



WHITE PAPER

Quality and value in the management of unstructured information: tools based on semantic analysis

Sponsored by: **Expert System**

Analyst: Fabio Rizzotto

September 2006



WHITE PAPER

Quality and value in the management of unstructured information: tools based on semantic analysis

IDC OPINION

Business analytics tools that focus on analyzing databases are not by themselves capable of supporting organizations in their efforts to leverage their full knowledge resources, in which unstructured information currently predominates.

It is estimated that about **80% of corporate information** is **unstructured** and contained in a variety of formats (electronic documents, Excel spreadsheets, presentations, .pdf files) that contain text and data that is not readily accessible but holds immeasurable value for understanding the internal and external dynamics of enterprises.

Information chaos is one of the primary sources of inefficiency and waste, especially if measured in terms of time spent on unproductive activities. A new direction in content management policies must start with the adoption of **content access tools** that facilitate accessing, analysis, extraction and visualization of information that is truly useful for knowledge workers.

Technological innovation has driven a shift from traditional search tools to systems that do not just "search for" and "find" information, presenting it in a normalized format, but go beyond that to **grasp the "meaning" of the content**, thanks to search methodologies based on **deep linguistic analysis (semantic engines.)**

This IDC White Paper puts these dynamics in context and describes the business value of IT solutions based on these linguistic analysis tools, including the **COGITO** platform by **Expert System**, conceived to bring "intelligence" to the search, extraction and classification of unstructured information for internal management purposes and for monitoring and analyzing external sources, such as the Internet.

Unstructured information: from “total chaos” to “reasoned chaos”

In an age characterized by change, globalization and growing competition in every sector, knowledge is undoubtedly the most important asset that an enterprise possesses for gaining a competitive advantage.

Starting in the mid-1990s, coinciding with the development of the Internet and exponentially more powerful computer technologies, significant investments were made in an attempt to transform the content of hundreds of files scattered throughout organizations into added value, or to integrate diverse solutions in order to permit more efficient use of a company’s knowledge base.

Despite this effort, substantial results were only achieved in situations where it was possible to arrange information as “data”, or rather in databases with objects (tables, columns, etc.) linked by a relational code. For the remaining information, which was only available in unstructured form and was estimated to represent about 80% of all significant corporate information, the **efforts undertaken produced only partial results**. These were constrained even further by the impact of developments in technologies, formats and sources that generated a range of problems, which can be grouped into two main types:

- ☒ **quantitative** problems associated with the growing volume of documents and content that make management more difficult both at the technology level (greater hardware capacity, loss of system efficiency and performance) and at the user function level, where physical and electronic volumes risk becoming unmanageable;
- ☒ **qualitative** problems linked to difficulties in automated information retrieval and management. This generates inefficiencies and costs that are not easily quantified but that have a major impact on productivity.

This **unstructured** information exists in a variety of formats such as **text, spreadsheets, presentations, e-mails and .pdf files** and represents not just the majority of business information, but is frequently the **repository of interpretative keys** that, together with the “data” itself, drive an enterprise’s business. In addition, text documents often contain real data that is buried within “unstructured” information and that loses its value and significance because it is not easily accessible or connected to a specific process.

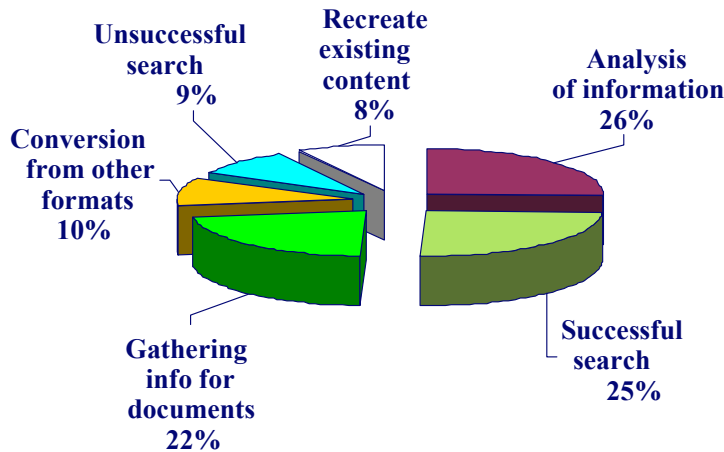
The parallel existence of a less-than-optimal organization of knowledge, with information silos arranged by departmental projects, multiple content repositories, etc. and a growing volume of important, strategic information available in an unstructured format makes the job of users even more complicated, since they need to adapt the contents to the workflow in order to build more structured processes. In addition, in the integration of internal and external information (for example, on Internet sites or external portals) traditional tools are not always appropriate for providing real “intelligence”, which efficiently identifies the desired information and extracts only the relevant information.

As shown in Figure 1, the state of content management leads users to invest time in inefficient activities. If the time spent by knowledge workers on activities with strong “informational” connotations is equal to 100, the estimates given below show how time is split among various procedures. About 50% of workers’ time is dedicated to

successfully searching for and analyzing information, while unsuccessful search activity takes up 9% of the time, which can cause a second distortion, namely re-writing existing material from scratch (8%). The conversion of existing content into new formats absorbs 10% of the time.

FIGURE 1

Information-related activities: how the knowledge worker uses time



Source: IDC, 2006

Tools for efficiently accessing unstructured information

Inadequate organizational mechanisms are one cause of the problems we have been addressing. Long-standing methodologies, non-integrated processes and lack of planning frequently make it hard for users to access relevant information. However, despite efforts at the organizational level, it is obvious that it is **difficult to overcome** these inefficiencies **without the help of suitable technological instruments**.

In the past, organizations invested primarily in the management of structured information, improving the infrastructure (databases) and adopting **business intelligence** applications to support data analysis processes.

These tools (for **business intelligence** or, more generally, **business analytics**) have shown to be of enormous value in supporting decision-making processes. Reporting, analyzing historic data and conducting forecasting simulations are only a few of the techniques that provide enterprises with information for understanding results and developing strategies.

However, the **growing weight** of unstructured content and the difficulty in **accessing** relevant information have underscored the need for a **broader view conception** of business intelligence, one that is not limited to just structured data but is also **capable of extracting** value from unstructured information.

Data analysis can indeed prove to be **incomplete** in explaining business dynamics and providing interpretive keys, the “why” that often can be found within unstructured

content that cannot be directly extracted from the numerical source. In turn, the data, if contained in a text document, for example, can go undetected by data analysis tools, making the need for unstructured information analysis tools even clearer (so-called "content access tools", Figure 2.)

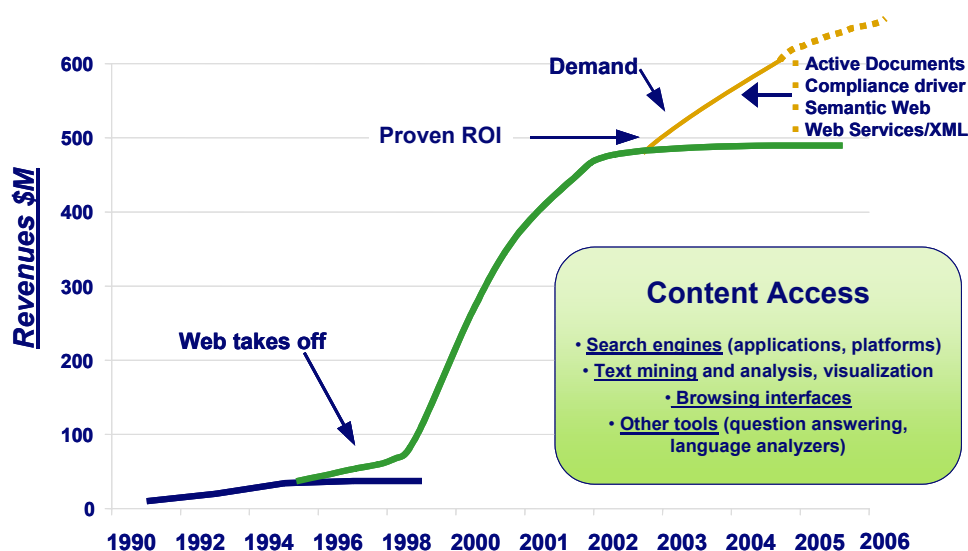
An expanded definition of content management that encompasses this perspective would allow unstructured information to be "normalized" and processed and methodologies to be automated, fostering growth in individual productivity.

Accessing unstructured content using search, visualization and automated categorization tools, or extracting this content (**text mining**) are not new concepts in the world of computing. However, they have received more focused attention over the last few years due to the increasing sophistication of the available technologies (Figure 2.) The boom created by the explosion of the Internet (which in the second half of the 1990s created the search engine business) was followed by a brief stabilization phase before the onset of the second wave of development, where we find ourselves today. Current demand for this type of solution is sustained by:

- ☒ **Internal drivers:** as noted, the growing awareness of the need to develop solutions for internal "chaos" and improving access to **external** information (for example, that residing on the Internet.)
- ☒ **External drivers:** expanding demand thanks to the introduction of more advanced information management technologies and to the need to ensure that processes are compliant with new regulatory requirements.

FIGURE 2

The "boom" in content access tools



Source: IDC, 2006

Text mining: semantic analysis as an effective intelligence tool

Unstructured information can provide **value to business processes** if the techniques used to make it accessible are effective. Over the years, a variety of

solutions to support information search and access have been introduced. There have been three main phases in this development:

- ☒ a **first generation** of tools, created before the arrival of the Internet, was based on full text search techniques using keywords. The databases and document collections that emerged towards the end of 1980s provided these functions.
- ☒ technological development then led to the refinement of traditional techniques in the direction of new functions: **search engines**, which are more advanced tools capable of combining keyword searches with statistical algorithms for more targeted access to information, gained ground.
- ☒ the **third generation** of tools goes beyond these approaches. In the last few years, a new class of products and tools have been introduced that are based on the use of **linguistics** as a search and analysis instrument. Of particular importance within this category of tools are those based on **deep linguistic analysis (semantic engines.)** Semantic technology adds logical and sentence analysis and the disambiguation of polysemic terms (namely identification of the correct meaning of all words in a language that could have a number of different meanings) to basic linguistic analysis (grammar and morphology.) We thereby obtain a univocal, superior understanding of the meaning of the content of a document, bringing computerized tools closer to the processes used by human beings.

Linguistic, especially semantics-based, technologies stand out for their ability to go beyond **the limitations** that afflict **traditional search instruments** in terms of:

- ☒ lack of **precision, namely their inability to identify ONLY documents that are relevant based upon the selection criteria** (just think of the impact of the various meanings associated with a single word.)
- ☒ limited recall, or the inability to extract **ALL** relevant documents. It is impossible with traditional systems, for example, to extract all documents that address the concept for which the search is being made unless all terms related to the concept are inserted one-by-one as keywords.

These problems make it **difficult to “find”** the information being sought after. In addition, where information is available, its treatment is complex since there is little or no comprehension of meaning and the results are **not presented in a structured format**. Tools based on linguistic analysis are designed to reduce these distortions. They allow text mining of content (providing the user with a “query code” that approximates mental processes) and the presentation of results in a structured manner. **Semantics-based tools go further** since they also improve handling of information, both in the case of automated categorization (i.e., the use of taxonomies and content classification criteria) and text mining. They are able to understand the precise meaning of each word based on the context in which the term is used.

The greater capacity to comprehend content also allows semantic linguistic technology to be used for **question answering** tools. These enable users to submit natural language questions to the system to obtain a result that is no longer dictated by frequency or chance (search for all information that contains the word or phrase entered), but rather one that responds specifically to the query posed by the user.

Linguistic analysis tools and the benefits of knowledge management

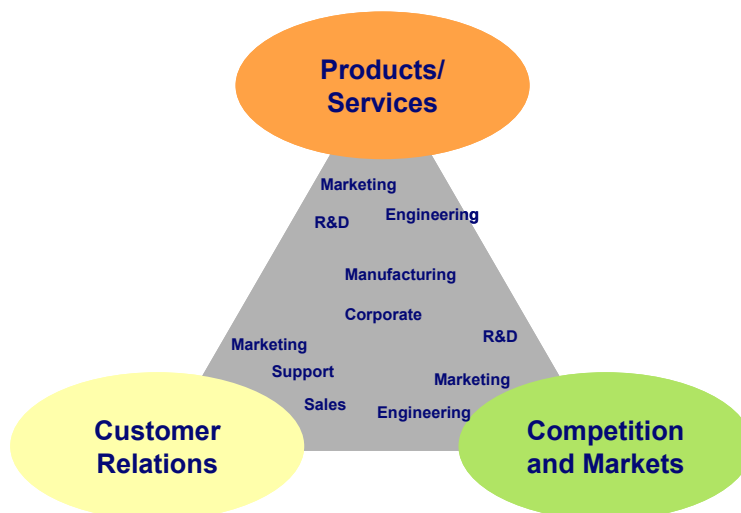
The third generation tools – based on deep linguistic analysis or semantic instruments – can **add value** by impacting three fundamental elements that make up the enterprise:

- ☒ **Technologies:** Thanks to their ability to understand, highlight and, if necessary, extract the most important unstructured content, these tools make it possible to exploit unstructured information. The effectiveness and productivity of investments in more traditional tools, such as business intelligence applications or customer relationship management tools, is thereby increased.
- ☒ **People:** the use of semantic linguistic analysis tools, permitting, for example, greater precision and recall in searches and more precise automated categorization of both internal and external information, increases the quality of work and the individual productivity of knowledge workers.
- ☒ **Processes:** the quality of the results of the analysis of unstructured information produced through the use of deep linguistic analysis tools permits greater automation of information management processes.

Further advantages include the possibility of increasing the number of users involved in processes and projects, inspiring cooperation and sharing of knowledge and keeping closer control over the market intelligence process. Figure 3 shows the organizational contexts in which significant returns can be gained from these types of initiatives.

FIGURE 3

The enterprise-wide scope of text mining tools



Source: IDC, 2006

Exploiting knowledge as a strategic asset obtained by the use of linguistic analysis tools can involve a variety of contexts:

- ☒ **Customer relations:** text mining tools can **improve** the management of customer relations. Organizations can adopt analysis techniques for unstructured information to improve the effectiveness of customer relationship management

(**CRM**) policies and tools. Take, for example, text-based notes produced as a result of call center conversations with customers: a text mining system could help identify problems or factors that are frequently overlooked by operational systems. In other cases, these tools can provide real-time solutions that the operator can offer the customer based upon the type of request made. Outside the company, semantic analysis of websites, blogs and other sources could provide information on specific customers, thereby improving pre- and post-sale processes.

- ☒ **Product/service excellence:** the processes that lead to the creation of a product or the delivery of a service can also benefit from the use of linguistic analysis tools. Efficient access to unstructured content and information can, for example, support the **study** and **design** phases of a product, allowing those in charge (of design or manufacturing) to exploit experience gained and to keep abreast of changes in the market. Generally, this sphere may encompass all advanced search and retrieval activities, such as monitoring internal sources of information, the search for and identification of correlations among legacy documents, etc.
- ☒ **Market intelligence:** text mining can support organizations in interpreting competitive **scenarios** and in adjusting business strategies. **Semantic web** tools, for example, enable enterprises to analyze the "invisible Internet" in order to glean useful information on competitors, markets, finance, etc. The intelligent analysis of external sources not only allows companies to tap into technological and sectoral dynamics, but also, more generally, to protect sensitive information more effectively, thereby safeguarding their **intellectual property**.

COGITO: Expert System's language technology for semantic intelligence

One of the available market solutions is the semantics-based **COGITO** platform developed by **Expert System**, which searches, categorizes, analyzes and extracts unstructured information and texts. **COGITO** is the product of Expert System's years of experience in applying linguistics to information technology. In the case of **COGITO**, the conceptual map - a typical part of this type of tool - is founded on a semantic network called "Sensigrafo". The **COGITO** modules cover very different areas in the world of information, cutting transversally across the internal content of organizations and external content in sources such as the Internet.

COGITO is adaptable to a wide variety of corporate situations and is designed to provide the benefits of "semantic intelligence" within the main areas described in Figure 3:

- ☒ The "**Semantic Search**", "**Discover**" and "**Categorizer**" modules represent the core of the **COGITO** platform and provide a series of "content management and access" functions such as indexing, searching, extracting, normalizing, categorizing and distributing information.
- ☒ The "**Contact**" module is dedicated to managing customer care services over the Internet or via text messaging, making the most of a company's knowledge base.
- ☒ The "**Intelligence**" module analyzes open external sources in real time in order to provide support to military defense, anti-terrorism and public safety as well as corporate strategic initiatives for analyzing and monitoring competitors and the market.

COGITO's features appear to meet the needs of a rapidly-changing market, which requires organizing and extracting benefits from the management of the unstructured information that fills internal IT systems and organizations' ecosystem of external relations.

In the three most critical areas for unstructured information management, the application of semantic intelligence generates clear advantages even compared with basic linguistic analysis tools. The benefits can be summarized as follows:

- ☒ **Semantic searching** makes it possible to limit the negative impact of the information explosion, thereby enabling users to bypass the problem of quantity and instead focus directly on **more qualitative issues**.
- ☒ Thanks to the greater comprehension of content and the disambiguation of terms, semantic-based categorization allows to reach a high-level of quality, comparable to that achieved by manual searches, even for classifications based on thousands of categories and sub-categories.
- ☒ Automated extraction activities also benefit from this greater understanding, allowing identification not only of specific entities but also relationships between concepts and information found in various parts of the same document or in different documents that, at first glance, are not obvious.

Case Study: EniTecnologie

EniTecnologie, as the Corporate Technology Company of the ENI Group, is one of the top industrial research centers. EniTecnologie engages in technological innovation in order to maintain its competitiveness over the short, medium and long term. To achieve these goals, it operates in a manner consistent with ENI's strategies in all aspects of the innovation process, from technological monitoring to scenario studies, applied research, technological development and evaluation, and the technology transfer of innovation to businesses.

Staying at the forefront of technology and generating value through innovative activity are an integral part of ENI's vision.

It is of fundamental importance for EniTecnologie to be able to monitor competitors and detect even the faintest signals of change in the relevant final markets. Of equal importance is the effective, efficient management of the data and knowledge products generated within the Group.

The principle of "**knowledge as an asset**" is therefore one of the pillars of EniTecnologie's knowledge management strategy, which for many years has focused on content analysis methodologies and techniques, and on the development of technological support platforms.

Towards the mid-1990s, the arrival of the Internet represented a discontinuity for EniTecnologie's content management policies, with EniTecnologie placing increasing emphasis on enriching its Web component. A second, more recent, key development was the arrival of linguistic analysis tools that revolutionized approaches to understanding and extracting information.

An important component of EniTecnologie's knowledge activities has been accessing and analyzing documents containing technological and patent information, a highly "**document-intensive**" activity.

With COGITO by Expert System, EniTecnologie found an especially incisive tool for handling the requirements of "unconventional" text-based searching. The application of a semantics-based search model made the tool much more user accessible, thanks to its capability of processing and presenting results in accordance with the conceptual maps principle.

With COGITO, a company can bring the tool "**closer to the mind**" of the user in order to unveil the links and relationships within unstructured information. For example, the system allows users to query patent documents by following an approach similar to their own mental processes, which are typically geared to identifying "solutions" to "problems" within the document. In turn, this logic fits the characteristics of patent documents which, by their very nature, address innovation using a "problem vs. solution" criterion.

In addition to its functional semantic and text mining features, the COGITO platform was also selected because of its flexibility and its adaptability to EniTecnologie's needs, unlike the products of competitors, which were more closed and could not be parameterized as extensively.

The **strong potential** of the COGITO tool prompted EniTecnologie to **quickly expand the project**. Originally set as an experiment including just a few users, the initiative evolved and rapidly expanded thanks to the positive feedback deriving from the practical and valuable "user experience."

The perceived benefits of the process led EniTecnologie to the subsequent development of an intranet **portal** that, among other things, enables functions such as collaboration, internal communication, filing and searching documents using metadata (keywords, author, etc.) and, especially, filing and searching based on knowledge classifications (taxonomies) and on recognition of concepts, meanings and complex linguistic structures.

The collaborative effort between the Expert System team and the EniTecnologie working group was a fruitful one throughout the entire process, which produced a strong internal commitment to problem-solving and to introducing a permanent learning environment, not only in the initial implementation of COGITO, but above all in the go-live and maintenance stages.

The **benefits** of adopting **COGITO** as the knowledge management platform range from the effectiveness of **semantic searching** – providing **targeted responses** and extracting "**meaning**" from text documents – to the more complete **organization** of the company's information assets, as well as ensuring security and compliance, fostering knowledge-sharing and cooperation among all users, and, especially, giving tangible support for **business processes** that are developed in the pursuit of the strategic mission.

Today, its internal knowledge assets are managed by the **COGITO Categorizer** and **COGITO Semantic Search** modules, which have already been incorporated in EniTecnologie's intranet portal.

Meanwhile, integration of the component for analyzing information from external sources has begun. The knowledge map created thanks to the COGITO platform was

completed with the **Intelligence** module, which monitors **external information**. **EniTecnologie's knowledge workers use COGITO Intelligence to gather information that is difficult to extract** without using the semantic approach, enabling rapid and complete organization of the company's information resources. External websites and portals can be examined using standard, parameterized or user-defined criteria drawn from specific queries based on the user's experience and judgment. This avoids the problem of "excess information" and focuses on content with value in order to analyze and process any information that could have impact on analyzing technology trends and identifying the technology and innovation strategies of its competitors.

Conclusions: capitalizing on knowledge

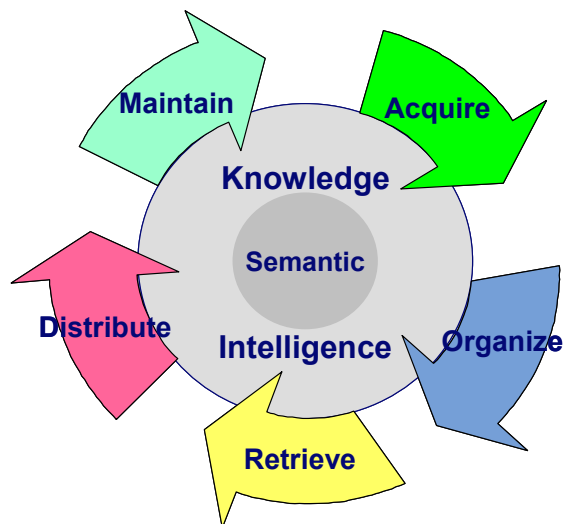
Effective corporate knowledge management, which must optimize the management of unstructured information, enables enterprises to create value by fully exploiting their resources. They create a “**learning environment**” that fuels a virtuous circle that facilitates and expands this value creation. New knowledge, in turn, will contribute to fueling existing, distributed knowledge.

Put simply, the operational elements of unstructured information management based upon the principle of “knowledge capitalization” can be summarized as follows (Figure 4):

- ☒ acquisition;
- ☒ organization and storage;
- ☒ access and retrieval;
- ☒ circulation and distribution;
- ☒ maintaining significance (from the perspective of value) and updating.

FIGURE 4

A semantic core fuels the virtuous circle of knowledge



Source: IDC, 2006

Solutions based on semantic linguistic analysis tools that optimize the management of unstructured information play an important role in triggering a learning process that extracts value from content distributed throughout various layers of IT architectures. This importance is certain to grow in the future in proportion to the quantity of unstructured information found in organizations. Finally, we must not neglect the process of **customization**, which allows tools to be adapted to fit the specific needs and features of the enterprise.

Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2006 IDC. Reproduction without written permission is completely forbidden.

