Technology White Paper

Autonomy®

Index

1. Autonomy Fundamentals

1.1 Introduction

More than 80% of all data in an enterprise is unstructured information. This surfeit encompasses telephone conversations, voicemails, emails, word documents, paper documents, images, Web pages, video, and hundreds of additional formats. Unfortunately, attempts to leverage this immense and strategic resource often fail because many businesses lack the requisite technology to understand and effectively utilize content that resides outside the scope of structured databases.

Similarly, unstructured processes are equally unwieldy yet comprise the bulk of business operations. While conventional business process management tools can tackle well-defined processes, they predictably falter under the weight of byzantine business rules and the complex logic required by dynamic processes prevalent in a multitude of business applications. Current trends anticipate the rapid proliferation of rich media, widespread adoption of VOIP, growing use of IPTV, and increased scrutiny of white collar crimes. This overwhelming profusion demands an automated solution that can effectively manage an unstructured digital morass.

These concerns necessitate an information infrastructure platform that addresses all classes of information in a manner analogous to well established methods for structured databases. Akin to the Relational Database Management System (RDBMS) that revolutionized the computing industry from the 1960s, this innovative platform would enable computers to process not only structured data, but also vast amounts of semi-structured and unstructured information using a global relational index.

Autonomy's ability to process all forms of digital information on a single platform offers a unique solution to a growing number of applications and devices that are increasingly dependent on utilizing unstructured information. Autonomy employs a unique combination of technologies to enable computers to form a contextual understanding of digital content. By automating key processes on unstructured information, Autonomy's technology enables the automation of business operations previously only performed manually. This represents substantial savings for every type of organization and industry, as evidenced by the significant penetration of Autonomy's technology in a diversity of vertical markets.

Autonomy's unique combination of technologies provides:

Autonomy is therefore able to power any application dependent upon unstructured information including:

- *Accuracy*
- *Speed and performance*
- *Scalability*
- *Security*

- *Enterprise Search*
- *Business Intelligence*
- *Customer Experience Management*
- *E-Commerce*

- *Language Independence*
- *Easy integration*
- *Support for any content format*
- *Support for future technologies*

- *E-Discovery*
- *Rich Media Management*
- *Knowledge Management*
- *Security and Surveillance*

1.2 Intellectual Foundations: An Open Philosophy

With an R&D spend that far exceeds it competition in the industry, Autonomy maintains an open philosophy to methods it uses and continually conducts research and evaluates all approaches to improve its technology. Accordingly, Autonomy embraces traditional or legacy methods such as keyword, Boolean, parametric and others. However, it is perhaps best known for its pioneering work in conceptual search based on computational pattern recognition (non-linear adaptive digital signal processing) and contextual linguistic analysis. Built upon the seminal mathematical works of Thomas Bayes and Claude Shannon, and by innovations that include 118 patents, Autonomy's technology identifies the patterns that naturally occur in text, video or voice files based on the usage and frequency of terms that correspond to specific concepts. By studying the preponderance of one pattern over another, Autonomy's technology understands that there is X% probability that the content in question deals with a specific subject. In this way, Autonomy extracts the content's digital essence, encodes the unique "signature" of the concepts, and enables a host of operations to be automatically performed on e-mails, phone conversations, video, documents, and even people's interests.

1.2.1 Bayesian Inference

The theoretical underpinnings for Autonomy's approach can be traced back to Thomas Bayes, an 18th century English cleric whose work has become a central tenet of modern statistical probability modeling. Bayes' efforts centered on calculating the probabilistic relationships between multiple variables and, when new information is obtained, determining the extent to which these relationships are affected.

Take the simple case of a coin toss. After how many tosses would one know, with high probability, that the coin is biased and not merely a reasonable streak of lucky flips? A traditional statistical argument posits that if a coin is tossed 100 times and comes up heads every time, it still has an even chance of coming up tails on the next throw. An alternative, Bayesian approach, is to say that 100 consecutive heads indicate a 99+% chance that the coin is biased, or perhaps has heads on both sides.

$$p\,(\theta\,|x) = \frac{p\,(x\,|\theta).p(\theta)}{\Sigma_{\theta' \in \Theta}\, p\,(x\,|\theta').p(\theta')}$$

Similarly, Bayes theorem also elucidates the seemingly paradoxical Monty Hall problem. Given three doors, one of which hides a car, the contestant must correctly guess the location of the car to win the prize. After the contestant guesses door #1, the host, who knows where the car really is, opens door #2 to reveal it is empty. The contestant then is given the choice of sticking with door #1, or switching the answer to door #3. Conventional wisdom would determine that the probability for either door hiding a car is each 50% and thus the odds of winning will not be affected by a switch; however, Bayes theorem reveals that switching the choice to door #3 would actually increase the chances of winning from 1/3 to 2/3.

What Bayes theorem clearly demonstrates in the previous two examples is that a) the more information you are given, the more accurate your view of the world will be, and b) prior experience should be used to inform new data. In a typical problem such as judging the relevance of content to a given query, Bayesian theory requires that this calculation be related to details that we already know. A good example of this theory at work is Autonomy's agent profile technology. Users can create agents to automatically track the latest information on their interests, and Autonomy determines the relevance of a document based on the model of the agent.

Extensions of the theory go further than relevance information for a given query against content. Adaptive Probabilistic Concept Modeling (APCM) analyzes the correlation between features found in digital information relevant to an agent profile, thus finding new concepts and content. Concepts important to sets of documents can be determined, allowing new content to be accurately classified. In a similar manner, knowledge about digital content deemed relevant by a user to an agent's profile can be used in judging the relevance of future documents. APCM allows this information to be "back propagated;" in other words, agents can be retrained by example.

While most other models start with a priori knowledge of the state of the system and apply training to it, Autonomy begins with a blank slate and allows incoming data to dictate the model. Instead of depending on grammar and linguistic rules as its prior state, the model mixes new information with a growing body of indexed content to refine and retrain the engine.

1.2.2 Shannon's Information Theory

Information Theory forms the mathematical foundation for all digital communications systems. Claude Shannon stated that "information" could be treated as a quantifiable value in communications. Consider the basic case where the units of communication (for example, words or phrases) are independent of each other. If p_i is the probability of the i^{th} unit of communication, the average quantity of information conveyed by a unit, Shannon's entropy or measure of uncertainty is:

$$\mathrm{H}{=}{-}\sum p_i.\log_2(p_i)$$

This formula reaches its maximum when the probabilities are all equal; in this case the resulting content would be random. In other cases the information conveyed by the content will be less than this maximum, which suggests some redundancy. This result is then extended, by more sophisticated mathematical arguments, to instances when units are related. Natural languages contain a high degree of redundancy, or nonessential content. For example, a conversation in a noisy room can be understood even when some of the words cannot be heard, and the essence of a news article can be grasped simply by skimming over the text. Information Theory provides a framework for extracting the concepts from this redundancy.

Autonomy's approach to concept modeling relies on Shannon's theory that the less frequently a unit of communication occurs, the more information it conveys. Therefore, ideas, which are rarer within the context of a communication, tend to be more indicative of its meaning. It is this theory that enables Autonomy's software to determine the most important (or informative) concepts within a document.

1.3 Conceptual Search

Built on a unique pattern-recognition technology, IDOL enables either a manual or fully automated means of matching and identifying the similarity between pieces of information. IDOL solutions offer higher degrees of accuracy and sophistication using a scalable technology that recognizes concepts.

1.3.1 Importance of Conceptual Search

When handling large amounts of data, legacy search methods such as keyword and Boolean search have proven to be both insufficient and inefficient. One way to measure effective search is the rate of precision and recall at which the technology performs. Precision is the proportion of documents in a query result that are relevant. A query result with perfect precision would consist of all returned documents that are relevant. Recall is a measure of a search result's ability to find all relevant documents. So for a search engine to achieve both high precision and high recall (an ideal scenario), it would be finding all relevant documents without returning along with them unrelated content.

Studies have shown that traditional Boolean search performs woefully according to this metric. For 50% recall (returning only half of relevant documents), users can expect 20% precision. This means that 80% of search results that the user must wade through is completely irrelevant to his/her purpose!

This failing becomes especially magnified and costly in an eDiscovery case, where the law dictates that all potentially relevant electronically stored information (ESI) be found. One can imagine that if a litigator were to search for the phrase “paint the town red,” the Boolean search engine would return not only documents relating to possible engagement of inappropriate behavior, but would also retrieve completely irrelevant information regarding art, color and cities that the busy litigator must spend his/her costly time culling.

Boolean search, moreover, is heavily dependent on the skill of the person performing the query. The user must pick the right key terms to search, or risks missing critical documents that are conceptually relevant. Criminals also understand how basic search engines work and purposely pick obscure code words to minimize their likelihood of getting caught.

Therefore it is essential for the enterprise to adopt a conceptual search technology which does not fall prey to the limitations of legacy methods, but performs effective search in a scalable manner. Since conceptual search can find information based on words not located in the document, investigators who don’t necessarily know all the details of the case can benefit from retrieving information they did not know even existed. By largely limiting the search results to only the relevant content, combined with advanced analytics that cluster the results into conceptual buckets, the user is saved enormous amounts of valuable time needlessly culling through unimportant documents. Instead of reviewing and conceptually categorizing the content themselves, he/she can concentrate his/her efforts into reviewing and analyzing the data at hand.

With conceptual search at the foundation of information access technology, enterprises can conduct content-based archiving and legal hold, as well as perform exact and near de-duplication in legal applications.

1.3.2 Performance of Conceptual Search

IDOL's conceptual query mechanism allows a seemingly simple query expression to be evaluated in complex ways; as well as the matching of the basic terms within documents using patented weighting algorithms, it is able to develop the terms to "read between the lines" and determine conceptual matches that legacy search engines would be unable to locate.

However, IDOL is able to perform these evaluations with surprisingly little overhead above the equivalent keyword query. The reasons for this are two-fold. Firstly, the majority of the work in the calculation and initialization of the conceptual matching is done at index time, as opposed to query time; the documents are analyzed while the data is being processed to form a statistical "pool" from which queries can draw key conceptual information, as well as an overlying Bayesian network in which apparently unrelated pieces of information are automatically linked via dynamic probabilities. The second reason is that the document-matching algorithm itself within IDOL uses widespread "short-circuiting" and iterative calculation to ensure that it only performs exactly as much calculation as is required. In essence, the key conceptual information is already available before the query has even started, and once it does begin, it feeds directly from the statistical core to load the information. The uniqueness of the query then forces the only truly complex step, a one-off calculation in which combination algorithms arrive at the most relevant set of documents to the query. These can then be returned without the need for looping through every potential match.

1.4 Avoiding Black Box Pitfalls

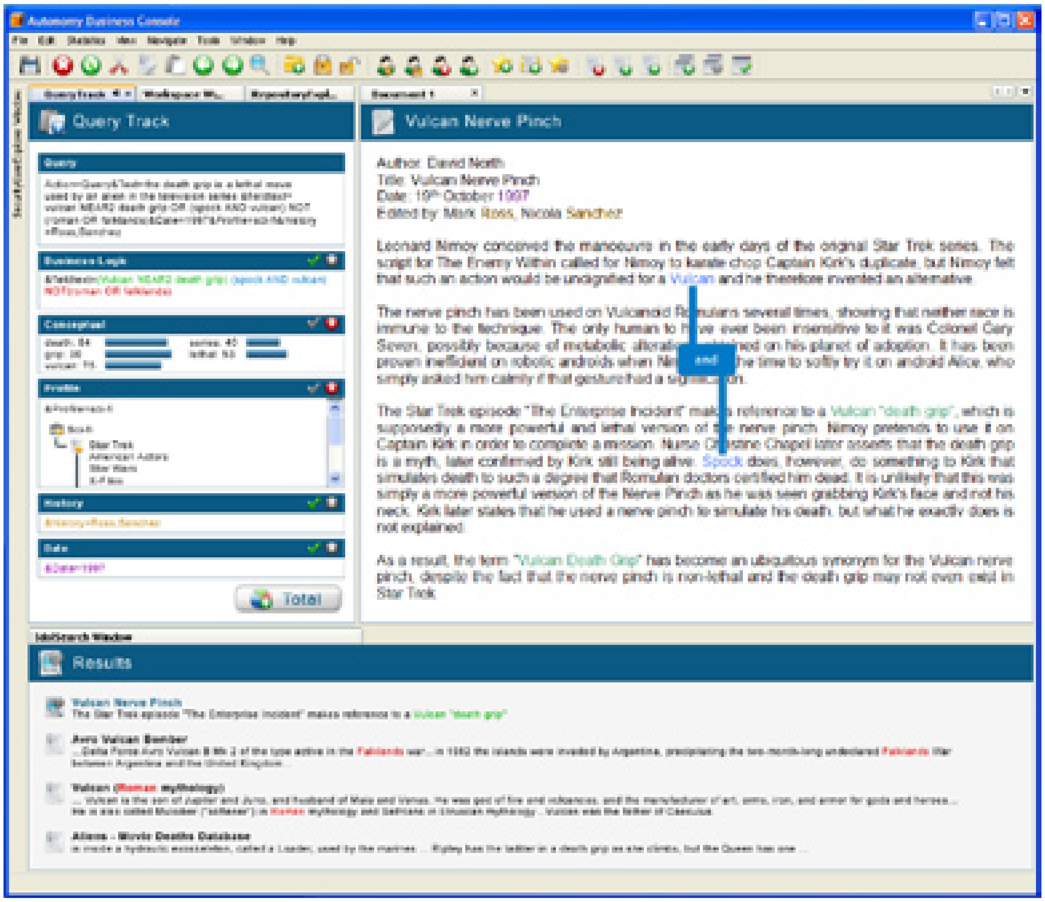
Some systems only operate “black boxes,” mistakenly believing that their technology can always provide the best answers with no tuning required. However, this idea belies a naïve understanding of enterprise demands, for not even the best automated systems can possibly anticipate the unique needs of each enterprise. These “black boxes” offer only a few, if any, relevancy tuning options and do not reveal to the user how the results were generated.

Autonomy’s technology, on the other hand, provides the best of both worlds—automatically retrieving the most accurate results using its conceptual understanding of content, and also offering the flexibility to modify the relevancy algorithm if needed. The computational process is fully transparent to the administrator. Autonomy reveals the basis for its determinations through easily understood representations such as dominant terms and idea distances.

Application administrators are provided with a full workbench to control and tune the relevancy of search results. Some unique advantages offered by Autonomy include:

- *WYSIWYG (What You See Is What You Get) user interface*
- *The ability to manipulate the weight of virtually every field (e.g. title, author); many operators are also available to alter relevancy*
- *Access to extensive range of pertinent information, including common queries, misspellings, query types—all presented in friendly visuals*
- *Full support for business modulated result “sponsoring” or “placement.” For example, a business user can elect to promote a certain result, set of results or object (such as an advertisement) to a defined position within a result set in response to a given query or input. If a user queries for “yellow Prius,” the administrator can define a rule to return the same set of results as a query for “gold Prius,” with a link to the advantages of hybrid cars being the first on the returned list.*
- *Autonomy Collaborative Classifier (ACC) module, which creates a workflow in which the subject matter experts and knowledge engineers, as identified by the organization, collaborate in real-time to create, modify, distribute, and manage taxonomy. As the classifications are created and managed by the people who actually use them, information is organized in ways that are specific and germane to the organization.*
- *Protection of user privacy by respecting entitlement rights and separating the administrator role from the “super-user,” thereby ensuring that the administrator will be restricted from information they are not privileged to view*

Autonomy enables an entire range of information processing options, both manual and automatic. With Autonomy it’s never an “either/or” choice. Autonomy supports legacy methods including keyword search and Boolean expressions, and existing thesauri can also be incorporated. The system can be configured to support as much or as little manual involvement as necessary, ensuring that Autonomy is not a “black box” where the running of the technology cannot be seen or adapted by administrators.



Query

Action=Query&Text=the death grip is a lethal move used by an alien in the television series &fieldtext=vulcan NEAR2 death grip OR (spock AND vulcan) NOT (roman OR falklands)&Date=1997&Profile=scifi&History=Ross,Sanchez

Business Logic

&fieldtext=(Vulcan NEAR2 death grip) (spock AND vulcan) NOT(roman OR falklands)

Conceptual

death: 84 series: 40
grip: 36 lethal: 53
vulcan: 75

Profile

&Profile=sci-fi
Sci-fi
Star Trek
American Actors
Star Wars
X-Files

History

&history=Ross,Sanchez

Date

&Date=1997

Vulcan Nerve Pinch

Author: David North
Title: Vulcan Nerve Pinch
Date: 10th October 1997
Edited by: Mark Ross, Nicola Sanchez

Leonard Nimoy conceived the manoeuvre in the early days of the original Star Trek series. The script for The Enemy Within called for Nimoy to karate chop Captain Kirk's duplicate, but Nimoy felt that such an action would be undignified for a Vulcan and he therefore invented an alternative.

The nerve pinch has been used on Vulcanoid Romulans several times, showing that neither race is immune to the technique. The only human to have ever been insensitive to it was Colonel Gary Seven, possibly because of metabolic alterations obtained on his planet of adoption. It has been proven inefficient on robotic androids when Ni[illegible] the time to softly try it on android Alice, who simply asked him calmly if that gesture had a signification.

The Star Trek episode "The Enterprise Incident" makes reference to a Vulcan "death grip", which is supposedly a more powerful and lethal version of the nerve pinch. Nimoy pretends to use it on Captain Kirk in order to complete a mission. Nurse Christine Chapel later asserts that the death grip is a myth, later confirmed by Kirk still being alive. Spock does, however, do something to Kirk that simulates death to such a degree that Romulan doctors certified him dead. It is unlikely that this was simply a more powerful version of the Nerve Pinch as he was seen grabbing Kirk's face and not his neck. Kirk later states that he used a nerve pinch to simulate his death, but what he exactly does is not explained.

As a result, the term "Vulcan Death Grip" has become an ubiquitous synonym for the Vulcan nerve pinch, despite the fact that the nerve pinch is non-lethal and the death grip may not even exist in Star Trek.

Results

Vulcan Nerve Pinch
The Star Trek episode "The Enterprise Incident" makes reference to a Vulcan "death grip"

Avro Vulcan Bomber
...Delta Force Avro Vulcan B Mk 2 of the type active in the Falklands war ... in 1982 the islands were invaded by Argentina, precipitating the two-month-long undeclared Falklands War between Argentina and the United Kingdom...

Vulcan (Roman mythology)
... Vulcan is the son of Jupiter and Juno, and husband of Maia and Venus. He was god of fire and volcanoes, and the manufacturer of art, arms, iron, and armor for gods and heroes... He is also called Mulciber ("softener") in Roman mythology and Sethlans in Etruscan mythology ... Vulcan was the father of Caeculus.

Aliens - Movie Deaths Database
is inside a hydraulic exoskeleton, called a Loader, used by the marines ... Ripley has the ladder in a death grip as she climbs, but the Queen has one ...

Autonomy Business Console fully reveals how results are generated from a query, including weighting of conceptual terms used, relevant user profile information and metadata

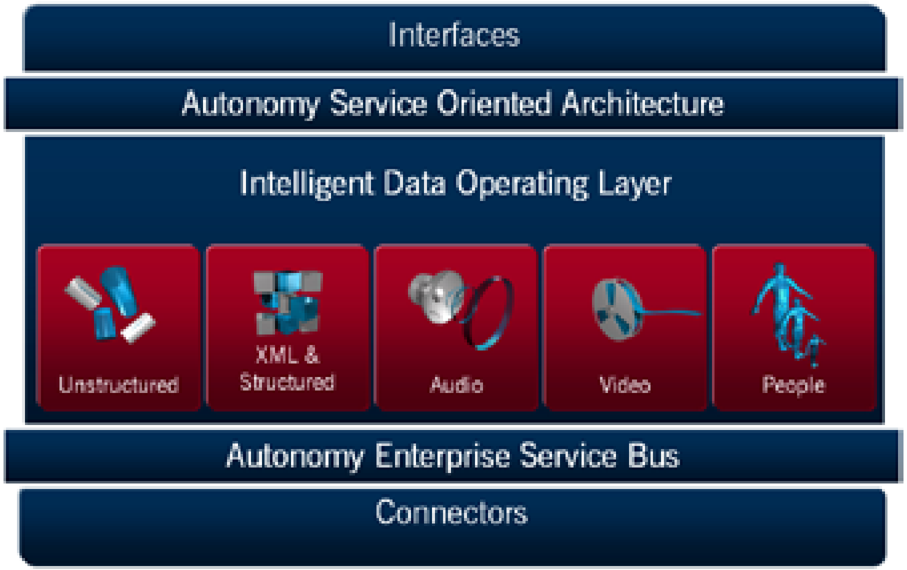
1.5 Avoiding Complexity

Autonomy offers the market's most powerful technology, wrapped in a simple interface. Featuring over 500 content operations and currently used by the world's largest and most mission-critical systems, the core of Autonomy's technology is as complex as its outer layer is simple. The system can be configured and administered by virtue of single point GUI-based interfaces with drag-and-drop functionality. Through an intuitive web interface, administrators can easily control and monitor all of Autonomy's modules and services, whether they are running locally or remotely.

2. Intelligent Data Operating Layer (IDOL) Server™

Autonomy's technology automates the management, processing, and delivery of structured and unstructured information from disparate internal and external sources. It integrates with all known legacy systems, eliminating the need for organizations to patch together multiple systems and support their distinct components.

2.1 Architectural overview


Interfaces
Autonomy Service Oriented Architecture
Intelligent Data Operating Layer
Unstructured
XML & Structured
Audio
Video
People
Autonomy Enterprise Service Bus
Connectors

2.1.1 Service Oriented Architecture (SOA)

Autonomy is built on an SOA model, maintaining open XML standards, and enabling easy integration to third party applications by design. Autonomy's infrastructure product is heavily predicated on the design principles of reuse, granularity, modularity, componentization, interoperability and performance. Autonomy provides a large number of meaning-based computing functions as a service, and all its modules are discoverable services that interface using Simple Object Access Protocol (SOAP) as a standard. Autonomy uses HTTP/XML technologies that can be used in any SOA platform.

Given the varying and potentially unpredictable nature of data communication and transport scenarios, Autonomy as an enterprise application has developed and utilizes the Autonomy Enterprise Messaging Bus (AEMB) to effectively manage all messaging across the Autonomy infrastructure, as well as delivering and building SOA. The AEMB is based on TCP/IP and maximizes available computational resources.

2.1.2 Connectors

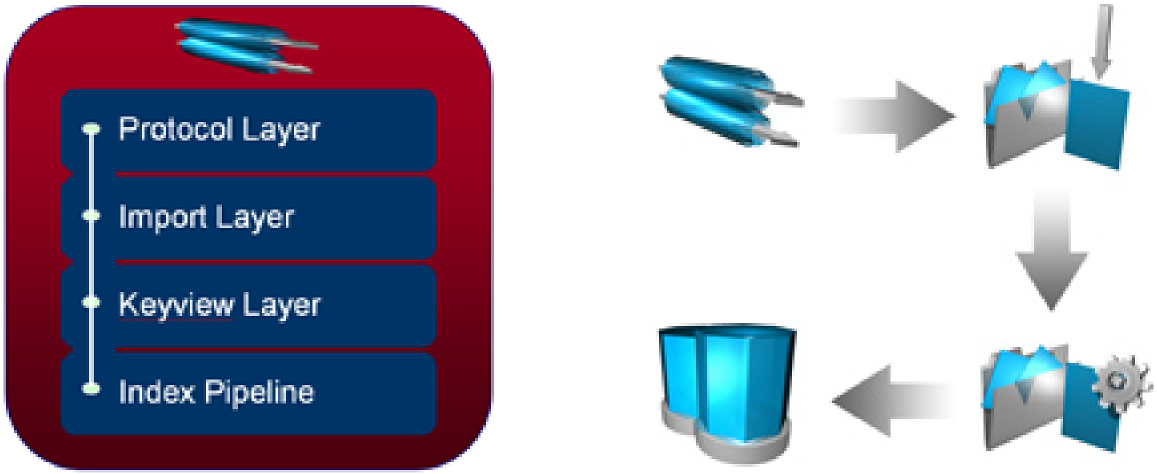
Connectors are responsible for extracting content from any type of local or remote repository, importing the data into an IDX or an XML file, and then indexing the data into IDOL for information processing. There are connectors for different content repositories such as Notes, Exchange, Documentum, SharePoint, Oracle, and over 400 others. Each connector extracts the entire content of a given file to ensure complete accuracy.

Powered by KeyView technology, the import module is fully capable of extracting all metadata and text from over 1,000 data formats, enabling the IDOL Server to provide a 360 degree view of the knowledge base. The KeyView Filter gives an application the ability to simultaneously filter multiple documents on a variety of platforms. The importing process is highly customizable for ingesting only the enterprise's desired data, with the import module accepting over 300 parameters for filtering out irrelevant content. For example, connectors can immediately reject 42 million known files in the NIST list as part of the pruning process, an especially valuable capability in eDiscovery cases. Users can manipulate the configuration file and create special fields and values, as well as perform real-time operations on the extracted text during data ingestion. Emails and their associated attachments that may be located separately can be linked together during the importing process, thereby preserving parent-child relationship. Moreover, users can plug in any number of modules (e.g. categorization, eduction) in the connector layer to further structure the ingested data during the importing process. Other benefits that Autonomy connectors provide include:

- *Comprehensive toolkit of parameterized filter operations, e.g. extract title*
- *Create logical fields*
- *Filtering rules: Min/Max of document*
- *Reference Manipulation*
- *Rendered HTML Copy*
- *Breaking Sections/Anchors*
- *Intelligently generated titles (prevents repetition)*
- *Intelligent extraction of summaries*

As an extension to the Connector Layer, Autonomy provides Legacy Compatibility Module (LCM) to enable organizations to deploy the IDOL Server, displacing existing legacy systems while maintaining existing system workflow. For instance, all IDOL output starts as XML and can be repurposed extensively through the use of templates and style sheets.

Following the importing process, the connector indexes the IDX or XML file into IDOL at a powerful rate of over 110 gigabytes per hour. All connectors keep an audit of the aggregated files, security entitlement recording modification, deletions and completion points, and their signaling mechanisms allow automatic data synchronization between Autonomy's infrastructure and the original data source. As each step in the importing and indexing process is done in parallel, IDOL maintains superior performance with immediate commit time.



The connector layer extracts the content from repositories, imports it into an IDX or XML file, and indexes it into IDOL Server

2.1.3 IDOL's Key Functions

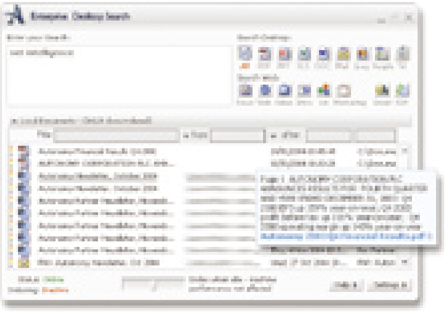
At the heart of Autonomy's infrastructure software lies IDOL. As the information processing layer, IDOL forms a conceptual and contextual understanding of all content in an enterprise, automatically analyzing any piece of information from over 1,000 different content formats and even people as data sets. Over 500 operations can be performed on digital content by IDOL, of which the following is a small sampling.

- ***Automatic Hyperlinking:*** *IDOL allows manual and fully automatic linking between related pieces of information regardless of format. These link to contextually similar content and can be used to recommend related articles, documents, affinity products or services, or concepts within voice and video mail.*
- ***Automatic Summarization:*** *Given a piece of content, IDOL returns a summary of the information containing the most salient concepts. Moreover, summaries can be generated that relate to the context of the original inquiry – allowing the most applicable dynamic summary to be provided in the results of a given query.*
- ***Automatic Clustering:*** *IDOL can take a large repository of data and automatically partition it so that similar information, even of varying formats, is clustered together. Each cluster represents a concept area within the knowledge base, making it easier for organizations to identify inherent themes.*
- ***Cluster Visualization:*** *IDOL provides three intuitive Java-based user interfaces (spectrograph, 2D Map, 3D Map) to make clusters visible and key trends easily identifiable over a specified time interval.*
- ***Automatic Categorization:*** *Users can derive precise categories through concepts found within unstructured content, ensuring that all data is classified in the correct context with the utmost accuracy. Existing legacy taxonomies can either be maintained or enriched with contextual understanding.*
- ***Eduction:*** *IDOL effectively identifies the relationships between documents that are stored on servers in different geographic locations, managed in a variety of business applications, and that are created by internal and external sources. It creates structure out of unstructured content by automatically propagating key content into a database. This tool can even identify patterns that entail actionable data. For instance, an e-mail that mentions "Bill Gates" could be flagged and be automatically forwarded to the CEO.*
- ***Interlinking:*** *IDOL automatically identifies sub-areas of the document for which there are relevant links and inserts hyperlinks into the document.*
- ***Automatic Taxonomy Generation:*** *By understanding the information in an enterprise, IDOL automatically generates taxonomies and instantly organizes the data into a familiar child/parent taxonomical structure. IDOL's data agnostic classification capabilities enable organizations to build taxonomies based on any type of data, unstructured or structured. Legacy taxonomies and manual processing are also supported.*
- ***Personalized Agents and Alerting:*** *IDOL allows users to set up Agents to monitor information 24x7 on specific topics from a range of data sources. An Agent represents a user's persistent interests and can be defined or trained explicitly, or, more powerfully, by example, simply by being shown a document, video, or other content that matches a user's interests. The Agent will then monitor information and instantly alert the user to anything new that matches the training.*
- ***Collaboration and Expertise Networks:*** *As IDOL can automatically understand individuals' interests based on browsing, content consumption, and/or content contribution, it can also automatically match users with common explicit interest agents or similar implicit profiles. This engenders a working environment that promotes collaboration and community. Furthermore, this profiling technology facilitates the recognition of highly focused experts (sorted by location, department, availability, etc) within the community and reduces the duplication of effort through teamwork.*
- ***Social Search:*** *IDOL provides a rich participatory platform for capturing unstructured, tacit knowledge created from Web 2.0, connecting people to related communities, projects and processes. Users can contribute and share tags, comments and votes, as well as share search folders and knowledge with their colleagues. All user activity, both explicit and implicit, can be incorporated to influence relevancy calculation.*
- ***Intent-Based Ranking:*** *IDOL determines a user's intent when querying and delivers highly targeted results based on the individual's profile and contextual factors*

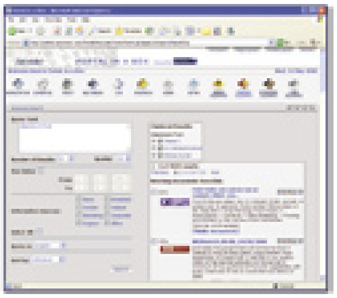
- ***Dynamic Query Building:*** *IDOL automatically generates query suggestions that best fit each user's intentions as the user types his/her query. This personalization is based on holistic user context and profile, making search more efficient and effective for the user.*
- ***Automatic Query Guidance:*** *Requiring no manual training, IDOL automatically provides keyword-oriented users with the suggestions they need to find the most relevant information. The suggestions are created dynamically based on the query and the time of query.*
- ***Automatic Query Federation:*** *IDOL delivers true query efficiency by federating a given query to only a subsystem or one specific search engine based on the user's intent, given the user's query pattern and the information contained within the query itself. Therefore it avoids bombarding all of the systems with the same query, and simply selects the one most likely to return the specialist result.*
- ***Implicit Query:*** *Used by the IDOL Enterprise Desktop Search module, this technology forms an understanding of the information on a user's screen and proactively links users to related information on their secured corporate network, on the web, or on their local machine. For instance, while the user is drafting an e-mail or a PowerPoint presentation, related hyperlinks to the employee's formulated content appear at the click of a button.*
- ***Directed Navigation:*** *Organizations can integrate valuable information from structured databases with unstructured content. For example, users can search and find products quickly and easily when information from structured fields (e.g. product name, size, manufacturer) is combined with unstructured content (e.g. product description).*
- ***Scene Detection Analysis:*** *Using superior neural networking techniques, IDOL can assist users of CCTV systems with the detection of important activities. This advanced solution can flag potential threats, illegal actions or other user-defined scenarios. It performs high quality image analysis in any environment, including feature extraction, blob analysis, and object tracking. IDOL integrates with any new or existing CCTV installation.*
- ***Web 2.0:*** *IDOL's conceptual understanding can be used to monitor blogs and circumvent legal issues by alerting users to content that discloses sensitive material. It can also facilitate navigating through the chaotic wiki environment by automatically generating hyperlinks to relevant material. IDOL can also be used in conjunction with user-generated taxonomies, or folksonomies, to produce a more reliable, consistent, and scalable solution.*
- ***Link Gap Analysis:*** *IDOL combines spider and ant technologies to ensures that links are always kept up-to-date with germane information and employees do not lose productivity by being led to websites that are no longer relevant or functional.*
- ***Business Console:*** *IDOL provides a robust set of widgets, wizards and tools to simplify the management of IDOL functions. IT administrators and managers can react quickly to changing usage patterns and provide highly targeted results to end-users without any programming.*

2.1.4 Interface

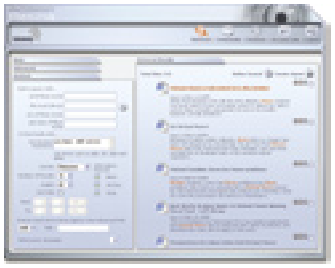
The interface layer of the IDOL architecture surfaces the powerful features of the IDOL Server in a front-end application. It can be experienced through Autonomy's own web applications such as Portal-in-a-Box, Retina, and IDOL Enterprise Desktop Search, or through existing business applications, CRM systems, and OEM partner applications.



Enterprise Desktop Search



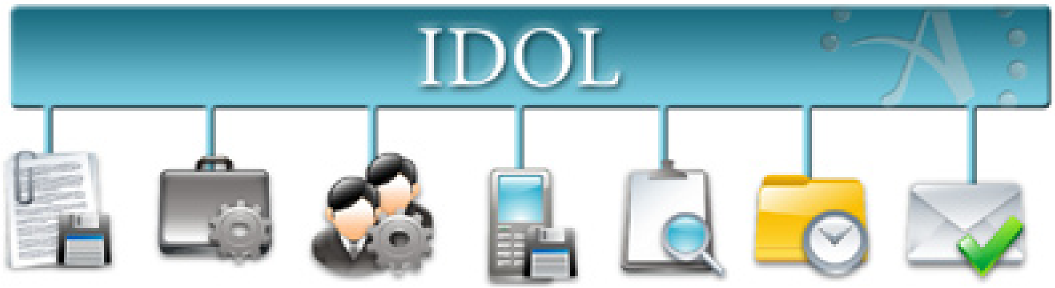
Autonomy Portal-in-a-Box™



Autonomy Retina™

3. Benefits of Autonomy's Technology

Autonomy offers a single platform (IDOL) for understanding and processing all digital information. Only IDOL can perform keyword and conceptual search, speech analytics, video search, email and Instant Messaging (IM) search, and categorization—all on the same platform. Alternative approaches require stitching together of different technologies with potentially conflicting formats. Consequently, they compromise stability, pose maintenance issues and may necessitate excessive technical support along the upgrade path. IDOL, as a single platform, eschews these problems, and its superior benefits are realized without any concessions in performance as each component of IDOL is best of breed.



IDOL aggregates content from different sources and makes information available at once to many applications

IDOL is data agnostic, language independent and fully scalable. It does not require complex programming, extensive integration, business rules or middleware. It also does not require information to be manually tagged, linked or categorized. IDOL-compliant applications are immediately compatible through their common understanding of digital information. All this is possible because IDOL understands information in a manner similar to humans—it directly relates concepts "read" from the portions of digital content that humans can process, not from rules that are dependent on synthetic tags.

IDOL also enables organizations to benefit from automation without sacrificing manual control. If circumstances should arise where users would want to apply traditional manual techniques to an information processing operation, IDOL supports this choice. This complementary approach allows automatic processing to be moderated by various manual overrides, offering the best of both worlds and never exacting an "either/or" choice.

Autonomy offers excellence in many critical facets, including accuracy, performance and security.

3.1 Performance & Scalability

The management of structured and unstructured content requires a platform that can meet the most rigorous performance requirements, and be easily resized commensurate to business needs. IDOL scales to support the largest enterprise-wide and portal deployments in the world, with presence in virtually every vertical market. Since IDOL's scalability is based on its modular architecture, it can handle massive amounts of data on commodity dual-CPU servers. For instance, only a few hundred entry-level enterprise machines are required to support ChoicePoint's 10-billion record footprint. By comparison, a competitor uses 150,000 machines to handle the same amount of data.

A single IDOL instance can:

- *Support over 470 million documents on 64-bit platforms*
- *Accurately index in excess of 110 GB/hour, with guaranteed index commit times (i.e. how fast an asset can be queried after it is indexed) of sub 5ms*
- *Execute over 2,600 queries per second, with sub-second response times on a single machine with 2 CPUs when used against 70 million pieces of content, while querying the entire index for relevant information*
- *Support hundreds of thousands of enterprise users or millions of web users, accessing hundreds of terabytes of data*
- *Save storage space with an overall footprint of less than 15% of the original file size*

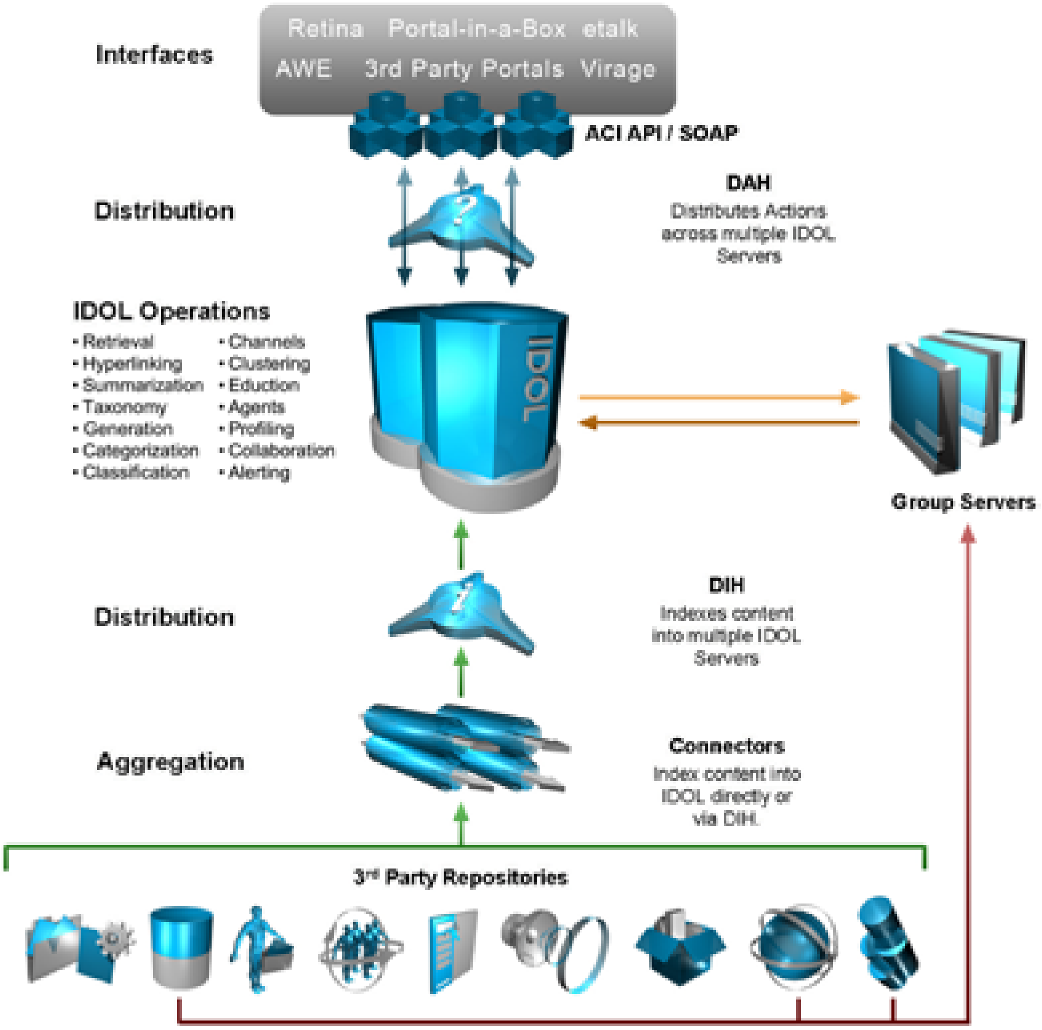
Real-world examples show IDOL satisfies high scalability requirements on various hardware/operating system platforms.

- *Bloomberg uses Autonomy across all 140 million, multi-lingual news articles, which enables it to send 1.2 billion emails and 1.26 million trading alerts per day*
- *Citigroup uses Autonomy to search across 1.3 petabytes of data*
- *Over 120,000 users within GlaxoSmithKline use IDOL to access 200 terabytes of data in multiple formats*
- *BBC Online uses IDOL to archive millions of documents in over 33 languages, and is used by 20 million people per day*

This enhanced scalability results in hardware cost-savings as well as the ability to address larger volumes of content. Though IDOL scales extremely well on commodity servers, its flexible architecture can take full advantage of massive parallelism, SMP processing capabilities, 64-bit environments (e.g. Intel Itanium 64-bit architecture), software platforms (e.g. Solaris 10, Linux 64, Win64), distributed server farms, and all common forms of external disk arrays (e.g. NAS, SAN) to further improve performance. This flexibility extends to being able to leverage one or a combination of these different environments.

IDOL delivers linear scalability through a multi-threaded, multi-instance approach with load-balancing to intelligently distribute the indexing and query workload, allowing additional machines, locations and indexes to appear as one. In addition, IDOL distribution components are uniquely "geo-efficient," permitting fault-tolerant national, trans-national and trans-global architecture assembly with ultimate flexibility in component placement. IDOL's distributed components—Distributed Index Handler (DIH) and Distributed Action Handler (DAH)—form a coherent layer with IDOL. DIH and DAH components can be placed inline or within fully nested topologies.

Content from various repositories is aggregated by connectors and then indexed into the IDOL Server or, for dissemination across multiple IDOL Servers, through the DIH. The DIH can efficiently split and index copious quantities of data into multiple IDOL Server instances, optimizing performance by batching data, replicating all index commands and invoking dynamic load distribution. The DIH can perform data-dependent operations, such as distributing the content by date, which allows for more efficient querying. Performance is augmented by the DAH, a distribution server that allows the user to distribute action commands, such as querying, to IDOL Servers. Multiple copies of IDOL Servers to which the DAH propagates actions further ensure uninterrupted service in the event of server failure. For flexibility, both the DAH and the DIH can be configured to run in mirroring mode (IDOL Servers are exact copies of each other) and non-mirroring mode (each IDOL Server is configured differently and contains different data). In addition, the Distributed Service Handler (DiSH) component allows effective auditing, monitoring and alerting of all other Autonomy components.



This linear scalability allows administrators to add servers as required, using any server platform in a homogeneous or heterogeneous environment. There is no theoretical limit to the number of servers and modules that can be interconnected in this way, and no central bottleneck, allowing scaling to support many thousands of queries per second and millions of users.

Moreover, Autonomy's modules can support RAID and EMC Symmetrix storage configurations, with extended caching and intelligent disk usage. Autonomy is extensively used in conjunction with the latest developments in SAN and NAS storage technologies, maximizing Autonomy's existing high alignment with enterprise technology trends.

3.1.1 Instruction-level parallelism

IDOL programmatically expresses itself as an expanding collection of operations. These operations can and are executed in serial pipeline form yet the inherent logic of simultaneously processing disparate forms of unstructured, semi-structured and structured data requires a high degree of parallelism. Not only does IDOL need to ingest multiple streams and types of data, it must also provide a real-time answer or decision against that data as it is indexed rather than force the user to wait an arbitrary period until serially accessed resource becomes available.

As a consequence IDOL has been designed with instruction-level parallelism (ILP) as the core of its process and operation model. ILP by definition is limited by the serial instruction model of scalar processors and thus Autonomy has been an extremely conscious early adopter of all forms of parallel architecture from multi-CPU, hyper-threading and now single die multi-core processing.

The engine's default process model is multi-threaded (with a configurable number of threads being utilized). IDOL operations can either be grouped by class, with indexing and querying performed by separate threads or for n-core models a single operation can be "atomized" into multiple threads. Concurrent querying and indexing is the default with no requirement whatsoever for "locking" any part of the indexes while querying takes place. All major multi-core manufactures are supported, including Intel, AMD and the latest Niagara offerings from Sun Microsystems.

Classic scalar models that rely on Moore's predicted doubling of transistor density over 18 months intervals have already demonstrated wire and memory access latencies in addition to heat sealing's. As a result, hardware manufactures such as Intel have declared multi-core strategies as key to crossing the consumer "teraflop" threshold and aim to produce n-core 32 billion transistor die within the next 10 years. Autonomy is actively pursuing a Tera computing R&D simulation program in anticipation of increasing transistor and core density and the declared aim of such manufacturers. Autonomy is currently performing "coalition" simulations of split thread IDOL operations against n-core "battalion" processor units that blend general-purpose cores with more specialist cores such as those dedicated to signal processing. These blended core units are predicted to be the first consumer teraflop chips. Autonomy is developing process thread models that dynamically co-opt different core types to act in "coalition" to perform the simultaneous deconstruction and analysis of unstructured sources such as video that commingle visual and auditory attributes.

3.2 Security

The world's largest and most secure intelligence organizations have deployed Autonomy to safeguard their most sensitive information assets. Autonomy provides all aspects of security management, including front-end user authentication, back-end entitlement checking and secure encrypted communication between the IDOL Server and its client applications with 128-bit Block Tiny Encryption Algorithm (BTEA).

There are three security models on offer within the market today – "Unmapped," "Cached" and "Mapped" security, with Autonomy being the only vendor to provide fully mapped security. Autonomy's mapped security model is the only empirically proven index security model that scales in the enterprise.

The challenge to any index-based technology is best illustrated by an example. If a user queries a common set of terms such as the name of the enterprise itself (i.e. "bigcorp inc"), the index will need to compare perhaps millions of legitimate matches to the user's own privileges. If the user has a common security setting and thus has an average level of permissions, he/she will only legitimately be able to see a fraction of all the potential hits.

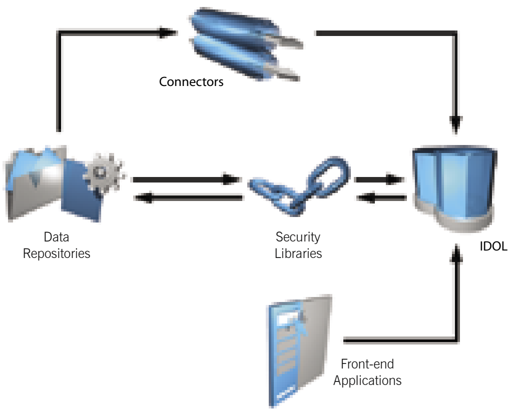
Legacy Security Models

In unmapped security, for every potential match to a given query, a call is made via the native repository's API to ascertain the access privileges for that particular document. A single query consequently floods the network and bombards the originating repository with document privilege requests as the retrieval system attempts to assemble a relevant results list from thousands of candidate hits. This method presents significant performance and scalability problems. A variant of the unmapped technique, cached security only marginally relieves the scalability issue by storing results for queries it has seen. Thus when a user repeats a query the result set can be retrieved from the cache and not trigger a network mediated request. However, this is still reliant on calling out across the network directly to the repository for each new query it sees. In addition it also misses potential results, as the results sets stored within its memory does not dynamically update new information.

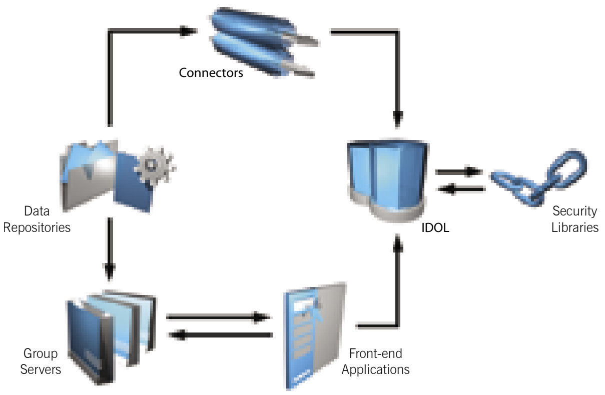
Autonomy's Mapped Security Model

In Autonomy's unique implementation of mapped security, IDOL maps the underlying security model in the form of ACL, group, role, protective markings, etc. from all of the underlying repositories directly into the kernel of the engine itself. As a consequence of the unique engineering solution, IDOL does not need to send any requests across the network to the data stores when building up a results list. What the user is allowed to see is assessed "inline" within the IDOL kernel at speeds that exceed the response times of the native repository. This response is so fast it can be contractually guaranteed at any scaling level.

Unlike other techniques, the model is never out-of-date as the transitional signaling mechanism within the connector layer informs IDOL in real-time of any updates or changes to permissions within the underlying content. In efficient signaling, the application identifies the user by sending user details along with the user's query. The server then calls the appropriate security libraries for each type of data repository at query time. The plug-in does *not* connect to the original repository but looks inside IDOL's structured and encrypted fields, which contain the ACLs.



Unmapped Security



Autonomy's Unique Mapped Security

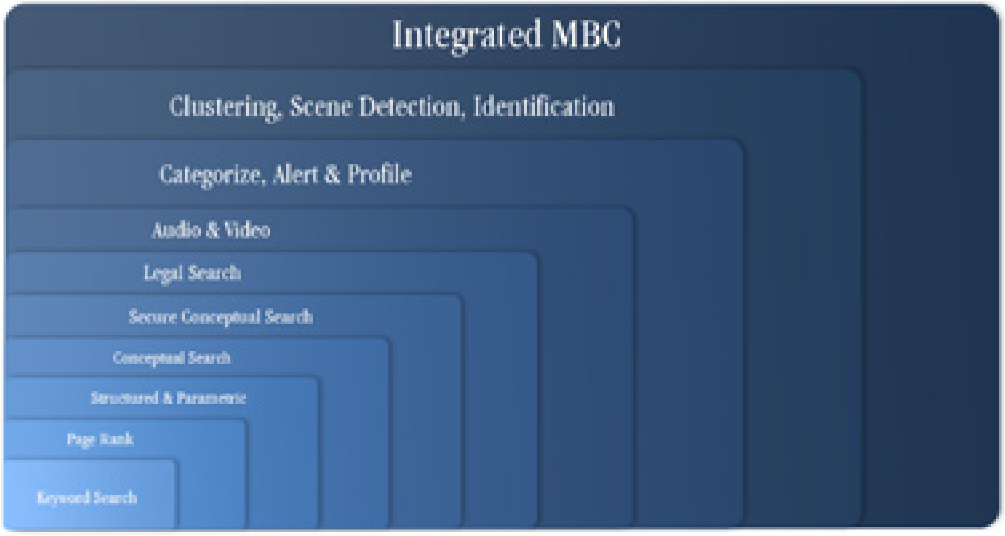
Autonomy also supports role-based segmentation of data access and metadata-based access control. The use of bitmapped entitlement fields is specific to Autonomy and allows entitlement information to be held and accessed extremely efficiently. This implementation is useful where a great number of users and documents need to be indexed and where access rights vary greatly.

Ability to Choose

Autonomy recommends mapped security but also offers the choice between mapped, unmapped and a hybrid solution. In addition, where an established solution is already in place, Autonomy servers can be integrated with the existing setup, providing full security compliance.

3.3 Accuracy

Autonomy's pioneering Meaning-Based Computing platform provides advanced and accurate enterprise search by encompassing and improving upon existing search mechanisms.



3.3.1 Legacy Search

Autonomy fully supports keyword and Boolean-based searches, as it does all legacy searches. Autonomy recognizes the value of keyword-based searches in environments with a large corpus of data, or when the user is looking for specific information and can tolerate the incomplete retrieval results endemic to this technology. However, Autonomy also recognizes their limitations. As the most basic of search technologies, keyword search only finds content that contains the words explicitly queried by the user, completely ignoring content that lack those specified keywords despite the presence of similar concepts, ideas or themes (see Section 4.1). Within an enterprise environment, however, where productivity is dependent upon the comprehensive and efficient retrieval of relevant content, keyword search often proves inadequate.

Equally ineffective in the enterprise world is page rank search, which tries to improve upon keyword search by ordering search results by relevancy. However, this relevancy is determined by the number of other web pages that link to a given page—a successful strategy in the widely linked world wide web, but not often so in the intranet, where manual hyperlinks are rare and unreliable.

Parametric search improves upon keyword search by allowing users to leverage the structure associated with business content. The user can search items by certain characteristics and their concomitant restrictions. For instance, when the customer specifies that the color of a shirt must be red, only the sizes available for red shirts would be available. Unfortunately, 80% of enterprise content is unstructured and therefore ill-suited for this technology. This type of search engine also does not understand the underlying concept motivating the search. However, recognizing its value in well-defined scenarios, Autonomy's Directed Navigation module enables users to navigate n-dimensional structured data sets and search across parametric data.

3.3.2 Parametric Retrieval / Directed Navigation

Directed Navigation offers a way for organizations to integrate valuable information from structured databases with unstructured content. Combining information found within structured fields such as product name, size or manufacturer, with unstructured content such as product description, allows users to search and find products quickly and easily.

For example, in an ecommerce environment, a customer looking for a specific wine can use a series of drop-down options, such as color, grape and year, to narrow down results and guide them to the product of their choice. Choosing “merlot” as a grape will immediately rule out all white wines. But what if the user then asked for a “strong, fruity Australian wine, under $10,” only to find that there were no Australian merlots in stock fitting the criteria? With other systems, the search would end there, but Autonomy uses its patented Bayesian statistical procedures to automatically identify other wines it feels will fit the requirements. For example, IDOL will already have statistically determined that the many Chilean merlots in stock not only fit the description but often fall within the specified price range, and would make this suggestion instead.

3.3.3 Advanced Conceptual Search

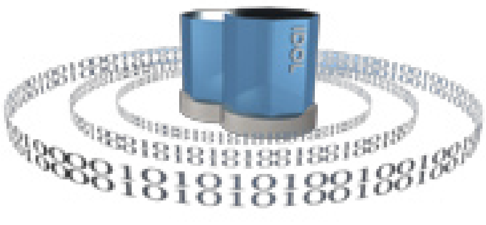
By looking for the idea defined by the given words, and not necessarily the words themselves, conceptual search avoids inaccuracy and retrieves the most relevant content, not merely the most popular. Parametric search is also combined with conceptual search, so that when the user enters a conceptual description of a clothing item (e.g. “loose red long-sleeve shirt that looks vintage”), IDOL uses both the parametric fields and conceptual matches in the specified description to cull eligible results. Importantly, IDOL’s mapped security ensures that all search results are vetted according to the user’s security rights.

3.3.4 Legal Search

IDOL furnishes the most complete set of relevant results through its seamless trifecta of keyword search, secure conceptual search, and its superior indexing engine. Enabled by its high-performing modular architecture and a powerful fetching rate of 4GB/hour per single instance of server, IDOL extracts the entire content of a file. Other technology vendors use partial indexing in order to mimic fast performance at the expense of accuracy. IDOL stores all of the content of a file along with its metadata and concepts, ensuring rapid and precise searching. During the search itself, IDOL presents a complete and uncompromised set of results (full discovery) with no premature cessation (jump-out); some search engines stop looking across an index as soon as enough results have been amassed, resulting in an arbitrary and incomplete set of documents. In today’s environment, where Gartner Research estimates that at least 75% of all global companies will be involved in legal or regulatory action that requires legal discovery,[1] a technology that provides the most complete set of results across all formats of information—including email, fax, audio, video and paper—is a business necessity.

3.3.4.1 Echo

IDOL Echo is a module that allows fully auditable and accountable monitoring of information use. Echo detection enables the enterprise to forensically account, track, and trace each piece of data that enters, leaves, is born or dies within the organization. Echo follows each content’s path and history and can report on who or what the asset has influenced. It can not only follow a traffic pattern such as the path of an email attachment or voice mail (i.e. the what and when of who read, heard, forwarded and retained it), but also detect the influence of its content (i.e. who in the enterprise has taken, re-purposed, been persuaded or even plagiarized what they have heard or seen).



[1] E-Discovery: What You Don’t Know Will Hurt You, Debra Logan, Gartner, October 24, 2006

The ripple effect of an idea can be measured as it jumps from mail to phone conversation to document: person A receives a call from person B who then relays it out of ear shot in the parking lot to person C who then e-mails these ideas to person D.

3.3.5 Rich Media Support

Autonomy provides a complete proprietary platform that is wholly data agnostic and compatible with rich media formats via broadcast, video conferencing, VOIP, CCTV and a multitude of other near-ubiquitous technologies. Legacy methods of searching rich media involve manually sifting through hours of content in order to apply metatags denoting the conceptual orientation of the content. IDOL integrates all rich media assets, automates the key processing operations and enables organizations to manage and manipulate multimedia as easily as text. Retrieval is based on an approach that analyzes the incorporated audio stream together with image recognition technology rather than on the output of a flawed transcript or imprecisely applied data labels.

3.3.5.1 Speech Analytics

Autonomy's speaker-independent technology, based on superior speech processing algorithms, enable live or recorded speech to be manipulated, edited, searched and hyperlinked as freely as text. It also develops a wide range of speech processing technologies, from audio segmentation and identification to real-time speech recognition and understanding from broadcast media and telephony. Every recorded voice transaction, as well as every element within those recordings, becomes easily searchable. By automating formerly manual data mining processes, IDOL represents a major enhancement in many practical areas, such as contact center quality control.

This language independent core technology is customized to the specific language and vocabulary required by the user, using specialist Language Packs, delivering dramatically more reliable recognition for multiple applications. Autonomy supports a virtually unlimited vocabulary size for an unrivalled range of single and multi-byte languages, including those traditionally considered more sensitive and problematic, such as Arabic, Russian and Japanese. New ideas are automatically trained into the system as they are encountered, delivering state-of-the-art accuracy while significantly reducing costs. The technology's extensive vocabulary speech system succeeds for any accent and rapidity of speech. It also provides various adapters to accommodate audio encodings across a gamut of quality and compression schemes, and offers the flexibility of a language model that can be tailored to industry jargon.

In addition, Autonomy includes advanced audio processing capabilities such as speaker and audio segmentation, speaker recognition and classification and audio/text synchronization. These unique capabilities can be combined with other IDOL functionalities to ensure unrivalled accuracy in retrieving and processing video and audio data. By combining complex mathematical models of natural language analysis with IDOL's meaning-based technology, Autonomy provides accurate, content-based retrieval of information. While legacy phonetic approaches merely use keywords and cannot contend with variant expressions of an idea, IDOL understands the language behind the spoken words, even distinguishing homophones and phrases. For instance, IDOL would not be ensnared by the following gaffe that typifies technologies dependent solely on phoneme processing:

"The oil spill will recognize speech"
"The computer will wreck a nice beach"

IDOL combines both phonetic and conceptual methods to disambiguate the limitations inherent in keyword and phoneme matching approaches. Although “recognize speech” and “wreck a nice beach” sounds similar enough to stumble phoneme technology, IDOL would realize that oil spill cannot recognize speech, but rather will wreck a nice beach, using the conceptual relations between words that determine their meaning.

Autonomy uniquely supports conceptual match, so if the audio contained the phrase “University of Cambridge,” the search would include the term “Cambridge University.” Again, legacy phonetic search would fail in this scenario. Autonomy also understands how context can alter the meaning of words. Presented with a news story involving the White House, for instance, Autonomy technology would predict that the word Bush is likely to follow George, and would aggregate information under the assumption that Bush refers to a person and not a plant.

3.3.5.2 Video Analytics

Autonomy performs concurrent encoding and indexing of video and other rich media. During the video encoding process, Autonomy’s advanced capture and analysis technology works in real-time to automatically generate a structured index. Time-synchronized to every encoded copy made, this index enables immediate and precise retrieval of assets. All rich media content is fully searchable, leveraging the power of IDOL to offer functionality such as automatic hyperlinking of related content, implicit and explicit profiling, clustering and alerting. In addition, video content can be automatically cross-referenced with any other forms of information such as PowerPoint presentations, word documents or web pages.

Autonomy offers rich media content management with its advanced archival system that stores, categorizes, manages, retrieves and distributes multimedia content. In addition to indexing and performing analysis, it can automatically clip and segment parts of the video using conceptual intersection. Autonomy also supports the creation of live, interactive webcasts and on-demand archived webcast events. It combines streaming audio and video with slides, documents, Q&As and polls for a full interactive experience, all controlled with a simple interface.

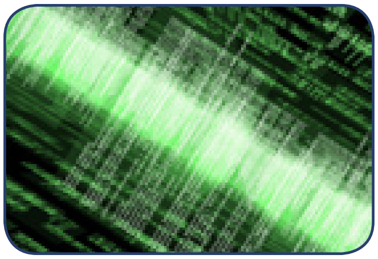
Autonomy’s open architecture and frictionless integration can conform to any workflow and be customized to suit specific workflow needs. Supplemental plug-ins work seamlessly with the main technology to layer additional information such as multi-lingual transcriptions, speaker identification, name extraction, story recognition, facial recognition, and on-screen character recognition. Manual data entry is now significantly reduced by practically eliminating the costly and labor-intensive annotation process traditionally associated with logging video.

Autonomy provides world-leading security monitoring solutions for a full range of environments including offices, airports, sports arenas, government buildings, oil and gas exploration sites and server rooms. Autonomy currently forms an integral component of some of the foremost government bodies and intelligence agencies such as the U.S. Department of Homeland Security.

Capabilities include:

- *Intelligent scene analysis with object tracking*
- *Counting and behavioral analysis*
- *Number/license plate recognition*
- *Container surveillance and management*
- *Electronic point of sale monitoring*
- *Digital/network video recording and transmission*
- *Video forensic analysis*
- *3-D face recognition*
- *Fingerprint analysis*

3.3.6 Advanced Analytics



As described in brief in section 2.1.3 IDOL's Key Functions, retrieval is just one of the 500 functions that IDOL can perform on a data set. Advanced heuristics such as user profiling and intent-based ranking ensure that IDOL orders results not just based on popularity or keyword matches or other legacy algorithms, but tailors the results to the user's purpose and to the specific query, making search more effective and efficient. From classification and categorization to content and people analytics, IDOL's value extends far beyond traditional search and has penetrated virtually every market segment. A few of its applications include:

- *Clustering allows litigators to rapidly prune large batches of conceptually similar content that are irrelevant to the case*
- *A profile generated through a user's interaction with PDA content can then recommend Internet content or news content via email or SMS*
- *Reputation Analysis allows politicians and product vendors alike to measure how they are viewed by the ever-active blogosphere over a given amount of time*
- *At content creation (or indexing) time, IDOL automatically detects whether a specific piece of information already exists in the system and can trigger an alert to the author to the existence of the other files, eliminating duplication of effort and providing the opportunity to aggregate the two sources*

3.4 Automation

By automating processes that were previously performed by costly and tedious labor, Autonomy's technology offers a direct path to substantial bottom-line savings. Cross-referencing of content is automatic as Autonomy's infrastructure identifies related material within the operating layer and determines significant relationships between information using multi-tiered relevancy modeling involving such algorithms as Natural Language Processing (NLP). Processes such as hyperlinking, information clustering, alerting and categorization of content can all be precisely automated with any document or set of documents.

For example, as an employee is reading or writing a corporate document, hyperlinks to contextually similar documents are dynamically generated for immediate viewing. Within a CRM environment, the system can present answers to frequently asked questions and related e-mails to customer service representatives through dynamically generated hyperlinks. Concurrently, automatic clustering provides instant high-level visibility of the knowledge base and identifies information trends and themes for effortless consumption. Automation of myriad processes translates into increased productivity.

3.5 Format Agnosticism

Autonomy is not reliant on any single file or data format. Autonomy's connectors penetrate the information silos in an enterprise by offering deep integration into over 400 repositories, including Documentum, Lotus Notes, Exchange, RDBMS, file servers and more. Autonomy can aggregate more than 1,000 file formats, including voice and video content, into a single index and thereby allow enterprises to make sense of information from the widest range of formats and sources available—from unstructured data (e.g. HTML pages, word processing documents, email, multi-media content) to semi-structured (e.g. XML) and structured data (e.g. Oracle, Lotus Notes, ODBC compliant material).

With its offering of KeyView Software Development Kit (SDK), Autonomy provides the requisite tools to successfully build applications to filter, export and view all unstructured information across all formats. KeyView enables high-fidelity viewing, highlighting and printing capabilities, as well as HTML file format emulation. Users can instantly view files and forgo the tedious installation of proprietary software for every respective format.

3.6 Language Independence

IDOL uses probabilistic modeling to extract meaning from digital content, and forgoes language-dependent parsing or dictionaries to form ideas. Because IDOL treats words merely as abstract symbols of meaning, it is completely language independent. It does not rely on an intimate knowledge of a language's grammatical structure, but rather derives its understanding through the context of the words' occurrence rather than through rigid definition of grammar. This highly mathematical logic yields great accuracy, and performance is further optimized through proprietary stemming algorithms, "sentence breaking" libraries, stoplists and n-grams.

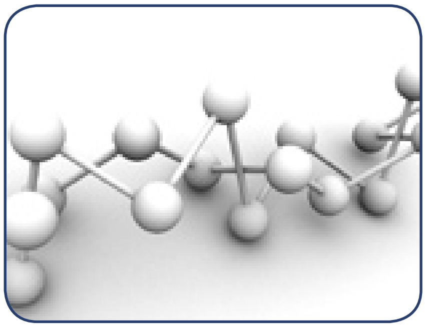
Although IDOL's fundamentals are predicated on a language independent model, it is still capable of using linguistic analysis to parse semantics to an intra-document level. A good example of this at work is the Sentiment Analysis functionality, where IDOL can determine the degree to which a sentiment is positive, negative or neutral for the entire content or a segment of the content. A blogger may have a positive opinion on the iPod, but a negative one on the iPhone, all within the same entry. By extracting information from every file processed, IDOL continually learns positive and negative language structures and concepts. Even out-of-the-box, IDOL's inherent ability to understand data performs highly accurate classification of tonality. In addition to this in-built learning ability, categories of what subjects are considered good or bad can be trained and administrators can apply multiple tagging functions and specific threshold cut-offs to determine the sensitivity of sentiment analysis.

IDOL analyzes units of word and not characters, so it also works well with double byte languages. Autonomy's software supports over 100 languages, including English, German, French, Italian, Chinese and Japanese, and can even be easily configured to auto-detect the language of incoming documents.

3.7 Cross-Lingual Functionality

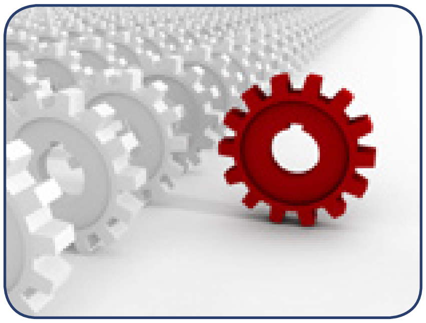
Autonomy's technology allows for cross-lingual search and data management. There is no compromise to the accuracy and concepts extracted regardless of the language used. For example, an employee based in New York and working in English can be provided with relevant documents and information from all over his company's global network, whatever language they are in, based on the similarity of concepts expressed within the documents. The software can also provide translations, if necessary. This language agnostic approach offers a significant benefit to any international business by enabling colleagues separated not only by miles, but by language as well, to collaborate and share knowledge.

3.8 Learning Ability



Due to the unique combination of Bayesian Inference and Shannon's Information Theory at the core of the technology, Autonomy software is able to continuously develop and learn. This learning ability significantly reduces the manual input required by other solutions and translates into large savings in time and money. Where other solutions need to be taught new words, phrases or concepts and shown how to categorize them, Autonomy can automatically deduce the significance of these new units of meaning, add them to relevant categories, and create new categories where necessary.

Autonomy's technology can also learn about its users by dynamically monitoring the content they view, and then deliver relevant content as it is added to the environment.



3.9 Easy Integration

IDOL enables organizations to seamlessly integrate with other systems across intranets, extranets and the Internet. Founded on a technology that is modular by design, Autonomy has developed a flexible infrastructure that allows optional use of the latest Web service standards, including SOAP and Web Services Description Language (WSDL), to enable organizations to build innovative e-business solutions. A cross-platform solution that supports, among others, Windows, Linux and Solaris, Autonomy has ensured easy integration with any environment. Autonomy's core action API uses HTTP to allow custom-built applications written in C, ActiveX, Java, J2EE, JSP, ASP, .Net, PHP, EJB, TCL, COM/+, C# or Perl to communicate with the IDOL Server. Providing full integration capabilities with all professional development languages and Web scripting technologies, Autonomy enables rapid access to information across all networks through any existing business application layer.

3.10 Expandability

Once IDOL is integrated into the enterprise system, it is trivial to add advanced functionality such as speech analytics and video search at a later time. Although the enterprise may currently not need some of the operations that IDOL offers, it helps to consider a platform that supports such facile addition of suitable technologies, especially since the enterprise cannot anticipate its future needs.

3.11 Support for Future Technologies

Automony's core meaning-based technology addresses such a fundamental problem that its solution is being used in virtually every market, in situations that would have been unimaginable when the technology was conceived ten years ago. From handling the latest rich media content to providing e-discovery support for current SEC compliance regulations, IDOL has demonstrated its extensibility by growing and adapting to new requirements and technologies. IDOL is flexible enough to accommodate future technologies while robust enough to deliver security, scalability, and performance in the present environment.

4. Alternative Approaches

Many companies claim to have solutions to the challenges of managing unstructured information, or have promised technologies to deliver personalized information services. However, most of these approaches have severe limitations, particularly in terms of scalability and cost.

4.1 Keyword Searching or Boolean Query

As the most common information retrieval methods, keyword and Boolean searches return a list of documents that only contain terms that were explicitly queried.

4.1.1 Limitations of Keyword Searching or Boolean Query

Accuracy and Context

Keyword and Boolean search are accurate tools when used against a large quantity of data and when the user has very specific knowledge of the unique information they are seeking. Autonomy fully supports this approach. However, while keyword search can match words and phrases within documents, it cannot tell how relevant the entire document is to the subject being researched.

Consider the following sentences:

`"I was walking down the street the other night. It was a long street, a dark street and at the end of the street I was attacked by a mugger."`

Although the word “street” is mentioned several times, the paragraph is really about a crime. A keyword and Boolean search for the word “street” that returned these sentences would be considered very inaccurate although the word is mentioned four times.

To improve this, keyword searches often rely on weighting to rank search results. If a keyword appears in a prominent place in the document (e.g. in the title), the document is given more importance, or higher weighting, than one which contains the keyword in a less obvious place (e.g. buried in the middle of the last paragraph). Higher weighting can also be given to a document that contains multiple occurrences of a keyword.

In well-formatted and consistent data (e.g. medical journals), weighting will improve results, and again Autonomy offers full support for this. On the other hand, this approach still does not take into account the context in which the word appears or the aspect of the topic discussed. Instead, it assumes that if a word is in the title or mentioned often, the document as a whole must be relevant. Autonomy overcomes this problem by using sophisticated pattern-matching techniques to form a contextual understanding of any document, and so suggests other documents that have matching concepts without relying on keywords.

Manual Refinement

Keyword search engines provide manual techniques to refine results, such as complex Boolean expressions, keyword tagging, librarian maintained keyword associations or categories. Again, Autonomy supports these techniques out of the box. In particular, the ACC module, as described in section 1.3 Avoiding Black Box Pitfalls, provides the workflow, user roles and automatic publishing options that are needed to distribute taxonomy and classification management to subject matter experts and knowledge engineers. As with all Autonomy applications, the ACC can be automatic, manual or both.

However, traditional keyword search engines do nothing more than look for a few specified words—a manually intensive process that requires humans to continually manage and update keyword associations or categories.

For example:

Keyword methodologies rely on the end user to author queries in complex and specific language (also known as Boolean form). This requires an ability to construct unwieldy search "rules." One initial rule may be:

```
<Israel AND Palestine> OR <Israel OR Palestine>
```

This may seem sufficient. However, if the user is particularly interested in the current escalation of tensions between Israel and Palestine, the above rule would yield documents that do not relate to their focus:

"Israel: The number of Jews in Palestine was small in the early 20th century; it increased from 12,000 in 1845 to nearly 85,000 in 1914."

And since many articles are written with the assumption of the readers' background knowledge, some relevant documents may not include the keywords "Israel" or "Palestine," e.g.:

"Mr. Arafat responded to the suicide attacks by declaring a state of emergency in the West Bank and Gaza and arresting 75 militants."

So the rule must be modified to include such documents:

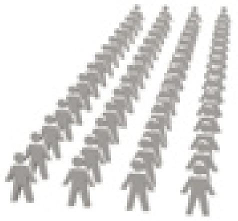
```
<Arafat OR "suicide attacks" OR "state of emergency" OR "West Bank">
```

This, however, may return results about other suicide attacks and potentially miss valuable documents on a directly related topic such as the reaction of other interested parties (e.g. The U.S. Government). The rule must be modified again and again in order to return only the relevant information:

```
<("Arafat" OR "Sharon" OR "Bush") AND ("suicide attacks" OR "state of emergency" OR
"bombings") AND ("Israel" OR "Palestine" OR "West Bank" OR "Jerusalem")>
```

This is only an initial rule without even taking into account ongoing maintenance of the categories involved.

The example rule requires a match from each of the categories of name, violence phrase and location, but not all the stories will contain all three and thus will be passed over by the query. This approach therefore requires extensive, detailed manual effort in order to bring back results that still may not be the most relevant available. In contrast, Autonomy can automatically deduce the main topics in a document, and channel related material to the user without requiring any manual input from them.



Ability to Learn

Keyword search engines cannot "learn" through use. It is also very difficult for keyword search systems to find content by being shown an example. Typically a "more like this..." function will increase the number of keywords in the query based on what terms appear most frequently in the example document. Documents are matched based on keywords and therefore on the categories into which they fall. While Autonomy can support this method, its technology is also able to match documents based on the concepts they contain, which is much more useful.

For example, someone interested in the financial dealings of the New York Yankees will be offered other articles on sports (golf, tennis, maybe baseball) by a traditional "more like this" function. Autonomy would be able to see that the focus of interest is financial in relation to a particular team and provide related documents accordingly.

4.1.2 Autonomy's Approach

Though Autonomy fully supports Keyword and Boolean search, it also has the unique ability to match concepts as well as simple keywords. The technology takes into account the context in which terms appear, eliminating many false hits while also retrieving documents that may not contain the specific term.

4.2 Page Ranking and Popularity-Based Internet Methods

The technology ranks documents or URLs based on the frequency of use or the frequency the document or the URL is quoted elsewhere, by virtue of being linked. This is then used in conjunction with a keyword entered by the user to conduct a search for the most relevant results.

4.2.1 Limitations of Page Ranking and Popularity-Based Internet Methods

Too General

As the pages that are linked the most are considered the most relevant, Page Rank will direct users to the most general pages first. For more specific information, however, a user must guess the correct keywords with which to conduct a successful search. This approach does not take into account users' idiosyncrasies or individual focus.

Ineffective in the Enterprise

This approach is ideal for the Web as there are a great number of hard-coded hyperlinks between Web pages. However, these links are absent and incredibly laborious to create in the enterprise as data sources are seldom Web information but rather a wide range of document management systems, file systems, databases and collaboration medium.

4.2.2 Autonomy's Approach

Autonomy fully allows manual control by the administrator to promote certain web pages given specific keywords. However, Autonomy understands the content of any given page and makes conceptual associations between pages, documents and search queries, ensuring that users are always provided with information that is particularly relevant to them. Autonomy does not rely on keyword search and does not need lists of links to tell it which pages are relevant to a query.

4.3 Explicit Thesauri

A thesaurus provides a list of industry-specific terms and their synonyms to a system so it can recognize these unusual words and phrases when it encounters them. This is useful in environments with a large corpus of industry-specific terms, abbreviations and jargon, such as the medical and scientific fields.

For example, a thesaurus can enable a system to understand that when someone is interested in cancer research, they may also wish to read documents that talk about oncology. However, thesauri are expensive and time consuming to create and definitions are often not accurate because meaning of words can change depending on context.

4.3.1 Limitations of Explicit Thesauri

Manual

An expert must compile synonym lists or thesauri before they can be added into the system - a highly manual and lengthy process. This approach comes with a significant risk of error because it relies on the ability of a person or group of people to predict and include any term that is likely to be used by anyone using the system. However, it can be rewarding for specific applications with a prior knowledge.

No Ability to Learn

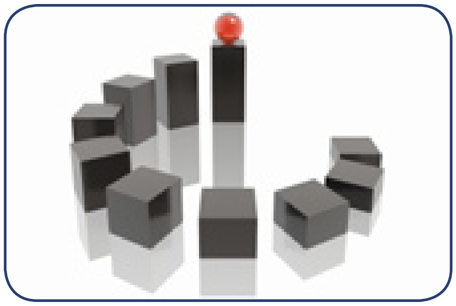
The lists are static once entered into the system. They are not able to learn as words change their meaning or as new words are added to the language. Therefore, when there is a change or addition, it needs to be added into the system by an administrator, making thesaurus maintenance another time-consuming manual activity. Autonomy will not only recommend words for static datasets, but will also deal differently with constantly changing datasets such as news stories.

Finally, words with more than one meaning can also cause problems. For example, a system that uses thesauri for understanding language would interpret the phrase “she’s a star” to have the same meaning as “she’s a cosmic gas ball.”

4.3.2 Autonomy’s Approach

Autonomy can load and fully utilize any existing thesauri. However, Autonomy can still function perfectly well without them. Since the technology is mathematically based and free from linguistic restraints, it need not use any language-dependent parsing or dictionaries to extract meaning. While many enterprise platforms rely on pre-existing knowledge of grammar and linguistic rules, IDOL allows indexed content to dictate the model, as it develops a statistical understanding of patterns that occur in the content over time. Hence even slang or industry-specific jargons do not pose a barrier for processing with IDOL. True to its Bayesian roots, the more content IDOL collects about that industry (e.g. legal terms, pharmaceutical developments, technology) the more understanding it will form about that domain. As new vocabulary and terms are introduced, or their meaning changes, they will automatically be updated in the system and the thesaurus does not need to be manually updated. An example is the overnight change in meaning of the term “Ground Zero” after September 11, 2001. Autonomy immediately understood that it was being used in relation to the terrorist attacks on the United States, and not in relation to a nuclear blast.

4.4 The Vector Method



The Vector method is concerned with the partitioning of data, or categorization. This is done by imagining documents as points in a multi-dimensional space, which are then divided into categories. Categories must be taught to the system; the more training that occurs, the more accurate the categorization needs to be. Many of today's search engines use a combination of Vector and Boolean methods.

4.4.1 Limitations of the Vector Method

Language Dependent

The system needs to be trained in its target language, and will only recognize words it has been taught. There is no inherent understanding of synonyms or related words. For example, it would be unable to deduce that "Creutzfeldt-Jakob" and "mad cow" are related terms.

Inaccurate

The Vector method is inaccurate because it is unable to perfectly divide categories and has particular trouble with documents that fit into more than one category. It will classify such documents under one category or another, but not both. There is also no notion of threshold or relevance, so if a document is put into a particular category, there is no indication of how relevant it is within that category - does it mention the topic only a few times, or is it entirely focused on it? The Vector Method is unable to tell.

Manual

All categories must be defined manually by administrators, and the system requires constant monitoring and maintenance to ensure its effective functioning. This approach cannot possibly scale, as any time there is a change in the categorization, the whole training process must begin anew because there is no ability to make updates to just one area of the system without compromising the integrity of the model.

Ranking Discrimination

The importance and relevance of one word compared to another is not understood. To combat this, common words can be ignored, and the focus placed on rare words, assuming they will give more insight into the theme of a document. However, this is not always accurate and can result in weight being placed on inappropriate words, resulting in categorization errors.

4.4.2 Autonomy's Approach

Autonomy's technology can understand the content of a document probabilistically, without depending on an understanding of a particular language, and create categories accordingly. Where necessary a document can be classified in more than one category. Autonomy's automatic categorization functionality ensures that taxonomies are created and maintained with as much or as little human interference as desired.

4.5 The “OneBox” Model

The OneBox model was designed to deliver a range of specialized results through a single search box. When the user submits a search query, the OneBox engine passes the terms to a number of third-party sub-systems, each one focusing on a particular information niche, such as a particular database or file-system, for example. It then displays the results from the various sub-systems alongside its own search results. This means that even for a very specific query such as a product code, the user only needs to type into one search box. Behind the scenes the search engine will pass the query to a sub-system provided by the manufacturer, and then deliver the resulting information. In theory, this saves the user time and effort, but there are notable disadvantages to this approach.

4.5.1 Limitations of the “OneBox” Model

Information Overload

This model is an unsophisticated approach that aims to circumvent some of the problems associated with keyword search. Since the engine has no way of knowing the conceptual origin of the search query, it systematically processes the terms through every possible information context available to it, in the hope that one of them will correspond to the context of the search query. Consequently, this approach unnecessarily bombards the network with requests to each individual sub-system, when only one query to the appropriate sub-system would have been sufficient. In the context of a global enterprise solution this represents a significant obstacle to scalability.

Inaccurate Results

Each of the third-party sub-systems uses a different scoring algorithm to determine the relevancy of the results. This means that in identifying the result that is most relevant to the query, the search engine receives conflicting reports from each of the sub-systems. For that reason this type of solution cannot be relied upon to yield the most accurate result every time.

4.5.2 Autonomy’s Approach

Autonomy can federate to any query-based system using Automatic Query Federation (AQF). AQF determines whether some queries are more suitable for federated search than indexed content. The AQF module federates the query to a system it has identified as the correct specialist, based on its conceptual understanding of the query. For example, if a user enters a shipment tracking number, IDOL forms a checksum from the string of numbers to realize the nature of the query and only federates to FedEx. In this way, unlike alternative approaches, IDOL avoids bombarding all of the systems with the same query, and simply selects the one most likely to return the specialist result. Moreover, search criteria can be pre-processed and results can be post-processed in nearly any way imaginable.

4.6 Parsing and Natural Language Analysis

For the last twenty years, much effort has been concentrated in parsing or semantic analysis to try to understand unstructured, textual information.

4.6.1 Limitations of Parsing and Natural Language Analysis

The Inherent Complexity of Language

In spite of more than two decades of research in semantic analysis, the approach is rarely used in real applications because its results and performance have yet to live up to expectations in real-world problems. The following cases illustrate the limitations of this approach—namely, its inability to handle ambiguity.

Example 1:

`“The dog came into the room; it was white.”`

It is unclear from the sentence whether it is the dog or the room that is white. On the other hand, a person would have little problem deciphering the following examples because of his or her familiarity with both rooms and dogs:

`“The dog came into the room; it was furry.” ”The dog came into the room; it was full of furniture.”`

The computer, however, would still be stumped. It lacks the understanding to solve such ambiguities. Some advanced systems will allow the construction of a set of rules for the machine to resolve these uncertainties. However, the instruction set would be cumbersome and difficult to maintain, and would significantly degrade the system’s performance.

Example 2:

`“The fly, it’s clear to me, can fly faster than the bee.”`

Firstly, the computer may be confused by the word “fly,” which is used in this sentence as both a subject and a verb. But that is an easy problem to solve. What about the word “it”? How does one parse a word that refers to abstract thought? These problems are exacerbated when a computer attempts to extract meaning by parsing full paragraphs.

Example 3:

`“The president arrived by car to meet the Chinese premier.”`

Like keyword-based approaches, semantic analysis cannot determine the relative importance of ideas. In other words, the computer will assign equal levels of importance to the President, his mode of transportation and the leader with whom he is meeting. Moreover, parsing is only designed to handle a few sentences and suffers great difficulty in extracting meaning from a full paragraph. In contrast, Autonomy can form an understanding of the concepts underlying a large corpus of information, from a paragraph to a whole document.

Reliability

Because semantic analysis is based on a true/false decision tree and rules structure, one incorrect decision or the occurrence of an unknown construct can derail the entire analysis.

Language Dependent

The semantic approach is language specific and reliant on grammar rules, making it vulnerable to slang or grammatically incorrect constructions. It is an non-scalable system that needs to be taught every new word or change in meaning. Generally, the system will only support a very limited subset of languages (e.g. English, German and Dutch), and adding a new and very different language, such as Chinese, can be problematic. Autonomy's technology is uniquely language independent.

Question and Answer Systems

An increasing number of search vendors now offer users the ability to retrieve information through natural language questions. While this approach may work well for one-sentence questions or queries concerning a known universe of information, the language model simply breaks down when employed on large documents with many concepts. This occurs because Q&A systems rely on the simple combination of manually defined "question forms" and a corresponding structured dataset that holds the relevant answers. As a result, these systems can only recognize precise questions and their matching answers that have been stored in their database. They cannot find concepts outside this manually defined structure that might supply relevant answers to users' questions. Equally, Q&A systems cannot understand questions that are worded slightly differently or are phrased using slang, even if these queries would make perfect sense to a human.

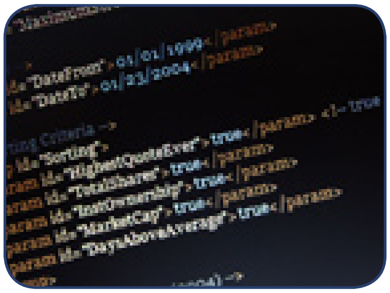
4.6.2 Autonomy's Approach

Autonomy's pattern-matching technology uses predictable statistical word patterns to represent concepts and functions independently of any given language. However, it can still perform sophisticated semantic analysis as needed to, for instance, detect sentiment within a given content (see 3.6 Language Independence).

4.7 Manual Tagging, Weighting and XML

With an upswing in enterprise portal use, manual tagging schemes are becoming an increasingly popular method of labeling digital material (including documents, structured data, HTML, XML and multimedia).

4.7.1 Limitations of Manual Tagging, Weighting, and XML



```
id="DateFrom">01/01/1999</param>
id="DateTo">01/23/2004</param>

ing Criteria -->
id="Sorting">
id="HighestQuoteEver">true</param> <!-- true
id="TotalShares">true</param>
id="InstOwnership">true</param>
id="MarketCap">true</param>
id="DaysAboveAverage">true</param>
```

Descriptive Inconsistency

Tagging is dependent on the author behind the tag, i.e. the person's mood, the person's background and the person's personality. Each person will categorize, or tag, a given document differently. Sometimes a person may even get lazy and tag much of the content under the category "General." This inconsistency that is intrinsic to the tagging process introduces further impreciseness and incompleteness to the retrieval process.

Additional complications arise when subjects incorporate multiple themes. Should an article about "technology development in Russia within the context of changing foreign policy" be classified as (i) Russian technology, (ii) Russian foreign policy, or (iii) Russian economics?

The decision process is both complex and time consuming and introduces yet more inconsistency, particularly when considering the sheer number of options available to a user. For example, over 800 tags are available for general newspaper subjects, presenting more challenges to developing an efficient and consistent tagging process.

Interoperability of Tagging

XML is not a set of standard tag definitions, but rather a set of definitions that allow users to define tags. Thus, if two organizations are going to interoperate and apply the same meaning to the same tags, they must explicitly agree upon their definitions in advance.

While this may prove possible for small groups of cooperating agents working over public networks, doubts remain as to whether this will scale to support an extended network of industry trading partners.

Idea Distancing

Tags also fail to highlight the relationships between subjects. There are often vital relationships between separately tagged subjects such as wing design/low drag and aerofoil/efficiency, a concept known as "Idea Distancing." Obviously, there will be a degree of overlap between these categories, and because of this a user may be interested in the contents of both. However, without understanding the meanings of the category names no clear correlation can be established between the two.

Not Scalable

The specific needs of the retrieval process require a high number of tags to be applied to documents. For example, tag numbers in a company such as Reuters run into the tens of thousands. However, as the number of tags increases, so does the likelihood of misclassification.

High Labor Costs

Taxonomy creation and tagging involve a predominantly manual effort that require input from librarians, users and IT staff.

4.7.2 Autonomy's Approach

In the presence of existing metadata, Autonomy extracts all such data and offers complete legacy support. Autonomy's commitment to supporting legacy taxonomy and workflow enables the enterprise to migrate to Autonomy and benefit from Autonomy's inventive technology without losing previous work. Autonomy adds a layer of intelligence to the management of XML and understands the content and purpose of either the tag itself, or related information, or both. However, Autonomy's technology understands the information itself and is therefore able to make decisions and perform operations dependent on the entire content—not just on a subset of words decided upon during an unaccountable and expensive manual process.

5. Conclusion

Proven by widespread penetration in diverse markets, Autonomy's technology addresses a pervasive enterprise need by providing a highly accurate, scalable and cost-effective alternative to existing approaches in information processing and management. With over 17,000 customers, over 300 OEM partners and 400 VARs and integrators, "Autonomy's search technology is becoming a de facto standard for companies."[2] With the proliferation of internet technologies and data formats, Autonomy's superior ability in handling structured and unstructured information provides solutions for modern enterprise hampered by information excess and legal demands. Businesses can now maximize their assets and reach their full potential by using a single technology platform to achieve automation across multiple formats and information silos.

Through its scalable and modular architecture, Autonomy's approach enables organizations to customize technology according to their unique and evolving business requirements. Respecting legacy systems and workflows and fully XML-compatible, Autonomy's easy integration solution makes it easier than ever for enterprises to transform data wealth into veritable riches of knowledge.

Autonomy adopts an open approach that embraces both legacy and advanced methods to continue to improve its technology. As demonstrated by its market-leading investment in R&D, enterprises can continue to rely on Autonomy to provide the best technology. Fully universal in its solution yet allowing unique customization, fully automatic in its processing yet allowing manual control, Autonomy offers a complete and flexible solution to any modern enterprise drowning in the sea of information.

6. Further Reading

- *“Interactive Technology Strategies - Personalize or Perish?”*
  The Forrester Report; Volume Two, Number Three, May 1997

- *“Bayesian Methods Applied to Neural Networks”*
  by Radford Neal, Springer Verlag

- *“Bayesian Forecasting”*
  by Mike West, Springer Verlag

- *“Applied Bayesian Forecasting and Times Series Analysis”*
  by Pole, West & Harrison, Chapman & Hall

- *“Bayesian Statistics”*
  by P. Lee, John Wiley & Sons

- *“Numerical Bayesian Methods Applied to Signal Processing”*
  by O-Ruinadah and W.J. Fitzgerald, Springer Verlag

2. Financial Times, 4th July 07

The information contained in this document represents the current opinion as of the date of publication of Autonomy Systems Ltd. regarding the issues discussed. Autonomy's opinion is based upon our review of competitor product information publicly available as of the date of this document.

Because Autonomy must respond to changing market conditions, it should not be interpreted to be commitment on the part of Autonomy, and Autonomy cannot attest to the accuracy of any information presented after the date of publication.

This document is for informational purposes only; Autonomy is not making warranties, express or implied, in this document.

(Autonomy Inc. and Autonomy Systems Limited are both subsidiaries of Autonomy Corporation plc.)