In Search of... the Good Search

Search is a problem. A more polite way to phrase this sentiment might be to say, “Search remains a challenging human-computer issue.” No, listen to one MBA's comment about the Louisville Free Public Library's online search functions. “Too many hits,” he said.

Accurate observation.

Such candor is generally in limited supply at conferences, in journal articles, and during sales presentations from vendors of search and retrieval. We have an elephant in our midst, and no one wants to ask, “What's this elephant doing here?” I suggest we make an attempt to acknowledge the situation and do what must be done to put the fellow back in the zoo.

Search Realities

Multiple extremely complex and costly search-and-retrieval systems are in use in many large organizations. A typical news-focused system supporting about 500 users costs one U.S. government agency more than $3 million per year. According to one expert close to the agency, “about 95 percent of the system's functionality is not used. People type one or two words and take whatever is provided. The users seem happy with good enough.” But there are different types of search engines and different “ecologies” for each.

Table 1: Search and Three Content Areas

Selected Attributes	Internet	Intranet	Special Domains
Point-and-click installation	No tuning and coding required.	Lower-cost or ASP services have simpler installations. More robust tools require customized installation and set up.	Custom work required. Most vendors support file types covered in Stellent’s “Outside In” tool via license or by custom code.
Automatic indexing	Standard feature. Usually statistical approach to eliminate recursive calculations for linguistic tools. Some engines support external knowledge bases.	Varies by vendor. Dedicated thesauri and classifications may be required.	Training or tuning required. Vendors may have to customize engine to handle certain content types.
Classification of content	A feature of certain “newer” engines; e.g., Vivisimo.	Varies by vendor. Third-party tools required depending upon customer requirements.	Custom work required.
Internet file types[1]	Support for HTML and XML “standard.” FAST and Google support Word, PowerPoint, PDF, and a handful other file types.	HTML, XML, Microsoft Office file types, plus common legacy file types; e.g., Rich Text format, Word Perfect, etc.[2]	Special filters and customer work may be required
Index refresh mechanism	Controlled by scripts	Lower cost packages provide minimal controls. More robust packages provide extensive controls.	Lower cost packages provide minimal controls. More robust packages provide extensive controls.

Table 1: Search and Three Content Areas

Selected Attributes	Internet	Intranet	Special Domains
Content removal	Reindexing may be required. Manual intervention required.	Tools vary by vendor. Most rely on manual intervention.	Customer development required if file type is not directly supported by search package.
Firewall functions	Commercial systems are tightly engineered to protect service. Spiders gnerally observe conventions of robot.txt file	Varies by vendor. Some engines cannot index through firewalls and update indexes without customization.	Varies by vendor. Some engines cannot index through firewalls and update indexes without customization.
Spidering depth	Script controlled	Varies by search engine vendor. “Depth” and “security flags” often require customization to ensure that sensitive content does not “leak” into the more generalized service.	Controlled by custom scripts.
Security controls	Robust protection of the core system.	Security assumed to be a function of the Intranet system, not the search system.	Controlled by custom scripts
Cross-server indexing	Supported. Some servers can be polled more frequently to maintain appearance of index freshness.	Varies by vendor. Support may require customization of the search system.	Controlled by custom scripts.
Cross-domain indexing	Supported. May require separate indexes as is the case with Google.	Varies by vendor. The “behind the firewall” index is often separate from the indexes of third-party content and “outside the firewall” Web sites. Complex issue affecting performance, security, and index freshness.	Custom code required depending on the content, domain, and content locations.
Support for Lotus Notes and Microsoft Exchange[3]	Not generally provided	Varies by vendor. May require a third-party product or custom code to provide a single search box to access the Notes and Exchange content.	Custom code required if data are to be accessible from a seinge search box. Third-party products often used to handle special domains; e.g., Data Beacon for data mining queries
Graphical interface	Not included. Interfaces coded by search vendor or customers	Templates provided. Customization required.	Customization or third party products required.
Selective Dissemination of Information functions	Varies by user’s level of access to the search functions. Not generally offered for “free” searching except inthe form of My Yahoo! or Google News which are variants of SDI technology.	Varies by vendor. If not inlcuded, a third-party tool such as BEA Systems WebLogic or WebSphere can be used to provide the function.	Customization required.

Table 1: Search and Three Content Areas

Selected Attributes	Internet	Intranet	Special Domains
User customizable interface	Application Programming Interface (API) provided	Templates can be changed by customer. API may support third-party services for visualization of results	Customization required
Field search	Limited	Varies. More robust packages support Boolean and field searching	Customization required
Support for SQL database content	Not generally available. Development underway at major vendors and services	Varies. Vendors offering SQL support may require one search box for "text" and one for "data"	Certain content domains may require a separate log on, authentication, and search process. Customization required
Multilingual support[4]	Most vendors provide support for multiple languages. More support coming but Arabic and Chinese continue to lag behind Romance language support	Low-cost tools may have no support for a language other than English. More robust tools provide greater support. API allows integration of third-party services.	Customization requires. Varies by data source. A database may require only row and field names to be translated; data are numeric and can be used as is.
Network load controls	Sophisticated and controlled through custom scripts or Web forms permitting values to be entered to control spidering threads and other load-centric functions	Varies. More robust packages provide control via scripts or Web forms. Lower-cost or ASP services may offer no or limited controls.	Custom integration required
Training	Varies. Focus is on selling managed services, not training clients	Varies. Even low cost packages focus on upselling service and support "bundles"	Depends on how the client approaches the problem of special domain content
System administration	Extensive tools relying on scripts and Web forms for most frequently "tweaked" values	Varies by vendor. More common is a two-tier approach. The client can handle basic functions like which folder to spider and when. More advanced services are part of the support "bundle."	Depends on how the client approaches the problem of special domain content.
Branded content	Biggest vendors can process branded content. Billing and rights management are issues.[5]		

1.Wireless content poses special challenges. These are not generally addressed by the high-visibility vendors of search. Specialized vendors such as Pinpoint in Durham, North Carolina, focus on this segment.
2.A list of the file types supported by the "Outside In" technology now owned by Stellent, Inc. is available at www.stellent.com. Some search engine vendors create their own filters in order to avoid paying license fees to a third party.
3.Both IBM and MIcrosoft offer search software to handle content in these proprietary software environments. The next release of Office will, for example, perform more robust searches of attachments for electronic mail.

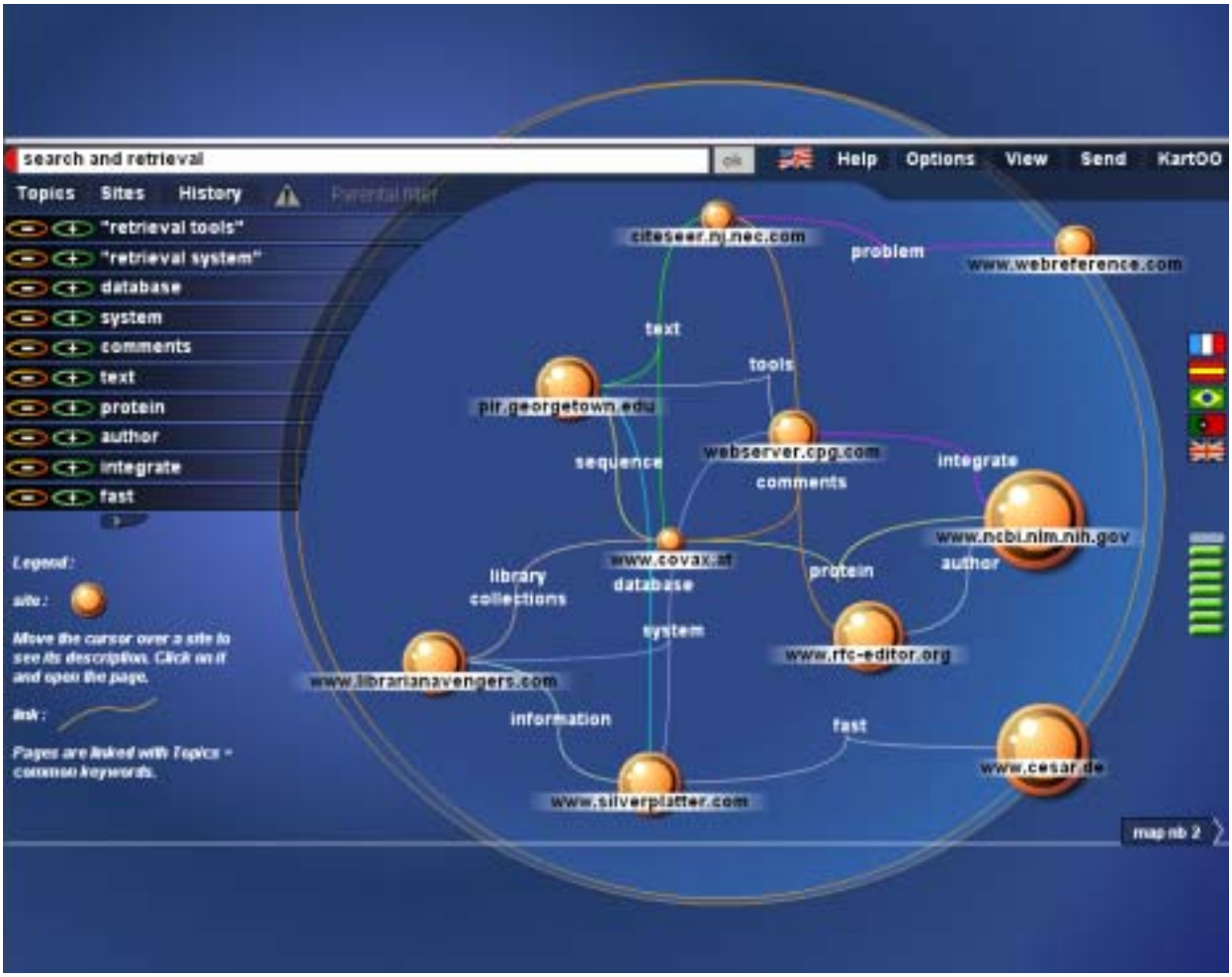
4.At this time, FAST Search & Retrieval and Google do the best job of supporting non-English searching of the public Internet. Their Intranet customers can use these services. Customization is typically required to meet the Intranet customers' needs. Pertimm, a French search engine, is one of the few products that supports a query in English against multilingual content returning hits across the languages in the corpus.
5.Copernic, a Canadian vendor of Intranet and personal search software, is working to sign up publishers so that their content appears in the list of Copernic "hits." This service will become available sometime in 2003. The branded content is not the challenge. The hurdles are keeping track of usage, billing, and reducing the risk of unauthorized reuse.

Meanwhile, ever more advanced linguistic-statistical, knowledge-based, adaptive search systems are showcased at trade shows, in impassioned sales presentations, and from often inscrutable Web sites. White papers explain and reexplain such concepts as "an ontology generation engine" and or "real-time linguistic analysis of diverse document types." Data are displayed in animated hyperbolic maps or relevancy ranked lists with the key concepts highlighted for the busy user. Some of the "advanced" systems create reports in the form of packets of Adobe Portable Document Format pages or, at the other extreme, collections of "key paragraphs" from longer documents. Dot points, extracts, and flagged items are supposed to make perusing a list of "hits" a more productive task.

Eyes glaze and potential buyers wanting to find information on computers in a marketing department wonder, sometimes out load, "What's all this jargon hiding? Does the system search, find results, work almost all the time, and fit my budget?" Not surprisingly, these are difficult questions to answer, and the answers are very, very hard to get. Search has become a digital form of roulette. The customer picks a product, spins the wheel, releases the ball, and hopes for a winner. Search and retrieval software is a similar bet. As in casinos, the customer—an information technology manager in charge of the search software acquisition—usually walks away disappointed.

Figure 1:The hot trends in search are the jargon generators for hundreds of software companies and thousands of service providers. Ontology generation. This is a bit of jargon used to activities ranging from creating a list of subject categories for a particular content collection to downloading the Library of Congress subject headings and making additions and deletions as required. Automatic ontology generation means no librarians, please. Real time is, of course, the term *de jour* for updating an index when new content becomes available. Real time is relative and, as a bit of testing on a breaking news story like the 9-11 attack, essentially untrue except for a handful of specialized services. Linguistic analysis is a slippery phrase. When used by a pitchman describing a new search engine, the listener is supposed to conclude that the software "understands" words and phrases in a manner similar to a human. Software remains rule based, and the fancier algorithms using "ant technology" or "swarming techniques" remain locked in research and development laboratories. Linguistics boil down to knowledge bases or statistical routines hooked together in a clever new way. Hyperbolic maps and other visualization techniques are increasingly available. The idea is that a list of "hits" for a query are displayed in a visual manner. For a look at what programmers can do with Macromedia's Flash technology, click to Kartoo. Are these techniques likely to tackle finding a single electronic mail message with a PowerPoint attachment. No. And not for quite a while.

Figure 2:Kartoo's Visual Presentation of Search Results



Anyone who has had to implement a large-scale system indexing content from 40,000 or more servers and processing 50 million or more documents knows that these search-and-retrieval software for a system of this size is going to be a bit different from the free search routine included with Windows XP or the "Find" option in Outlook Express.

The surprising truth is that there are a very small number of companies with products that can handle a 50 million document baseline, keep it up and running 99 percent of the time, and update the index in less than 48 hours.

There is a reason why Google grows its index in chunks, jumping from one million to three million documents over a period of years. That reasons include planning the growth, engineering the many subsystems that make up search, and accruing the money to extent the infrastructure to add content while maintaining response time.

Anyone making it to the third or fourth year of a computer science program at a major university can implement the software to find, index, and make searchable content. What computer science classes and a quick course in Excel macros do not cover is how the costs work in a large-scale Internet or commercial Intranet search-and-retrieval product. Costs, not technology, account for much of the attrition in the search and retrieval sector.

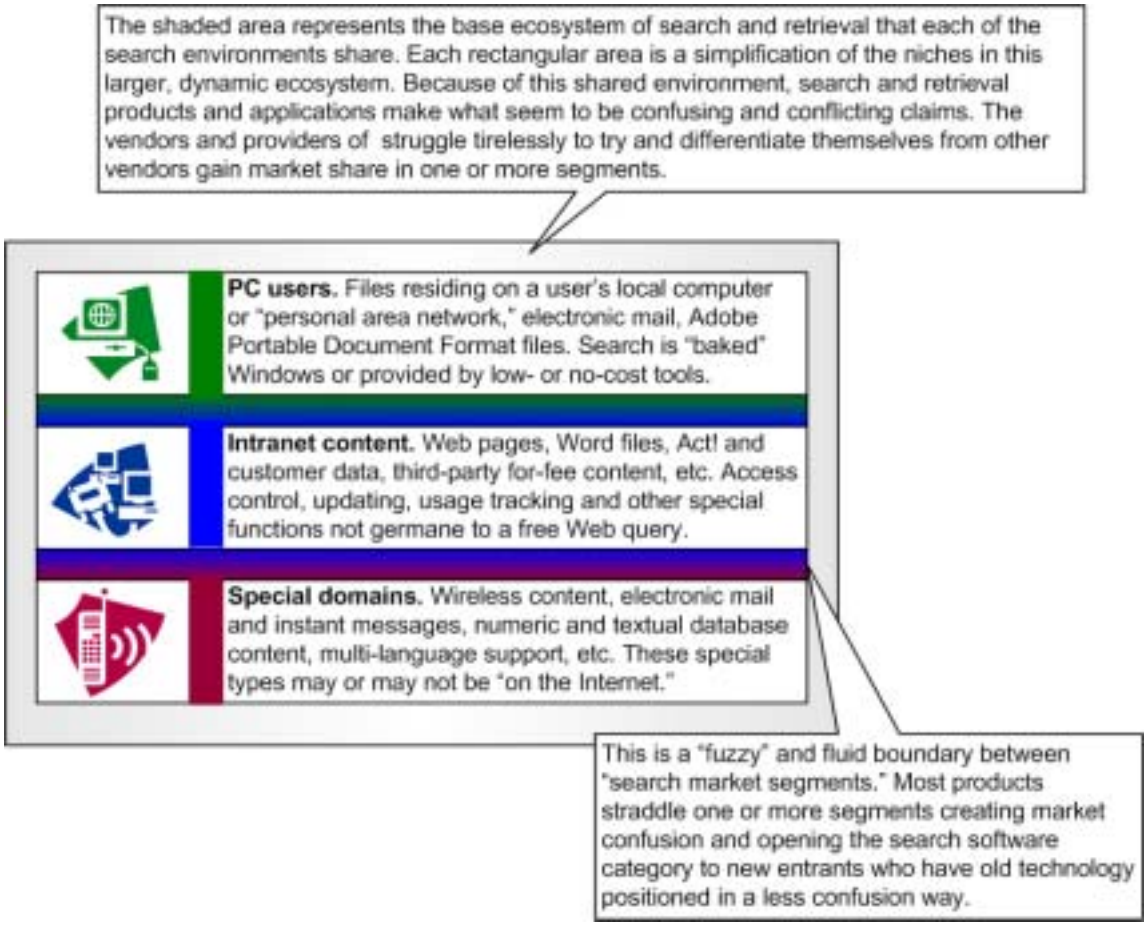
The Darwinian nature of the search business allows boutique search companies to appear and often disappear as quickly. Among the companies whose investors have considerable optimism are MyAmigo

(Australia), Pertimm (France), ClearForest (United States), and iPhrase (United States). Hopefully most of these companies will survive and thrive. It is, however, doubtful that any of the newcomers are going to challenge in the near term the dominance of a handful of search and retrieval companies. Vertiy, Inc., PC Docs, Autonomy Ltd, and a few others dominate the commercial market. Google and FAST Search & Retrieval are best positioned for a run at the market leaders. Overture, a company that has quietly transformed search to a form of advertising, has revenues and earnings that dwarf virtually all search and retrieval companies. Overture, however, downplays its search technology and focuses on its revenues of more than $500 million. Only Google with an estimated 2002 revenue of $300 million seems positioned to mount a threat. The rest of the thousands of search vendors are finding themselves new homes nestled inside of such fuzzy product packages as customer relationship management, knowledge management, and content management software.

Search and retrieval is at once everywhere in the form of free Web searching via Google, FAST's "alltheweb.com", and Yahoo!'s Inktomi service. It is also nowhere because search is part of the fabric of Windows XP, e-mail programs, and the ubiquitous "search" box on Intranet and Extranet Web pages. Search is ubiquitous, so most users do not see it as a separate function. It is a handy and necessary tool.

There are distinct niches or market segments for search. The diagram entitled "Search and Soft Market Segment Boundaries" provides a simplified view of customer and user clusters. First, every computer or mobile device has some type of search function. In a mobile phone, search may be hitting a number key and seeing the name and phone number stored at that location in the phone's memory. Search may be using the built-in tools in commercial application software. Even Excel has a "find" command. Within this segment, complete micro-ecologies of search software exist. For chemists, Reed Elsevier and Chemical Abstracts offer specialized tools that meet these users' needs in their laboratory or in an organization's Intranet. An "Intranet" is a network that operates within an entity and access requires a user name and a password.

Figure 3:Search Ecologies



Second, there is the Internet. Internet search and retrieval has been free, although the financial model for monetizing search has been cobbled together from the failure of many early entrants. Google, Alta Vista, Yahoo!, and newcomers such as Bitpipe offer "free searches." The thirst of search engines for money is unslakeable, so revenue is generated by selling advertising, selling a "hit" when a user does a search on a particular topic, or reselling searches of content to other Web sites. The variations for monetizing search are proliferating. The issue with monetizing is, of course, objectivity. In the Internet "free" search segment, objectivity is not the common coin of the realm. Paying for clicks and traffic is more important than relevance. When Google goes public in the next year or so, the need for revenue means that objectivity takes a back seat to monetizing.

The third segment is what I call "special domains." These are the collections of content that defy the mainstream text-centric search engine. Music, videos, computer aid drafting diagrams with a database of parts and prices, medical images, and audio content are not searchable with the software that falls in the purview of librarians or expert searchers. These special domains account for as much as 90 percent of the digital content produced at this time, based on a study we conducted over a period of six months in 2002 for a major technology firm. Chinese language Web pages, an electronic mail message with an Excel attachment in a forwarded message, purchase order information in CICS system files, and streaming audio from radio stations are three examples of content that is plentiful and difficult if not impossible to search.

The key point in the diagram boils down to the boundaries among and between segments. These boundaries are like those of a paramecium's. The boundary is semi-rigid, permeable, and subject to its environment. Search, therefore, can be explained in an infinite number of ways. One consequence of this interesting property of search is that comparisons are difficult. What was true yesterday of Google may not

be true today. Google's catalog service was essentially unusable. However, the Froogle service is a useful, high-value service. Consultant analyses and comparisons of search software are the intellectual Twinkies of this software sector. One can eat many Twinkies and go into sugar shock, but the essential nutrients are simply not there and the growling stomach is not satisfied. Fuzzy boundaries make comparing search software in an "apples to apples" way difficult if not impossible.

This diagram explains in part why the search landscape and the dominant companies in the search business change over time. Consider the Canadian search company Fulcrum. Fulcrum's software is quite good as Intranet tools exist at this time. Several years ago, was bought by another Canadian company (Hummingbird). Hummingbird provided software to permit a PC user to access data on a corporate mainframe via a screen scraping program. Hummingbird was, in turn, acquired by another Canadian company (PC Docs). This outfit was a document management company and wanted to upgrade its search and retrieval functions and leverage Hummingbird's customer base. Now, Fulcrum search is a facet of PC Docs product suite. A similar tale can be told about Open Text. Originally a Web indexing and SGML database with a search function, Open Text now consists of pieces of Tim Bray's search engine and the BASIS database search tool plus other search functions to handle the collaborative content in Live Link. Inktomi has sold its Intranet search business to Verity and then allowed itself to be purchased for about $250 million by Yahoo! Other companies have simply retreated from search, repositioned themselves, and emerged as taxonomy and ontology companies. Examples of this include Semio (France and California) and Applied Linguistics (formerly Oingo, operating in Los Angeles). There is more horse swapping and cattle rustling in these three segments than almost any other software sector. Confusing. Absolutely.

Table 2: Snapshot of Key Players

Company	Snapshot	Secret Sauce
Applied Semantics Inc.	Originally Oingo, this Los Angeles-based company offers automatic classification and human-edited "ontology" services. www.appliedsemantics.com	Company has found a lucrative market applying its technology to suggesting new domain names. See Register.com for an example.
Autonomy Ltd	Originally based in Cambridge, England, Autonomy has become the poster child for the European software agency. With Verity, one of the dominant Intranet indexing engines. www.autonomy.com	Made Bayesian algorithms the solution to Intranet search. New initiatives include search and retrieval of audio voice mail messages.
ClearForest Corp.	A sophisticated classification and indexing engine. The product is aimed at corporations and intelligence agency applications. Plan on a six figure price tag. www.clearforest.com	Features automatic bound phrase extraction
Overture Services, Inc.	Formerly GoTo, this service incorporates string matching and a number of other sophisticated technologies. The company has transformed search by monetizing the hits that are displayed based on who buys a word.	The company generates revenues that are roughly six times the revenue of Verity. The financial winner in search. Overture will be increasingly challenged by Google's listing business.
Pertimm SA	A product of French scientists, the Pertimm engine provides a suite of technology that supports Web services and handles queries in one language across content in any of the dozen languages the engine supports. www.pertimm.com	Software returns hits based on automatic query expansion and point-and-click navigation of "glimpses" or relevant extracts from a corpus.

Table 2: Snapshot of Key Players

Company	Snapshot	Secret Sauce
Stratify, Inc.	Funded by the U.S. government, Stratify performs a range of classification and indexing functions. www.stratify.com	Positioned as one of the first “content discovery” tools. Human-assisted indexing added when software-only solutions needed “tweaking.”
Verity	Verity holds with Autonomy the lion’s share of the US search and retrieval business. Customers include Adobe and metasearch provider Bull’s Eye. www.verity.com	Owns Inktomi’s Intranet customers and the Ultraseek text retrieval engine used to provide search and retrieval for Bitpipe.com
Yahoo!	Yahoo! has shifted from a directory service although Web site owners are encouraged to pay for listings to a search model. Yahoo! left Inktomi for Google and in 2002 bought Inktomi.	Inktomi provides custom Web spidering that can be costly to scale and refresh.

One interesting twist in the search business has been morphing search and retrieval into a system that *discovers* what information a company has. As silly as this sounds, there are organizations that simply do not know what information is on the organization’s own servers. (If this sounds like a commercial for knowledge management, it is not.) Search and retrieval software has been packaged as a way for a security officer at a large company to know what the Des Moines, Iowa office put on the Internet.

Most organizations have allowed their Web server population grow like Manchester, England's in the height of the industrial revolution. In our security-crazed post 9-11 and post-Enron world, the boards of directors have to know what information exists, in what form, who has access, and, of course, what information is available to whom. As more companies reinvent themselves as “knowledge organizations” or “information companies,” few—if any employees—know such basics as:

- What information is in digital form
- Which information is the most recent or “correct” version
- Where a particular piece of information is.

If this reminds the reader of “content management,” it is an easy mental leap to the role of search and retrieval in this market sector. In the pre-digital age, people could stay late and look through stacks of paper. Today, not even the most caffinated Type A can browse hundreds or thousands of files on different machines in many different formats. The job is too onerous, too tedious. With a few deft marketing professionals' help, a search engine can be paraded as an information discovery engine.

The idea is that the search-and-retrieval system looks at a company's information objects, figures out what each object is “about”, and then clusters the objects in a Dewey Decimal type of scheme. There is a word for this type of work, and that word is *indexing*. Indexing professionals, librarians, and content specialists working for the National Library of Medicine, cite a few examples, used to do this work. Now that such individuals are deemed non-essential or are simply too expensive, software is supposed to do the job. Not a chance. Verity, the current industry leader, makes it very clear that part of the firm's professional service fees include payment for humans who “tune” and “train” the Verity system. For those who can't afford Verity, there are the transformed search companies or specialist firms who can deliver software that indexes and classifies so someone, somewhere knows what is on a corporate Intranet. Clear Forest is one company that has been identified as a leader in this 'discovery' niche. For military intelligence and government security applications, i2 Ltd. (Cambridge, England) provides an tightly integrated suite of

tools that allow discovery to run as a process with the results depicted in with icons, connector lines, and "hooks" to non-text objects.

The companies that have done the most effective job of getting their technology embedded in content management, customer relationship management, and -- my favorite meaningless discipline -- knowledge management, are Verity (Mountain View, California) and Autonomy Ltd. (Cambridge, England). These "M" businesses—document management, customer relationship management, knowledge management, and content management—need search that is reliable.

The segment leaders with about 70 percent of the US and European corporate and government market are Verity and Autonomy. Both companies' products "work". The precise meaning of "work" is somewhat difficult to define because an inspection of the lists of organizations each has as customers reveals an overlap of about one-third. For basic search and retrieval, these companies are market leaders. Unlike Overture, the business model for Verity and Autonomy is to license software and then sell support, customization, and services. Both firms will provide the services required to satisfy the customer. The price for search that works can reach seven figures.

The strengths of Verity and Autonomy are not the firms' respective technologies. (Verity relies on thesauri and what might be called traditional indexing by extracting terms. Newer algorithms have been added and the company can process database files in the recently upgrade K2 engine. Autonomy relies on statistical techniques originally based on Bayesian statistics. Like Verity, Autonomy has embraced other approaches and acquired companies to gain customers and technologies in speech recognition.) Both Verity and Autonomy can support corporate customers. Smaller companies with lower fees usually find that the juicy accounts go to Verity or Autonomy because of the firms' ability to install, support, and service enterprise clients. One systems manager said in a focus group in 2002, "No one gets fired for licensing Verity or Autonomy."

Most commercial search software with an Intranet version works at what might be called the 70 percent level. For a query, more than two-thirds of the content will be available when the query is passed. The results will be about 70 percent on the topic. The very best engines push into the 80 percent range. It is very difficult with today's technology to get consistently high scores unless the content domain is tightly restricted, updates are frozen, and correctly formed fielded queries are crafted. The reader familiar with SQL queries or Dialog Boolean queries will immediately see why typing one or two terms, hitting the enter key, and looking at a list of hundreds of results requires considerable manual filtering.[1]

To the question, "Do commercial search engines work?"

The answer is, "Yes … effective search and retrieval software gets about 80 percent of the relevant material. Stated another way, the most effective searches usually miss at least 20 percent of the content that could be highly pertinent to the user's query. Verity delivers this type of search effectiveness when the company's software is properly set up. But, as many Verity customers have discovered, this means that considerable human, manual effort is needed. Out of the box, the best search software tracks with the results of the TREC competition. Searches in limited domains with tightly controlled word lists are more satisfying than searches run across heterogeneous domains of content. For most users, precision and recall at or near the B minus or C plus level is "good enough."

"Good enough", in fact, describes how most search and retrieval engines work. Google is "good enough" because the results are ranked by a voting algorithm that weights pages with links over pages with few links. What if a page does not have links, but the page has outstanding content? Google may index the

1.The reader interested in the performance of various search engines may wish to review the results of Text Retrieval Conference (TREC), co-sponsored by the National Insti6tute of Science & Technology, Information Technology Laboratory. http://trec.nist.gov.

page, but unless the query for that page is well-formed, the page without links may be buried deep in the list of results or not displayed. Most users follow the Alexis de Toqueville rule that when the majority votes, the result is mediocrity. Google receives such praise that it is heretical to suggest that as Google's popularity grows, potentially useful pages disappear beneath pages that may be more popular and, hence, less excellent.

Back to Basics

In the early days of search and retrieval, there was only one way to find information. A proprietary system offered a command line interface. To find a document, the *searcher* (certainly not a pejorative term in most Fortune 500 companies or at NASA where search began in the late 1960s) crafted a query.

The query required a reference interview with the person wanting the information or a conversation with a colleague who understood a particular domain's jargon and the context of the query for a particular client. The search query was assembled using appropriate terms, usually selected from a printed controlled vocabulary. (In the early days of search, it was considered a point of professional pride to have professional indexers assign key words using a thesaurus to the documents or entries in a database. The word list—called a *controlled vocabulary*— was a road map to the information in the database.

If synonyms were common, as they were in medical, technical, general business or news databases, 'use for' and 'see also' references were inserted into the thesaurus. The searcher then crafted a well-formed Boolean query using the syntax of the online system. (*Well forme*d means that the logic of the query would return a narrow set of results or hits. It was these precise, on-point result sets were the proof that an expert searcher was at the controls.)

The expert research then selected specific databases (now called a *corpus* in today's search jargon).[1] The well-formed query was then run against the appropriate databases. At first there were a handful of these online databases in 1970. The number grew to about 2,000 by 1985. Today, the Web has given rise to database proliferation where a single Web page counts as a database and there are more than three billion Web pages indexed by Google alone. The results were reviewed by the searcher, and the most relevant were selected by the searcher. Additional queries were run by constructing search syntax that essentially told the online system "give me more like this."

How far we have come since 1970?

Not far. Today software is supposed to look at the user's query and automatically expand the terms. Software is to bring back relevant documents and rank them, highlighting the most important sections. The blunt truth is that for most online users today, looking for information boils down to a pretty straightforward set of actions. Hold on to your hat and fasten your seat belt; users today do one or more of three things:

- Type one or two terms into a Google-style search box and pick a likely "hit" from the first page of results. (The most searched term on Google, I have been told, is Yahoo! so see item 2 in this list.)

- Go to a site with a Yahoo!-style *taxonomy* or *ontology* and click on a likely heading and keep clicking until a "hit" looks promising. (This is the point-and-click approach much beloved by millions and millions of Web users each day. It meets the "I'll know

1.(The Citeseer Web site provides useful information about corpus. A shorthand definition is "the body of material being examined." See http://citeseer.nj.nec.com/hawking99results.html. Links were updated in 2002.

it when I see it" criteria so important in research.)

- Look at a pre-formed page and use what's there. (This is the digital equivalent of grabbing trade magazines in one's in box and flipping pages until a fact or table that answers a question catches one's eye. Believe it or not, a 1981 Booz, Allen & Hamilton study found that this was the second most popular way to answer a question among polled executives. The most popular way was to ask a colleague. When the study was updated, asking was still first but looking at a Web page was the second most popular way to get information.)

Why Search and Retrieval Is Difficult

Search-and-retrieval is a much more complex problem than most information professionals, systems engineers, and even MBA-equipped presidents of search engine companies grasp. Search brushes up against some problems that are what computer scientists call *intractable*. An intractable problem is one that cannot be solved given the present state of computing resources available to solve the problem. Let me highlight a few to put the search challenge in context.

First, language or languages. The "answer" may not be stated explicitly. Years ago, Autonomy Ltd. pitched its search engine by saying that its Bayesian approach would find information about *penguins* if the user entered, "Black and white birds that cannot fly." I think the demonstration worked, but in the real world, the Autonomy system performs best on closed content domains of homogeneous information. Language is a problem because of metaphor, structure, and neologisms, and it becomes intractable when one tries to support, say, a French query delivered against content consisting of Arabic, Chinese, and Korean material.To be fair, most people looking for a pizza in Cleveland, Ohio, want to use English and get the information with a single click. Interface and presentation of search must balance power and ease of use.

Second, most people doing searches don't know what the answer is. The human mind can synthesize and recognize, but is less adept at foretelling the future. Searching then requires looking at information and exploring. Much of the Web's popularity was a direct result of the browsing, exploring, and wandering in interesting content spaces. One difficult is that clicking on a list of "hits" that numbers 100,000 or more is mind numbing. User behavior is predictable. Give me something useful, and I go on with life. Search and retrieval systems must permit a chance encounter with information that may illuminate a problem. Showing 10 hits may be inappropriate in some cases and just right in others.

Third, as demographics change and thumb-typing young people join the work force, new types of search systems are needed. It is difficult to tell what the long term impact of Napster's peer-to-peer model will have on information retrieval. One pundit (Howard Rheingold) opines that swarm behavior will become the norm, not solitary search and retrieval. I think of this approach to answering questions as Google's popularity algorithm on steroids. The answer is what people believe the answer to be. One part of my mind wants to stop this type of information retrieval in its tracks. The other part says, "Maybe swarm searching is good enough - let many flowers bloom as a famous resident of China once said." The idea is that one asks a question and passes it among many system users. The answers that come back reflect a swarming process. Swarm technology has been replicated to a degree in the search and retrieval technology developed by NuTech Solutions in Charlotte, North Carolina.[1] NuTech uses the term *mereology* to describes its approach.

Fourth, the emergence of an ambient computing environment supports the pushing of information to individuals. With IPv6, every digital gizmo can have a Web address. Personalization technology is

1.See NuTech Solutions description of its technology at www.nutechsolutions.com. The search product is marketed as Excavio. A demonstration is located at http://www.excavio.com.

becoming sufficiently robust to deliver on-point information to an individual's mobile phone without that "user" having to trigger any query. In Scott McNealy's vision, an automobile that needs fuel will query a database for the nearest gasoline station. When a station comes in range, that data will be pushed to the digital display in the automobile along with a map of where to turn to get fuel. Search in this model reverts to what used to be called Selective Dissemination of Information. Today, the words used to describe the SDI approach range from text mining to agent-based search and into even more esoteric word crafting. Search will mesh with decision support or Amazon-like recommendation systems. Most of the people looking for information today seem to open their arms to environments that wake up when the "searcher" turns on a wireless device. The screen says, in effect, "Hello, here's what you need to know right now." The Research in Motion Blackberry device showed that e-mail and pushed stock quotes was a potent online combination for go-go executives in financial services and management consulting.

Fifth, in some search and retrieval situations source identification and verification -- or what art dealers call *provenance*—is difficult. Few point-and-click Google searchers or employees browsing filtered news on a personalized portal page know or care what the editorial guidelines are for a commercial database. If a consulting firm's table of statistics appears on a Web site, it "must be" accurate. Some pundits have winced when thinking about Enron executives making decisions based on a casual Web search or television talk show. Bad information and loose ethical boundaries are combustible.[1] Most search engines for Intranets drag in whatever they find. My dog often brings me a dead ground hog. Thoughtful of the dog, but not germane to my needs. I am not sure software alone can address this challenge, but it warrants thought.

This list of challenges can be extended almost indefinitely. And we haven't even touched the cost of bandwidth to index large content domains, the size and computational capability of the indexing environment that must process data, make judgments, and deliver results often to thousands users who hit a system simultaneously, or the performance issues associated with making updates and results display before the user clicks away in frustration over slow response. The costs associated with search are often difficult to control, and when search firms run out of money, they close. Bang.

Approach Search's Weaknesses Objectively

The search vendors are scrupulously polite about their competitors' technologies. That *politesse* stems from the results of large-scale tests of search engines. Look at three or four years of TREC results. Most of the technologies perform in a clump. Precision and recall scores are essentially the same. What's more interesting is that the scores top out in the 80 percent range for precision and recall, and have done so for several years.

Two observations are warranted by the TREC data and actual experience with brand-name search "systems." First, search has run into a wall when it comes to finding relevant documents and extracting all of the potentially relevant documents from a corpus. Despite the best efforts of statisticians, linguists, and computer scientists of every stripe, improving the TREC score or satisfying the busy professional looking for a Word document works less than 100 percent of the time. As noted, the use of voting algorithms has created a self0fulfilling prophecy whereby users are thrilled with a C or B-minus performance. The more people who find this level of performance satisfactory triggers a feedback loop that guarantees mediocrity. Second, the compound noun neologisms of marketers cannot change the fact that commercial search systems work on text. Most commercial and university think tank software of the search engines—including the ones wrapped in secrecy at the Defense Advanced Research Projects Agency—cannot handle video, audio, still images, and compound files (a Word document that includes an OLE object like an Excel

1.I received a round of applause at the Library of Congress during my talk on wireless computing when I said, "Where should information quality decisions be made. In the boardroom of companies like Arthur Andersen and Enron or in management meetings where trained information professionals vet data."

spreadsheet or a video clip). There are search engines for these files. Just ask any 13-year-old with an MP3 player or your college student living in a dormitory with a DVD recorder, ripping software, and a broadband connection.

Multilingual material complicates text search in certain circumstances. Accessing information in other languages is gaining importance in some organizations. Search engines, when carefully set up, can handle the major languages, but running search queries across the multilingual search engines perform less well than search engines running on a corpus composed of text files in a single language. This means that finding associated or relevant documents across corpuses, each in a different language, is essentially a job for human analysts. Said another way, search produces manual work. In the post 9-11 world, the inability to address Arabic, Farsi, Chinese, and other “difficult” languages from a single interface is a problem for intelligence analysts in some countries. Toss in digital content with corpuses composed of audio clips, digitized video of newscasts, electronic mail, and legacy system file types, and we have a significant opportunity for search innovation. From the point of view of small company, solving the problem of searching electronic mail might be enough to make 2003 a better year.

Search is serious, and it is a baseline function that must become better quickly. Search will not improve as long as buyers and users are happy with “good enough.” A handful of information professionals are aware of the problem. In the rush of everyday business voices are not heard when questions are asked about purchased relevance versus content relevance, bait-and-switch tactics, and the thick “cloud of unknowing” that swirls around data provenance, accuracy, completeness, and freshness.

New technology acts like a magnet. Novelty and the hope is that the newest search technology will be the silver bullet for search problems. The pursuit of the novel and the word-spinning the purveyors of new search technology use is attractive but no one steps back and asks hard questions about search—free, Intranet, Internet, peer-to-peer, or wireless.

An example of the how the snappy can befuddle understanding of the limitations in present search and retrieval technology is the positive reception given Kartoo (Paris, France) and Groxin (Sausalito, California). Strictly speaking these two companies have interesting and closely-related technology. A query is launched and the “hits” are grouped into colorful balls. Each ball represents hits related in some way to a particular concept. The links among the balls show relationships among the concepts. Sounds useful, and the technology is appropriate for certain types of content and users. Visualization of results in clusters, of course, relies on underlying clustering technology which must be sufficiently acute to “understand” extreme nuance. To get a sense of how well that technology works, run a query on Kartoo in an area where you think you know the subject matter well. Now explore the balls. Are the “hits” clustered correctly? In our tests of Kartoo, we found that more than half the balls contained some information that was useful. But Kartoo and Groxim return results that are too coarse to be of value to an expert in a particular domain.[1]

Results: Biased and Sometimes Useless

Search and retrieval is believed to be unbiased. It is not. Virtually all search systems come with knobs and dials that can be adjusted to improve precision or adjust recall in a commercially-successful search engine such as FAST Search & Retrieval (Wellesley, Massachusetts and Oslo and Trondheim, Norway). The company can make adjustments to the many algorithms that dictate how much or how little a particular algorithm can affect search results. Yahoo!-Inktomi's, Open Text's, and Alta Vista's search engines have similar knobs and dials. Getting the settings “just right” is a major part of a software deployment. For Intranet search, Verity is the equivalent of the control room of a nuclear submarine.

1.Kartoo’s engine is located at www.kartoo.com. Groxim requires that the user download a software module and run the program on the user’s machine. Groxim’s software is located at www.groxim.com.

The flip side of the knobs and dials is the question, “Is it possible to set the knobs and dials to bias or weight the precision and recall a certain way so when I type *airline* I display the link of the company paying the most money for the word *airline*?”

The answer is, “Absolutely. Would you like to buy *hotel, travel, trip, vacation, rental car*?” One can see this type of adjustment operating in Google when the little blue boxes with the green relevancy line appear on a page of hits. Indeed, the very heart of Google is to use weighting that places emphasis on popular sites. “Popularity” is defined by an algorithm that considers the number of links pointing to a site. A different view of Google’s controls can be seen on the BBC’s Web site.[1] Enter the word *travel* in the search box. The hits for both the BBC Web site and the “entire Web” are BBC affiliate sites. Coincidence. No search bias.

The clever reader will ask, “What about sites that have great content, no links pointing in or out, and relatively modest traffic? These sites are handled in an objective manner, aren’t they?” Go to Ixquick, a metasearch site with a combination of links and traffic popularity algorithms. Enter the term *mereology.* No hits on www.mereology.org. No hits for the NuTech Solutions Web site whose founder brought mereology from obscurity to the front lines of advanced numerical analysis. Serious omissions. Absolutely. Such problems are typical among specialist resources for very advanced fields in physics, mathematics, and other disciplines. However, similar problems surface when search tools are used on Intranet content. The research and development content and probably most of the data residing in accounting are black holes in many organizations.

For expert searchers, locating the right information pivots on Boolean queries and highly precise terms. This assumes, of course, that the desired content resides in the index at all. Verity’s PDF search engine stumbles over PDF files for one good reason[2]. The content of PDF files is not designed for search and retrieval. PDF files are designed for page rendering. Textual information runs across columns, not up and down columns. PDF search and retrieval requires deft programming and canny searchers. For Intranets, indexing corporate content is somewhat less problematic than indexing the pages on a public Web server or the billions of pages on the hundreds of thousands of public Web servers, but comprehensive and accurate indexing of even small bodies of content should not be taken for granted.

A common example of deliberately biased search results may be found in the display of for-fee “hits”. Companies selling traffic allow a buyer - essentially, an advertiser - to “buy” a word or phrase. When a search involves that word or phrase, the hits feature the Web site of the buyer. Such featured results are usually segregated from the “other results” but searchers may not take notice. Google and Overture are locked in a fierce battle for the pay-for-click markets. FindWhat.com is a lesser player. In the U.K., eSpotting.com is a strong contender in the “we will bring you interested clients” arena.

What about sites that offer to priority index a Web page when the Web master pays a submission fee? Alta Vista offers a pay-for-favorable-indexing option. Yahoo! offers a similar service as well even as the company shifts “search” from its directory listings to spidered search results. In fact, most of the search engines have spawned an ecology of services that provide tricks and tactics for Web masters who want to get their pages prominently indexed in a public search engine. Not surprisingly, there are discussions about the use of weighting algorithms is public sector search services as well. An example of how this might work is to use “knobs and dials” to ensure that income tax information is pushed to the top of results lists in the month before taxes are due. (I hasten to stress that this is a hypothetical example only.)

1.The BBC is located at www.bbc.co.uk

2.PDF is the acronym for Adobe’s Portable Document Format. Adobe has placed the PDF specification in the public domain to ensure wide adoption. Like PostScript, the PDF file focuses on rendering a page for rasterization, not search and retrieval.

Innovation Checklist: The Ideal Search Engine's Functions

Over the last two years, my colleagues and I have compiled a list of what I call yet-to-be-mined gold ore in search. The list includes functions that are not available in the software on the market today and could be viewed as a checklist for innovators. I view it as a reminder of how much work remains to be done in search and retrieval.

The table below provides a summary of what search-and-retrieval systems cannot do at this time:

Table 3: Search and Retrieval's Challenges for 2003

Function	Comment
Harmonization	The search engine must recognize, convert, index, and provide pointers to multiple types of PDF files, file formats, database files, etc.
Auto indexing	Novices and experts can search the index using terms common to their training and experience and find comparable results
Clustering	The system groups like objects that meet the needs of lay people and professionals alike
Learning-centric	The system monitors users' actions and adapts to those patterns in order to return optimal results for each particular user
Administrative	The system manager interacts with the search and retrieval subsystems via interfaces that make explicit the consequences of settings and minimizes or eliminates the need for programming
Scaling	The system can handle the number and type of documents it is asked to index and make available without reengineering subsystems when thresholds are crossed
Distributed	The search system is modular and can be distributed to make the best use of available resources and to minimize adverse effects on system response time from heavy indexing
Explainable	When the search systems make a decision about placing an object in a cluster, an audit trail or some other type of concrete explanation for the action taken must be available
Change-aware	The search system and its subsystems must be able to acquire new content, recognize changes to existing content, and identify available but unchanged content and index the objects accordingly
Date-aware	The search system must be able to handle various type of date and time information and use each in the appropriate context for a particular query; specifically, date and time stamp assigned by the system, file creation date, file change date, and implicit dates extracted from the content cues in the object
Adaptive	As new content objects are discovered, the search system is able to identify the object, intelligently process the object, or notify the system administrator of the new object and request guidance for handling that object
Language	The search system can recognize languages, index them, support a query in the user's language, return results from any object in the corpus in either the original language or in the user's language. The user makes a decision and the search system behaves the way a particular user requires.
Trainable	The search system supports input and guidance from humans via interfaces explicitly designed to accept existing terms, new concepts, ontologies, taxonomies, dictionaries, thesauri, or knowledge bases as required
Application Programming Interfaces	The search system provides a documented set of APIs or hooks so that authorized users can make use of the search system or a particular subsystem from another program or in support of another process
Security	The search system integrates seamless with existing security systems so that results intended for a user with a particular level of access displays only content matching that access level

Table 3: Search and Retrieval's Challenges for 2003

Function	Comment
Usage tracking	The search system provides a native usage tracking subsystem giving detailed information on a cycle set by the system administration. No third-party tools need be used to determine usage patterns.
Multiobject	The search system can handle database, text, and proprietary file forms in a structured, flat, or compound form
Query flexible	The user can query the system using a single term, bound phrase, or free text entry. These functions may be exposed by the system administrator making use of portlets DID YOU MEAN APPLETS? (tiny prewritten routines) that activate a particular search and retrieval function
Point-and-click	The search system should generate a Yahoo!-style directory when the system administrator activates that function. An administrative interface (as noted in this table) allows modification or "training" of the search system to handle certain objects in a manner specified by the administrator.

When will these functions be available? Progress will be made each year, but ultimate resolution of most of these challenges will follow be keeping engineers busy for the forseeable future.

What's ahead? The safe and correct answer is, "Change." Search is needed. The challenge of "getting it right" and reaping untold riches is a powerful lure. Good-enough searching is likely to be the driver for the foreseeable future; that is, three months, maybe four. Perhaps a breakthrough will occur that provides enhanced searching at an affordable cost. Wouldn't that be nice?

Actions Information Professionals Must Take

The focus should shift to what trained information professionals, expert searchers, and the search engine providers themselves must do. This is the equivalent of practicing good hygiene. It may be tilting at windmills but the action items I have identified are:

1. Explain, demonstrate, teach by example the basic principles in thinking critically about information.

2. Emphasize that the source of the information-its provenance-is more important than the convenience of the fact the source provides

3. Be realistic about what can be indexed with a given budget and articulate the strengths and weaknesses of a particular approach to search and retrieval. (If you doesn't know what these strengths and weaknesses are, digging into the underpinnings of the search engine software's technology is the only way to bridge this knowledge gap.

4. Do not believe that the newest search technology will solve the difficult problems. The performance of the leading search engines is roughly the same. Unproven engines first have to prove that they can do better than the engine we now happen to have in place. This means that pilots, testing, and analysis are needed. Signing a purchase order for the latest product is expedient but it usually does not address the underlying problems of search.

5. Debunk search solutions that are embedded in larger software solutions. Every content management system, every knowledge management system, every XP installation, and every customer relationship management system has a search function. So what? Too often these "baby search systems" are assumed to be mature. They aren't; they

won't be; and more importantly they can't be.

It is important that professional information associations become proactive with regard to content standards for online resources. Within one own organization, asking questions and speaking out for trials and pilot projects can be noncontentious actions.

Search has been a problem and will remain a problem. Professionals must locate information in order to learn and grow. The learning curve is sufficiently steep that a few Web search university sessions or scanning the most recent issue of *Search Engine Watch* are not enough. We have to turn our attention to the instruction in library schools, computer science programs, and information systems courses. Progress comes with putting hard data in front of those who are interested in the field. Loading a copy of Groxim's software won't do much more than turn most off-point "hits" into an interesting picture. Intellectual integrity deserves more. Let's deliver it.

Stephen E. Arnold

President, Arnold Information Technology

Postal Box 320

Harrod's Creek, Kentucky 40027

Voice: 502 228 1966

E mail: sa@arnoldit.com

Web site: www.arnoldit.com

January 2, 2003