R20 CONSULTANCY

Extending Business Intelligence with Text Exploration Technology

A Whitepaper

Rick F. van der Lans
Independent Business Intelligence Analyst
R20/Consultancy

June 2013

Sponsored by

INTERSYSTEMS

Table of Contents

1 Management Summary

An excerpt from the book *Mining the Talk* by Spangler and Keulen[1]:

> *"People are talking about your business every day. Are you listening?*
> *Your customers are talking. They're talking about you to your face and behind your back. They're saying how much they like you, and how much they hate you. They're describing what they wish you would do for them, and what the competition is already doing for them. They are writing emails to you, posting blogs about you, and discussing you endlessly in public forums. Are you listening?"*

A wealth of information is hidden in the vast amounts of data being created every day. The challenge for every organization is to extract valuable business insights from this mountain of data that allows it to, for example, optimize its business processes, improve the level of customer care it offers, personalize products, and improve product development.

Almost every organization knows how to turn their own structured data that has been collected by business processes through the years into valuable insights. Countless reporting and analytical tools are available to assist them. But what about the textual data that has been gathered in emails, document management systems, call center log files, chat or instant messaging transcripts, and voice transcripts from customer calls? And what about all the external textual data, such as blogs, tweets, and Facebook messages? Most organizations have barely scratched the surface with respect to analyzing textual data. This is a missed opportunity.

Most organizations have barely scratched the surface with respect to analyzing textual data.

This whitepaper describes the latest technology for analyzing textual data, so-called *text exploration technology*. Text exploration enriches the palette of technologies already deployed in business intelligence environments.

Text exploration enriches the palette of technologies deployed in BI environments.

Text exploration technology can be summarized with three terms: no advance preparations, unguided analysis, and self-service.

- **No advance preparations:** There is no need to develop thesauri or ontologies in advance of the analysis work. If the texts are available, it's possible to start the analysis straightaway without any preparations. Even if these are new texts and cover a new domain.

- **Unguided analysis:** Analysts are able to invoke the text analysis technology without having to specify a goal in advance. This text exploration technology is able to analyze the text in an unguided style.

[1] S. Spangler and J. Keulen, *Mining the Talk, Unlocking the Business Value in Unstructured Information*, IBM Press, 2008.

- Self-service: Analysts can invoke text analysis techniques without help from IT experts.

Almost every industry can benefit from deploying text exploration, especially those industries in where storing text is crucial for business operations, such as advertising, healthcare, legal, pharmaceuticals, publishing, and real estate.

Besides explaining text exploration technology, this whitepaper introduces the text exploration technology of InterSystems called iKnow. iKnow is truly text exploration technology that uses a unique approach to text analysis. Instead of identifying words and two- or three-word phrases, it breaks texts into sentences, and discovers the concepts and relations in the sentences.

iKnow is a truly text exploration technology with a unique approach to text analysis.

It's openness makes it possible for most reporting and analytical tool to invoke iKnow's text exploration technology. In addition, it allows home-made applications to access iKnow. It's integration with InterSystems DeepSee allows business analysts to enrich their more classic analytical capabilities with text analysis techniques in a fully integrated way. With Caché as data storage mechanism and as gateway to other data sources, such as SQL databases and big data stores, iKnow has access to a wide range of data sources and potentially massive amounts of text.

2 How Business Intelligence can Benefit From Text Exploration

In the world of business intelligence a wide range of tools and technologies exists that allow business analysts to study data in many different ways. This section explains the unique and distinguishing features of *text exploration technology* and how it enriches the palette of tools already deployed in business intelligence.

What is Business Intelligence? – Business intelligence is about supporting and improving the decision making processes of an organization. Boris Evelson of Forrester Research[2] defines business intelligence as follows:

> *Business intelligence is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making.*

From this definition can be derived that business intelligence is not a tool, not a technology, or some design technique, but it's everything needed to transform and present the right data in such a form that it leads to valuable *business insights* and improves the decision-making processes of an organization.

[2] B. Evelson, *Topic Overview: Business Intelligence*, November 21, 2008.

Reporting Tools Versus Analytical Tools – The tools used by decision makers to study and analyze data can be classified in different ways. A traditional classification is by dividing all the tools in *reporting tools* and *analytical tools*. Reporting tools show what *has* happened and possibly what *is* happening. Although the shown data may have been transformed, processed, and aggregated, still, the data shows the past and current situation. Typical examples of questions answered with reporting tools are "Show the total revenue per sales region for the last two weeks" and "Present a 360° report of a particular customer." Dash boards are also examples of reports; OLAP tools, with which users can look at data from every angle and at every level of detail, belong to this category as well, as do batch reports.

Reporting tools show what has happened and what is happening.

Analytical tools, on the other hand, are used to find out what may or can happen. They use techniques such as predictive modeling, simulation, and forecasting to come to valuable insights. The result of analytics is usually not (aggregated) data, but a set of rules. Examples of such rules are "When a customer buys Cola and chips, there is a 75% chance he buys dip sauce as well" and "The most efficient route to deliver goods to a particular set of shops is the following."

Analytical tools show what can or may happen.

Self-Service Tools – Another way to classify tools is by indicating whether they're designed for self-service or not? With *self-service tools* (sometimes called *data discovery tools*) users can develop their own reports, hence the name. The user-friendly, intuitive, and graphical interfaces make it possible that users develop and change reports very quickly, without having to get technical help from the IT department or a BICC department.

With self-service tools users can develop their own reports without technical help from the IT department.

With *non-self-service tools*, reports are pre-built by IT experts. Before decision makers can work with the reports, these experts have to design and develop the reports. They will interview users to understand what they need and create a fitting solution. Users are still able to interactively play with the data, but there are usually restrictions to what they can do.

Self-service tools are often associated with the term *agility*. The reason is that with these tools users can develop and change their reports themselves, allowing users to react more quickly to new and urgent demands. More and more reporting and analytical tools are becoming available that belong to the category of self-service tools.

Already in 2011, Gartner[3] predicted a rosy future for self-service tools:

> *Vocal, demanding and influential business users are increasingly driving BI purchasing decisions. They're choosing easier to use data discovery tools [their term for self-service tools] over traditional BI platforms—with or without IT's consent.*

Self-service tools look to be the next big wave in business intelligence. In a January 2010 report, Gartner pointed to the growing demand among organizations for a "data discovery tool architecture" that provides end users with data and reports and enables them to navigate and

[3] Gartner, *Gartner Forecasts Global Business Intelligence Market to Grow 9.7 Percent in 2011*, February 2011.

visualize data in a "surf and save" mode. Data views can be stored for reuse or sharing. The self-service tools bring BI information to non-technical users; high-level analysts who need ad-hoc reports straightaway can benefit as well.

In an interview in 2011, James Kobielus[4] of (at that time) Forrester Research described the reason for the success of self-service tools as follows:

> *… many IT staffs face growing backlogs of information requests from increasingly frustrated end users. But self-service features … can help. They buffer less-technical end users from the complexities of the underlying data infrastructure. This frees up IT professionals from having to spend "an inordinate amount of time" responding to requests for new data, new views or updated report formats*

Exploration Tools – Most tools impose analytical and reporting restrictions on users. There can be a limit to the data sources they can use, or the level of detail they can see. In addition, even organizing data in a particular star or snowflake schema limits the reporting capabilities, because these structures are designed for a certain set of queries and reports. If no relationships between tables exist, users can't use them for analysis. With such tools, unexpected or unplanned user queries may be hard or impossible to answer. Note that these limitations apply to self-service tools as well.

With exploration tools users can roam all the data freely-without any restrictions.

Normally, what tools allow users to do is sufficient to satisfy their data and reporting needs. All these tools work fine if the information, reporting, and analytical needs of the users are well defined. However, there are users whose questions and information needs are not that crystal clear and well-defined. This group of users need tools that allow them to roam all the data freely—without any restrictions with respect to how data can be viewed, reported, or analyzed. Such tools are called *exploration tools* (sometimes called investigative tools).

Many organizations can benefit from exploration functionality. Let's use an example to illustrate where an exploration tool can be deployed usefully. Imagine that a truck owned by a retail company was supposed to deliver several pallets of a particular type of soda at several stores in Boston before opening time. Unfortunately, the truck has engine troubles and has been parked at the roadside. The challenge for the manager on duty is to find an alternative way of getting the soda to these stores. A solution could be to send an empty truck to the stranded truck, rack the pallets, and bring them to the stores. But are there empty trucks available in the area? Another alternative might be to check whether there are other stores in the area from which soda can be retrieved? Or is it better to just send a new shipment? Whatever the solution, this manager needs access to all kinds of data. In addition, because the solution can be anything, a manager must have access to an exploration tool that can help to find the best solution. He should be able to freely query the available data on current truck positions, stock levels, shipments, orders, and so on. Hopefully, somewhere in all that data, the manager on duty discovers the best solution.

There are many more examples of situations where exploration can be used. Think about a hospital environment where a patient is brought in who urgently needs a particular type of

[4] Elisabeth Horwitt, *Self-Service BI Catches On*, Computerworld US, January 2011.

operation. Unfortunately, all operation rooms are occupied. What to do, find another hospital? At what time do the current operations end? Another example is an airport that has to close temporarily because of heavy fog conditions. What do you do with the airplanes that are supposed to land there? Or what do you do when the cargo doors of a recently landed airplane don't open? What do you do with the luggage? In both situations, users should be able to freely navigate and explore all available data.

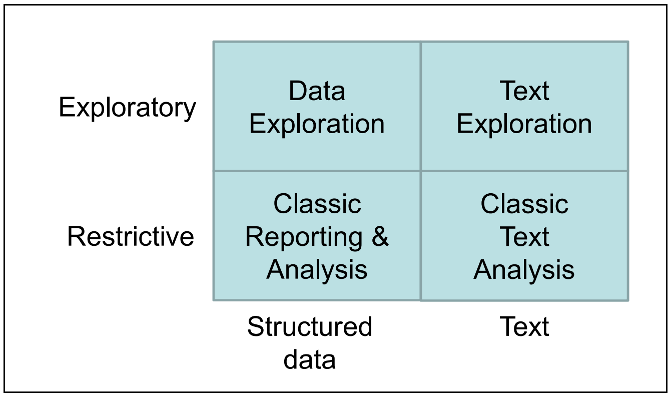
To summarize, in all above situations, what's needed is a tool that users can use to freely analyze data without the IT department having to predefine reports, tables, or relationships. Exploration tools can be summed up with the terms: self-service, easy-to-use, no advance work, no restrictions, and minimal IT involvement.

Text Exploration Tools – Most of the existing reporting and analytical tools, including the exploration tools, do a great job when analyzing data that consists of numbers, strings, and dates. Unfortunately, they can't really handle text intelligently. The problem is that text cannot be easily aggregated, summarized, and filtered, nor can regression analysis be applied to it. Before text can be analyzed and studied, it must be processed. It must be decomposed in sentences and words.

Tools and technologies for analyzing text have been around for quite some time. In fact, research into automatic analysis of text goes back to the 1950s. Today, many of the tools and technologies, such as Search, are being used daily by millions of people.

All tools and technologies for text analysis can also be characterized as being exploratory or not. However, most of them are non-exploratory, because before text can be analyzed, so-called thesauri, ontologies, and lists of synonyms have to be setup. These are required to improve the results of text analysis; in fact, the better the thesauri and ontologies, the better the analysis results. Unfortunately, developing them takes quite some time and has to be done by domain experts. This limits the free exploratory capabilities of such text analysis tools.

As there is a need for exploration tools for structured data, there is a need for tools that analyze text in an exploratory way: the *text exploration tools*. Figure 1 shows how these text exploration can be positioned.



Exploratory	Data Exploration	Text Exploration
Restrictive	Classic Reporting & Analysis	Classic Text Analysis
	Structured data	Text

Figure 1 *Positioning of text exploration tools.*

Many tools are available for three of the four quadrants in Figure 1. The text exploration quadrant is still somewhat vacant. Currently, most tools for analyzing text (right-hand bottom side) require a lot of work in advance. As indicated, ontologies, thesauri, and so on, have to be setup. For many situations this isn't a big issue. For example, if we want to study the frequency of use of certain words in historical books, there is probably no need for urgency. Yet, there are many situations in which users do need to analyze text right then and there. The same applies to new text for which there is no time to develop an ontology or thesaurus. This is where text exploration comes in.

With text exploration tools text can be freely analyzed without any preparations in advance.

Now that organizations have to deal with more and more information in the form of text, the interest for data exploration is growing proportionately. The coming sections dive deeper into the world of text exploration.

3 Textual Data is Available in Abundance

In the early days of computing, there wasn't that much text stored in enterprise systems. Nowadays, textual data is largely responsible for the explosion of new data stored every day. This is not only because of all the textual data that's flooding the internet, it's also because of the textual data available in enterprise systems, and there is probably more of that than most people think. Here are some examples of internal enterprise systems that store and manage text:

- Email systems; studies show that 144 billion emails are sent per day
- Contract management systems
- Document management systems that contain scanned correspondence and other documents
- Customer visit reviews
- Call center log files
- Chat or instant messaging transcripts
- Voice transcripts from customer calls

As indicated, besides all the textual data stored in the internal enterprise systems, externally massive amounts of textual are being produced and stored. Consider the following:

- Facebook: over 200 million messages are sent per day.
- Twitter: more than 340 million tweets are sent per day.
- SMS: 7.8 trillion SMS text messages were sent in 2011.

Next, there are all the blogs and websites on which text is published, and the list goes on. In fact, it's on the internet where the amount of new textual data is staggering.

Note: One of the hottest trend in the IT industry is undoubtedly *big data*. Textual data is a clear example of big data. In fact, textual data is one of the key contributors that makes big data big.

Textual data is one of the key contributors that makes big data big.

Conclusion, textual data is available in abundance. In other words, quantitatively textual data is ok. However, the big challenge is how to extract useful information from it? How can it be analyzed and turned into valuable business insights? If this isn't done properly, valuable information will go missing and organizations won't profit from the potential benefits.

4 Text Analysis Explained

Use of Text Analysis Today – Organizations can benefit from analyzing textual data. For example, an insurance company may want to analyze all the contracts (textual documents) to find out how many of them expire within one year. A hospital may be interested in analyzing the descriptions written by specialists and included in patient files to discover patterns with respect to allergic reactions to medications. An electronic company may want to analyze messages on Twitter to find out if their products are mentioned and whether they're positive or not—sentiment analysis. Consider also all the text in emails sent out and received by an organization, all the information on the web, and in social media networks. Transcripts of call center log files can be analyzed to determine whether there are popular questions, or whether the last couple of weeks specific products have been mentioned more than usual.

Organizations are analyzing text today. For example, on May 4, 2011, the USA Today reported that Wall Street traders mine tweets for investment clues. They monitor and decode the words, opinions, rants, and even keyboard-generated smiley faces posted on social-media sites. In other words, they're analyzing text.

In many countries during political elections, the media analyzes tweets and other social media sources to determine whether and how certain political parties or frontrunners are discussed. Besides showing scores, these studies even influence voting behavior—who wants to vote for a loser?

IBM's Watson[5] can be seen as one of the cutting-edge technologies for text analysis. Watson is an artificially intelligent computer system that can answer questions posed in natural language. Its claim to fame is that in 2011 it competed in the game show called Jeopardy! on US television and beat two former winners. Before the game, 200 million pages of structured and unstructured content were loaded.

Note: On April 24, 2013, USA Today reported a hoax. A hacked Twitter account of a major news organization was used to inform the world that there had been two explosions in the White House injuring Barack Obama. The effect was that Wall Street went into panic mode, sending the Dow Jones Industrial Average into free-fall and erasing nearly $200 billion off the broader market's value. The lesson to be learned here is that when external text is analyzed, we have to be careful with the quality and authenticity.

What Exactly Do We Mean With Analyzing Text? – It's easy if we only want to know how many words appear in a text, or how often a word appears in a text. This can be solved with a simple algorithm. But what if we want to apply more complex forms of text analysis, such as:

[5] Wikipedia, *Watson*, see http://en.wikipedia.org/wiki/Watson_(computer), May 2013.

- How often do particular symptoms and medications appear together in patient files?
- Does a text have a positive or negative attitude?
- How many texts deal with the bankruptcy of Bank X?
- For each month, show the number of texts dealing with brain surgery.
- Which concept appears most often in texts together with the concept credit card fraud?
- Which book is, with respect to concepts used on the content of the book, most equal to Jeff Shaara's *Gods and Generals*, and which one is most different?
- How do we identify the characteristics of customer calls that resulted in escalation.

These questions are much harder to handle. For example, how do we determine whether a text has a positive sentiment? How do you "measure" the difference between the contents of two books? This is the type of question for which text analysis is used.

Analyzing text can also be defined as deriving structured data from unstructured data (in this case the misnomer unstructured data is used for text). For example, when a text is analyzed based on whether it's positive or not, the result is a structured data value: the value yes or no. The answers to the first and fourth questions above also lead to structured data. The advantage of deriving structured data is that this newly created structured data can be combined easily with other structured data sources.

Note: In this whitepaper, the term text analysis is used. This is done to stay in line with the world of business intelligence. For text analysis other terms do exist, such as text mining, natural language processing, information retrieval, and search technology. Whether these terms should be seen as synonyms or as overlapping technologies, falls outside the scope of this whitepaper. What can be said is that, whether they are synonyms are not, they are very much related. They all focus on trying to process text.

The Prehistory of Text Analysis – The history of analyzing text goes way back. An article by Dr. Vannevar Bush[6] on information retrieval, which appeared in 1945 (more than 65 years ago) may be considered the beginning of literature on mechanized information retrieval. In this article Bush describes an imaginary machine, the "Memex", in which research workers could store their personal library together with other documents and from which they would be able to select all references relevant to the information desired.

It's interesting to know that one of the more well-known researchers and experts in text analysis in the 1960s, Hans Peter Luhn, is also regarded as the man who first coined the term business intelligence[7]. His article on the automatic derivation of information retrieval encodements from machine readable texts[8] is still considered a landmark. This means that business intelligence and text analysis have always belonged together.

Business intelligence and text analysis have always belonged together.

Indexes and Thesauri – The first developments in trying to understand texts were based on *indexing*. Indexing texts means that terms are selected from the documents that provide a sufficient

[6] V. Bush, *As We May Think*, Atlantic Monthly, Volume 176, 1945.

[7] H.P. Luhn, *A Business Intelligence System*, IBM Journal, Volume 2, Issue 4, October 1958.

[8] H.P. Luhn, *The Automatic Derivation of Information Retrieval Encodements from Machine Readable Texts*, in Information retrieval and machine translation, Editor A. Kent, Volume 3, 1961.

indication of the subject-matter of that document to ensure that it can be retrieved using a specific query. Indexing, however, has its limitations. First, developing an index is time-consuming—which words should be indexed? Second, if the right terms are not indexed, some important and relevant texts may not be found, and even incorrect texts may be found.

Quite early, the concept of a *thesaurus* was introduced. With a thesaurus the relations between terms are defined. In a way, a thesaurus can be seen as an intelligent index. The result of deploying a thesaurus is that a more accurate set of texts is found.

But building up and managing a thesaurus is also time-consuming. For this reason, already in 1961, Luhn introduced the notion of automatic derivation of thesauri. His proposal was a technique based on systematic reduction of text based on statistical properties. Besides the fact that this is time-consuming, a thesaurus must be kept up to date. New words are introduced, new domains are introduced, and so on.

Techniques for Stripping Text – Besides the use of thesauri, many other techniques are used to analyze texts:

- Eliminating non-content-bearing words, called *stop words*, such as *the* and *an*.
- Combining words using lists with synonyms. For example, bicycle and bike should be seen as synonyms.
- Identifying objects and subjects. For example, the words Obama and the president of the USA refer to the same subject.
- Removing repetitive phrases. These are pieces of text that appear on every page, such as copyright notices, chapter titles, and author names.
- Removing terms that occur infrequently, because they have little value, and removing terms that occur too frequently—the ones that appear in every document.
- Suffix stripping algorithms that bring terms back to one with a common stem. For example, the terms connect, connected, connecting, connection, and connections, all have the same stem, namely connect. Suffix stripping is not as easy as it sounds. For example, an sloppy algorithm may strip the terms relate and relativity erroneously to the same term.

5 Text Exploration Technology

The Need for Text Exploration – As indicated in the previous section, most of the text analysis techniques require work in advance. For example, thesauri and ontologies must be developed before the analysis can start. Such tools are very useful, but they can only be used if there is time to do all this work. What if a new and urgent question arises and in the thesaurus this hasn't been catered for? Or what if new texts become available for analysis and questions have to be asked right away?

Also, with most text analysis techniques the goal of the analysis exercise must be clear in advance. In other words, the tool is guided by the analyst. For example, search technology requires that one or more words are entered first. Another example is when patient files are analyzed to discover new insights with respect to the effect a particular medication has on patients with diabetes. As can be imagined, even when the same patient files are analyzed, a different thesaurus may be needed when the goal is to look for historical patterns in side effects after surgery. A thesaurus limits the analytical freedom and thus limits potential outcomes. The flat box in Figure 2 represents the amount of work to be done to analyze text plus it shows the time needed for the preparations, which is considerable in many projects.

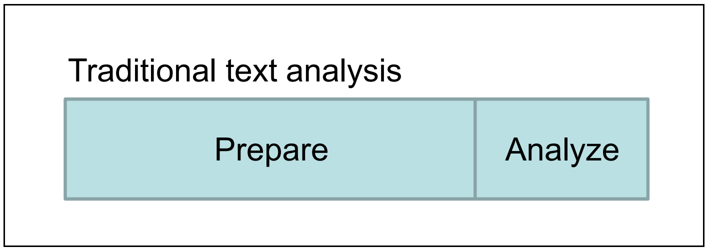


Figure 2 *Most text analysis tools require a considerable amount of preparations in advance before analysis can start.*

Technology is needed that allows text to be analyzed without the need to do all that work in advance. This is what we call *text exploration*.

With text exploration text can be analyzed without work in advance.

A hospital environment is a good example of where data exploration can be used. Imagine a patient is brought to the emergency room. If doctors must act quickly, they probably don't have time to read the full patient file. What they want is a summary that shows all the important aspects related to the patient. Is he diabetic? Does he usually have a high blood pressure? What kind of medication is he taking? Has he been here before? This requires text analysis on the spot. The analysis should also be unguided, because the doctors know nothing of this patient.

Another example is analyzing tweets. Everyday new words (abbreviations in many cases) and hash tags are invented. It would be undoable to constantly update a thesaurus. Furthermore, is there time to develop one?

Many situations exist in which there is no time for all this preparation work. Here, text exploration is needed to give the desirable business insights; see Figure 3. Compare the size of the box called Prepare in this figure with that in Figure 2. Basically, the difference in size is the difference between more classic forms of text analysis and text exploration.

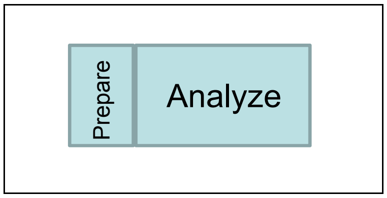


Figure 3 *Text exploration tools require less preparation work than more traditional text analysis tools.*

The Three Requirements for Text Exploration – To summarize, data exploration is a form of text analysis that meets the following three requirements:

- **No advance preparations:** There should be no need to develop thesauri or ontologies before the analysis work can be started. It should be possible to start text analysis straightaway without any preparations. Even if this is a new text covering a new domain.

- **Unguided analysis:** Analysts should be able to invoke the text analysis technology without having to specify a goal in advance. The text analysis technology must be able to analyze the text in an unguided style.

- **Self-service:** Analysts are able to invoke the text analysis without help from IT experts, although connecting the tool to particular data sources may require some assistance.

6 InterSystems Corporation and the InterSystems Platform

InterSystems Corporation – InterSystems was founded in 1978. After first focusing on Mumps applications, they released a popular database server called Caché in 1997. Later on, in 2003, they added an integration tool called Ensemble. These modules still form the core of the InterSystems software.

In 2011 InterSystems acquired iKnow. This product was developed in Belgium by a small startup founded by Michael Brands and Dirk van Hyfte. The goal of the developers was to create technology for analyzing text in a new and revolutionary way. Their psychiatric, medical artificial intelligence, and linguistic background allowed them to think out of the box. The result was the product iKnow. Very soon after this product was commercialized by the startup, it was acquired by InterSystems that now distributes and promotes the product worldwide as an integral part of the InterSystems Platform.

The InterSystems Platform – The InterSystems Platform consists of the following set of integrated modules:

- **InterSystems Caché®** is a hybrid, object-relational database server that supports SQL and an object-oriented language called ObjectScript.

- **InterSystems Ensemble®** is an application integration platform that supports business process modeling and built-in dashboards to view the status of running business processes.

- **InterSystems DeepSee®** is an interactive OLAP tools for analyzing data.

- **InterSystems iKnow™** is the module for analyzing text and is fully integrated within the InterSystems Platform.

7 How Does iKnow Work?

The Classic Approach of Text Analysis – Tools for analyzing text usually try to identify important concepts in sentences. For example, in the sentence *The enterprise search market is being reshaped by new consumer experiences*, the key concepts are *enterprise search market* and *new consumer experiences*. Most text analysis tools try to locate these concepts by looking at individual words, which results in the words *consumer, enterprise, experience, market, and search*. They are considered key concepts in this text.

Some tools search for two-word phrases and even three-word phrases. The result of this approach, however, can be that words are "connected" that should not be connected. Take the following sentence as an example: *Michael Phelps breaks a world record*. If two-word phrases are identified, the result contains the concepts *Michael Phelps* and *Phelps breaks*. Now, the first one is probably a useful one, but the second isn't. And if we would search for all two-word phrases in the first sentence, we get *enterprise search* and *search market*, but not *enterprise search market.* This more classic approach doesn't guarantee that the words that are linked together form the right concept.

In addition, to make sense of the sentences, developers have to build up thesauri and ontologies. This can be quite a lot of work and requires domain knowledge. For every domain a new thesaurus and ontology must to be setup. In most situations, work on this will never end, because the use of words changes over time. New terms are being introduced and the meaning of words can change. As an example take tweets, every day new important hash tags are being introduced. Also in the BI domain new terms are introduced. Who had heard of the term big data a few years ago?

iKnow's Approach and the Three Requirements for Text Exploration – The approach taken by iKnow to analyze text is different from many other approaches. iKnow breaks texts into sentences, and sentences into concepts and relations. Breaking sentences is done by first trying to identify the relations in a sentence. Verbs can represent relations between concepts in that sentence, but other language constructs can signify relations as well.

iKnow analyzes text by identifying concepts and relations in sentences.

By starting to identify the relations, iKnow has a better chance of discovering the desired concepts. For example, in the sentence *The programmer found bugs*, iKnow considers the verb *found* to be a relation between the concepts *programmer* and *bugs*. In iKnow this is called a *concept-relation-concept sequence* (CRC). Note that iKnow automatically discards all the stop words from sentences, such as *the*, *an*, and *he*.

As indicated, other language constructs can indicate a relation. For example, in the sentence snippet *Mammals, such as elephants and lions …* a relation exists between *mammals* and *elephants* and one between *mammals* and *lions*. Another example is the sentence *I like the car in the showroom*. Here, the word *in* represents a relation between the concepts *car* and *showroom*. iKnow has been designed to recognize many different language constructs that can identify relations.

If the concepts and relations consist of multiple words, iKnow still recognizes them. For example, in the sentence *The enterprise search market is being reshaped by new consumer* *experiences,* iKnow discovers that the verb clause *is being reshaped by* represents the relation between the two concepts *enterprise search market* and *new consumer experiences*. These two concepts are called *concept-concept pairs* (CCs).

iKnow is not restricted to analyze simple sentences consisting of CCs and CRCs. It can handle more complex sentence structures consisting of multiple CRCs. These are called CRC sequences.

Note: This approach to text analysis works for several languages, including Dutch, English, French, German, Portuguese, and Spanish. Japanese and Russian are in development.

Results of iKnow's Text Exploration Approach – When iKnow has analyzed all the texts and sentences, and has identified the concepts and relations, it returns the following set of measures that can give valuable insights in the texts:

- **Frequency of a concept:** The frequency is the number of times a concept appears in a text. Note that this is not the same as the frequency of a word, because a concept can consist of multiple words.

- **Dominance of a concept:** The dominance of a concept indicates how important it is in the text. This measure takes into account the frequencies of the words of the concept in the text, their positions in the concept, and the position of the concept in the sentence.

- **Proximity:** This is a measure of the "semantic distance" between two concepts in a text. iKnow builds a proximity "cluster" for each concept in the corpus. The cluster lists the other concepts that appear in sentences within this cluster, with a proximity score for each of the concepts. This gives a measure of how closely related a concept is to every other concept in the text.

Note: Although iKnow doesn't need thesauri, ontologies, and lists of synonyms to be able to analyze text, the above returned measures can support in the creation and development of them.

iKnow for Text Exploration – Why is iKnow suited for text exploration? Section 5 describes the three key requirements for text exploration: no advance preparations, unguided analysis, and self-service. iKnow supports these three requirements:

- **No advance preparations:** iKnow doesn't require nor support the development of thesauri and ontologies. It can analyze text coming from a domain or industry it has never analyzed before, and is still able to discover the important concepts.

- **Unguided analysis:** iKnow does not need a goal. It does not, like search technology for example, need a search term before it can analyze the text. iKnow can analyze text in an unguided style. The result can be studied by the analysts and those results can trigger them to search in a certain direction.

- **Self-service:** Analysts can use the intuitive DeepSee module (more on this module in Section 8) to invoke all the text analytical features of iKnow. DeepSee can be categorized as a self-service tool that allows users to develop their own reports and do their own analysis without the help from IT experts.

8 The Architecture of iKnow and the InterSystems Platform

Figure 4 shows the overall architecture of the InterSystems Platform and the position of iKnow in this architecture. iKnow operates as a module within the InterSystems Platform. Note that the InterSystems Platform should not be seen as a set of loosely coupled modules that form some kind of software stack in which each module can be installed independently. On the contrary, together the modules form one integrated tool. In other words, developers experience the InterSystems Platform as one tool.

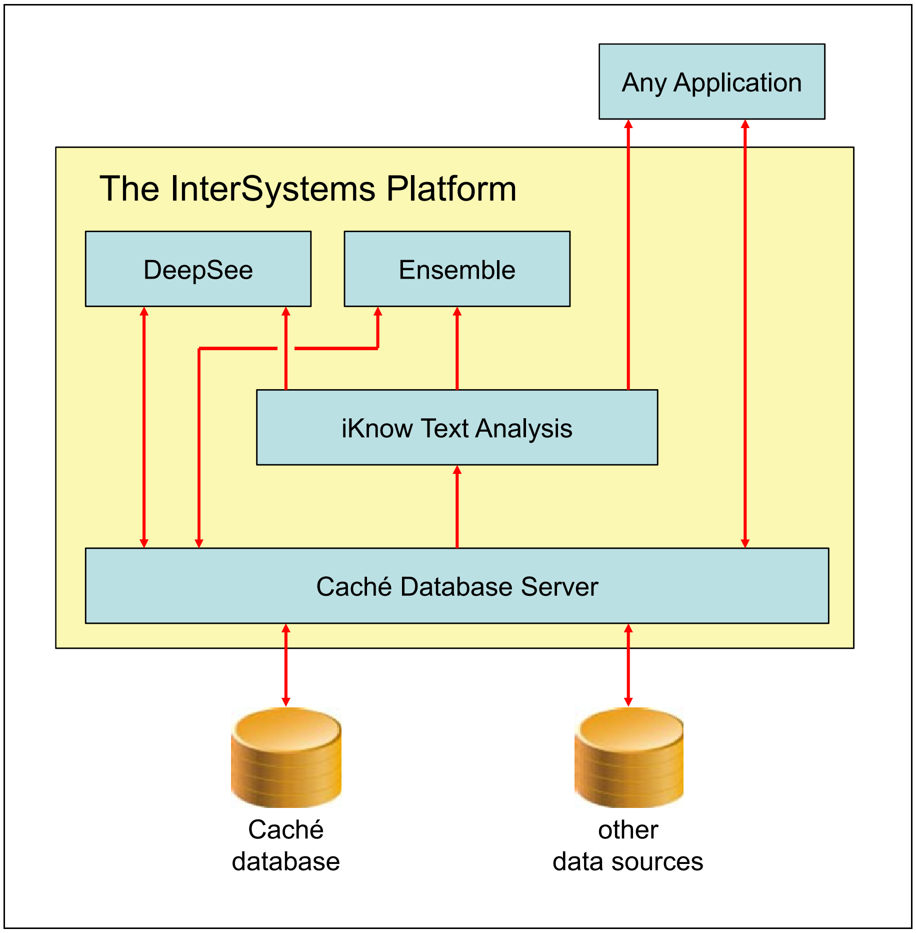


Figure 4 *The architecture of the InterSystems Platform showing the relationships between the various modules.*

Caché – Caché is a hybrid database server supporting the standard SQL language for querying and manipulating data as well as an object-oriented language called ObjectScript. All data in the Caché database can be accessed by both languages interchangeably. The ObjectScript language is very efficient for high-end transactional applications, whereas the SQL language is ideal for reporting and makes data available for any reporting and analytical tool that supports SQL.

The product has proven its performance and scalability in many applications. For example, for the European Space Agency a system has been developed that can insert 5 billion discrete objects of approximately 600 bytes each in 12 hours and 18 minutes, at an average insertion rate of 112,000 objects per second.

iKnow – iKnow uses the Caché database server to store all the texts. This means that texts that must be analyzed, have to be loaded in a Caché database first.

The iKnow text analysis technology can be invoked through SQL and ObjectScript. This means that any application and any reporting and analysis tool can invoke the text analysis features of iKnow. For developers that use SQL, the iKnow features have been implemented as table functions. In other words, iKnow technology has an open and well-defined interface.

iKnow uses the Caché database server for storage of text.

The advantage of integrating iKnow with Caché so tightly is that any application can invoke iKnow and thus analyze text and is able to analyze text together with structured data. For example, in one SQL statement structured data stored in tables can be joined with unstructured data stored in the same or other tables. A straightforward join is sufficient. In addition, all the statistical and grouping capabilities offered by SQL can be used.

DeepSee – DeepSee, also part of the InterSystems Platform, has been integrated with iKnow as well. Like all OLAP tools, DeepSee organizes data in hierarchically structured dimensions and measures and allows for drill-downs and roll-ups of the data. Concepts discovered by iKnow are treated by DeepSee as elements of a dimension. This allows for drill-downs and roll-ups of the results of a text analysis exercise. Because DeepSee treats the iKnow concepts as first-rate citizens, it doesn't make a distinction between dimensions on structured data and dimensions developed on the concepts coming from an iKnow analysis.

Every operation on structured data that DeepSee supports, can be executed on text. This enriches the exploration capabilities offered by iKnow. For example, this allows for time travel analysis. If date-time data is available for each text, analysts can see whether certain concepts are starting to appear more and more often. They can also explore whether two concepts appear together in one and the same text often. Or, a medical environment can use the combined strength of DeepSee and iKnow to analyze whether certain patient symptoms and certain medical prescriptions appear together often. It's also possible to determine which texts are most different from others based on the included concepts. This can be useful for website environments to allow visitors to search for books that are totally different from the ones they have already bought.

iKnow enriches the analytical capabilities of DeepSee.

One of the many features of iKnow is that it can create summaries of texts by returning the full sentences coming from those texts that contain the most important concepts. iKnow allows the level of summarization to be set. This is done by specifying the importance level of the concepts appearing in those sentences. In medical environments this can be used to quickly get a high-level overview of a large patient file. In an emergency department where time can be crucial, it can be useful to quickly get critical information on a new patient. To show the summarizing capabilities of iKnow, Appendix A contains a 1% and a 5% summary of this whitepaper.

Ensemble – Like any application, Ensemble can access the Caché database server, so it can also invoke iKnow. This makes it possible to integrate text analysis within business process definitions. For example, it can be used for automated content-based routing of emails; depending on the content the emails are routed to different steps of the process or different actors of the process.

Importing Text from Other Sources – Caché offers the functionality to access data in external data sources. So, applications that need access to data stored in those external data sources, can query them as if they are common Caché tables. The fact that Caché retrieves the data from external sources is completely hidden. If a data source can be accessed through an ODBC or JDBC interface, Caché can get to it.

A lot of new textual data is stored in the new generation of so-called NoSQL systems, of which Hadoop with its HDFS file system is clearly the most popular one. For HDFS, various SQL interfaces are available, including Hive, Cloudera's Impala, Hortonworks' Stinger, and MapR's Drill. Caché can use these interfaces to access textual data in HDFS. Therefore, iKnow can analyze texts stored in HDFS as if they're stored in Caché itself. Analysts can even combine structured data stored in, for example, Oracle or DB2 with data stored in the Caché database, and with texts in Hadoop HDFS. Being able to do text analysis on a heterogeneous set of data sources this easily, enriches the text exploration capabilities.

9 Two Case Studies

This section contains two brief case studies of organizations using the iKnow text exploration technology of InterSystems.

PCS and Social Knowledge – UK-based application vendor PCS has developed an application called *Social Knowledge*. This application uses iKnow to analyze and monitor social media conversations. Today, social media conversations have the power to shape and steer how brands are perceived. Knowing the limitations of the more classic text analysis technologies to provide real-time insight on issues that matter to brand owners, they adopted iKnow.

The Social Knowledge application enables brand owners and their marketing agencies to monitor and analyze what's being said about their brands across popular social media channels. It shows from hour to hour what is said about their brands and helps in selecting the messages to react to. Key in this process is that there is an immediate comparison between what is important in these messages about a brand and what is closely related to that brand in those messages. This way, one can immediately detect how central the brand is in the conversation and how well it's connected to the key conversation topics. Social Knowledge also allows users to respond to and shape those conversations in a relevant and timely manner. Important to note is that there is no need for developing thesauri or ontology in advance.

According to Philip Walker, Managing Director at PCS, iKnow's data exploration capabilities allows Social Knowledge to deliver actionable insights about social media conversations as they are happening.

Parnassia – Parnassia is an organization based in The Netherlands that helps adults with psychotic diseases. For various reasons, it may be necessary to place patients in seclusion temporarily, and for some time, Parnassia has been investigating how seclusion of patients could be reduced. The advent of an electronic patient record (EPR) provides a wealth of digital data and thus offers a unique opportunity for clinical research. Parnassia decided to start a study to search for variables in these EPRs. The goal of the study was to determine whether certain words or phrases typically precede seclusion, and thus may have a predictive value.

The nursing and medical records were collected from patients admitted to a closed acute admission ward between October 2010 and April 2011. These records were analyzed with iKnow to discover those concepts in the text that distinguish isolated from non-isolated patients. Records of 524 patients were examined, of which 21% were isolated patients. 1,261 reports were assessed, of which 278 of the isolated patients and 983 of non-isolated ones. When analyzing the texts, the following terminology was found significantly more frequently with patients who were isolated: chaotic, claiming, conflict, manic psychotic state, motorial restless, non-empathetic impression, psychotic impression, dysphoric, discussion, demanding, enforcing, resistance, mania, and aggressive.

The conclusion from this study was that there are indications that this method of text analysis offers the possibility to predict from these reports which patients are at risk to be isolated eventually. The method seems to be a promising, complementary, and labor-saving way to increase patient-safety and reduce seclusion.

About the Author Rick F. van der Lans

Rick F. van der Lans is an independent analyst, consultant, author, and lecturer specializing in data warehousing, business intelligence, data virtualization, and database technology. He works for R20/Consultancy (www.r20.nl), a consultancy company he founded in 1987.

Rick is chairman of the annual European Data Warehouse and Business Intelligence Conference (organized in London). He writes for the eminent B-eye-Network[9] and other websites. He introduced the business intelligence architecture called the *Data Delivery Platform* in 2009 in a number of articles[10] all published at BeyeNetwork.com.

He has written several books on SQL. Published in 1987, his popular *Introduction to SQL*[11] was the first English book on the market devoted entirely to SQL. After more than twenty years, this book is still being sold, and has been translated in several languages, including Chinese, German, and Italian. His latest book[12] *Data Virtualization for Business Intelligence Systems* was published in 2012.

For more information please visit www.r20.nl, or email to rick@r20.nl. You can also get in touch with him via LinkedIn and via Twitter @Rick_vanderlans.

About InterSystems Corporation

Founded in 1978, InterSystems Corporation is a US$446,000,000 privately held software company with offices in 25 countries and corporate headquarters in Cambridge, Massachusetts. They provide the premier platform for connected healthcare, and their innovative products are widely used in other industries that demand the highest software performance and reliability. Clients include TD Ameritrade, European Space Agency, U.S. Department of Veteran Affairs, Johns Hopkins Hospital, Belgium Police, Mediterranean Shipping Company, and thousands of other successful organizations.

Leading application providers also leverage the high performance and reliability of InterSystems' advanced technology in their own products. These organizations include Epic Systems, Fiserv, GE Healthcare, and hundreds of others.

[9] See http://www.b-eye-network.com/channels/5087/articles/

[10] See http://www.b-eye-network.com/channels/5087/view/12495

[11] R.F. van der Lans, *Introduction to SQL; Mastering the Relational Database Language*, fourth edition, Addison-Wesley, 2007.

[12] R.F. van der Lans, *Data Virtualization for Business Intelligence Systems*, Morgan Kaufmann Publishers, 2012.

Appendix A Summaries Generated by iKnow of the Whitepaper

This appendix contains two summaries generated by iKnow of this whitepaper. The first text is a summary of 1% and the second one a summary of 1%. Note that both are automatically generated.

Summary of 1%

Besides explaining text exploration technology, this whitepaper introduces the text exploration technology of InterSystems called iKnow.
With Caché as data storage mechanism and as gateway to other data sources, such as SQL databases and big data stores, iKnow has access to a wide range of data sources and potentially massive amounts of text.
Analyzing text can also be defined as deriving structured data from unstructured data (in this case the misnomer unstructured data is used for text).
Text exploration tools require less preparation work than more traditional text analysis tools.
The advantage of integrating iKnow with Caché so tightly is that any application can invoke iKnow and thus analyze text and is able to analyze text together with structured data.

Summary of 5%

This text exploration technology is able to analyze the text in an unguided style.
Almost every industry can benefit from deploying text exploration, especially those industries in where storing text is crucial for business operations, such as advertising, healthcare, legal, pharmaceuticals, publishing, and real estate.
Besides explaining text exploration technology, this whitepaper introduces the text exploration technology of InterSystems called iKnow.
iKnow is truly text exploration technology that uses a unique approach to text analysis.
With Caché as data storage mechanism and as gateway to other data sources, such as SQL databases and big data stores, iKnow has access to a wide range of data sources and potentially massive amounts of text.
iKnow is a truly text exploration technology with a unique approach to text analysis.
This frees up IT professionals from having to spend "an inordinate amount of time" responding to requests for new data, new views or updated report formats Exploration Tools – Most tools impose analytical and reporting restrictions on users.
As there is a need for exploration tools for structured data, there is a need for tools that analyze text in an exploratory way: the text exploration tools.
With text exploration tools text can be freely analyzed without any preparations in advance.
Use of Text Analysis Today – Organizations can benefit from analyzing textual data.
Analyzing text can also be defined as deriving structured data from unstructured data (in this case the misnomer unstructured data is used for text).
For text analysis other terms do exist, such as text mining, natural language processing, information retrieval, and search technology.
The Prehistory of Text Analysis – The history of analyzing text goes way back.
The Need for Text Exploration – As indicated in the previous section, most of the text analysis techniques require work in advance.

Basically, the difference in size is the difference between more classic forms of text analysis and text exploration.

Text exploration tools require less preparation work than more traditional text analysis tools.

To summarize, data exploration is a form of text analysis that meets the following three requirements:

The text analysis technology must be able to analyze the text in an unguided style.

The Classic Approach of Text Analysis – Tools for analyzing text usually try to identify important concepts in sentences.

Most text analysis tools try to locate these concepts by looking at individual words, which results in the words consumer, enterprise, experience, market, and search.

iKnow's Approach and the Three Requirements for Text Exploration – The approach taken by iKnow to analyze text is different from many other approaches.

Results of iKnow's Text Exploration Approach – When iKnow has analyzed all the texts and sentences, and has identified the concepts and relations, it returns the following set of measures that can give valuable insights in the texts:

iKnow for Text Exploration – Why is iKnow suited for text exploration?

The advantage of integrating iKnow with Caché so tightly is that any application can invoke iKnow and thus analyze text and is able to analyze text together with structured data.

This makes it possible to integrate text analysis within business process iKnow uses the Caché database server for storage of text.

Importing Text from Other Sources – Caché offers the functionality to access data in external data sources.

Being able to do text analysis on a heterogeneous set of data sources this easily, enriches the text exploration capabilities.