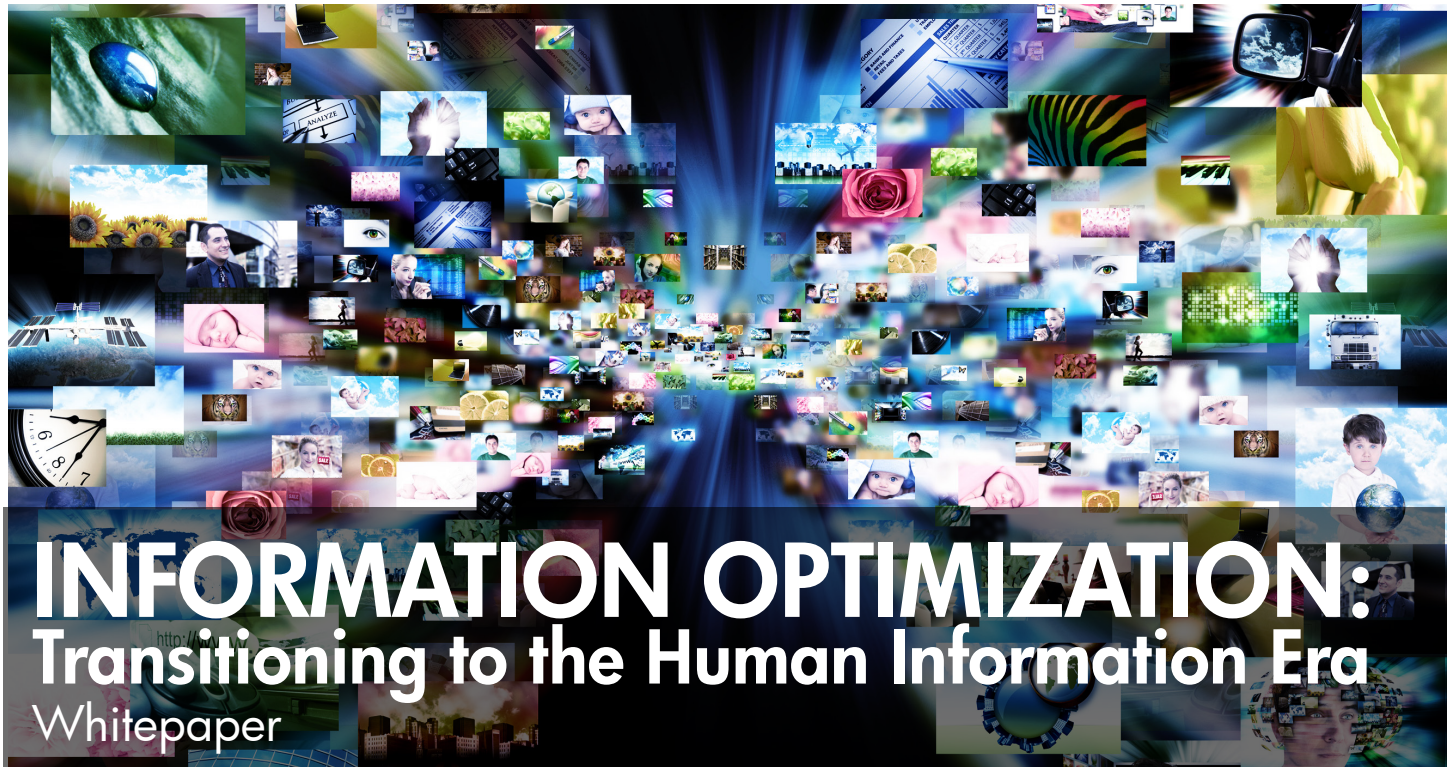


INFORMATION OPTIMIZATION:
Transitioning to the Human Information Era
Whitepaper

ALTHOUGH TODAY'S "INFORMATION" IS AT THE FOREFRONT, THE REAL VALUE IS IN THE ABILITY TO UNDERSTAND AND LEVERAGE IT.

In the whitepaper, *Information Optimization: Turning Information into Enterprise Business Decisions*, we discussed the critical connection between enterprise success and the ability to understand and leverage ALL forms of information. In this paper, we take a closer look at the largest and fastest-growing component of enterprise-relevant information being created and consumed: unstructured, 'Human Information.' More importantly, we cover the technology breakthroughs that enable the optimization of Human Information, taking it from a 'desirable theory' to a real-time, competitive reality.

WE NOW LIVE IN AN ERA OF HUMAN INFORMATION.

"Human-friendly" information makes up about 85 percent of all data and includes emails, audio, video, social networking, blogs, call center conversations, machine-generated sensor data, and more. It grows at a breathtaking rate: 62 percent CAGR. This is the future of information computing and it represents a fundamental shift in the way people and businesses interact with information.

Beyond its sheer size, unstructured information is where all the interesting, differentiating, and vital things happen. When processing information looking to uncover a crime, investigators look for incriminating emails. When trying to understand their customer base, marketers look for information on their customers. But, unfortunately, customers don't send you databases; they tweet or blog. And this is only becoming more complicated with the explosion of social media activity.

WHY IS HUMAN-FRIENDLY INFORMATION DIFFERENT?

When discussing unstructured information, which dominates today's enterprise, people typically assume it has to do with search. This is only because of the historic inability of computers to process this type of information, so the seemingly obvious way to deal with it is to search it. However, this still requires human intervention, and more importantly, the computer can't actually do anything with the information because it can't understand it.

When searching, all a computer does is find every instance that a particular combination of words occurs. A search for "D-O-G" does not understand what a dog is and can generate millions of results. A user then has to sift through all of those results to find the desired context. To improve this process, rules, popularity ranking, federation, and other basic functions are added, all of which have their limitations.

THE ABILITY TO UNDERSTAND CONCEPTS

The leap forward comes with conceptual search. When a computer can understand that the letters "D-O-G" mean a dog, man's best friend, a Labrador, an animal that likes to go for walks, the process becomes more human.

Yet the lack of structure in Human Information makes the search process challenging for the simple reason that people search or analyze data using an attribute of the data, such as the date a video was taken, who is in a photo, or whether or not a blog gives a positive view of a product. This requires some form of metadata (data about data) to be tagged to the data item or generated on the fly as the item is saved.

If no such metadata exists, users will have difficulty finding it or may not be able to find it at all. This is not an easy issue to solve without human involvement. For example, it is not easy for software to tell if a picture is of a yellow rose, a yellow Labrador named Rose, or a girl named Rose in a yellow dress.

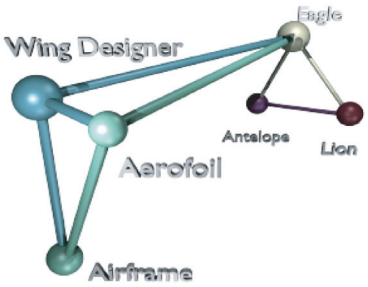
Compounding these challenges, Human Information is also often more difficult to manage than structured or semi-structured information in terms of size, organization, and availability. Human Information comes in two categories:

1. **UNSTRUCTURED TEXT DATA.** This category includes unstructured text data such as content posted to blogs, news feeds, documents, and social media outlets such as Twitter and Facebook.
2. **OTHER UNSTRUCTURED DATA**. Under this category, we include photos, videos, sound files, and other forms of data that by default DO NOT HAVE ANY TEXT INFORMATION on their subject.

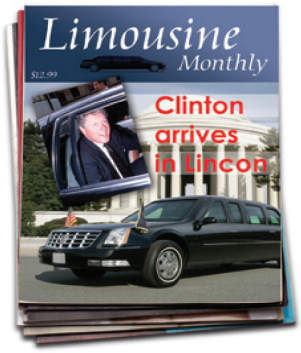
To realize the importance of understanding the concepts contained in information, one must recognize the unique challenges posed by human-friendly information.

- **Information is diverse**. Human-friendly, or unstructured, information is not one file type or from one source of information. It represents all types of information that does not fit neatly into a structured database. It includes text in the form of emails, documents, IMs, social media and SMS, audio in the form of speech and sounds, video, XML, and images.

- **Ideas do not match, they have a distance**. No two ideas are exactly the same, but they have degrees of similarity based on how close they are to each other conceptually. Consider the description "low-drag wing design expert" versus "high-efficiency aerofoil designer." These words do not match, but the ideas are very similar. They are only slightly different from one another. These similar ideas are, in turn, very different from other ideas, such as safari animals



Distances between ideas also change with the context around them. When the story "Clinton Arrives By Car to Meet the Chinese Premier, Drives Up in Black Lincoln" appears, the main point changes based on who reads it. For most people, the news is that Clinton has met with the Chinese Premier. For the subscribers to Limousine, Charter & Tour Magazine, the real news is that Clinton arrived in a black Lincoln. When analyzing human-friendly information, the context must be understood to grasp the meaning of the information.



- **Information does not exactly match**. There is a definitional problem when dealing with human-friendly information. When a user poses a query, information never matches exactly the way structured information would. The question "Is Snoopy a dog?" does not have a simple answer, as there are many ways to define Snoopy. You must take into account why he would or would not be considered a dog. The answer to a question like "Is Snoopy a dog?" is also dependant on other pieces of information. For instance, if the answer was "No, he is a cartoon character," then Snoopy would not be a dog. This demonstrates the relative nature of information.

- **Meaning is dynamic**. This is especially true in the age of social media, where new slang is continually emerging. Even within the same phrase, a single word can have multiple meanings based on the intent behind it. Take the tweet "Saw Red Riding Hood, the wicked wolf got boiled - it was really wicked." The word wicked can either mean bad or good, based on where it appears in the post. The ever-changing nature of a word's meaning makes it especially difficult to understand and process human-friendly information.
- **Meaning is multi-layered**. Within the same set of phrases or words, there can be multiple levels or layers of meaning. This principle is best seen in poetry, where complex metaphors can run through a set of text, building on each other and adding depth.
- **Meaning is relative**. What something means is heavily informed by one's own perspective, whether historical, cultural, or other. Two opposing cultural groups will view a set of results very differently. Meaning also changes over time and is subject to historical perspective.

AT THE MEANING LEVEL – IT'S NOT JUST ABOUT SEARCH

Although unstructured information presents complex challenges, once you are at the meaning level, information is interchangeable effortlessly. For example, once you can process a phone call and understand it is about Peruvian gold futures, you can then relate that phone call to conceptually similar emails, documents, or instant messages. The actual understanding is much more difficult to get, but once you get it, it is possible to relate it to any other piece of information, regardless of the original format.

This is not possible with database information because the structure must be aligned with the specific database it comes from. For example, information from one database cannot be related to another piece of information from a different database. Multiple copies of the same type of data can exist in different locations, causing great inefficiencies because there is no communication between applications or databases. Enterprise application integration (EAI) attempts to link these information silos to improve communication and facilitate automation of business processes. EAI still focuses on linking databases, translating between computer languages, and even bridging the gap to legacy systems—instead of understanding the content within them and relating them conceptually.

The ability to understand the meaning of information is not merely about searching or storing; it is about processing unstructured information. Searching for a term and returning similar, conceptually linked documents or files is remarkably useful, but only if you already know what you are looking for. Very rarely do we already know what to look for when faced with a corpus of unstructured data, which is why powerful, automatic processing is so valuable.

Clustering is one example of automatic processing, and enables a mass of information to be assembled into groups based on what the data actually contains, not just what you think you should look for. For instance, with clustering, the organization could classify articles in a newspaper automatically by topic, alert someone in a call center when you have an angry customer, decide if a blog talking about your company is positive or negative, and respond to it accordingly. This is where the value is. Understanding meaning is about processing and being able to do that effortlessly across all different types of information.

Fundamentally, the ability to understand meaning and automatically process information is about distance, probabilities, relativeness, definitions, slang, and more factors. It is an overwhelming and continually growing problem that requires advanced technology to solve.

COMPUTERS HAVE FINALLY CAUGHT UP, AND THE RACE HAS BEGUN BETWEEN YOU AND YOUR COMPETITORS

In looking at information's history and advancements in the last 50 years, the databases of the 1960s were run on computers that were not powerful enough to understand 'real-world', rich information as humans could. To handle this, we had to create a new machine-friendly paradigm, which gave rise to a structured data world.

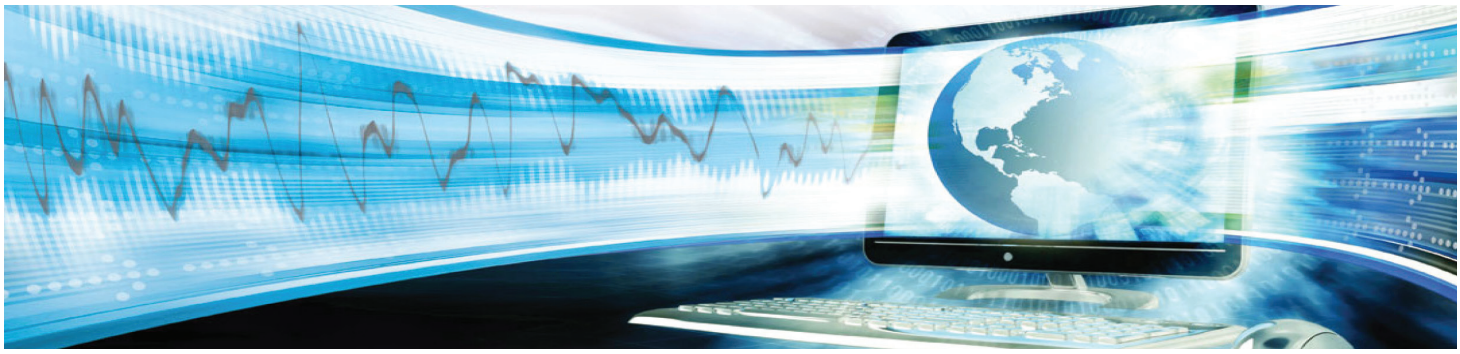
As computer processing power increased, this machine-friendly paradigm proved very useful to business. When information was structured, its location gave it meaning. For instance, if a column stood to represent the 'number of teddy bears in the warehouse', when that column went to 0, an amazing thing would happen: the computer would automatically order more. The value here came from the computer's ability to process the information, not merely to locate or retrieve it.

As the years continued, up to and through the advent of the modern Internet, the focus in IT was on changing the "T" through the introduction of mainframes, client servers, IP, cloud computing, and more. Meanwhile, the "I" in IT began to morph dramatically.

The example of video information is a vivid one. Outside the enterprise, the volume of consumer content creation and consumption has increased significantly over the last five years. Simply looking at the data from popular websites shows that the rate at which data is created and consumed is also increasing. For instance, since YouTube was founded at the beginning of 2005, the rate of data used and consumed on the site has grown rapidly. Consider the following:

- Users upload 35 hours of video every **minute.**
- Unstructured data will grow to over 35 Zettabytes by 2020.
- Videos on YouTube were viewed 2 billion times per day, 20 times more than in 2006.
- In the last four years, video uploads have seen an eight-fold increase.

Growth in video use on YouTube is not an isolated case, as Apple's iTunes and Facebook have also shown increased use relating to music and photos. Organizations are following consumers, increasing their online presence, and attempting to leverage the data published by consumers. YouTube and Internet data only further show the proliferation of Human Information in today's culture. Human Information is exactly as it sounds: information that is created by people and easily understood by people.



HOW DID COMPUTERS CATCH UP?

The primary breakthroughs have come in the areas of software that can derive meaning, spot patterns, 'connect the dots,' and take automatic actions. These developments have combined with advances in hardware and software that allow massive amounts of constantly created and updated multi-media information to be analyzed real-time.

A pioneer in the area of deriving insight, intuition, and ideas—and better enterprise decisions—from unstructured Human Information is the technology from Autonomy Corporation, which recently has become a part of HP. Autonomy's technology understands any form of unstructured information, whether text, voice, or video, and based on that understanding, performs automatic operations such as, but not limited to, "if you like that, you would like this" on the information. Autonomy's core technology, the 'Intelligent Data Operating Layer' (IDOL), allows text to be searched and processed from database, audio, video, text files, or streams. Autonomy refers to the processing of such information by IDOL as Meaning-Based Computing.

THE ARCHITECTURE OF STRUCTURED INFORMATION AND THE UNSTRUCTURED DATA ENTERPRISE

With structured information, each structured application has a set of accompanying information. This information must be paired with the application because its meaning is conferred by its location within the application. In this approach, information is not interchangeable; it only goes with the context it is found in, such as column 3, row 4.

When faced with how to access and leverage structured information, most of the IT industry still takes a stovepipe approach. Each application has a custom connector to access the information. Every data set has its own connection to the application. For instance, this DATA 1 connects to APP 1, this DATA 2 connects to APP 2. This approach results in a segmented world with little to no overlaps across information or applications. Business Intelligence is cut across a few repositories, but not in a meaningful way. In a world of very structured information this provides some value, but to deal with the influx of unstructured information and derive meaning across separate silos, this approach falls short.

THE UNSTRUCTURED DATA ENTERPRISE

Methods of dealing with unstructured information began in the same way as structured. Emails would go with the email server, documents went with Documentum, etc., and this tactic very rapidly became a whole lot more complex than the structured world. In the unstructured world, a piece of information does not exist without any relationship to other pieces of information. They may have varying distances or degrees of similarity, but are still linked. When looking at one piece of data, it is necessary to look at lots of others. For example, customer information may come in the form of call center calls, tweets, emails, or website comments. Data relevant to discovery may include voicemails, emails, documents, SMS messages or more. This very quickly becomes a rat's nest, as every application has a separate connection to every data type. As soon as any data type or source is changed, all the connections must also be changed. This is the same problem with operating systems; when one chip is changed, all of the software must be re-written.

Figure 1: Architecture of Structure Information and the Unstructured Data Enterprise

TODAY'S APPROACH: SINGLE LAYER ACCESS

The solution to accessing and processing all structured and unstructured information is a single layer that goes across the enterprise—one system that is able to process both structured and unstructured information together. The next-generation information platform, IDOL 10, is designed to understand and act on 100 percent of enterprise information in real-time. This new platform promises dramatic business impact, as organizations can develop new applications that leverage the diversity and richness of Human Information combined with extreme structured data.

- IDOL 10 provides a single processing layer for forming a conceptual, contextual, and real-time understanding of all forms of data, both inside and outside an enterprise.
- The new platform combines Autonomy's infrastructure software for automatically processing and understanding unstructured data with Vertica's high-performance real-time analytics engine for extreme structured data.

A PLATFORM BUILT FOR THE HUMAN INFORMATION ERA

To succeed in the new Human Information era, organizations require a fundamentally new approach, with technology that delivers insight, ideas, and intuition into the constantly proliferating set of Human Information generated every day. Technology that cannot account for shades of grey is limited to historical data and is dependent on manual methods for tagging and cataloguing data. These methods are insufficient in the dynamic, complex, and nuanced era of Human Information.

CONCLUSION

This shift towards human-friendly information represents the biggest change in the IT industry, and is a once-in-a-generation opportunity. People do not live in rows and columns. Customers don't send you databases, but instead they call, email, and tweet. Policies, regulations, and governance practices are built on human meaning and intent, not programming languages like SQL. For the first time, the "I" in "IT" is changing, not the "T." Up until this point:

- The majority of the industry has been built on only 15% of the digital universe.
- Systems can only answer the questions you already know to ask.
- Analytics on historical data could produce pretty charts.
- Humans have had to fit the machine.

Now, for the first time, it is possible to address the whole problem, to have machines fit the human. It is now possible to run analytics across all information types, including structured, unstructured, audio, video, and more, with real-time meaning-based analysis and the ability to produce actionable outcomes, not just charts.

Now, a customer's call center call can be linked to their website activity, to their entry in the database, to their purchase history, in real time. Now, policies can be implemented across emails, documents, voicemails, social media, SMS messages, and transaction histories, to not only flag non-compliant materials, but also stop non-compliant posts or even transactions before they occur.

Now, it is possible to answer the questions you didn't even know to ask.

Human Information: The Next Evolution of IT.

ABOUT AUTONOMY

Autonomy Corporation, an HP Company, is a global leader in software that processes Human Information, or unstructured data, including social media, email, video, audio, text, web pages and more, enabling companies to leverage their data assets.

ABOUT HP

HP creates new possibilities for technology to have a meaningful impact on people, businesses, governments and society. The world's largest technology company, HP brings together a portfolio that spans printing, personal computing, software, services and IT infrastructure to solve customer problems. More information about HP (NYSE: HPQ) is available at http://www.hp.com.

4AA3-8560ENW November 2011