spider and file conversion system that identifies concepts, classifies people and other objects, creates summaries of source documents, and assigns a sentiment value to objects. XML-tagged outputs can be pushed to a database for additional analysis or into a search system’s index.
- Salience Engine—This is a subsystem that: performs entity extraction (people, companies, places, products, dates); discovers entity relationships among people and jobs, their titles, etc.; generates document summaries; calculates sentiment / tone extraction by document, paragraph, summary, or entity.
- Sentiment Toolkit—The Sentiment Toolkit is an add-on module for the Salience Server that provides users with the ability to enhance Lexalytics core sentiment database to work better with their content, or allows users to build entirely new sentiment databases tuned to particular vertical markets like threat detection, or political satire. The toolkit allows inexperienced users to jump in and make simple enhancements to the system tailored to their content, and also allows those users wishing to build out defensible intellectual property with the option to design and deploy a unique sentiment database that will set them apart in their vertical market.

- Classifier Toolkit—The system uses multiple classification mechanisms for each node in the user's taxonomy. Included are tag and keyword matches, query based matches, and signature or training-based matches.
- Analytics Toolkit —this is an application that processes Web log and other content to determine how a brand is perceived. The stories contributing to the sentiment score can be accessed by clicking the graph, which Fast Search uses in its Marketrac product.

Examples of the System in Use

A good example of Lexalytics' functions appears in Fast Search & Transfer's implementations of Salience. To illustrate: Fast Search's Marketrac function converts a list of results into a ready-to-distribute report. The plumbing for this system pivots on Lexalytics' technology.

You can also navigate to <http://www.politicaltrends.info/> and explore the Salience system's ability to provide near real-time insight into user perceptions of key topics. The demonstration focuses on US presidential candidates, but the firm plans to provide other timely interactive examples of its technology when the campaign ends in November 2008.



Figure 50: FAST Marketrac Report Using Lexalytics

Fast Search & Transfer uses the Lexalytics' content processing technology in Marketrac to generate this publication ready report.

Key Features

The most interesting feature the the Salience Engine provides is sentiment and tone at an entity level, which allows customers to discover, understand and act on the information available in Web logs, email, and other unstructured content. In order to help marketing and PR professionals understand where to drive their brand, companies need to understand and organize what's being said about their products and services. This means marketing professionals need the ability to digest thousands or even tens of thousands of messages concerning their product or brand. Analysts, police, and financial services professionals need the same type of data.

Sentiment Analysis

The "secret sauce" in the company's product is its sentiment engine, which can compute the sentiment or tone in, not just a document, but the key entities (People or Products or etc...) in the document, as well. This allows the system to cope with the natural "compare and contrast" writing style often seen in product reviews. Entity sentiment is measured by only considering the toned phrases that occur in close proximity to the

entity being measured. The Sentiment Engine, which ships as a standard component of the Salience Engine, comes with a base dictionary of 250,000 tonal phrases that are used to measure tone. The system also comes standard with a user configurable sentiment file that can be used to tune the engine to different vertical markets.

Sentiment Toolkit

An add-on component of Salience is the Sentiment Toolkit, which extracts tonal phrases from a domain specific corpus, so that users can tune the sentiment database to their particular vertical market. The tool identifies candidate tone phrases (typically adjective/noun phrases) and presents them to the user for inclusion in their “hand scored dictionary” or HSD file. The user has total control over the inclusion or exclusion of these phrases and the actual score of the phrase (e.g. “rotten day” = -0.85). User built HSD files can be integrated into Salience to improve its performance in a particular domain like network security or consumer product reviews.

New Features

Enhancements available in Salience Engine 3.2 include snappier performance when extracting entities and the ability to designate specific directories for a variety of entities including companies, people, products and brands. This modified directory structure allows users to access multiple data directories simultaneously. The new structure allows more control when accessing the content analyzed by the Salience Engine.

Normalization

Salience Engine 3.2 has improved how it can normalize the information collected by associating key names and brands with each other. For example, understanding that *General Motors* and *GM* are the same entity is a key function when analyzing content and returning accurate sentiment results.

Analytics Module

The new Analytics Toolkit 1.0 (ATK) allows users to create graphs and charts for presentations to their clients. The ATK pulls raw data from within content sources and automatically generates presentation quality graphs or charts. A point-and-click interfaces allows anyone with a feed of content – be it RSS or news feeds – to create live content components that can be placed in a report.

These ATK widgets do not require programming and can be used to track sentiment and tone of a client's product or brand. ATK can also extract themes or categories surrounding a release or message. ATK components automatically update if desired. Lexalytics' ATK can sit on top of the current Lexalytics Text Analytics Suite as an add-on tool, or can work with a company's SQL-based knowledge silo of information.

Feature	Beyond Search Comment
Knowledgebase Support	Supports knowledgebases and ontologies. These can be supported using the “go to Web” for guidance innovation
Query Types	Keyword, natural language, Boolean, SQL Query
Visualization	Query Trac provides pie, bar, and other graphic functions
Entity Extraction	Discovery and controlled term lists
Platforms Supported	Linux, Unix, and Windows
Export	Exports tabular reports and presentation-quality documents
Third-Party Support	Can be integrated with such enterprise search systems as those from Autonomy or Fast Search & Transfer, among others
Vertical Support	Versions of the product are available for customer support, brand management, and political analysis
Analytic Functions	Third-party tools may be integrated via an API

Table 33: Technical Highlights for Salience 3.2

Upside

The company has combined proven text mining techniques with the clever twist of using a query passed against another index to resolve ambiguities regarding unknown terms and phrases. The focus on a value proposition that shows a non-technical marketing person the value of an ad campaign via sentiment is an intelligent and valuable capability. The company continues to generate buzz in consumer product companies and PR firms, but it is less well known in what might be called the traditional search and text mining sector. Lexalytics is important because its approach quantifies an area of business that has been difficult to quantify. A desperate marketing manager is one who needs to provide the value of a multimillion dollar campaign and has no data or data that looks like high-value data. Lexalytics can deliver charts that “prove” what are the most appreciated brand and similar fuzzy notions.

Downside

One possible drawback is that the current version of the Acquisition Engine only runs in Windows because it requires Windows libraries to handle Office documents.

The company’s self-funding also limits financial resources for marketing options to compete with other text processing companies.

Another possible issue is that Mike Marshall, one of the principals in the company resides in Scotland.

Net-Net

If you want to provide users of a behind-the-firewall system with reports instead of laundry lists of results, consider Lexalytics. Like many of the companies profiled in Beyond Search, the firm’s profile is lower than better known competitors such as SAS,

Beyond Search: Lexalytics Inc.

SPSS, and other business intelligence-centric vendors. The company's technical team is eager to work with customers to implement the firm's content processing system into existing enterprise applications or incorporate Lexalytics' technology into other behind-the-firewall solutions. The interest in analyzing information for customer preferences and sentiment is increasing. Lexalytics warrants a closer look.

16. Linguamatics Ltd.

www.linguamatics.com

Linguamatics' text mining technology pivots on natural language processing (NLP) and text search. The union of these two techniques is novel, and the company engineers have embraced this challenging "shotgun marriage" with the added functionality of "experimentation" - interactive querying and results exploration for both expert and more casual users. The idea is that the system can understand source materials and present the user with more than a list of documents in which a user's query terms appear. The result is providing the user with a tabular view of documents that match the query with a drill-down option to enable the user to explore the supporting evidence.

The core of the firm's technology is the work of Dr. David Milward, who co-founded Linguamatics in the UK in 2001. Dr. Milward holds a Ph.D. in Computer Science from Cambridge University. In addition to text mining, Dr. Milward has had an interest in spoken dialogue systems. The user interacts with the system via a dialogue--essentially a question-and-answer session--with the system and the user exchanging information.

Item	Quick Facts
Product	I2E, a NLP-based search and text mining system
Price	\$100,000. Custom price quote required
Key Feature	Real-time, agile NLP-based querying able to leverage domain knowledgebases/ontologies, controlled vocabularies, and specialized entities such as gene sequences
Purpose	Allow user to ask questions, interactively extract hidden facts and relationships, and present results in tabular form
Clients	Include Astra Zeneca, Bayer, Biogen-Idec, Hoffman-La Roche, and Pfizer
Company	Privately held
Contact	info@linguamatics.com

Table 34: Quick Look at Linguamatics Ltd.

Examples of the System in Use

Consider a researcher looking for drug receptor interactions. Using a traditional search-and-retrieval system a user would be able to identify only the document in which the protein and the interaction appear. They would be required to come up with a list of proteins and then review each document, looking for the interaction information. Linguamatics I2E helps to eliminate this tedious manual process.

More interesting, Linguamatics asserts that it answers more general questions, which are typically more difficult than technical queries using a very precise set of jargon. For example, a user could send this query to the Linguamatics' system: *John Smith is the*

chairman of which company? or a brand manager can ask: *What are these physicians saying about drug x in therapeutic area y?*

Human Genes	biomarker	breast cancer	Evidence	Link	Score
SNCG SNCG	biomarker is expected to be a useful marker for	breast cancer	SNCG is expected to be a useful marker for breast cancer progression and a potential target for breast cancer treatment.	source 16821081	100
TRIM25 Efp immunoreactivity	biomarker is a significant prognostic factor in	breast cancer	CONCLUSIONS: Our data suggest that Efp immunoreactivity is a significant prognostic factor in breast cancer patients.	source 16144914	100
TP53 the TP53 gene	biomarker are a well-documented strong prognostic factor in	breast cancer	Mutations in the TP53 gene are a well-documented strong prognostic factor in breast cancer.	source 16864176	100
ERBB2 HER-2	biomarker is an important prognostic factor in	breast cancer	HER-2 is an important prognostic factor in breast cancer, and its overexpression is observed in 20-60% of cases with micrometastases in the bone marrow.	source 16685382	100
MCAT MT	biomarker is a potential prognostic biomarker for	breast cancer	Evidence that MT is a potential prognostic biomarker for breast cancer is supported by many reports in the literature.	source 17018874	100
CTCFL BORIS	biomarker can be a valuable early blood marker of	breast cancer	Detection of BORIS in a high proportion of patients with various types of breast tumors indicates that BORIS can be a valuable early blood marker of breast cancer.	source 17062669	100

Figure 51: Linguamatics Search Term Report

Linguamatics displays a report with the search terms highlighted.

Linguamatics provides advanced, interactive natural language processing solutions for organizations, and customized services for clients with unique text processing requirements. The company positions its system as “accessible” text mining for all decision-makers”. The company’s engineers have placed considerable emphasis on running a query, scanning results, and then interactively exploring the results.

The company says:

Our... examples point to a whole extra dimension beyond simply retrieving documents. Put simply: businesses benefit from being able to cope with the vast quantities of information available to them, but the smartest businesses want to do more than just cope--they seek to get extra value from these resources and turn them into competitive advantage.

Key Features

The Linguamatics’ software provides a user-centric, “advanced linguistic analysis”. Features include:

- Grouping words into meaningful units such as relationships and entities, allowing queries with “linguistic wildcards”. This feature allows a user to ask such questions as “What are the side effects of compound X?”
- Proximity controls for phrases and sentences
- Increased recall with support for variant forms of words; for example, *binds* and *bounds* and *binding*.
- Ability to find contextual information such as interactions between proteins in documents about breast cancer
- Search-with-words function to look for sequences, chemical entities, parts of formulae, using substring, wildcard, or regular expression queries.

- The licensee can provide ontologies, controlled vocabularies, and other specialized lists such as names of people or companies. This enables the system to utilize synonyms - a specific set of words or terms referring to the same concept - and classes -- term lists with thousands of terms, for example, gene lists, lists of adverse drug reactions, or proprietary lists.
- Targeted fact extraction from “specific regions” of a document
- Quantitative information extraction to find numeric values, such as dosages, concentrations, times, etc.
- Tabular results output plus options for other formats, including integration with network visualizers like Cytoscape
- A user can save and share queries, merge multiple queries or result sets, and can compare and combine the results of different queries

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	Supports ontologies, controlled vocabularies, and lists of people, places, and things. Entity references such as standard gene identifiers are also supported
Query Types	Keyword, natural language, Boolean
Visualization	Support for third-party tools including relationship maps and mind maps
Entity Extraction	Discovery and controlled term lists
Platforms Supported	Linux, Unix, and Windows
Export	Exports tabular results to Excel, Xml, tsv, csv, sif for post processing
Third-Party Support	Can be integrated with third-party systems; for example, work flow technology from InforSense
Vertical Support	Primary focus is pharmaceuticals, bioscience, and related fields
Analytic Functions	None

Table 35: Technical Highlights for Linguamatics

Technology

To make good on the promises of “interactive information extraction” or I2E, Linguamatics makes use of a wide range of technologies to “understand” both the content and the user’s query. The foundation of the system is natural language processing (NLP). The system processes linguistic information, that is, identification and indexing of parts of speech, phrases, and names of people, places and things in the documents. I2E works through each document to parse, index, tag and link concepts via syntactic and term look ups. The system then “reasons” mathematically about these tagged items.

Within I2E a knowledgebase can be plugged-in to provide information about the content domain.

Linguamatics refers to this knowledgebase as an ontology or thesaurus. This knowledgebase typically provides vocabulary and hierarchical relationships so that the document processing function can identify concepts not necessarily expressed as key words in a document.

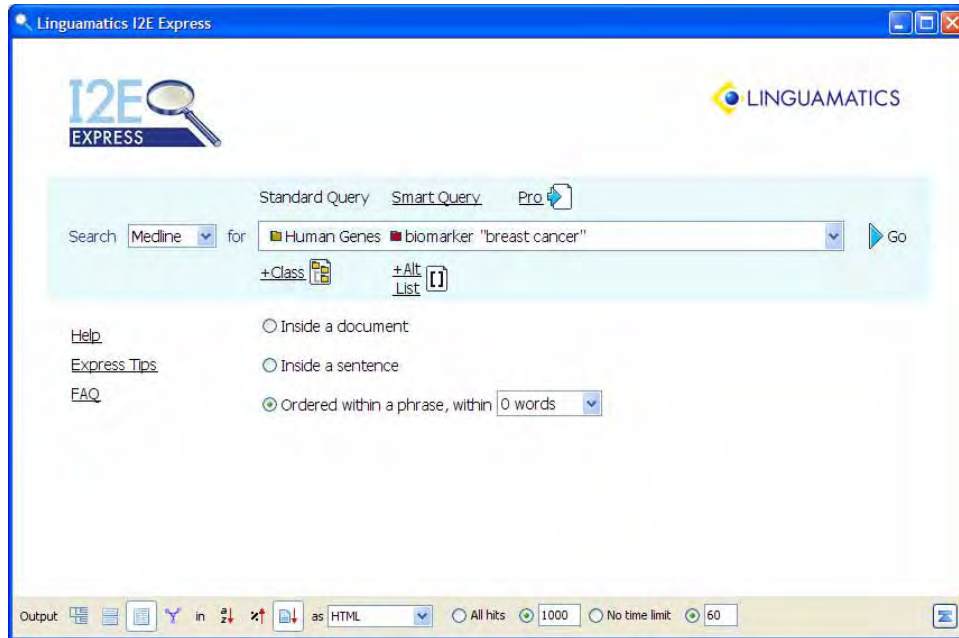


Figure 52: Linguamatics Express

Linguamatics' Express presents a clean, simple interface. Users have a choice of a key word query or a professional interface.

Texts or content processed by the system are tokenized and parsed. Then the system performs entity identification; that is, it looks for the names of people, places, and things. During the entity identification phase of text processing, I2E processes the content against its knowledgebase and controlled term lists. For example, in life sciences, the system can accommodate terminology such as chemical names, protein or gene sequences, and other highly specialized information.

In an example of document indexing, a user might import a set of documents with their authors. The system identifies concepts or entities in the documents, which Linguamatics calls classes. These entities are then used to tag or index documents authored by these individuals; they can also identify specific areas of expertise for and author and correlate those competencies to other content processed by the system.

Ontologies typically provide relationships; a company General Electric or GE may have a subsidiary NBC Universal or NBCU. The NBCU entity may have a program called "30 Rock". With an explicit ontology containing the hierarchy, Linguamatics can allow a user to search for financial results of companies in a business sector or identify that a particular company seems to be targeting its merger and acquisitions' activity toward a specific market sector.

The system converts source documents into a representation of the processed content to permit queries, interaction, and content exploration among the metadata for the

processed content. I2E is domain independent, but the licensee can tailor the system to handle specific domains in chemistry, bioscience, and homeland security, for example.

New Features

The company's flagship product is I2E 3.0. New features included in the January 2008 update are:

- Pre-defined smart queries for sophisticated searching with minimum effort
- More powerful querying capabilities including disambiguation, negation and optional elements
- Enhanced results reporting for rapid analysis of extracted information plus click-through to supporting evidence
- Streamlined system management with the new admin GUI

Upside

The upside for Linguamatics' system includes:

- Strong support for knowledgebases
- Tabular results output is well-suited to scientists and analysts
- Rich configuration options allow the system to be tuned to handle uniquely challenging document collections such as medical research or patent documents.
- Versatility to answer questions in any domain without hard-coding queries
- Queries can be refined by the user to provide the appropriate balance of precision and recall.

Downside

Considerations for the Linguamatics' approach include:

- Quality of knowledgebase content has a significant impact on the system's processing.
- Although the Express interface provides ready access to search tools, some training or familiarization is required to make full use of the Pro interface
- On-going editorial work is required to keep the knowledgebases synchronized with terminology and other domain issues such as sequences, formulae, and other information that is fluid.

Net-Net

The company can make a strong case that users experience increased productivity gained by rapid extraction of new insights and hidden relationships from text, particularly challenging content collections such as scientific, medical, and technical information.

Though controlled vocabularies are not a pre-requisite, Linguamatics may not be the best choice for content domains that lack controlled vocabularies or well-formed ontologies. The cost of the editorial support required to maintain general business information may be prohibitive. However this can be reduced by using the I2E system itself to help develop and maintain vocabularies.

Content that is in well-formed XML or in structured database tables lends itself to the Linguamatics' system. As with other NLP-based systems, unstructured text, particularly in terabyte-sized chunks, may require a significant investment in servers, storage, and infrastructure.

Although the company's marketing collateral and presentations assert that the search and NLP functions are suitable for competitive intelligence, Linguamatics' major focus is currently the pharma industry.

17. Microsoft Corporation

<http://office.microsoft.com/en-us/sharepointserver/FX100492001033.aspx>

www.fastsearch.com

As *Beyond Search* goes to press, Microsoft's acquisition of Fast Search & Transfer appears likely to conclude successfully. This profile is organized in two parts. The first part will discuss briefly the new SharePoint 2007 search. I won't be digging into the free versions of Microsoft search in order to have some space to describe the Fast Enterprise Search Platform (hereinafter ESP).²⁴

Item	Quick Facts
Products	Microsoft SharePoint 2007 Search <i>Fast Enterprise Search Platform</i>
Price	SharePoint Search is included in the SharePoint installation. Fees vary. <i>Fast ESP begins at ~\$175,000; a hosted option is available. A customer price quote is recommended.</i>
Key Feature	SharePoint Search indexes content in a SharePoint environment; other Microsoft servers may be required to access other content types; e.g., SQLServer, Office Server, Internet Information Server <i>Extensive customization is possible.</i>
Purpose	Microsoft SharePoint Search indexes information managed by SharePoint, Microsoft's content management system <i>Fast ESP processes structured and unstructured information</i>
Clients	SharePoint has more than 65 million users in thousands of organizations worldwide; most Fortune 1000 firms <i>Fast Search: Cnet, Dell Computer, Factiva, Reed Elsevier,</i>
Company	Publicly traded Privately held
Contact	+1-703-793-3270

Table 36: Quick Look at Microsoft and Fast Search

The assessment section of this profile will not focus on specific issues with either search platform. Instead, I will identify some high-level considerations to assist you in determining what questions to ask. My inquiries to the two companies about future plans for each firm's enterprise search system have gone unanswered. The lack of

²⁴ For detailed descriptions of the Microsoft SharePoint Search system and the Fast Search & Transfer Enterprise Search Platform (ESP), please see the third edition of the *Enterprise Search Report*. The discussion of these two approaches covers most technical aspects of each system and includes screen shots of the different interface options, including the Microsoft "blue" and the "green" interface in Microsoft Office SharePoint Search (MOSS) and other versions of the system.

information is to some degree understandable, but it does create in my mind significant uncertainty about each platform's role in a single enterprise.

SharePoint Server Search

SharePoint Server 2007 implements search as a shared service. The system collects and indexes content. The current version of the service supports full-text searching using Structured Query Language (SQL)-based query syntax and provides a new key word syntax to support key word searches.

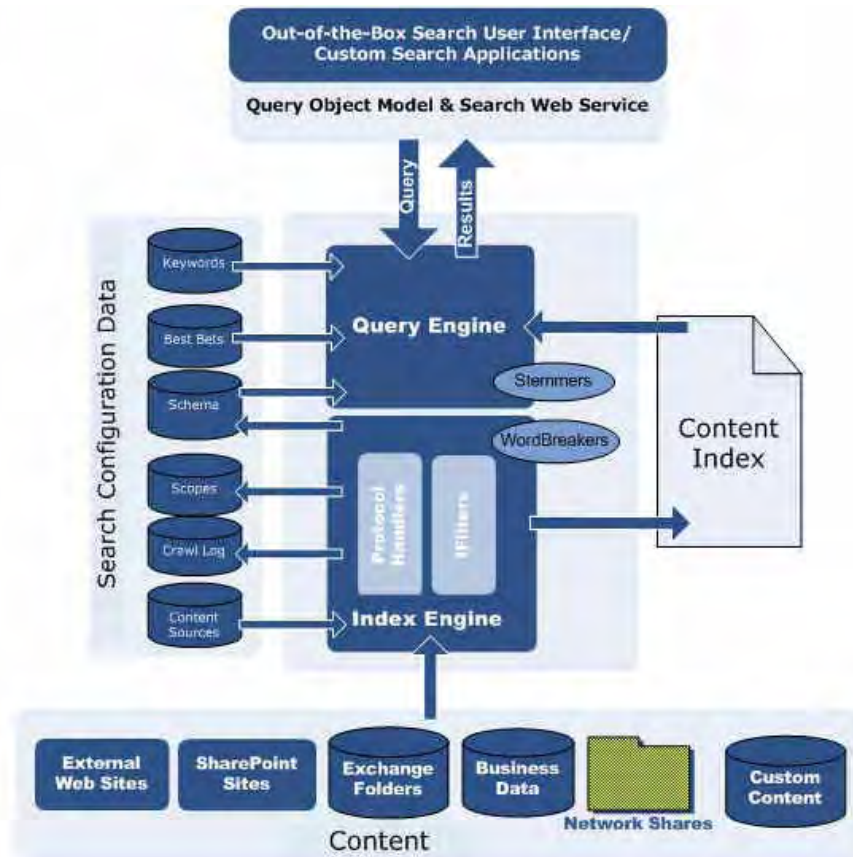


Figure 53: SharePoint Search Architecture

The architecture of SharePoint Search relies on SQLServer and the SQL search syntax. Performance issues can become apparent if the SharePoint search system is not properly resourced; that is, clustering, high-speed storage, and sufficient bandwidth.

The system consists of several components. Among these are:

- Index Engine – Processes the chunks of text and properties filtered from content sources, filing them in the content index and property store.
- Query Engine – Executes key word and SQL syntax queries against the content index and search configuration data.
- Protocol Handlers – Opens content sources in their native protocols and exposes documents and other items to be filtered.

- IFilters – Opens documents and other content source items in their native formats and filters them into chunks of text and properties.
- Content Index – Stores information about words and their location in a content item.
- Property Store – Stores a table of properties and associated values.
- Search Configuration Data – Stores information used by the Search service, including crawl configuration, property schema, scopes, and so on.
- Wordbreakers – Used by the query and index engines to break compound words and phrases into individual words or tokens.

The architecture for the current version of SharePoint Search is similar to previous versions of SharePoint Search.

The key feature of the system is its ability to make use of available metadata for a processed file. For example, the system identifies and tags the author of a document, the date, and other information available in the file system. These metadata plus document indexes allow key word queries, sorting by date, and other useful manipulations.

The default interface can be used as is or customized. If you want to add a pre-coded function to a search interface, you first use the graphical editor included with SharePoint. Then to output the functional code for the system, you open the interface in VisualStudio.Net and save (compile) the files.

Although this two-step process seems complicated, it is a continuation of Microsoft's new effort to separate design from coding.

In the last few years, there has been a surge of search vendors developing versions of their products to replace or supplement Microsoft's own search. The reason for this boom in "snap in" solutions is the large number of SharePoint installations. At the end of 2007, I learned that there were more than 65 million SharePoint installations worldwide. A large market means that some of the SharePoint customers will want additional functionality or features that are not included with SharePoint.

It's not possible to create a comprehensive list of search and content processing vendors who offer "plug compatible" SharePoint Search alternatives. I do want to give you a sense of the range of product offerings. You will want to consult Microsoft's own list of Certified Partners and, if possible, look at vendors on that list. It is located at <https://solutionfinder.microsoft.com/>. Be aware that if you install software into a SharePoint environment that is not certified, you may invalidate your Microsoft support license.

Among the vendors providing "plug compatible" search and content processing systems are:

- Autonomy plc. Autonomy is a relative newcomer with a build of IDOL specifically for SharePoint. IDOL, in a broad sense, competes with the Windows Server family and allows a licensee to develop enterprise information

applications. You can locate more information about this Autonomy solution at <http://www.autonomy.com>, then search for SharePoint. Note: You will have to register to obtain this information. Autonomy's embrace of Microsoft may suggest to some that Autonomy's platform push for IDOL has not been successful in converting some organizations from Microsoft's solutions to Autonomy's.

- Coveo. One of the first search vendors to recognize the opportunities created by SharePoint's wide adoption is the Coveo system. The firm has offices in San Francisco, California. Technical development is located near Montréal, Québec. The system provides key word search, classification, and some other enhanced metatagging functions. As I write this, Coveo is in the midst of a public relations campaign designed to give the company a higher profile. Information is available at <http://www.coveo.com>.
- dtSearch. Located in Bethesda, Maryland, dtSearch is a key word search system that runs in Windows. The system can process information in a SharePoint environment. For advanced content processing, you can integrate the Bitext natural language processing system with dtSearch. The combination provides key word, Boolean, and NLP functions. More information is located at <http://www.dtsearch.com>. Pricing for dtSearch is competitive, and the company has a good profile among developers who want to include a fast, key word system in a third-party application.
- Interse. This little-known Danish vendor offers several enhancements for SharePoint. The company provides a content processing component that automatically categorizes and tags documents in a SharePoint environment. You can learn more at <http://www.interse.com>. When I visited with the company's management team, I learned that Interse had plans for rapid growth with an office in the Washington, D.C., area.
- ISYS Search System. Profiled elsewhere in this report, the ISYS technology provides an easy-to-deploy alternative to the standard SharePoint search. The enhanced content processing functions are available with no fiddling with the existing SharePoint system. ISYS can often deploy its system in less than a day.

The \$1.2 billion dollar offer for Fast Search & Transfer speaks volumes about the compelling need Microsoft had to strengthen its enterprise search product. Let's look quickly at Fast ESP.

Fast ESP

Shortly after Autonomy positioned IDOL as a platform capable of supporting enterprise applications, Fast Search & Transfer responded with its platform. The name Fast Search chose was a memorable one. *ESP* evoked "extra sensory perception" and the notion that a Fast Search installation could deliver on-point information in a supernatural fashion.

Like Autonomy IDOL, ESP is not a single software component. Fast ESP consists of various systems and subsystems that process content, generate results, and perform various value-added functions.

One characteristic of Fast Search that sets it apart from Autonomy is that the software included in ESP comes from different sources. For example, the core content acquisition and indexing engine has roots deep in spidering the public Internet. Thus, the high-speed content acquisition and indexing functions may be seen today at Yahoo!'s AllTheWeb.com service at <http://www.alltheweb.com>.

Onto this core, Fast Search built additional functionality to better serve the enterprise market. Several years ago, Fast Search dropped out of the online advertising and Web search business to focus on behind-the-firewall search. At the same time, Google narrowed its focus to ad-supported search. In the last 36 months, Google's revenue soared to nearly \$16 billion. Fast Search peaked in the \$160 million to \$200 million revenue range and found itself in an intensely competitive market. Autonomy marketed aggressively against Fast Search, and in the last year, Autonomy pulled ahead of Fast Search with its acquisition of Zantaz, an e-mail compliance company.

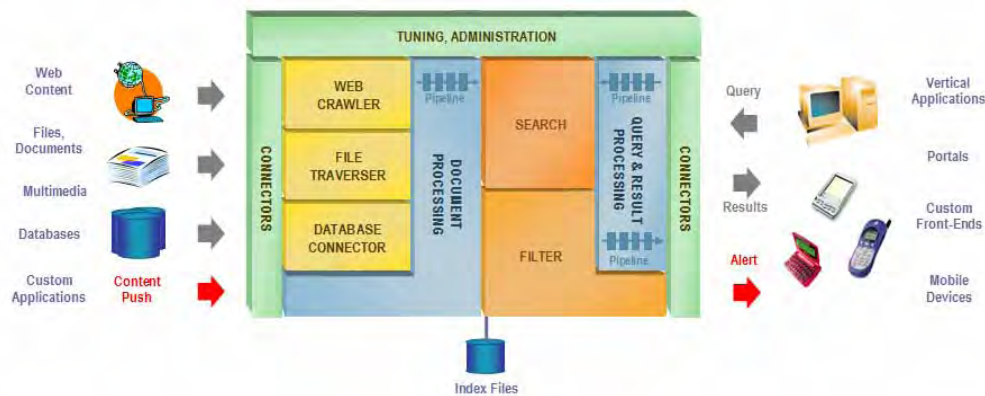


Figure 54: FAST ESP Components

This overview of the ESP system shows its principal components. Note that the system can support on-the-fly output for different devices used by a single user.

Against this background of corporate maneuvering, Fast Search's engineers embraced some open source technology. Blending open source and Fast Search's own proprietary software made it possible for Fast Search to capture a number of high-profile customers; for example, the U.S. government's portal search, the *Financial Times*, and Reed Elsevier's SCIRUS service.

As interest in enhanced content processing rose, Fast Search licensed technology from such companies as Teragram, a little-known content processing systems vendor in the Boston, Massachusetts, area. Fast Search cut a deal with Lexalytics, also in Massachusetts, for technology that would allow sentiment analysis to be included with Fast Search.

Fast Search also acquired companies for specific functionality and to get engineers with expertise in search and related fields. Among Fast Search's acquisitions were Platefood, a provider of online search and search-based advertising services to media firms. I saw this acquisition as a reversal of the firm's earlier strategy of abandoning the online search market. Fast Search also acquired Agent Arts, a vendor of personalization tools.

To recap, Fast Search is a heterogeneous platform consisting of:

- Proprietary code originally developed for a Linux platform. I learned several years ago that Fast Search's engineers employed similar techniques to those in use at Google
- Open source code developed by a community. Fast Search engineers created "middleware" to link the open source components with the proprietary Fast Search engine
- Licensed technology, which Fast Search integrates into its other software components. This approach is used by other behind-the-firewall vendors, and it is a way to extend functionality and obtain some for-fee customization work from licensees.
- Acquired technology, which, if my research is correct, retains its identity. Integration is handled through Web services, adaptors, and middleware.

Other vendors use a similar approach to building out their enterprise offerings. There are two points that one may wish to consider when investigating a Fast Search solution. First, the ESP platform is somewhat heterogeneous. Instead of loading a function, you may have to create a script, fiddle with configuration files, and involve a Fast Search engineer. In some situations, the fiddly bits can slow down a licensee's ability to deploy a needed function quickly. Second, it is not clear how the Fast Search ESP will be handled by Microsoft. At this time, Fast Search offers an adaptor for SharePoint, but major portions of the Fast Search system, in my experience, deliver their best performance in Linux or UNIX environments.

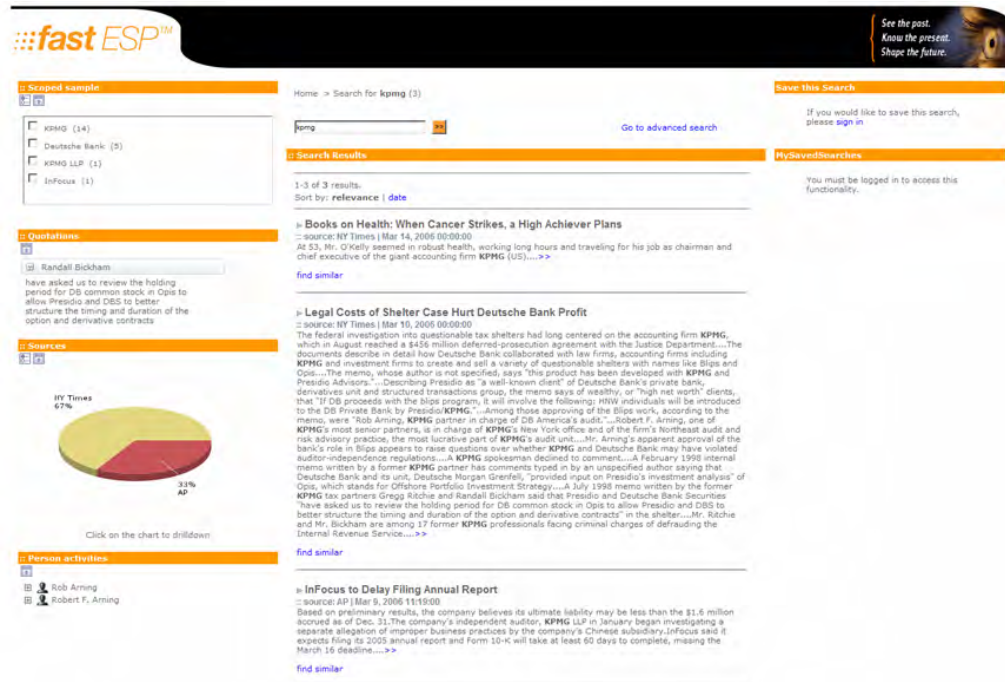


Figure 55: Extracting "Emotion" with FAST

Fast Search uses the Lexalytic technology to extract and process content for its "emotion" or "feeling." These metatags make it possible to provide a brand manager or a customer support supervisor with a way to identify hot spots or broader trends in perception.

Tools, Not Toasters

Beyond Search believes there are two key aspects of the Microsoft – Fast Search deal relevant to you. Before looking at these similarities, let's step back. Microsoft has mounted a number of search initiatives over the years. Prior to the Fast Search announcement, Microsoft has made available:

- Various "flavors" of search for documents on a machine running the Windows operating system. Two familiar to you will be the "search" box on the Windows start button and the "find" feature in Microsoft's e-mail programs. Both work, and both implementations have engendered a number of alternative options, ranging from X1 Technologies to shareware products like Brothersoft Search
- The desktop search "toolbar". Available without charge, this search system indexes documents on a user's PC and the documents to which the user has access on a network. Windows Desktop Search is available without charge.²⁵
- XP users can use the Windows Live Toolbar to initiate searches and find specific e-mail messages from within Outlook. In addition, the toolbar offers protection from online ID theft scams, pop-ups, and viruses. With the Windows Live Toolbar you also gain a rich set of features for organizing your content, including smart search tools and personalization options.

²⁵ <http://www.microsoft.com/windows/products/winfamily/desktopsearch/choose/windowsdesktopsearch.msp>

- You have the previously mentioned SQLServer “search” function and the SharePoint search service, plus other search features embedded in various Microsoft products; for example, the Xbox search system which has been licensed from a third party.

Feature	Beyond Search Comment
Knowledgebase Support	SharePoint: Can be supported. Coding required <i>ESP: Can use existing taxonomies and word lists</i>
Query Types	SharePoint: Key word and Boolean <i>ESP: Key word, Boolean, assisted navigation, and ready-to-distribute reports</i>
Visualization	SharePoint: Microsoft Graph, but custom scripting required <i>ESP: Optional modules can generate graphs and other graphic elements</i>
Entity Extraction	SharePoint: Metadata attached to a document <i>ESP: Depending upon features licensed, entity extraction is supported</i>
Platforms Supported	SharePoint: Windows server only <i>ESP: Linux, Unix, and Windows via an “adaptor”</i>
Export	SharePoint: Microsoft XML and other Office formats <i>ESP: XML. Custom formats can be created via the API</i>
Third-Party Support	SharePoint: Certified Partners can integrate SharePoint into most third-party enterprise applications. Note: CRM Live may require the use of non-standard programming languages <i>ESP: Native support for Documentum and other content management systems; the API permits integration with most third-party enterprise applications</i>
Vertical Support	SharePoint: none <i>ESP: Publishing, customer support, and government</i>
Analytic Functions	SharePoint: Can be set up to use Excel and Microsoft Analysis Services <i>ESP: Analytic components are available from Fast Search</i>

Table 37: Technical Highlights of Microsoft SharePoint Search and Fast ESP

For developers, Microsoft makes it possible to create customized implementations of search, integrate third-party search systems or components into Microsoft environments, or code your own search engine with VisualStudio.Net.

The point is that Microsoft’s approach is to offer many types of tools and products to its users, partners, and developers. If anything, Microsoft offers so many options that I find it difficult to keep them properly separated. In our tests, I have learned when installing several of these products on one machine, the indexing processes slow down the test platform. I have learned that it is important to pick a Microsoft search technology and learn to get the best out of it.

I find it interesting that Fast Search has a complementary approach. The company offers its licensees many options from which to choose. For example, you can select from AdVisor (structured data indexing), Folio Publisher (a content vending system), InPerspective (a relevancy tuning product), LiveAnalytics (a report generation component), and ProPublish (content repurposing and rights management), among others. Each of these products can be integrated into the Fast ESP and further customized with the company's Application Programming Interface (API).

The Similarities

Microsoft and Fast Search are platforms that may be extensively customized. Both companies offer a number of products that can be difficult to differentiate. The implementation of a search system boils down to using different components in order to assemble a search solution. These tasks can be completed successfully by engineers who have a solid grasp of each company's technical approach. In most cases, both Microsoft and Fast stand ready to provide the engineering and technical services necessary to:

- Select the particular components needed to meet your requirements
- Integrate the necessary components into the platform
- Tune, debug, and deploy the system

We see, therefore, that the business and technical approach of Microsoft and Fast Search are similar.

The Differences

There are a few significant differences between the two companies. You may want to consider these as you chart your course for next-generation search and content processing.

First, Microsoft is new to the Linux/Unix operating system. Fast Search does support Microsoft SharePoint and Windows servers. The core of Fast Search is Linux, and in many ways, the company's culture has some similarities with other Web indexing systems originating in the late 1990s.

Second, Microsoft relies on its product management approach to products. For many years, Microsoft targeted an upgrade path, made specific engineering changes, and then offered the upgrade to its customers. Microsoft retains this "push" approach. Fast Search, on the other hand, has been a reactive company. For example, Fast Search has emulated the actions of Autonomy. Some of the animosity that exists between these two companies is due in part to this imitative or "me too" behavior.

Finally, success is defined differently at each company. Microsoft wants to increase its market share for search in the enterprise. The number one job is to stop the defection of SharePoint licensees to non-Microsoft third-party solutions. Fast Search wants to win search customers from its arch rival Autonomy and become the dominant provider of search in the enterprise. These two goals are sufficiently different to make for some challenges to integrate the technologies and make deployments work.

The Company

When the deal goes through, *Beyond Search* believes that operations will continue without much change for a period of six to nine months. After the first six to 12 months of the two partners' honeymoon, it is difficult to predict what the organizational structure of the search unit will be.

Upside

The uncertainties of an acquisition are numerous. *Beyond Search* is not a business book; it is a review of advanced content processing technology. The benefits of a successful merger of Microsoft and Fast Search can be multiplied because of the financial resources and market reach of Microsoft. Fast Search contributes a customer base of about 2,000 enterprise licensees of ESP, and a number of content-processing components that are, based on my research, loosely integrated.

My work has given me access to information that suggests three upsides for readers of this study:

1. Microsoft's marketing will, if the Fast "brand" remains distinct, increase inquiries, proposals, and ultimately sales. Autonomy and Google are likely to feel this impact. But the companies most likely to experience slower or more difficult sales into Microsoft-centric organizations will Certified Partners offering their own proprietary search and content processing solutions. For pure-Microsoft shops, getting software directly from Microsoft means that Microsoft will at some point resolve certain issues and stand behind the terms of the license.
2. Microsoft has an opportunity to offer a homogeneous enterprise search solution. I have to assume that the political and technical issues will be successfully resolved. Once done, Microsoft can offer a system that eliminates the confusing, Balkanized systems now deployed.
3. Microsoft gets an injection of search expertise. Despite Fast Search's management missteps, portions of the company's technology are quite good; namely, the Web spidering and indexing components. Despite the cloudy financial environment, people with search and content processing expertise are in short supply. Fast Search has some excellent engineers who can make an immediate and direct contribution to its likely new owner.

Are these benefits enough to make the \$1.2 billion acquisition pay off? Over the long haul, if Microsoft can manage its other business interests wisely, the Fast Search deal will benefit Microsoft. It is difficult to determine how the deal will unfold for existing Fast Search licensees two or three years in the future.

Downside

Generating a long list of negatives for this proposed deal is easy. Let me identify the two major negatives and leave it to the reader to make his/her own decision about the validity of the negatives identified by Web log authors and various search experts.

First, an acquisition is complicated. Compared to issues with jurisdiction and philosophy, the technology is challenging but manageable. As Microsoft meshes Fast Search into the corporate polity, licensees are likely to receive strong reassurances that it is “business as usual.” These reassurances are a matter of form. Anyone who has worked with a vendor acquired by another company knows that some things (usually short-term projects) are unaffected. Other parts of a license for software and services will change, often dramatically and with little warning. In my experience, fees and license terms can shift. Notification of a change may arrive by e-mail sent by a nameless, faceless person in contracts or accounting. The account manager may not know that a change was in the works or made. Communications within behemoths like Microsoft are often mired in bureaucracy or simply go unread because of the tyranny of the schedule. Microsoft executives are busy, often scheduling routine meetings weeks or months in advance, if you are already a customer. You will want to keep in close contact with your Fast Search representative and work to get a Microsoft contact as well. Contingency planning is a wise licensee’s first order of business.

Second, Microsoft has edged closer to the Linux/Unix community. The deal with Novell is an encouraging step because it helps ensure interoperability. Based on my experience with Fast Search since 2000, I have found the company to be more like an Internet start up. Microsoft is no longer a start up. Fast Search has been quick to react to the actions of its immediate competitors and opportunistic in its push into publishing. The cultures of the two companies, therefore, are sufficiently different for me to believe that two situations may arise. Both of these may affect you if you are considering the Microsoft-Fast Search “platforms” or are now a Fast Search licensee:

1. Integration of Fast Search’s components with SharePoint may focus on the use of Windows server technology, not Linux/Unix. If you want to run a 100-percent Linux operation, you may find yourself having to bite the bullet and license Windows server technology to get the features you want. Alternatively, you may have to do without a desirable feature and look for a third-party solution despite the uncertainties of integrating the various software components.
2. New features may be available only for Windows operating systems. If you are looking for a system that will be in operation for three or four years, you will, in effect, be making a commitment to the Windows platform. The likelihood of Microsoft “going Linux” is increasing among certain factions at Microsoft. The revenue power of the Office and server units is sufficiently strong at this time to keep Linux/Unix as a second-class operating system option for the foreseeable future.

Net-Net

The Microsoft-Fast search acquisition is a significant development in the behind-the-firewall search market. However, I want to keep the deal in perspective. Microsoft’s offer is approximately six times Fast Search’s estimated 2008 revenue, which I peg at about \$200 million. For reference, I estimate that Autonomy will generate about \$330 million in calendar year 2008 and Google about \$400 million in the same time period

from its Google Search Appliance. In terms of behind-the-firewall search, Microsoft has an opportunity to gain market share and freeze third-party vendors' sales into SharePoint installations.

In terms of making a licensing decision about Fast Search, I would move forward with close investigation of Fast ESP and similar products. If you are looking at an installation life of one to two years, I don't think the Microsoft buyout will have a significant operational impact in that time frame. If you are looking farther into the future, you will want think through your licensing terms and your organization's interest in becoming increasingly dependent on the Windows server technologies.

How will this deal affect the more than 100 vendors offering behind-the-firewall search and content processing systems? For smaller vendors with highly specialized systems, I don't think the deal will make much difference. However, for vendors now competing against Fast Search in companies with more than \$500 million in revenue, the presence of Microsoft will make life more difficult. Microsoft's market presence and its existing software installations give Microsoft an advantage only IBM, Oracle, and a few other software and system vendors have – Microsoft can “give away” search as part of a higher-value system such as Live CRM or adoption of Office and Vista on desktops with SharePoint as the search and content management system. Therefore, what would have been a separate procurement is now bundled into an existing or larger deal. Autonomy, Endeca, IBM, and Oracle are the vendors most likely to be directly affected by the acquisition.

Google and up-and-coming vendors like ISYS Search Software and Siderean Software will be able to make sales due to their systems' respective features and price-performance ratio.

Behind-the-firewall search is becoming a utility, if not a commodity. Firms wanting to save money and move away from Microsoft systems can use Lucene or a lower-cost key word search solution such as Tesuji. Content processing remains a more complex and, at this time, a non-commoditized function. I believe that the firms offering value-added content processing will have an opportunity to increase their sales because licensees of Microsoft-Fast will want to add certain features without the cost and uncertainties inherent in the buyout.

Bottom line – some vendors will face a difficult time in certain markets. Other vendors will find a way to flourish. The more interesting issue looms in the future when Google and Microsoft clash in the behind-the-firewall search market's most lucrative segments. But we'll have to wait a year or two for that struggle.

18. PolySpot SAS

www.PolySpot.com

For lovers of architecture, PolySpot's office in Paris, France, is a short walk from the 18th-century Beaux-Arts style façade of Gare Saint-Lazare. PolySpot is anchored in the 21st-century despite the historicity of its location immortalized by Claude Monet in his 1877 painting of the neighborhood. Olivier Lefassy, the dapper CEO of PolySpot told Beyond Search in December 2007:

The area is very convenient, but we are so busy creating new features and functions for our 360 degree search engine, I don't think too much about the past, just about the customers who want to break out of the chains of the search vendors who promise more than their systems can deliver.

Item	Quick Facts
Product	PolySpot Enterprise Search
Price	\$50,000. Custom quote required
Key Feature	Rich text processing and collaboration tools with assisted navigation and key word search
Purpose	Provide a 360 degree view of information
Clients	BNP Paribas, SchlumbergerSuez Environment, Belgium Police
Company	PolySpot SAS
Contact	sales@PolySpot.com or info@PolySpot.com

Table 38: Quick Look at PolySpot SAS

Paris is a hotbed of innovation. You can read elsewhere in this report about Exalead, but there are Datops, Lingway, and the eclectic Kartoo, among others. French technology influenced Groxis, a company doing business in San Francisco.

The cause of this efflorescence in French search technology are French universities. The quest for a better search solution occupies entrepreneurs from Saint Lazare to Montpellier.

Mr. Lefassy said:

We French think there is a better way to search. The key word is useful in some situations, but now we have users who know that search systems must do more than generate a long list of results that make the user do even more work opening documents, scanning them, hunting like bird dogs for the elusive information needed to do a job.

PolySpot is a private-held firm. Investors in the firm include CIC Banque de Vizilla. With 24 full time employees, we estimate that the firm's revenue in 2008 will be in the \$5 to \$7 million range.

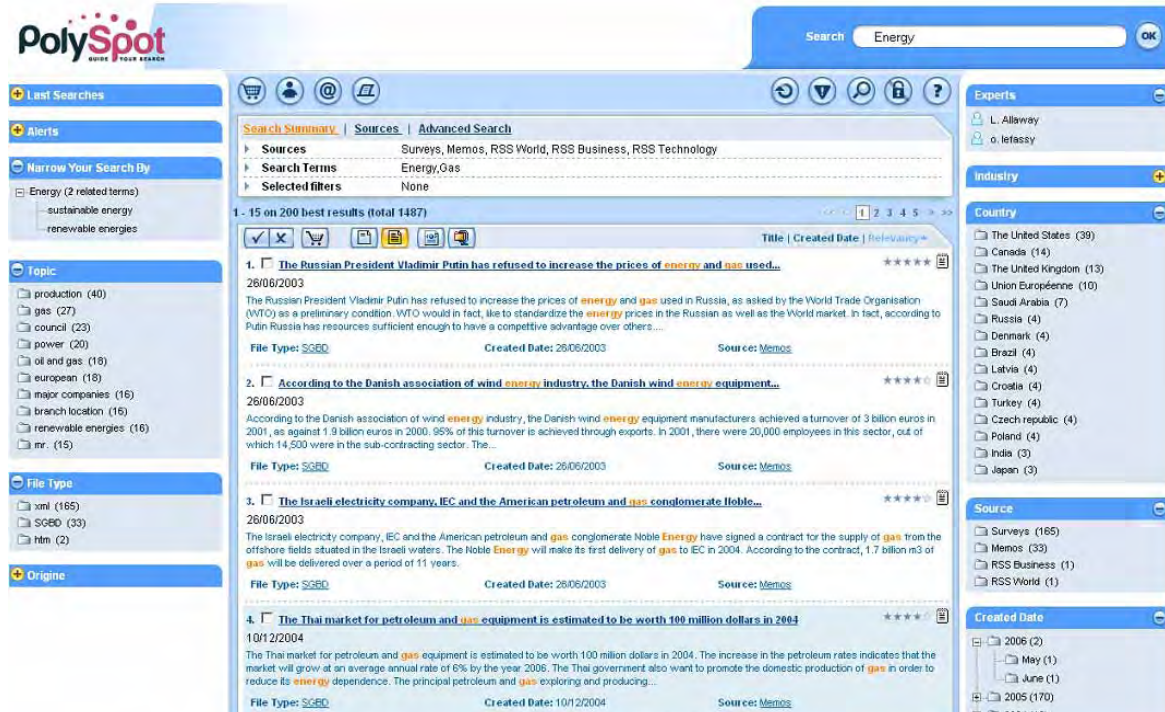


Figure 56: The PolySpot Interface

The PolySpot interface uses a multi-pane display with hot links in the outer two panes and the results in the center of the display.

Hybrid Technology

PolySpot Enterprise Search combines both a statistical and a semantic approach to handling information. Not only does PolySpot have powerful automation features but it also supplies a Terminology Manager, which is included in the Web administration tool. Customers can also upload their own corporate thesauri into the Terminology Manager. This provides customers with the flexibility and control they demand. The solution is definitely not a “black box”, which is the opposite of Autonomy’s well-known *Neurodynamics*’ approach.

PolySpot offers customers the ability to automatically generate specific terms that are derived exclusively from the corpus of customer information. As part of the unique technology developed by PolySpot there is an embedded terminology extractor which automatically provides suggested terms such as synonyms and hyponyms (specific terms) to each community of users. Thus a user who has searched the term *energy* will also be able to retrieve documents including the terms *gas* or *petrol*. Furthermore PolySpot can also suggest searching on specific terms such as “renewable energy” or “sustainable energies”. This comprehensive search functionality provides customers with a flexible infrastructure for their search strategy. PolySpot asserts that their customers not only require great search functionality, but also PolySpot’s commitment to developing product providing applications that take their customers beyond key word retrieval.

Company CEO Olivier Lefassy told Beyond Search:

By giving our customers extended information discovery, navigation and collaboration tools. We are improving information access with intuitive features our customers insist they must have to deal with today's information challenge.

The PolySpot system incorporates a distributed architecture. The modular architecture of the system makes it possible to customize the system, tune its performance, and add or delete sources.

A metadata browsing function is included with the system. The system manager can include existing knowledgebases such as thesauri.

PolySpot's Enterprise Search & Retrieval solution offers each user within the company a 360 degree view of the available information by searching simultaneously the internal and external sources. These include: shared directories, Intranets, Electronic Document Management Systems (EDMS), Content Management Systems (CMS), Relational Databases (RDBMS), proprietary databases, personal files, mail boxes, Web sites, search engines, on-line services, RSS feeds, Audio & Rich media files and more.

PolySpot provides an information retrieval infrastructure that enables companies to automate their operations and processes on all types of information.

PolySpot has designed a highly efficient dynamic drilling-down technology. The technology is based on a unique software development utilizing advanced probability algorithms based on the theory of *Information Gain*. The more a term or a phrase is rare within a document or a corpus, the more it is relevant. In order to suggest to end users relevant search refinements, PolySpot has developed a technology that can retrieve specific terms that are part of the company content and that are associated with a given query.

Other features of PolySpot include:

- "More like this" feature so a user can mark a document and locate similar documents
- "Push" newly indexed content to a user based on his/her interests
- "Find an expert" function

The PolySpot platform provides a Search Service API and uses Java standard language for better customization for business application integration needs.

Rich Text Processing

Those people who have responsibilities for managing corporate information can also enhance the quality of retrieved information by developing their own taxonomies (or folksonomies). PolySpot offers a complete, comprehensive range of terminology and taxonomy management tools from those enabling automatic classification and categorization of information to those enabling users and communities to tag and further improve the quality of their own corporate knowledge.

Analytics

PolySpot provides a market leading analytical tool as part of its portfolio. PolySpot can generate a trend analysis across selected collections of information over a given time-frame. The goal here is to enable users to visualize the behavior of connected terms to a given input over a certain period of time. This tool dramatically helps end-users to detect and react to new and emerging issues that are hidden within the vast quantities of information. The business benefits for such functionality are numerous. They include examples such as dynamic monitoring of online customer interaction on web sites and across call-centers, product research & development, financial analyst information; the potential uses are endless.

Federation

PolySpot is a federation engine (see the discussion of Dieselpoint for more about federated search). PolySpot can process and present content in a single interface. The federating services include:

- Access to specific content on Intranet depending on a user's access rights. Users can search across the company's content and access only information where rights are granted.
- Indexing of content in structured databases such as Oracle, SQL Server, MySQL, XML, and others

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	Can use existing thesauri and taxonomies
Query Types	Free text, assisted navigation, and Boolean
Visualization	No
Entity Extraction	Yes
Platforms Supported	Windows, Linux, Unix
Export	Export the result set in Zip, Xml, Excel
Third-Party Support	API allows third-party integration
Vertical Support	No
Analytic Functions	System and term use metrics; analytics may be extended via the API

Table 39: Technical Highlights for PolySpot Enterprise Search

- PolySpot includes a tool that can analyze and understand each different structure of a given database. Users can then search and filter through all the available data or fields.
- Access to documents in Content Management Systems (CMS) PolySpot includes adaptors to access Documentum, E-Room, SharePoint, Vignette, Basis, FileNet, DocsOpen, iManage (Interwoven), Tridion, among others.

- Web sites and RSS feeds can be aggregated from an unlimited number of URLs. Processed content is automatically categorized by user-definable criteria. The automatic categorization and classification process can also be fed into the corporate taxonomies or user controlled schema. This ability to automatically update PolySpot also updates the taxonomy knowledgebase. If intervention is required, an administrative interface is provided.
- Single-sign on. The system can manage access to content from commercial, third-party services such as Factiva, Lexis-Nexis, Gartner, and Elsevier, among others. Users do not need to sign in to execute each query.



Figure 57: PolySpot Search Term Highlighting

PolySpot highlights the user's search terms in retrieved document.

Big News: Collaboration

Users can easily e-mail, print or download the retrieved documents, which they can quickly export as a list of documents in XML or as compressed (zip) files; they can store any queries or documents either in user directories or in shared file repositories. These can be made accessible by the members of the same network or community. Furthermore, users within networks or communities, or indeed across the enterprise, can fully interact with the information they have found by adding notes or commentaries to the documents and then sharing this added value with one another.

Other features include practical user controls. For example, unlike other information retrieval vendors PolySpot offers intuitive document control and collaboration features as standard. Documents that are of general interest can be easily stored for later reference in user-definable folders. There is also the facility to submit these documents

into Public or Group folders so that colleagues can share, review and collaborate on them.

Users can also define their personal preferences with respect to their search needs. User can explicitly define their own personal areas of expertise and interests in order for the system to be tuned to an individual's requirements. Each user can develop a Personal Tracker, which can launch customized searches at regular intervals (that is, each week, every four days, every day, in real time). The results can be made available through the personalized section of the user's interface, or e-mailed direct to the user at appropriate intervals.

Examples of the System in Use

PolySpot has several customers who are enthusiastic about the system.

BNP Paribas, one of the principal financial institutions in France, uses PolySpot to access its myriad internal data sources. Michel Benadina, a executive with the bank, said: "PolySpot was able to index our internal content as well as repositories outside of the company. No other search system we had licensed had been able to give us this access."

SUEZ Environment a branch of the SUEZ Group, provides equipment and services in energy and the environment. The company replaced its former search system with PolySpot. The organization had several terabytes of information and wanted to add assisted navigation and collaboration to the search service. PolySpot implemented "unified search" across the content, and provided federated access to documents, Lotus Notes email, and content from external Web sites. The system provides access to these data to more than 72,000 employees with an annual turnover of \$12 billion.

Upside

The upside for PolySpot's system includes:

- A large number of features, ranging from on-the-fly classification, to assisted navigation in a pleasing default interface
- SDK and an API make it possible to integrate, customize, and extend the core rich text processing functions
- A federated search solution that may be used instead of a portal solution from IBM or BEA Systems
- Support for a wide range of file types and third-party applications without requiring the licensee to pay for extra adaptors for Lotus Notes, for example

Downside

The downside for PolySpot's solution includes:

- A modest presence in the North American market, so customer and technical support is available from London or Paris

Beyond Search: PolySpot SAS

- The multiple functions of PolySpot make it difficult for a potential licensee to determine what specific functionality is optimal for a particular search installation. PolySpot can deliver a wide range of functionality, which may be more difficult to evaluate than an appliance solution from Google or EPI Thunderstone, for example.
- The low profile of the company in North America may make some procurement teams give the PolySpot solution less attention than a more well-known vendor's system.

Net-Net

Beyond Search has been favorably impressed with French linguistic and search technology. PolySpot delivers on a number of rich text processing features. Its most interesting addition is a well-conceived approach to collaboration and sharing certain types of information functions from the basic interface. PolySpot matches up well against offerings from companies with a higher profile. If you are looking for a way to deliver search in a portal or dashboard package, PolySpot delivers.

19. Recommind

www.recommind.com

Recommind is an information management software company that helps organizations index, retrieve, categorize, review and analyze information. Recommind's applications include an accurate and automated search engine, categorization, expertise location, email management, eDiscovery review and analysis, taxonomy management/development applications, personalization and recommendation functionality and intelligent software agents functionality.

The big push in search is to do more than index key words. Recommind is one of the search systems that is pushing into indexing documents by the concepts in them. Unlike some enterprise search and retrieval systems, Recommind's approach is automatic. Eliminating the manual labor associated with classification allows organizations to scale information management at the same rate of growth of electronic information. Recommind told *Beyond Search*, "With the tremendous explosion of electronic information, organizations are struggling to organize, file, access, retain, and when appropriate, delete information. The only way to keep up with this is to automate the categorization of information, so that organizations know what information they have, who can access what information, and what information they must produce or delete according to government regulations."

Item	Quick Facts
Product	MindServer Enterprise Search 5.2
Price	Pricing is on a per seat basis, or approximately \$50,000 per processor
Key Feature	A proprietary statistical algorithm named PLSA (Probabilistic Latent Semantic Indexing), "Smart Filters" to enable deep faceted search, and federated search across internal and external sources
Purpose	Provide access to both structured and unstructured information in more than 30 languages in common file types, Web sites, and data repositories
Clients	Bayer, Bertlesmann, Eversheds, DLA Piper, Novartis, Shearman & Sterling
Company	Recommind
Contact	info@recommind.com

Table 40: Quick Look at Recommind

The current release of Recommind's core technology is MindServer Enterprise. Over the last three years, the product family has undergone substantial growth. All of the company's products, including its vertical market bundles for law firms and companies, are anchored in the company's patented PLSA technology. Customers can license enterprise search, categorization (including entity extraction and taxonomy management), or select a vertical market combination of search and categorization specifically tailored for legal matter, ediscovery review or news content. MindServer technology crawls and indexes text from sources including document management

Beyond Search: Recommind

systems, records management systems, email, intranets, Web sites, CRM applications, databases, and file systems and repositories.

Recommind's approach to search relies on making sense out of groups of documents by categorizing them automatically into concepts. Its technology can recognize the ideas behind a search and break down the results with greater specificity than keyword-based search engines.

The key feature of Recommind is the ability of its systems to automatically discover concepts in the processed document set. Recommind's technology can be installed and aimed at content in multiple data repositories – all accessible from one intuitive search interface. After the processing, the licensee can search the information by key word, phrase, or the discovered concepts. Pre-built connectors are available for any JDBC-compliant database, including: Anacomp Case Logistix, CA Filesurf, CMS, EMC Documentum, IBM Commonstore, IBM Lotus Domino, IBM Websphere, Interwoven iManage, Interwoven WorkSite, Lexis Nexis Applied Discovery, Lexis Nexis Concordance, LexisNexis Interaction, Microsoft Exchange, Microsoft Sharepoint 2003/7, nMatrix, OpenText DOCOpen, Thompson Elite, XML archives, Oracle, DB2, and SQL Server.

Recommind sees its products as enabling users to quickly and easily organize and find information, without the irrelevant noise returned by more basic search engines.

Customers

Recommind's customers consist of a number of high-profile licensees including the Novartis pharmaceuticals and the German publishing giant Bertelsmann AG.

The company's core market is in large law firms and legal departments. The company has more than 100 firms using the MindServer Legal product, making Recommind one of the largest vendors of search technology in this market segment.

Technology

PLSA is a machine learning technique that can automatically identify and structure relevant concepts and topics from a document collection.

PLSA is a patented algorithm that performs a statistical analysis of word co-occurrences in documents. The algorithm then identifies repeatable contexts, topics or concepts in which a certain group of words occur.

A search-and-retrieval system based on PLSA does not require any manual input in the form of lexicons, thesauri, or topic annotations. The system is completely automatic, operating with what information retrieval professionals call unsupervised learning. PLSA generates its own representation of the content in a compressed, quantitative form. The small size of the description and its numerical properties allows Recommind systems to deliver brisk performance for most search-and-retrieval tasks.

Feature	Beyond Search Comment
Knowledgebase Support	Can obtain documents on Intranets, from databases, third-party content, document and records management systems, email and Web pages
Query Types	Key word, natural language, Boolean, fuzzy queries
Visualization	N/A
Entity Extraction	Yes
Platforms Supported	Windows, Unix, Linux
Export	N/A
Third-Party Support	Integrates with CMS and other enterprise applications
Vertical Support	Yes – MindServer Legal, MindServer Media
Analytic Functions	

Table 41: Technical Highlights for MindServer 5.2

Recommind's identification of concepts or topics serves two purposes.

- The process eliminates most of the ambiguity associated with words. A reference to *jaguar* may refer to the animal, the automobile brand, or any number of other meanings. Recommind can determine which specific meaning applies in a specific use of the word.
- The PLSA-based system learns about synonyms and semantically related words; that is, words that are likely to occur in a common context. No language-specific or domain-specific thesaurus or dictionary is required.

How PLSA Works

PLSA is, on the surface, similar to the technology of a few vendors in that it utilizes a statistical approach to providing its concept-search functionality, yet there are important differences. One vendor who claims to have conceptual search, utilizes an approach that only takes the “fingerprint” of the query the user typed in and matches it to documents. This approach is useful for finding documents that are similar if you already have a document you prefer. However, for typical short queries most users rely upon, this style of concept searching can lose predictive power. For one-word queries this boils down to a simple key-word search.

PLSA is different in that the concepts are generated from the corpus itself, so that the identification of the concepts is independent of the query. Even short queries get the benefit of the concept search.

The principal differences are buried in the algorithms that enable the document processing, indexing, and query processing subsystems. In general, the PLSA routines perform a statistical analysis of each document in the collection, rolling up data so that metadata about the collection are generated as part of the process. The relationships

among documents and concepts are used to provide point-and-click access to related documents.

PLSA organizes its findings accordingly, arranging each document by category. The technology makes it possible to find related documents that do not necessarily contain the specific word in a user's query. Run a query on *Java* and the system can separate documents about terrorism in Indonesia, on coffee brands, or on Sun Microsystems' virtual machine technology.

Because PLSA does not require linguistic rules or language-specific word lists, the system supports retrieval on any language that can be tokenized or broken into words. For organizations working in an industry with specific jargon or technical terminology, the Recommind system processes these terms without need for training sets or manual preparation of specialized word lists.

A PLSA system “learns” directly from the unstructured content the system processes. Recommind argues that its approach offers several advantages to a licensee. These are:

- PLSA is applicable to any language, as long as the language can be tokenized (broken into words). Languages commonly dealt with in Recommind systems include all major European languages, Russian, Chinese, Japanese, and Korean.
- The extracted concepts are specific to the given document collection and have been automatically adapted to the language, technical terms, and specific jargon of that collection. Building such a thesaurus manually would be costly and time consuming.
- PLSA also generates a “numerical” model in which each word has some probability to occur in a certain concept. The model allows a PLSA-based system to quantify the relationships among words. Recommind is not aware of any other thesauri or linguistic resources providing this type of quantitative information.

The PLSA-based system uses the numerical model to estimate the probability that a certain word will be used as a query term for a document. The result is that the results will include words that did not explicitly occur in a document, but are semantically implied by related words that did occur. For example, a document containing the terms *car*, *accident*, *traffic*, etc. may not contain the word *automobile*. Documents containing these terms would be a reasonable match for the query *automobile*. Because PLSA identifies the *car* topic, it can associate related terms like *automobile* with the document. Hence, the document will have a high probability to be relevant to a query like *automobile*.

SDK

Recommind MindServer offers a Software Developer's Kit (SDK), which provides programmatic hooks to integrate Recommind's technology into third-party software applications. The SDK includes programming tools and code samples as well as documentation and code snippets.

The SDK supports:

- Automated categorization of information using Recommind's patented technology
- Automated generation of metadata for documents
- PLSA-based concept-based retrieval
- Integration and incorporation of data from multiple repositories
- Content filtering and extraction
- Automated indexing and linking of structured and unstructured information
- Automated hyper-linking of documents
- Automated mapping of documents into XML

Figure 58: Recommind's Advanced Query Interface

The advanced query interface allows point-and-click narrowing and access to categories discovered by the system when documents were processed. Specific types of content can be selected so that the user can limit the query to specific information resources.

Upside

The MindServer system delivers concept-based retrieval that does not require costly and time consuming training and tuning steps. In one, mid-range package, Recommind provides its licensees with a tool that can access most content repositories in an organization from a single interface.

Other benefits of the MindServer system include:

- PLSA is language independent, so the system can operate on documents in any language
- Drop down boxes allow the user to filter and access the full result set by the categories discovered by PLSA

Beyond Search: Recommind

- Federated search enables search across multiple locations through one interface
- Expertise location is automatically derived by joining information across different information repositories to provide a complete snapshot of a person's expertise
- "More like this search" allows a user to use a sample document as the basis for another search

Downside

The company's visibility is strongest in the legal market in the United States. Although the company has some blue-chip clients in Germany, the firm's marketing has not created a high profile for the company among Fortune 1000 firms in the U.S. Other considerations are:

- The system relies on algorithms, not linguistic and semantic features which are included in other vendors' systems.
- The system has a lower profile than other search systems. As a result, Recommind's capabilities are not as widely known in some markets.
- Under Firefox 1.5 opening a new window caused a fluttering in the display that was correctable by dragging the windows away from one another. This is a bug that is likely to be fixed when the company ships updates to the current version of the product.

Net-Net

Recommind provides a solid, customizable solution for organizations wanting to process terabytes of information automatically. MindServer provides conceptual, Boolean, smart filtering, and "more like this" searching features that deliver solid results.

The highlighting feature is particularly welcome when a "more like this" search has been launched. The MindServer system automatically formulates a query based on the document's key words. The highlighting of search terms makes it easy to see how the "more like this" query was constructed.

Lawyers, researchers and knowledge workers will find the ability to slice and dice the data by the discovered categories particularly useful. For example, in a result set, a user can scan the categories and subcategories and one click displays only the documents assigned to those specific categories. Using the deep smart-filtering feature, the result set can be sliced by facets including person, company, date, or any other concept or system assigned metatag. This "slicing and dicing" feature makes access to both structured and unstructured information quick, relevant and painless. Unlike keyword search, PLSA automatically discovers concepts in unstructured content so that email, documents, and Web page content can be organized and accessed by concept, phrase, author, industry, time, geographic location, or any other "field" discovered and assigned by the system.

20. SchemaLogic Inc.

www.schemalogic.com

This company provides a metadata management system. Think of it as a content management system specifically designed to keep metadata in a single management system. The idea is that metadata controlled from a SchemaLogic server avoids a situation in which an employee must learn two or more ways to locate certain information.

Item	Quick Facts
Product	SchemaLogic Server
Price	~\$150,000. Custom quote required
Technology	Proprietary metadata management system
Key Feature	Provides an “air traffic controller” function for an organization’s metadata
Purpose	Keep an organization’s metadata consistent across different enterprise applications
Clients	Boeing, Chevron, Corbis
Company	Privately-held, Kirkland, Washington
Contact	+1 425 885 9695

Table 42: Quick Look at SchemaLogic Inc.

SchemaLogic eliminates the manual hassles of using the same metadata in a SharePoint environment and a Documentum content management system. In addition to making life easier for users of search and retrieval systems, SchemaLogic reduces the cost of managing metadata.

The Company

SchemaLogic was founded in 2003 by Breanna Anderson and Teveor Traina. Mr. Traina sold Compare.Net to Microsoft in 1999, recruited Ms. Anderson, and launched SchemaLogic, the pioneer in the metadata management niche. Not surprisingly, the firm has competency in Microsoft SharePoint, Microsoft’s still-evolving content management and collaboration platform and Microsoft enterprise search technologies.

Breanna Anderson retired in November 2007 as SchemaLogic’s chief technical officer. Prior to SchemaLogic, Ms. Anderson was a software architect and program manager at Microsoft from 1995 to 2001. She was the architect of SchemaLogic’s complete suite of enterprise metadata management solutions and authored the firm’s key patent, “Schema Server Object Model”, US 2004/0181544 A1 (patent application publication, Sept. 16, 2004) .

The company shipped its first product in November 2003, and in the last four years has refined the company’s metadata content management technology.

The CEO is Jeff Dirks, who supplemented Andrei Ovchinnikov June 2003, prior to the “official” launch of the company. In March 2007, Mr. Dirks helped the company obtain an additional \$14.7 million in financing led by Goldman Sachs with participation from Chevron Technology Ventures and Madrona Venture Partners, among others.

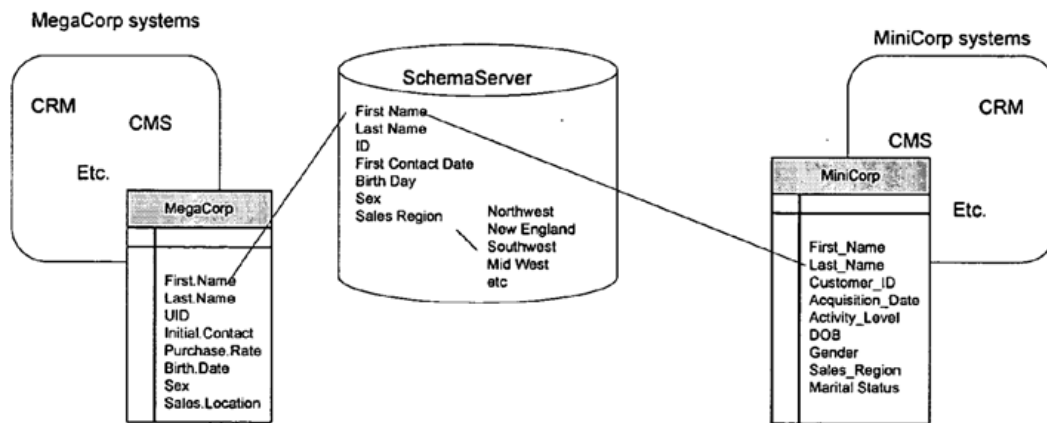


FIG. 10b

Figure 59: SchemaLogic's Metadata Management

The SchemaLogic architecture implements a separate metadata management system. [From Patent US 2004/0181544 A1]

The SchemaLogic server provides a number of built-in controls.

This series C round of funding keeps SchemaLogic at the red line of investor’s metadata fever. SchemaLogic’s initial funding in 2003, is estimated to be about \$5.0 million from Seattle-based Phoenix Partners and several other investors. Then, in December 2004, the company completed a \$4.6 million series B round in, led by Seattle-based Madrona Venture Group. With the March 2007 cash infusion, Beyond Search estimates that SchemaLogic’s total funding is in the \$26.0 million range, a significant bet on metadata management.

We estimate that SchemaLogic’s revenues are in the \$15 million range. Privately-held SchemaLogic does not reveal its financials, but investors have confidence and great expectations for strong growth in the metadata management niche created by SchemaLogic.

SchemaLogic told *Beyond Search*, “Enterprise search is becoming the de facto way of finding information in corporations today. By providing consistent metadata structure and meaning via aggregated taxonomies and controlled vocabularies, we enhance SharePoint and other systems’ functionality. Text mining analyses gain an immediate boost in accuracy due to the use of standard metadata across all systems.”

SchemaLogic says that its software can increase “findability”, efficiency and agility using the company’s “master metadata” framework while reducing metadata management costs.

What SchemaLogic Does

SchemaLogic is one of a small number of companies offering a system that complements selected search-and-retrieval and content processing systems. SchemaLogic told Beyond Search:

Our company provides the only enterprise scalable platform and collaboration and management of a metadata plan, installation of the plan in SharePoint, and synchronization of the evolving metadata throughout the global SharePoint environment.

The problem SchemaLogic “solves” is the different versions of metadata that separate systems create and use. Users expect one set of terms and nomenclature, and systems use variations of metadata.

For example, in an organization with SharePoint and Documentum, some users interact with both systems. SchemaLogic provides a unified information management system. Documentum manages metadata as part of its “managed change process”. SharePoint uses a more relaxed and dynamic approach to metadata. SchemaLogic provides a mechanism to perform metadata mapping between the SharePoint and the Documentum metadata. In addition to providing consistent terms for the users of these systems, the enterprise can maximize the utility of the separate systems.



Figure 60: SchemaLogic's Architecture

The SchemaLogic system provides an enterprise-wide knowledgebase. Other enterprise applications tap into the SchemaLogic taxonomy repository. A user will be able to use standardized concepts regardless of the application exposing concepts and terms.

The Approach

The core product is the SchemaLogic Enterprise Suite. It provides a framework that provides the functionality required to map and manage semantic standards underpinning controlled term lists, taxonomies, and knowledgebase content.

SchemaServer is the active enterprise governance repository for enterprise vocabularies and taxonomies. Think of it as a content management system for metadata. The SchemaServer feeds the normalized metadata to other enterprise software systems requiring metadata. In a sense, the SchemaServer is a subject matter expert/editor and workflow engine for metadata.

The SchemaLogic Workshop is the desktop graphical application. Authorized users can manipulate the knowledgebase and interact with the various mapping functions used in the system. SchemaLogic also offers a Workshop Web product. This browser-based tool allows users to view, manage, and collaborate to develop what SchemaLogic calls “enterprise-wide semantic models” or the rules the SchemaLogic system applies to metadata. The idea is that users of the system are in the best position to enter a new term or modify a concept mapping. The SchemaLogic system handles the updating across the various enterprise systems tapping into the metadata repository, thus metadata are “in sync” without further manual intervention.

Technology

SchemaLogic employs a partner program to meet the needs of its customers. SchemaLogic offers consultants and resellers a partner program. Partners receive access to technology and training from SchemaLogic. SchemaLogic and its partners cooperate with marketing the SchemaLogic server. The idea is that partners contribute specialized expertise in vertical markets, content management, XML, data integration or information architecture to design and implement solutions using SchemaLogic technology. SchemaLogic partners include Microsoft, IBM, EMC2/Documentum, Fast and Metalogix.

Jeff Dirks said in Dan Keldsen’s Web log in 2006:

We really see collaboration and participation at the heart of what we are calling business semantics management and frankly a key differentiator in the SchemaLogic solution because it allows for us to control the process of how a group, an individual, a community or even an enterprise strives toward the common version of definitions, knowledge, know-how and corporate memory.

How It Works

The operation of the system is similar in some ways to content management systems. With SchemaLogic, organizations create and manage taxonomies centrally, and capture multiple perspectives by mapping back to the central definitions. For example, different departments might use IBM, I.B.M. and IBM Corp - these variants will resolve to IBM Corporation, and pull all relevant information.

The “models” used by SchemaLogic are a combination of rules and term lists. The system can range from handling tagging sales regions to complex multi-faceted taxonomies with thousands of terms.

SchemaServer describes the structural models used to store and exchange information as a hierarchy of information classes. This logical modeling capability allows a licensee to capture a consistent, easy-to-understand model of the information systems in the enterprise. The model specifies how these systems interrelate with each other. The system accommodates:

- Relational models; that is, traditional database schema
- Object-oriented models; that is, a class such as Products is used as a framework to instantiate an instance of a class such as a specific product.
- XML; that is, structured documents
- Service-oriented architectures; that is, metadata applied to Web services description language

How these different models' metadata are applied and what metadata to use are functions handled within the SchemaLogic system.

SchemaLogic can capture relationships between semantic and structural models.

One way to think of SchemaLogic is to visualize the system as what Ms. Anderson calls "a digital metadata librarian". The difference is that the SchemaLogic server automates most of the time-consuming work needed to keep metadata synchronized across different enterprise software systems. To get around the bottleneck of manual updates, SchemaLogic allows users to make adjustments to the term lists, thus reducing the cost of maintaining the knowledgebase.

Keep in mind that the SchemaLogic system must be configured, its basic rules tweaked for your specific organizational requirements, and the workflow and other rules must be tweaked. Once set up, the SchemaLogic system can operate silently and with modest manual intervention and tuning.

Concept

SchemaLogic's server is a centralized repository for defining and maintaining content type metadata definitions. It also houses list values that can be distributed and monitored for compliance across the distribution content systems, data farms, and site collections in an organization.

SchemaServer allows information architects to assign semantic relationships (such as "author of," "related to," "component of," "skills needed for" and "cause of") and other descriptions, making it easier from a system-independent perspective to mine and analyze the most relevant information.

SchemaLogic supports most Java-compliant environments. The current release adds certification for AIX, HP-UX and Sun operating systems to the existing support for Windows. SchemaLogic supports a variety of Application servers, including IBM WebSphere and BEA Web Logic. Databases certified for this release include DB2, Oracle, and SQL Server. Taxonomy and metadata terms include global language support. Structural metadata is expressed in XSD, shorthand for XML schema definition files, and taxonomic metadata comes in the form of taxonomies. Authorized

users can interact directly with the SchemaServer. When end-users are permitted to add or modify metadata housed in the SchemaServer, they can do so through the system supported Web-based forms.

SchemaServer describes the structural models used to store and exchange information as a hierarchy of information classes. SchemaServer is a database and content management system that manipulates:

- Content classes
- Elements
- Vocabularies
- Terms
- Vocabulary views

Product Line Up

The principal products available from SchemaLogic include the Suite and the Server. Other offerings include:

Workshop

The Workshop is a user interface designed to allow users who need to define and govern the data models. Workshop is a desktop graphical application with functions required to import, model, rationalize, and manage the synchronization of metadata models, schemas, and business semantics.

The Workshop provides a user-friendly way to interact with the object-oriented data modeling environment used by SchemaLogic. A user can manipulate relationship types, extension property fields, import templates, and data views.

SDK

The SDK (software development kit) provides developers with a set of service utilities, documentation and sample code. The SDK allows customers to integrate the semantic and structural models with other enterprise systems. The SDK allows licensees to create import filters or “adapters” to manipulate file types not supported by the system’s built in filters. Also, SchemaLogic offers professional services to customers working with the SchemaLogic Enterprise Suite SDK.

SOAP API

The SOAP application programmers interface is accessible from Java, Dot Net, JavaScript, and other SOA-compatible languages. The API can make calls to the principal features and services of the SchemaLogic components.

The System in Use

SchemaLogic’s customers include the Chevron, Associated Press, and Boeing, among others.

Chevron uses SchemaLogic in its Global Information Link (GIL3) initiative. Chevron has deployed Microsoft Office SharePoint Server 2007 as its core search engine. SchemaLogic makes it possible that Chevron's digital information assets are described in a consistent way globally across Chevron's many operations. The SchemaLogic system is used to synchronize MOSS and the SharePoint installations within the Chevron organization.

The Associated Press uses SchemaLogic to index its content in five languages and repurpose its information. In addition to classifying news stories consistently, the AP relies on SchemaLogic to reduce indexing and knowledgebase maintenance costs.

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	Allows licensees to create and use knowledgebases
Query Types	Key word
Visualization	No
Entity Extraction	Identifies and extracts people, places, and things
Platforms Supported	Linux, Unix and Windows
Export	XML format
Third-Party Support	Microsoft SharePoint, IBM WebSphere, Oracle, Tomcat, and SQL Server
Vertical Support	Integrates with most enterprise search and text mining systems
Analytic Functions	Reports about index term and metadata usage in the licensee's organization

Table 43: Technical Highlights for SchemaLogic Inc.

Upside

SchemaLogic's master metadata framework enables simpler access, integration and delivery of distributed information via enterprise software. When properly configured and resourced, SchemaLogic can reduce information management and programming costs.

SchemaLogic is a sound choice when you want to:

- Deploy a consistent, cross-system metadata repository
- Reconcile differing views of what categories are needed under a topic heading in a taxonomy
- Enforce tagging and nomenclature standards
- Make taxonomic dependencies visible and editable
- Offer a collaborative process to normalize metadata

The workflow and integrator components of the SchemaLogic solution provide process and technology that enables IT workers to synchronize metadata across different enterprise systems.

Downside

SchemaLogic occupies an interesting position in the market. On one hand, the company's technology makes it possible to synchronize metadata in a cost-effective way. On the other hand, some organizations may be unaware of metadata and the problems that inconsistent metadata create for users. Consequently, SchemaLogic like some other vendors discussed in this report find that sales cycles are long and often require "missionary marketing".

Other issues associated with SchemaLogic include:

- You will need to determine which specific component of SchemaLogic functionality best meets your needs (e.g. taxonomy development and management).
- A SchemaLogic installation requires dedicated hardware and a careful configuration and deployment process. A short cut can create a trouble-shooting headache for overworked information technology professionals.

Your team, SchemaLogic, or a third-party integrator will have to dot the "i's" and cross the "t's" to ensure that you get the functionality you want with a minimum of custom scripting.

Net-Net

Beyond Search has found that inconsistent metadata plagues most enterprise systems. The problem is that awareness of the cost-to-fix problem and the headache for users is low. On the bright side, metadata awareness is increasing.

For prescient information technology managers, SchemaLogic helps reduce manual reindexing and metadata maintenance. But some managers may wonder if the costs of special purpose software, dedicated servers, and an additional burden on the existing information technology will offset the six-figure cost of a SchemaLogic deployment.

With a \$26 million bet on SchemaLogic, the investors will be riding herd on this promising company's technology and market success. The company is one of the first to create a niche by applying content management discipline to the problem of keeping metadata in sync across an organization. If you are wrestling with metadata heterogeneity, SchemaLogic's system may provide the solution you need.

21. Siderean Software Inc.

www.siderean.com

Siderean founder and chief technology officer, Bradley Allen, told *Beyond Search*, "It's time for computers to allow you to find information the way you think."

Siderean's Seamark Navigator system dynamically organizes the available data to leverage a person's ability to recognize the information that's needed. "Key word search is a barrier," Mr. Allen continued. "We've tried to give busy people a way to obtain needed information by scanning suggested resources and by point-and-clicking on potentially relevant suggestions in a conversational interface. The key word search is available at any time. If the user finds one path less useful, a single click returns the user to a previous point in his or her research."

Item	Quick Facts
Product	Seamark Navigator
Price	Hosted option \$3,000 - \$5,000 Monthly. Perpetual license \$150,000 - \$500,000. Custom price quote recommended.
Technology	Proprietary implementation of Semantic Web standards to permit relational navigation over internal and external information stores
Key Feature	Permits assisted navigation and handles structured and unstructured information; find and follow relationships between information objects; low-latency throughput
Purpose	Automatically generates indexes for entities, categories; perform automatic classification
Clients	Media and Publishing firms, Technology companies, and Marketing Organizations
Company	Siderean Software, privately-held
Contact	sales@siderean.com

Table 44: Quick Look at Siderean Software Inc.

At some point in the near future, most Web pages and standard office documents will have "Semantic Web" tags that identify the structure of each document and other vendors' systems will be able to generate these tags. For now, Siderean is one of a small number of companies able to generate and assign these enhanced tags.

The Seamark Navigator system has been developed to exploit metadata associated with documents that comply with the RDF (Resource Description Format) standard. Since most data does not currently comply with this standard, Seamark Navigator is able to ingest a wide variety of information sources such as: syndicated content (RSS/OPML/Atom), databases, file systems, XML and more, and convert their aggregated and inferred metadata into RDF.

RDF may not be as well known as XML, but RDF is synonymous with the functionality of the Semantic Web. An information object tagged by Siderean can be manipulated in many useful ways.

What Seamark does is systematically examine the various data sources to which it is introduced, perform sophisticated content analytics on the data, and produce a metadata description of its content and characteristics. The system automatically generates a browse-able, prototype application based upon that description.

The easiest way to grasp how Siderean's system can make content visible is to look at an implementation for Oracle Corporation for their marketing events site. Notice that a user can scan available events and jump directly to events of interest without having to formulate a query. In addition, the conversational interface which includes navigation suggestions, map mashups, and related web events is updated with every search or navigation click.

The screenshot displays the Oracle Events website interface. At the top, there's a navigation bar with 'Events' highlighted. Below it, a featured event 'Oracle CIO Executive Summit' is shown with a date range of March 31 - April 2, 2008. A 'FIND EVENTS' section follows, with filters for 'In-Person Events' and 'Web Events', and a date range from 2008-02-07 to 2008-05-08. A map of Europe is shown with red pins indicating event locations. To the left of the map are filters for 'Date' (This Week, This Month, etc.) and 'Location' (Germany, United Kingdom, etc.). Below the map, a list of events is displayed, including 'Oracle Database Forum, 7 februari 2008, Stockholm' and 'Oracle at GSMA Mobile World Congress 2008'. On the right side, there are sections for 'Most Popular' (Oracle at COLLABORATE 08, Nordic AppsDay 2008, etc.) and 'Downloads' (Oracle Database Lite 10g, Oracle Audit Vault, etc.).

Figure 61: Oracle's Use of Siderean

Justin Kestelyn, OTN (Oracle Technology Network) Editor-in-Chief says, "I, for one, think this is the coolest app ever to appear with an Oracle.com header on it - by far." The idea is to increase Oracle event registrations and exposure to web events by making events easier to find and by showing related web events as the user navigates thru in-person events.

How Seamark Navigator Works

Siderean makes use of two complementary functions to index a document.

First, the system identifies the words and phrases in a document. The system generates a “traditional” searchable index from these. The basic system is bundled with Lucene, an open source search engine, to perform key word queries but can also be integrated with the Google Search Appliance, Oracle’s Secure Enterprise Search, Yahoo’s API and other third-party search engines.

Second, Seamark Navigator processes metadata attached to a document and generates a representation of that data. According to Mr. Allen, “Our index is somewhat like double entry bookkeeping. You can look at data and easily cross reference information.” With these indexes, Siderean allows data to be sliced and diced in many different ways. “It’s similar to an Excel pivot table except the system does the complex part. Users can pivot to different information views thru a simple point and click interface,” he added.

Technology

Seamark Navigator is a java-based suite of software. Documents processed by Seamark Navigator are transformed into tables. Seamark Navigator can be delivered as software as a service (SaaS) or installed as a turn-key system. Siderean can configure the system and deliver it ready to process content on the licensee’s premises to streamline deployment. The basic Seamark Navigator consists of one or more servers that perform the functions needed to make content accessible. The Seamark Navigator may consist of multiple servers or clusters. The configuration depends upon the volume of content and the frequency of changes to that content.

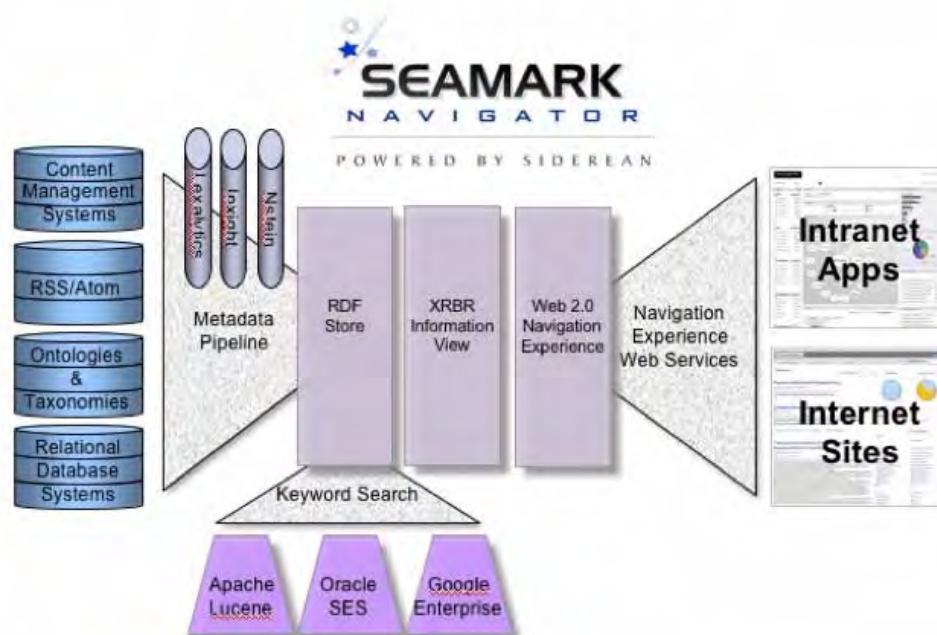


Figure 62: Seamark Navigator

The Siderean system consists of four primary subsystems. The first component is the metadata pipeline. The metadata pipeline processes information objects (documents, database tables, html, etc), applies metadata to content, and indexes the words and phrases. The system automatically seeks and finds relationships among documents.

The second component is the metadata (RDF) store. The metadata components exist as tables along with an inverted index. Categories, subcategories, and document counts are generated from the data in the metadata repository.

The third component, an Information View creation, starts from an automatically-generated “best guess” by Seamark as to the appropriate navigational interface given the metadata processed and the class of object to search and navigate. The executable query specification, XRBR (XML for Retrieval by Reformulation), can be tweaked by the administrator to specify facet suggestions, features of results, text searchable fields and other variables to inform Seamark Navigator’s multi-dimensional user experience.

The fourth component implements the Navigation Experience or what you may want to think of as a query processor. In the Siderean system, this component handles interaction with your browser and performs filtering, visualizations, user participation features like tagging and sharing and other post-processing chores.

You can extend the functionality of any of these modules via the Siderean API. For example, you can integrate a metadata management system such as SchemaLogic’s or add/daisy chain additional entity extraction software such as InXight (now SAP) and Lexalytics.

Features

The Seamark system integrates various and disparate data sources (both structured and unstructured from both inside and outside the enterprise). When new content becomes available to the system, the indexes are updated and the interface automatically reflects the new categories, subcategories, and content counts. The Web-ready, Seamark-generated application can be used “as is,” refined as necessary for look, feel or function, incorporated into a Web page, or linked to other applications as a Web service. If knowledgebases or taxonomies are available, these can be used by the system as it generates metadata.

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	Yes, the system can use available controlled term lists and knowledgebases
Query Types	Key word, Boolean, assisted navigation
Visualization	Supports a wide variety of visualizations including relationship explorers, date sliders, charts and graphs, map mashups, and more. Third-party tools may also be used to implement graphic display of content
Entity Extraction	System can discover entities and bound phrases via built in processes. Word lists and dictionaries are supported.
Platforms Supported	Linux, Unix, and Windows

Export	XML; other formats possible via the API and custom scripts
Third-Party Support	IBM's UIMA specification; any RDBMS, commercial content vendors providing News XML and almost any other enterprise application exposing SOAP or Java Server Pages
Vertical Support	Not required
Analytic Functions	Built in analytic functions. Additional applications can be integrated via the API

Table 45: Technical Highlights for Siderean Seamark Navigator

An organization that has a word list that includes *See Also* and *Use For* references can integrate these connections in Seamark Navigator. The system administrator can set up the system to allow a user to annotate and tag data, thus adding a social or folksomic dimension to the Siderean system. Metatags are added from the Dublin Core and SKOS vocabularies. These documents are then made navigable in the Seamark system using the dc:subject (tag), dc:creator, dc:publisher (site), dc:moderator (feed) and dc:date as the facets. [Note: dc=Dublin Core Standard metadata tagging]

XRBR

Siderean uses XRBR (XML for Retrieval by Reformulation) to facilitate its slicing and dicing of processed content. This query language makes it possible for a text-centric system to manipulate concepts the way an online analytical processing system like Cognos manipulates numeric data. XRBR gives Siderean's system a way to show information to a user from different vantage points. Another advantage of this XML format is that it sidesteps the scaling issues associated with RDF. One additional advantage of this approach is that results can be generated in a page layout form specified by the user.

Customers

The company has a number of high-profile customers. These include Oracle Corporation, Jet Propulsion Laboratory (JPL), and The Financial Times.

Other customers include:

- Environmental Health News archives from Environmental Health Services (<http://www.environmentalhealthnews.org/archives.jsp>)
- Librarian's Internet Index (www.lii.org)
- Chipworks (www.chipworks.com)
- SpendMatters (www.spendmatters.com)

Upside

If you need to address indexing problems or user demands for assisted navigation as well as a search box, you may want to take Siderean's system for a test drive. Seamark Navigator can inject digital content navigation into existing applications or Web sites, or be used through its own navigation user interface.

Based on our tests, Siderean's approach makes it possible to be up and running with assisted navigation and such features as personalizing views for content within the last five or six days. If you need even faster content processing, the company can offer a hosted solution. You receive an XML stream of tagged content ready for use in your existing search system.

Other benefits of Siderean's approach include:

- Full compliance with Semantic Web and related World Wide Web Consortium (3WC) standards
- Includes an RSS data feed so that one of the data sources could alert the user that new information has become available, and that new categories of information are available
- An assisted navigation tool kit that makes it possible for anyone familiar with a browser to navigate upwards, across, and downwards through the relationships discovered within processed content
- A turn-key approach that offers more flexibility and faster installation/deployment than either mainstream search systems or some of the search appliances currently available.

Downside

The demand for point-and-click interfaces with suggestions, categories, and assisted navigation is increasing. Organizations looking for these next-generation interfaces often have a difficult time figuring out if a vendor's system is delivering the interface or if the interface is a cosmetic layer on top of unsophisticated indexing procedures.

Siderean delivers "100% beef" when it comes to advanced metatagging and supporting enhanced interfaces, rich with Use For and See Also references. The problem is that Siderean has to overcome some prospects' perceptions that other vendors deliver exactly what Siderean offers. *Beyond Search* strongly recommends that an organization interested in assisted navigation invest the time to understand the differences between a cosmetic solution and a robust implementation of enhanced content processing.

Other considerations include:

- Siderean does not have the type of profile enjoyed by such companies as Autonomy and Endeca
- The firm's "fast deployment" approach runs against the lengthy deployment times required for other systems. Not surprisingly, those unfamiliar with Siderean's approach may express skepticism
- Plan on spending some time planning your assisted-navigation interface. Siderean offers a basic layout, and you will want to tailor this to meet the needs of your users. Interface design requires effort. You can deploy the default Siderean interface as you work on a more tailored version.

Net-Net

Siderean combines the type of interface made popular by Endeca with the robust content processing of such companies as Attensity or SRA International.

Assisted navigation and integrated classification of content, entities, and concepts is a way to break free of the restraints imposed by a “naked” search box.

Users respond positively to hierarchical browsing and searching with the search box available when it is needed. Siderean’s system allows users to backtrack or jump laterally in an information space. In addition, the Siderean system gives you the option to display the number of results in each category, a feature that makes it easy to pinpoint when the content depth is greatest.

Siderean’s challenge is to increase its profile and position the firm’s technology as a way to make existing systems “smarter” and easier to use. At the same time, Siderean will have to overcome skeptics who think that a robust content processing system can take months, not a week, to deploy.

Beyond Search continues to be impressed with the Siderean system. You owe it to your users to take a close look at Seamark Navigator.

22. Thetus Corporation

www.thetus.com

Danielle Forsyth CEO, who along with Roy Hall founded Thetus Corporation, recognized the gap between the explosion of high-value non-text data and the ability of organizations to cost-effectively leverage that data. She told *Beyond Search*, “The evolution from non-text data to knowledge discovery requires intelligent, automated systems that streamline the knowledge management process and provide intuitive tools for searching and accessing information.”

Ms. Forsyth has over 20 years of 3D graphics engineering, marketing, and management experience at Hewlett-Packard, Microsoft, @Last Software, Digimarc, and Wavefront. She is the co-author (with her co-founder Roy Hall) of *Interactive 3D Graphics in Windows*. In April 2007, the *Portland Business Journal* named Ms. Forsyth as Woman Entrepreneur of the Year.

Item	Quick Facts
Product	Thetus Publisher and Fusion Portal
Price	\$250,000. Custom price quote required
Key Feature	Allows federated search of different content types
Purpose	Perform intelligence analysis
Clients	Central Intelligence Agency, petrochemical companies, financial services firm
Company	Privately held
Contact	sales@thetus.com

Table 46: Quick Look at Thetus Corporation

Mr. Hall, prior to co-founding Thetus, was the CTO at Crisis in Perspective where he designed and developed software systems for Currenex, @Last software, Driveway, Microsoft, Microsoft Research and many others. He was the chief architect of the first commercial 3D animation system at Wavefront Technologies and worked for Robert Abel and Associates in Hollywood, California.

Thetus Inc. offers a framework in which licensees can explore, discover, and analyze data in an interactive manner.

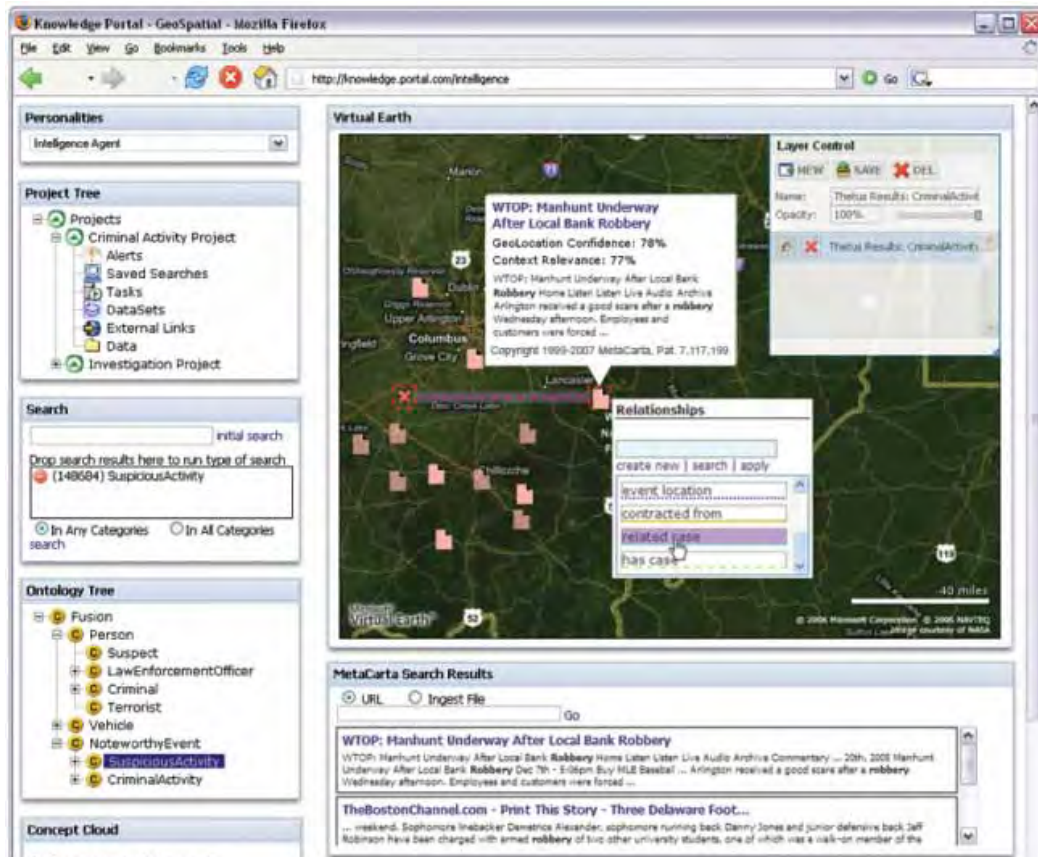


Figure 63: Thetus Search and Data Display

The Thetus system can support a “knowledge portal”. In addition to search, the system can integrate data and display them on a map.

Semantic Search

The Thetus system implements semantic search; that is, a user can search based on meaning, not key words. To illustrate: a user can search for *terminal* according to whether the user wanted information about a transportation hub or a computer peripheral. Thetus makes it possible to explore links among processed content and from many different content collections, automatically correlating information among these sources. One useful feature is that a user can search text, images, geographical data, and other types of data not normally available in a single search system.

The inspiration for the system was the need to query vast and disparate data sets in scientific research. After 9/11, Thetus was embraced by the US intelligence community. In-Q-Tel, the investment arm of the Central Intelligence Agency, pumped more than \$1.0 million to the company. Then in 2006, Thetus secured an additional \$3.6 million in venture capital. At the end of FY2006, Thetus had revenues of about \$2 million and about 30 employees in its Portland, Oregon, headquarters.

Thetus, therefore, is not a search company. The firm’s software and systems deliver “intelligence fusion”, a clever metaphor for making words, images, and data accessible.

Restricting analysis to text, at least from Thetus' viewpoint, is tantamount to piloting an aircraft with most of the instruments inoperable.

The firm's technology addresses the problem that arises when there is no single language to define certain data such as geospatial information, research data, or non-text content like a video. The amount and importance of these diverse data types mean that most organizations cannot exploit their high-value information assets. Employees, therefore, face a gap between data and users' ability to transform it into actionable knowledge.

Ontology Centric

The Thetus system uses ontologies that are tightly integrated into the system's functions. Ontologies, unlike the rigid XML schemas used by some systems, are more flexible, permitting fuzzy logic or soft logic to be applied in Thetus' algorithms. An authorized user, for example, can create an overlapping category to handle certain types of relationships such as Same As. The company's notion of ontologies is more suggestive, not prescriptive. Thetus' categorization operations permit "non-exclusive property association", which makes it possible for the system to make judgments about how to tag data. One nice touch is that Thetus' approach to ontologies permits user annotations. Unlike rule-based systems, Thetus generates inferred relationship properties, including transitive, symmetric, and inverse tie-ups. The approach makes Thetus' content processing more expressive.

Properties

The Thetus system allows the licensee to spell out properties for information objects. Properties are defined in dictionaries, which may be custom generated, pre-defined, or assembled from multiple sources.

A typical property supports such concepts as the type, domain, and range of a particular property. An example is Thetus' ability to process content with an *IsMarriedTo* tag. The Thetus property, like Siderean's descriptor function, is generally described as a triple. In addition to the relationship that "Joe Wilson IsMarriedTo Valerie Plame," Thetus tags this as a collection. The collection tag adds a fourth dimension to the relationships to help Thetus generate a lineage that allows a user to see where an item originated. Collections make it possible for Thetus to assert properties for each collection; for example, one collection may be public documents and another may be commercial database documents.

Thetus in Use

The Army Corps of Engineers is engaged in research focused on creating a GIS-based decision support tool that will provide military planners with situational awareness of how the local population operates in time and space, and how people move through and make use of the built environment in certain culturally proscribed ways. This requires understanding the cultural influences on how people interact with their built

environment, using this knowledge to capture and model the rhythm and flow of daily life in an urban environment.

The overall modeling approach focused on using three distinct yet interconnected ontologies: a geo-cultural ontology that describes geo-cultural elements and behaviors; a schedule ontology which describes recurring things that happen in embedded cycles (for example, a school has an annual schedule, a semester schedule, a daily schedule, etc.); a traversal ontology which describes the traversal path and distances in seeking the “closest” relevant data source. Users are able to obtain overviews of pertinent information and display specific information items on a map, in a results list, or on a link diagram.

Thetus Publisher

The core of Thetus' system is its proprietary Publisher server. The server houses a knowledge model for the licensee's organization. This model is a representation of known facts and concepts and relationships. In an enterprise, the model captures, retains and disseminates intellectual capital to maximize operational advantage

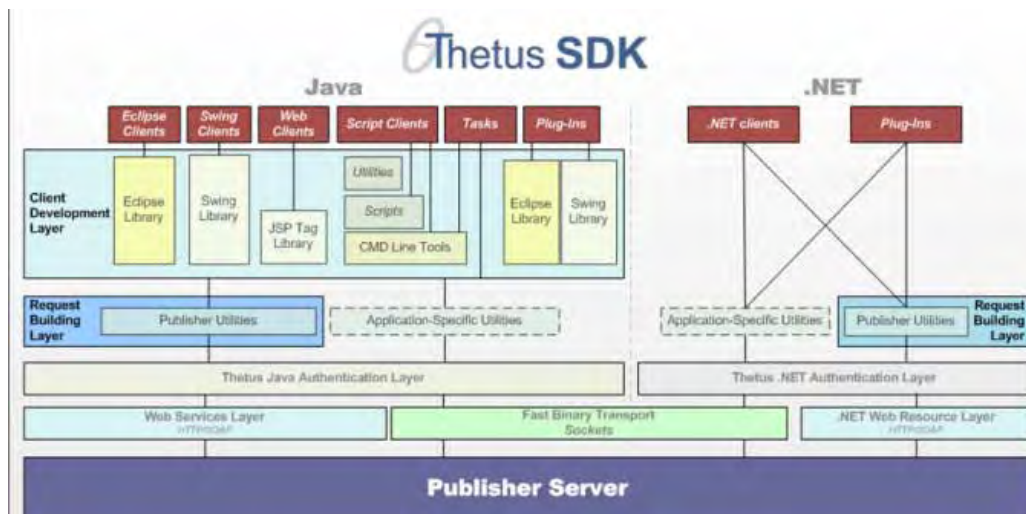


Figure 64: The Thetus SDK

The Thetus SDK gives licensees a comprehensive way to deploy Thetus across different content repositories, develop customized client interfaces, and perform content processing.

This component of the Thetus system performs the core content processing function. This proprietary and complex system provides what Thetus calls “an ontology for everything – knowledge, users, policy, lineage.”

The Thetus Publisher makes it possible to search, discover, share, and reuse available information. Thetus’ integrates with third-party text analytics to expose non-obvious relationships between information in text and non-text form.

Search Management

Publisher includes management tools for the search-and-retrieval process. A licensee can also use a third-party tool such as Metacarta for search.

Task Management

Publisher also includes administrative controls to manage tasks the system is to perform. Task management consists of four interlocking processes: [a] routing to transfer data to other systems for processing [b] filtering a data stream and normalizing the data [c] classifying of data to enhance it by adding categories and properties; [d] notifying users or processes to send alerts or information based on identified events or data changes. Publisher also includes administrative tools to perform policy and user management, federation management, and ontology management.

The Publisher system enables modeling, discovery, sharing and reuse of data, metadata, and knowledge across sources, applications, disciplines and objectives.

Features of Publisher

Other interesting system features are a user’s ability to:

- Access a persistent work, space which can be shared
- Filter information views and tools by specific user function and objective
- Visualize, discover, and explore explicit and inferred relationships
- View integrated information in geographic context
- Annotate mapped items and create new relationships on-the-fly

Third-Party Text Analytics Integration

Attensity plugs into the Thetus Analytics Pipeline, which is a set of generic workflow components. The Attensity application looks for ‘facts’ or actor-action-object groups that can be derived from unstructured text. With Thetus integration, the extracted entities are situated in the knowledge model using inferencing capabilities. If an extracted concept does not exist in the model, it is created dynamically. Similarly, relationships in the extracted entities are expressed using the connections defined in the knowledge model or they are generated on-the-fly. Together, the Thetus Publisher and Attensity allow users to quickly turn unstructured text into a set of concepts and facts that can be queried and reasoned across. Typically customers use the Thetus

Knowledge Portal to inspect the results of the extractions, verify them and augment them with tacit knowledge of observations.

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	Supports knowledgebases and ontologies. Includes editorial tools for knowledgebase management
Query Types	Keyword, natural language, Boolean, SQL query, various visual representations for point-and-click discovery
Visualization	Relationship displays built in. Third party add-ins supported
Entity Extraction	Knowledgebase used for entity identification
Platforms Supported	Java-based servers for Linux or Windows
Export	Exports data and generates user-configurable reports
Third-Party Support	Supports third party subsystems from Attensity, Metacarta and ESRI, among others
Vertical Support	A basic version of the system can be customized for vertical applications
Analytic Functions	Third-party tools may be integrated via an API

Table 47: Technical Highlights for Thetus

Fusion

The company's newest product is the "intelligence fusion platform". Thetus has tackled a complicated problem in information retrieval--querying a large number of disparate sources of information and giving the user the ability to see where an answer came from. In search jargon, Fusion allows the user to perform federated search and have instant access to the lineage of the data in the result. Lineage is the rough equivalent of knowing the provenance of a valuable art work.

Fusion, as the name implies, is a portal service that creates an index allowing access to content processed by the system. Other features offered in Fusion are:

- Work flow tools for automating and routing information
- Enhanced access control functions to ensure that only people authorized to view content enjoy that privilege
- Enhanced historical tracking and reporting; that is, the lineage function to allow users to analyze leading indicators of a trend.

The Fusion approach takes the notion of search and retrieval and embeds it into a knowledge portal. The informing vision for Fusion is that a user needs a mechanism for accessing and interacting with personalized, filterable views of information. The portal framework adheres to industry standards, enabling rapid deployment on a broad range of enterprise servers.

Searching with Thetus

Querying Thetus is similar to searching Google or Yahoo. A user can enter a word or phrase. The system also accepts natural language and phrases. A user can enter a query via the Thetus syntax to search for relationships or point-and-click through a graphic display. The system supports an interesting range of querying and interacting methods:

- Search for relationships among multiple entities
- Access direct links to source documents
- View and interact with information using familiar, domain-specific terms
- On-the-fly changes to search parameters

Upside

The upside of using Thetus is access to disparate data types. The system allows a licensee to use an evolving, dynamic data model— allowing knowledge to grow organically. For organizations wanting to know “where information comes from”, Thetus’ lineage function makes shorter work of determining the credibility of certain information.

Downside

The Thetus system demands significant hardware, memory, and bandwidth. Its core component, Thetus Publisher, is among the most complex content processing engines that is available today. A dedicated system administrator is strongly recommended. One or more subject matter experts may be required, and these subject matter experts will require specialized training in the Thetus system.

The company does state that customers with less demanding model demands can run the system on reasonably modest hardware. Also they claim that system maintenance is similar to that required for a database systems needs for a DBA, but that this can also be a cross-functional person dedicating only part of their time to maintenance of the Thetus system.

Net-Net

For an organization that wants to manipulate text and non-text content, the Thetus system is one of a handful of military-grade content processing systems in the commercial channel at this time. Thetus warrants a closer look when an organization places a considerable emphasis on intelligence, not search.

23. Vivisimo Corporation

www.vivisimo.com

Vivisimo, Inc. is a privately-held corporation founded in 2000 and headquartered in Pittsburgh, Pennsylvania. Vivisimo has moved from an invention (at Carnegie Mellon University's Computer Science Department) into a growing software company with an international reputation.

When I first met its CEO and co-founder many years ago, Raul Peres-Valdez, he said, "Overlook. Users need information overlook."

The phrase stuck in my mind, and I have used it, sometimes inadvertently without attribution, because he was right. Vivisimo's first product snapped into existing search systems, intercepting results, and on-the-fly automatically classified them into clusters.

The automatic classification, or clustering is still available as a feature in its product today, but Vivisimo has grown beyond a utility. Mr. Peres-Valdez told Beyond Search, "In the past year, business search has successfully made the transition from departmental uses to enterprise-wide adoption across many organizations we serve. We take great pride in our role in its evolution and more importantly, for the success of the Velocity Search Platform, which has evolved from an interesting classification feature to a robust, complete search solution."

Item	Quick Facts
Product	Vivisimo Velocity 6.0
Price	\$35,000 and up
Key Feature	Clustering, social search, faceted navigation and federation
Purpose	System allows users to access information from one search "box". System clusters results for faster search and discovery. Information assets include internal and external (licensed content, Web content, etc.) sources
Clients	Cisco, Eli Lilly, Fidelity, Tyco Electronics, Organon NV, USA.gov, National Library of Medicine, Government of New Zealand
Company	Vivisimo Corporation
Contact	(866) 294-8484

Table 48: Quick Look at Vivisimo Corp.

Less than a decade ago, an organization with information in a database located at headquarters, a records management system located in the firm's engineering department, and a dedicated server receiving news from Dow Jones had an all-too-familiar problem. A person looking for information about a particular topic would have to find someone to run an SQL query to pull the data from the database, log in to the server with the records management index, to run a query on that system, and then 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 this type of functionality. 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.

A number of companies emulated the Verity approach with varying success. However, until Vivisimo entered the market in 2000, most solutions to this problem of federated search were less than elegant. Federated search refers to a search system’s ability to index diverse content, file types, and repositories which may contain copies, then remove the duplicates, relevance rank the results, and display the results list, grouped in categories. This provided a type of guided navigation or faceted search while solving some of the more complex challenges associated with enterprise search.

The Carnegie-Mellon University computer scientists responsible for Vivisimo have crafted a search system that has remarkable versatility, runs on commodity hardware, and supports Web services and standards. One other key point is that Vivisimo makes it possible for a licensee to integrate results from Web search engines such as Google, Yahoo, or Microsoft with content from specific Web sites, and licensed subscription feeds in addition to crawling and indexing internal data repositories. In short, Vivisimo is a capable, flexible search system that provides a significant bang for each licensing dollar.

A “New Breed” of Search System

Vivisimo is one of the “new breed” of search systems that can play different roles, depending upon the customer’s requirements. The general characteristics of the Vivisimo search system are a blend of traditional word-and-phrase search systems and faceted or guided navigation search systems.

Other key features of the Vivisimo technology include:

- Needs no maintenance unless a licensee wishes to use dictionaries and chooses to update these manually
- No pre-processing of documents or collections
- Installation can be accomplished in as little as one hour and be delivering search results to Intranet users in less than one day
- Support for clustering results from standard databases such as Oracle and SQL Server among others. Vivisimo also supports Lotus Notes repositories.
- Clustering Engine now featuring a Web-based administration tool
- Useful, detailed tutorials

An example of the “federating” and clustering function is the [FirstGov.gov](http://firstgov.gov) system. The user does not need to know that some of the content searched resides on U.S.

government servers, Microsoft's MSN servers, or on the Vivisimo servers. The user goes to one place and searches the content as if it were in one index accessible from a standard browser. Unlike Google, Vivisimo does not require that the user select a collection before running a query. Google, for example, does not provide a single search across news, Usenet postings, and Web site content while Vivisimo does.

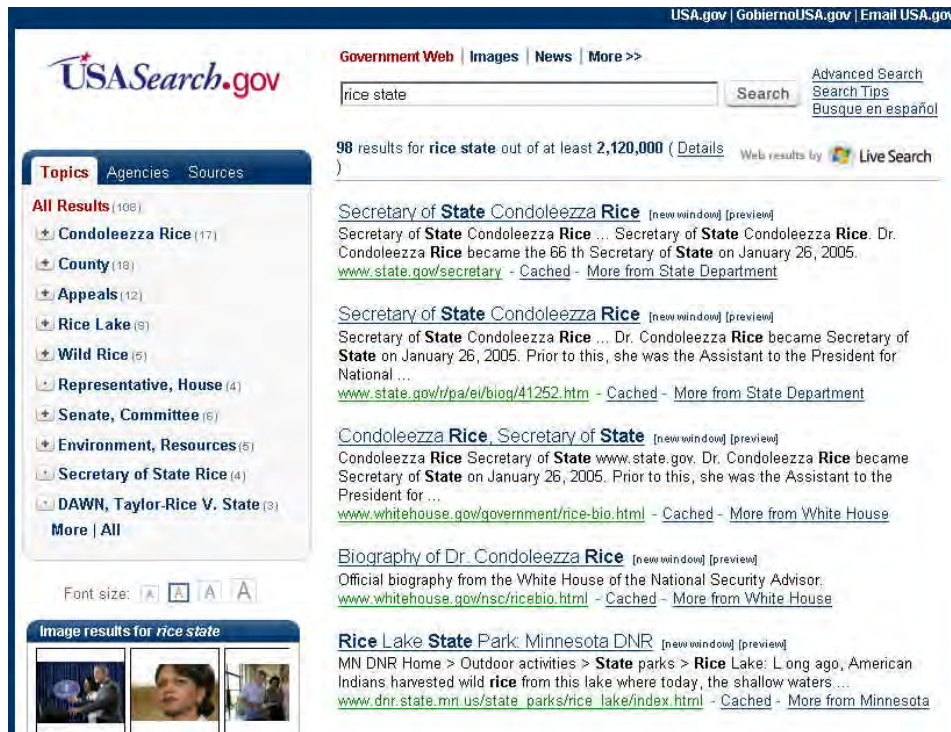


Figure 65: USA.gov's Use of Vivisimo

Vivisimo provides the search technology for the US government's portal, USA.gov. This system processes queries, retrieves information from Microsoft's MSN servers., and then integrates information spidered and maintained by Vivisimo. The results are clustered. Duplicates are removed, and information is segmented into collections.

Technology

Vivisimo supports multiple platforms. The system uses XML configuration files. The search system's CGI script for the federated search uses a C library that has been ported to Windows, Linux, FreeBSD, Sun Solaris and other Unix "flavors". However, Vivisimo delivers the most bang for the enterprise search dollar when run on Linux systems.

The software is distributed so that the most common scripting languages can be used to integrate the particular Vivisimo module into their applications or Intranet. Organizations can also use Vivisimo Velocity as a hosted service, so an enterprise can offer their users a turnkey portal in a matter of days.

Components

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 located 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.

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 (e. g. title versus abstract), globally-important words, lexical stemming, and others. Large sites will want to invest some time in this activity.

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	Knowledgebase module accepts company and industry-specific knowledge such as synonyms, acronyms, spelling variants, taxonomies, etc.
Query Types	Keyword, natural language, Boolean, automatic fuzzification of queries
Visualization	Yes – mashups and graphical representations are provided
Entity Extraction	Yes
Platforms Supported	Windows, Unix(Solaris), Linux
Export	Yes – text, HTML, XML, Endnotes, Procite, Reference Manager, email
Third-Party Support	Yes – connectors provided for third-party applications and databases, such as Documentum, Microsoft Sharepoint & Exchange and Lotus Notes
Vertical Support	No
Analytic Functions	Third party through API

Table 49: Technical Highlights for Velocity

For licensees needing customization, a knowledgebase 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 Text can be disabled in order to speed up the indexing process within Oracle's Text.

Licensees should note carefully that clustering 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 your search engine's result set.

Dictionaryes

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.

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 languages: Danish, Dutch, French, German, Italian, Norwegian, Portuguese, Russian, Spanish, and Swedish. Vivisimo products also embed other semantic and syntactic knowledge, but the company declines to provide details about these technologies. Vivisimo management told *Enterprise Search Report* that a Chinese version is now available, and a Japanese version was planned for release later in 2006.

Customizing Velocity

A licensee can customize most Vivisimo functions in two ways:

- The administrative screens
- Templates.

Vivisimo's clustering subsystem is made up of CGI scripts and 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.

Velocity Search Platform

Search Engine

Vivisimo Velocity's search component is built with a unique architecture in that enterprises can search their content as it is, without requiring any 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 with no standards for creation, organization, or management. 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 goes beyond this by offering a search engine that can generate several independent results from a single page, such as a blog's front page in which each entry is a unique result. Velocity parses the resulting XML feed or HTML output with XSL to provide clustered search results.

The Vivisimo Velocity search is also able to leverage metadata. Unlike many other search solutions that require that metadata 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 includes 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.

Content Integration

Content integration lets companies integrate search or database query results from multiple sources and deliver dynamic integrated content - a "metasearch." A metasearch is a query that is run automatically across multiple indexes. Regardless of where the content is stored (internal or external) or the number of disparate sources, metasearching will present a single unified view.

Other vendors sell metasearch tools. End-user software such as the desktop application Bull's Eye and Copernic are metasearch tools. Open Text's QueryServer.com is a metasearch service that demonstrates the Canadian company's metasearch technology. In addition, the Swiss company Albert S.A. provides a suite of metasearch tools for Intranet and Web searching. However, Vivisimo has emerged as the apparent front runner in this type of search software application.

Clustering and Performance

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. Vivisimo's 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 both 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 uses XML search engine output or parses its default HTML/Text output. As a result, it avoids 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 who have already developed a taxonomy. That means that Vivisimo can configure the clustering topics to be based on an organizations existing taxonomy and software will place the relevant search results in the pre-existing topic listing. Organizations can also have both static topic 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.

Upside

Vivisimo Velocity offers rapid deployments into any type of search application.

The benefits of the Vivisimo approach include:

- Ability to leverage pre-existing search and information assets
- Allows users to search all content from one search box
- Allows for guided navigation and passive information discovery

Downside

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. Vivisimo provides the information needed to handle relatively simple indexing jobs and the more complicated ones as well.

The documentation is useful, but the solution to certain configuration settings is in the sample code Vivisimo provides. Finding the half dozen lines that are needed can be difficult for someone not used to reading code for a solution to a configuration or set up issue. However, Vivisimo's current documentation is much more thorough and user friendly than the information accompanying earlier versions of the system.

Other considerations include:

- Configuration files are used to control certain system functions. While not overly complex, you will want to have familiarity with editing these files. Beyond Search does not recommend learning by experimentation unless you have a development server on which to test your scripts.
- Federation is a powerful tool. However, you will want to make certain that you have verified the security settings on the servers and systems from which you will pull content.
- For some users, the categories may not be intuitively obvious.

Net-Net

Vivisimo Velocity has moved beyond being only a clustering “add on utility” to a robust enterprise search and clustering platform. Keep in mind that you can use Vivisimo to “fix” a problem search system. Vivisimo can post-process search results and federate content, thus squeezing more from an existing search installation.

Vivisimo's approach is now being emulated by some of the vendors profiled in this study. The company's content processing solution warrants a close look.

24. ZyLAB

www.zylab.com

ZyLAB has a long history in text processing, and it was one of the first text search products. The US company was acquired by a Dutch holding company and is now managed by the indefatigable Dr. Johannes Scholtes, a former naval officer. ZyLAB and Mr. Scholtes serve customers throughout the world.

ZyLAB was among the first text processing companies to offer what might be called enterprise content management with fuzzy search; that is, an algorithm that allows the system to find variations of the user's query terms.

The company has a strong presence in government agencies, helped with the clever marketing ploy of Texas "hold 'em poker nights." An invitation-only, night-out for bonding and betting has helped the firm expand its reseller network and land important OEM licensing deals.

Item	Quick Facts
Product	ZyFIND with E-Discovery
Price	Begins at \$20,000
Key Feature	Provides case management and discovery tools in a search-centric interface
Purpose	Process, explore, and repurpose structured and unstructured data
Clients	Halliburton, US Department of Defense, Avon, Superior Court of Delaware
Company	Privately-held
Contact	info@zylab.com

Table 50: Quick Look at ZyLAB

Technology

The company offers a full range of text processing tools. These range from software that can scan, process, and structure paper and digital content. The company calls its ability to retrieve information from scanned documents *W-Y-H-I-W-Y-G*, short for *What You Had Is What You Get*. The idea is that ZyLAB creates a repository for processed content. A user, therefore, can access that information as documents, snippets, and facts.

The XML repository offers ZyLAB users an important advantage. The metadata and the keyword indexes allow the company to offer text mining functions as well as keyword searching. ZyLAB text mining technology incorporates best-of-breed visualization technology to manipulate the results of user query or a stored process that automatically scans new content for matches. Alerts, therefore, keep an analyst informed of new information known to the system.

Beyond Search: ZyLAB

ZyLAB uses what it calls “advanced linguistic technology” to process text, our tests reveal that ZyLAB is competitive with systems that cost more and have a higher profile. The system supports knowledgebases of known named entities and uses rule-based regular expressions in its parsing subsystem.

Entity extraction can identify and tag more than two dozen entity types from documents and then classify these entities by company, people, dates, places, and currencies. Furthermore, ZyLAB discovers new entities, using technology licensed from Inight (now a unit of Business Objects) and extended by ZyLAB engineers.

ZyLAB incorporates a range of display technologies in its product. For example, the company can present a list of results in a table view. If the user wants to see a representation of the data, ZyLAB incorporates hyperbolic map technology developed at Xerox’s Palo Alto Research Center. The weakness of hyperbolic maps is that a user may be able to “see” only a limited number of nodes. ZyLAB has incorporated what is called embedded menus or illuminated links in a tree-map. A tree map is a space-constrained, visual representation of statistical information that automatically re-sizes to show proportion. The tree map is designed for illustrating complex hierarchical structures. This type of mapping uses size variation, color-coding, and individual pop-up tags to provide an overview of the results. ZyLAB’s implementation allows a user to compare nodes and sub-trees at different points within the tree. Exceptions and discontinuities become easier to identify and explore.

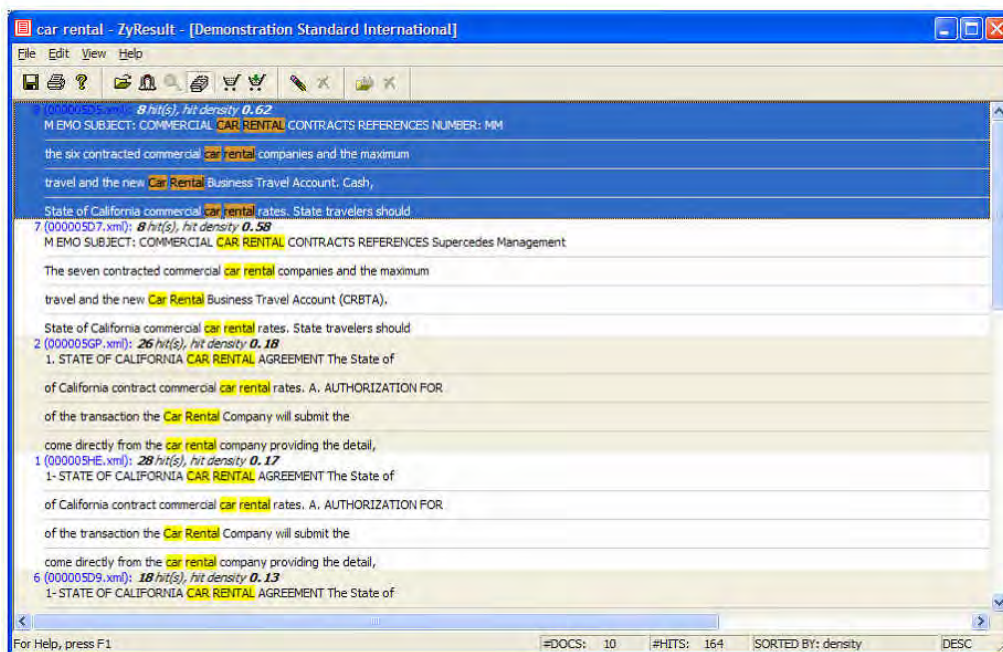


Figure 66: ZyLAB's Search Result Interface

ZyLAB's interface highlights the user's search terms. This in the default result list display for a wiki search. The display may be customized.

Examples of the System in Use

ZyLAB has a large number of intelligence agency and law firm clients. Not surprisingly, the company provides few details about the use of the ZyLAB system in these organizations. Some details do leak out, including:

- A major law firm uses the ZyLAB system to process hard copy documents that can run into the hundreds of papers per deposition. In addition, the law firm analyzes electronic mail from Microsoft, Lotus Notes, and Novell GroupWise, among others.
- An intelligence agency uses ZyLAB to process a range of structured and unstructured content. Using knowledgebases of persons of interest, ZyLAB generates alerts when new information about these individuals becomes available.

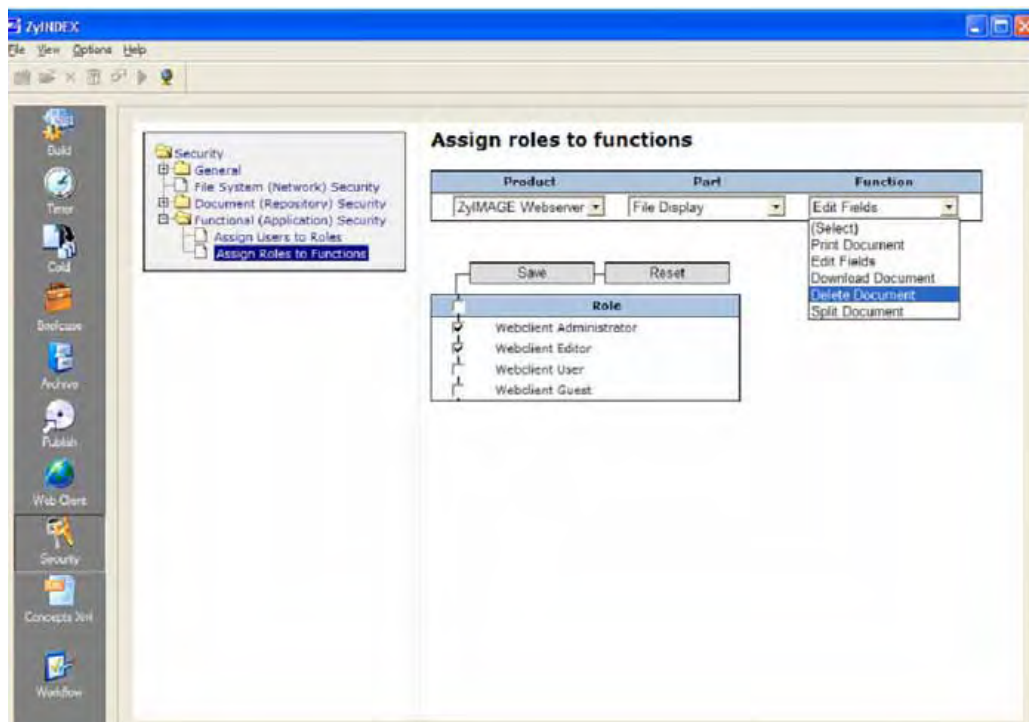


Figure 67: ZyLAB's Administrative Tools

ZyLAB's graphical administrative tools reduce the complexity and time required to set up and maintain security settings.

Basic Functions

ZyLAB packs a number of features in its basic system. Note that the company can unbundle certain functions—for example, document scanning—from the search and text mining modules.

- ZyLAB supports more than 370 different file formats eliminating the need to code custom scripts to filter unusual file types.
- ZyLAB supports 64-bit systems; thus, there is no practical limit on an index size. However, the system supports “index series”. The function allows a user to

create multiple indexes for certain data sets. New indexes can be created as required based on volume or date and time parameter restrictions. This allows new indexes to be generated when certain file sizes are reached

- A summary of documents can be automatically generated. ZyLAB relies on “sentence picking” so the summary contains the original text found in the source document.
- Workflow functionality is included with the system. Once configured, result sets, analyses, and document collections via the bookmark feature may be automatically routed to individual users or groups of users. A digital signature service is included with the workflow engine. The work flow allows a licensee to display a “live” document next to the indexed document. Police, investigators, and researchers can then provide additional information related to a case or activity of interest.

Search

ZyLAB includes a full range of search functions. In addition to Boolean and fuzzy search, the system permits phrase searching, positional operators, number range operators, and commands to limit a query to a specific range in a document.

Results are relevance ranked. When results appear in tabular form, the system permits sorting or viewing the document in KWIC or key-words-in-context view. The full document for any result may be displayed at any time with a mouse click.

Categorization

When content is processed by the ZyLAB engine, the system supports manual and automated categorization. A user or system administrator can specify a basic tree structure based on an existing taxonomy or classification system.

In automatic mode, The ZyLAB tagging system can process email. It can process more than 100 email properties such as author, title, company, dates, and attachments. It can identify, extract, and tag personal names, countries, addresses, telephone numbers, Internet addresses, social security numbers, and monetary amounts. Three types of automatic categorization are supported:

- Rule-based. Rules are created by ZyLAB engineers or the licensee and permit fine-grained control over specific tagging functions.
- Machine-learned based categorization. ZyLAB uses multi-pass technology to “learn” from a content training set. This technique works well on content that is “about” a specific topic such as those found in scientific research.
- Ontology-based categorization. The system processes a provided word list. ZyLAB’s algorithms identify the similarity among items in the knowledgebase and the processed content.

The key point is that ZyLAB can automatically tag content, but it offers an unusually rich, flexible means of handling specific types of document tagging. A case management

Beyond Search: ZyLAB

function allows documents to be retrieved by user-defined tags so a “case” can be assembled automatically and shared with other authorized users.

In manual mode, a user can drag and drop a document to a node on a hierarchical display of the results. The system supports bookmarks, which are user definable notes within pages of a document.

For example, a lawyer using ZyLAB for analysis of legal documents can “tag” a document page and locate it later without rerunning a query. Bookmarks permit user-definable categories, dates, and short notes of searchable free text. ZyLAB permits a user-level bookmark as well as a network-level bookmark.

Users or dedicated editorial staff can add tags to any ZyLAB document. These tags may be used to identify shared concepts within a work group or department.

Feature	<i>Beyond Search</i> Comment
Knowledgebase Support	The system can use a commercial thesaurus or ontology such as MeSH as well as customized term lists
Query Types	Boolean, free text, and options for point-and-click discovery
Visualization	Includes hyperbolic graphs, hierarchical displays, and treeleaf services
Entity Extraction	The system supports persons, places, and things. Custom entities can be identified and controlled with rule-based scripts
Platforms Supported	Supports data on Linux, Unix, and Windows. Software runs under Windows only.
Export	Multiple export options including a “case” format which gathers documents for a legal proceeding into one “package”
Third-Party Support	Application programming interfaces are provided
Vertical Support	No vertical builds of the product are available
Analytic Functions	Includes a standard tabular display with counts

Table 51: Technical Highlights for ZyFind

New Features

The company's current release supports a number of enhanced features; for example, [a] facets for one-click expanding or narrowing of a query; [b] term highlighting in result lists and displayed documents; and [c] a redaction feature to allow a user to mask out certain portions of a document. ZyLAB now includes a graphical interface for these security features.

Upside

The upside for ZyLAB's system includes:

- A content processing system that can be extended to perform repository services and text mining
- Inclusion of third-party visualization tools makes it easy for end users to explore large datasets
- Support for hard copy combined with case management and annotation tools makes ZyLAB useful for audit, legal, and intelligence applications
- Extensive support for email
- Fully XML based

Downside

Considerations for the ZyLAB's approach include:

- ZyLAB lacks the profile enjoyed by other, often less robust systems. Procurement teams may be faced with the question, "Who is ZyLAB and what does the company do?"
- ZyLAB's technology performs well on standard servers; however, for processing terabytes of content, a dedicated system administrator is needed to handle optimization and customization.

Net-Net

ZyLAB is a dark horse in text mining. Most companies overlook ZyLAB because it does not limit itself to a single niche. The firm has customers worldwide, and it continues to grow at a double-digit pace. The ZyLAB system is worth a hard look, particularly if text mining and case management services are needed.

Glossary

These definitions are designed to make more easily understandable some of the terms used in the search-and-retrieval industry. The definitions are not academic. Instead I have tried to make certain concepts clear and mostly jargon free.

Term	Definition
"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
adaptors	A device used to effect operative compatibility between different parts of one or more pieces of apparatus.
appliance	An instrument, apparatus, or device for a particular purpose or use.
appliance vendors	Businesses that sell appliances; for example, a search "toaster" like Google's Search Appliance
application platforms	The basic technology of a computer system's hardware and software that defines how a computer is operated and determines what other kinds of software can be used.
application programming interface	Equipment or programs designed to communicate information from one system of computing devices or programs to another to operate application platforms.
assisted navigation	A point-and-click interface for exploring information or performing a search on a topic by clicking on a hyperlink.
automatic classification	Delivering classification of results without use of a human subject matter expert.
autonomic servers	A self-optimizing computer that independently makes services, as access to data files, programs, and peripheral devices, available to workstations on a network.
backup device	A piece of hardware to 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 organized for employees who need information for work purposes.
business intelligence	High-value solutions and information on demand based on information about specific businesses.
certification procedure	Verification of hardware or software to make certain its meets specific requirements
classification	Categorization; that is, a function to distribute things into categories of the same type

Term	Definition
cloud-based service	A service that delivers applications via the Internet.
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 specific group of content; for example, the proposals created by an organization
computational linguistics	A content processing technique based on analysis of language syntax.
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 subsystems; the files may be edited.
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
content processing expert	A person with knowledge about content processing
content processing system	A computer or group of computers that perform content processing
custom scripts	Programming code lines developed for a specific purpose in running software.
cyclic redundancy check	A method used to set a value which if change indicates that the content used to derive the value has changed
dashboard interface	A graphical representation of various search or information functions and operations
daspaces	A representation of information and information about information derived from multiple sources such as databases and content collections
enterprise	A commercial enterprise; hence, enterprise software as distinct from software used on a single user's laptop computer
entity extraction	Identifying, indexing, and extracting the names of people, places, things, and such values as dates from a document
ExaScript	Exalead's Java-like scripting language
faceted navigation	A term coined by Endeca to describe point-and-click navigation via categories and other hot links; a synonym is point-and-click navigation
fat client solution	A solution designed to run on the user's own machine to increase performance
federated search	The simultaneous search of multiple online databases
filter	A function used to specify criteria for selecting or rejecting data. A filter can add or exclude addresses that do not respect the same pattern as the entry point address defined for the information source.

Term	Definition
firewall	Hardware or software running on a computer which inspects network traffic running through it and allows or denies passage based on a set of rules
FPGA	Field-programmable Gate Array is a logic chip that can be programmed
Googleplex	A term referring to the hardware and software infrastructure deployed at Google
graphical editors	An editor interface that allows the display of data in logical graphical objects and schemas
GSA	Google Search Appliance
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"
high-speed persistent cache	A method for retaining information in a high-speed storage area to increase system performance
hosted solution	Managed solution or application handled and managed by the vendor
hybrid display	Combines text, hot links, and graphics on one screen
hybrid interface	Synonym for a dashboard interface
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
intelligent search agent	A search system that makes use of algorithms that 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.
internal search system	A program that indexes or searches information that is not publicly available on a company's network behind protective measures.
internet search systems	A phrase used to describe the publicly-accessible search systems offered by Google, Microsoft and other vendors
inverted index of the words	An index of the words in the text of a set of documents accessed by a search method. Each index entry gives the word and a list of texts where it occurs, possibly with locations within the text.
JDBC	Java Database Connectivity, a Java API that enables Java programs to execute SQL statements.
key word search	Entering a keyword or term into a text field, such as "dog". The keyword search engine then searches through its index for documents that contain that word. To improve precision of search results, keyword search engines utilize lists of keyword associations or "topics "such as dog =hound =canine or dog is 90%canine, 10%furry. Association lists like these require manual maintenance.

Term	Definition
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
latent semantic indexing (LSI)	An algebraic model of document retrieval based on mathematical techniques that represent a document as a series of values.
lemma	A proved proposition used as a foundation for a larger result
lemmatization	A term used to refer to the process of dropping prefixes and suffixes to obtain the root of a word
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
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
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.
MapReduce	Google's proprietary method for performing look ups and matching operations across a distributed system
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	Often called meta-information, this is 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.
Natural Language Processing (NLP)	Also known as NLP, it uses 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.
Neurodynamics	Autonomy's coinage for a company developed to refine the statistical engine in the IDOL platform
OneBox API	The name of Google's application programming interface for the Google Search Appliance
ontology	The study of the categories of items that exist or may exist in some domain with an emphasis on "knowledge representations."
open-source search	A name give to products such as Lucene, a non-commercial, user-community supported search system
open-source search system	A system incorporating Lucene, Flax, or other non-commercial, user-community supported search system

Term	Definition
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.
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.
PLSA	The acronym used by Recommind to describe its statistical search engine. PLA stands for Probabilistic Latent Semantic Indexing)
query processing	The process used by a search system to convert the user's query into a form suitable for identifying and retrieving matching information
relevance ranking	A method to determine the order of importance of each item in the result list
response time	The 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
search	One or more methods of locating information in digital form
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 administrators	An individual who is responsible for a search system within an organization
semantic search	A method of searching for information using concepts which may not be expressed in the text of a document
semantics	The study of meaning
slipstream code updates	A vendor connects to a search system via the Internet and automatically copies new or changed instructions to the search system
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.
statistical systems	A search system that makes use of mathematical routines for processing text. A Bayesian system is a statistical system

Term	Definition
stemming	Identifying the root form of words so a document can be searched for all forms of the word. This extends to words with irregular pluralization and tenses. For example, in English "university" is stemmed to include "universities". It is important to note that stemming is language-specific.
structured data	Information that can be stored in a table where a document is a row and the column heading is a field name. The data within the table appear in the structure.
structured query language (SQL)	An industry-standard programming language for creating, updating and, querying relational database management systems.
taxonomy	Taxonomy's first meaning as a strict type 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. Web taxonomy is an ontology that does not explicitly define the nature of relationship between its concepts.
term mapping	A phrase used to describe the process of instructing a search system that "IBM" is equivalent to "IBM Corporation"
text mining	A process or series of processes that processes text in order to identify items that can be counted or otherwise analyzed
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.
troubleshooting	The process of determining what caused an error and fixing that error
unstructured data	Information contained in 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.
Unstructured Information Management Architecture (UIMA)	IBM's standard for information exchange
Use For	A term that means a specific terms should be used instead of another term in a query; e.g., use "focus group" for "discussion group"
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
work flow alerts	A specific event causes a search system to take an action; for example, when new information is indexed that contains the word "IBM", the search system generates an email message and a bibliographic entry, a hot link, and a summary of the newly arrived document
XRBR	An acronym coined by Siderean to describe its proprietary XML; XRBR stands for XML for Retrieval by Reformulation