IDS WHITE PAPER

# Integrating Unstructured Text Into The Structured Environment: The Value Proposition

UNLOCKING THE MAGIC OF UNSTRUCTURED CONTENT

INMON DATA SYSTEMS

The world of computing has grown from a small, unsophisticated world in the early 1960‘s to a world today of massive size and sophistication. Nearly every person on the globe – in one way or the other – is affected by, or directly uses, computation on a daily basis. Nothing less than international productivity from the 1960’s to the present has been profoundly and positively affected by the growth of the use of the computer.

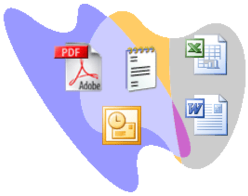
The growth of computing can be measured in two ways – growth in what is termed structured systems and growth in what is termed unstructured systems.

## STRUCTURED SYSTEMS

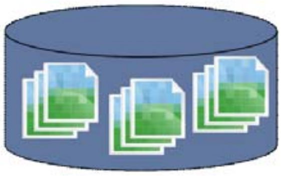
Structured systems are those where the data and the computing activity is predetermined and well-defined. Structured systems are designed by, built by, and operated by the IT department. ATM transactions, airline reservations, manufacturing inventory control systems, and point of sale systems (POS) are all forms of structured systems.

## UNSTRUCTURED SYSTEMS

Unstructured systems are those that have no predetermined form or structure and are usually full of textual data. Typical unstructured systems include email, reports, contracts, transcribed telephone conversations, and other communications. The person doing the communication can structure the message in whatever form is desired, using language in any form desired.  There are few, if any, rules for the content of unstructured systems.



**Unstructured Systems**



**Structured Systems**

From the beginning, the worlds of structured systems and unstructured systems have grown separately and apart and yet in parallel with each other. It is no surprise that today each environment is separate from the other in many crucial ways:

- Technical
- Organizational
- Structural
- Functional

In truth there is little overlap or connection between the two worlds.

However, imagine what the world would look like if there were synergy between the two environments and if they were able to be connected in an effective and meaningful way?  Imagine the new types of systems that could be built and the enhancement to existing systems that could be created in ways that are not possible using today’s technology.  The good news is that we need not accept the limitations of today’s technology.

If a bridge is to be built between the two environments, it makes sense to bring the unstructured text into the structured environment. By doing so, the decision support analyst can take advantage of the analytical processing capabilities that exist in the structured environment.

The possibilities for new systems and major benefits blossom when the gap between unstructured data and structured data is crossed.

### INTEGRATING UNSTRUCTURED TEXT

The key to the crossing of the bridge between the two environments is the integration of unstructured text. Raw unstructured text can't simply be placed into the structured world and still be meaningful and useful. Stated differently, unstructured text placed directly into a structured environment tends to create a mess. There is too much data, there is data with the same name that has different meanings, there are alternate spellings, there are extraneous words, there are documents that have no relevance for business, and so forth. All of these limitations of unstructured text become apparent when unstructured data is moved into the structured environment without proper conditioning.

### UNSTRUCTURED INTEGRATION – THE ISSUES

Following are the issues that must be addressed in the integration of unstructured text into the structured environment:

- A determination if the unstructured document has any relevance to the business itself. If the unstructured document is not relevant to the business being conducted, then it does not belong in the structured environment.
- Removal of stop words from the unstructured environment which are extraneous to the meaning of the text. Typical stop words are "a", "and", "the", "is", "was", ″which", and so forth.
- Reduction of words to their Greek or Latin stems. By reducing the words found in unstructured text to a common stem, the commonality of words can be recognized when the words are literally not the same.
- Resolution of synonyms. Where there are synonyms, the reduction of the synonym to a common foundation allows for a common vocabulary, which allow for meaningful searches to be accomplished.
- Resolution of homographs. In the case of words that have multiple meanings, the correct and unique term is replaced with the non-unique common term. This is a key ingredient for the establishment of as common vocabulary.
- The ability to handle both words and phrases.

- Allowance for multiple spellings of the same name or word.
- Negativity exclusion. Where there is a negative, the words that follow the negative expression are removed from any indexing or other reference.

These activities of integrating unstructured text are the minimum subset of processes that need to occur. There are many other related processes that can be applied to unstructured text as it is prepared for movement to the structured environment.

## EXTERNAL CATEGORIZATION

The creation of external categorizations (or "indirect indexes") is a critical activity in the preparation for movement of unstructured data into the structured environment.

The starting point for external categorization is identifying when a topic is associated with multiple words. The unstructured text is examined within the context of the associated words and phrases. As a simple example consider the topic "Sarbanes Oxley".

### Topic - Sarbanes Oxley

Associated words and phrases –
Promise to deliver
Contingency sale
Revenue recognition
Contract terms

The unstructured data is examined and wherever a word or phrase is found that is indirectly related to Sarbanes Oxley, a reference is made to and from that document to Sarbanes Oxley. This is an example of an indirect index. Now when a search of unstructured data is made on the term "Sarbanes Oxley", all the references to terms that relate to Sarbanes Oxley are found.

Contrast an indirect search to a direct search. In a direct search, if the search were made on Sarbanes Oxley, then only text where the term Sarbanes Oxley is found would be referenced. Such a direct reference has limited usefulness.

As data is placed in the structured environment from the unstructured environment, not only is the unstructured text integrated, but external categorization of the data is performed simultaneously.

### Application Example One: Simple Integration

The first step in creating an integrated textual environment is to access and gather the unstructured documents to be processed. If an organization has many unstructured documents in many different places, then the ability to find, access, and gather those documents into a single location is an important feature.

The simplest kind of application that can be created from integrated unstructured text is one where data is simply integrated into the structured environment, then accessed and analyzed. Once the raw documents have been gathered, they are integrated into the structured environment. Once text has been integrated, there are many uses of the data and standard business intelligence (BI) tools can be used to create queries against the textual data.

As an example of a simple application, suppose there is a toxic chemical which has just been unearthed as a threat. The integrated text can be used as a basis for a search. It can take a matter of seconds to find out what information there is about this newly uncovered chemical.

A second simple use of integrated textual data in the structured environment is that of doing an indirect search. For example suppose there is a need to do an indirect search on the term “toxic chemical”. All of the different kinds of toxic chemicals and all of the information about these toxic chemicals can be accessed and organized very quickly.

A third valuable use of unstructured data that has been integrated and placed in the structured environment is the ability to link that integrated text to other structured data. As a simple example, suppose there is text about the word “nitroglycerine”. This text can be connected to and related to other occurrences of “nitroglycerine” in the structured environment, forming a comprehensive and robust query.

In short, there is widespread applicability of simple unstructured content integration and merely by integrating text and placing it in the structured environment a wealth of information is tapped.

### Application Example Two: Examining Existing Stores of Unstructured Content

Most organizations collect vast amounts unstructured data in the form of emails. And other corporations collect unstructured text in content management systems (e.g. Documentum). Over time, these collections of unstructured text grow large. And as the collections of data grow large, the content becomes stale and unintelligible. Stated differently, there is so much content and the content is so scattered and disparate that it becomes difficult to find anything in the files of unstructured data.

By first integrating the large volumes of unstructured text, then bringing the text over to the structured environment, the unstructured text is able to be meaningfully read and examined. All corporations that have large stores of unstructured data – emails, documents, etc. – need this capability.

### Application Example Three: Unstructured Contact File – Creating the 360 Degree View of the Customer

An unstructured customer contact file is a record of every contact or communication the customer has had with the corporation. This can include emails, letters, text messages, and other documents. The unstructured customer contact file is an index of the date of the contact, the nature of the contact, to whom the contact was made, and so forth. One of the most powerful uses of the customer contact file is in terms of supplementing a CRM system.

The essence of the CRM system is to create what can be termed a 360 degree view of the customer. By creating a 360 degree view of the customer many avenues and opportunities are opened, including:

- cross selling. If you understand a lot about the customer in one arena, the opportunity to sell to the same customer in another arena will materialize.
- prospecting. The more you understand about a customer, the better you can qualify a sales or sales prospect list.
- anticipation. By understanding a lot about the customer, you can anticipate future needs, and so forth.

One of the basic tenets of CRM is that it is much easier to sell into an established customer base than it is to bring in new customers. In this regard, creating a long term and genuine relationship with the customer is a worthwhile (and profitable) objective.

So exactly how is this relationship established? The basis of the relationship is the integrated knowledge about the customer. The integrated knowledge includes many different facets about the customer –

- age
- education
- occupation
- marital status
- address
- net worth
- income
- spending habits
- children
- type of car driven
- cost of home, and so forth.

The idea behind creating the 360 degree view of the customer is to bring together data from many different places in order to integrate the data and achieve a truly cohesive and comprehensive view of the customer.

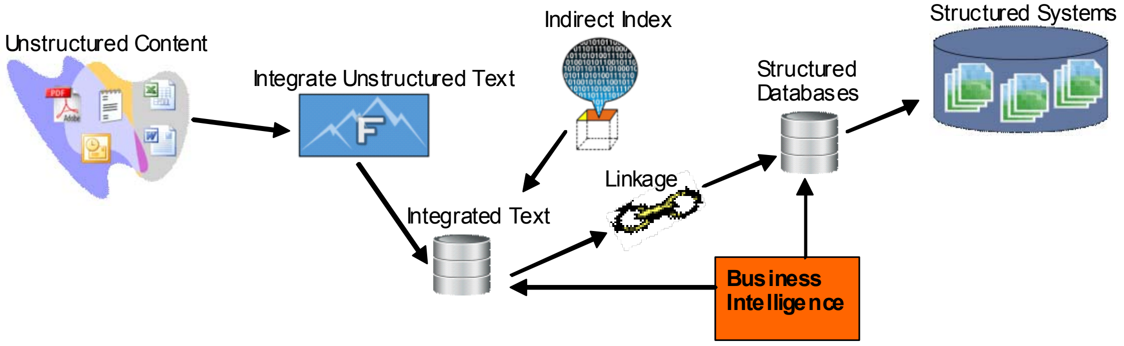
The unstructured contact file adds the dimension of communication to the existing structured content. Now, instead of just knowing odd facts about the customer, the corporation can know what the customer has been saying – what communications have transpired. The corporation can know if the customer has been irritated, pleased, or anything in between. In short, the unstructured contact file allows the corporation to know the recent state of mind of the customer.

### Application Example Four: Enterprise Metadata Repository

There is a great need for the ability to locate, gather and integrate enterprise wide metadata. Metadata – wherever it is found, is nothing but unstructured text. The same tools that can enter the unstructured environment and integrate content of text can also be used for enterprise wide metadata gathering. Once gathered, the enterprise wide metadata can be used for the creation and population of a repository.

## THE ARCHITECTURE FOR INTEGRATING UNSTRUCTURED CONTENT

The architecture for text integration looks like:



By integrating unstructured text and organizing into the structure as shown –

- the unstructured data in and of itself can be analyzed.
- the unstructured text can be accessed by direct or indirect searches.
- the unstructured text can be linked to structured databases and a composite query can be created.

## About Inmon Data Systems

Inmon Data Systems (IDS) was formed in 2003 to develop and distribute software to facilitate the merger of unstructured and structured content environments. We believe that a powerful relationship is formed when unstructured data is merged with structured data. IDS has proven that corporations and public sector organizations can achieve tremendous value from adding unstructured content to the knowledge pool, and creating a new infrastructure for reporting and business intelligence that is available to the end user.

Contact us at 303-681-0474 or sales@inmondatasystems.com for further information.