RANDOM THOUGHTS

Is Text Really Unstructured?

by W H Inmon, Forest Rim Technology

Structured data is repetitive data that occurs over and over. Banking transactions, airline reservations, retail sku sales, telephone calls are all classical examples of what is known as structured data. In most cases structured data is created as a result of the execution of a transaction.

Structured data fits nicely and neatly inside a standard data base management system (dbms).

And then there is text. Text is commonly referred to as unstructured data. And – prior to textual disambiguation – text did not fit comfortably and conveniently into a standard data base management system.

But is text really unstructured?

What does the term “Unstructured” really mean?

In general the term “unstructured” refers to a lack of structure. And if text were really unstructured we wouldn’t be able to understand each other when we have a conversation. But we do understand each other when we speak. People do understand books when they read them. So what is going on here?

There is definitely structure behind text. There is proper spelling. There is proper punctuation. There is proper sentence construction. There is proper thought development. Ask any English teacher and you will find out just how much structure is behind text. Lots of it.

So indeed there really is structure behind text. But the structure behind text is quite complex. Language is taught in school from the first grade on. Parents start teaching their children

language at a very young age. It takes a long time for a human to learn how to speak properly and also to learn to understand speech. And the deeper you go into language the more arcane and complex it becomes. Indeed, you can get a Phd in language and make it your life's work.

So there really is then a structure behind text.

But does the structure that language has allow the text to be considered to be structured in the eyes of the computer? The answer is no, because even though text is structured, that structure is so vast, so complex, so arcane that the computer cannot understand the structure of language. Stated differently, the computer is capable of understanding only the simplest of structures. Language is simply beyond the pale for the computer. Therefore, in the eyes of the computer, text is unstructured, even though there really is an underlying structure to text.

So when the computer professional refers to unstructured processing the computer professional is referring to something different than the general understanding of what is meant by unstructured. Stated differently, there is the dictionary understanding of unstructured and there is the computer professional's understanding of unstructured and these two understandings are very different things.

To make matters even more complex (as if they need to be!), unstructured data in the computer senseincludes a lot more than text. Unstructured computer data includes all sorts of other data – image data, sound data, log tape data, and meteorological data, to name a few.

Now why does the computerized definition of what is structured d what is unstructured make a difference? The difference is made because the computer was made to handle structured data and NOT unstructured data. The computer expects records to be in nice neat little piles called records. Each record has a key and other attributes. Once data is organized into a structured format, the computer rips through the data, much like bullets flying through a machine gun. But if there is a bullet that is out of place the machine gun jams and the machine gun no longer is a military asset but a military liability.

So the structure and organization of the data makes a big difference when it comes to efficient processing inside the computer. One of the interesting questions becomes if the computer cannot handle unstructured efficiently, then can unstructured be turned into a structured format? The answer is yes – there is technology that can be used to turn textual data into a structured format and maintain the unstructured flavor of the data. That technology is known as textual disambiguation.

It is the role of textual disambiguation to ingest raw, unstructured text and to transform the important parts of unstructured text into a structured format while maintaining the essence of the unstructured data. It is sort of like riding a bicycle across a tightrope stretched across Niagara Falls while juggling monkeys that dash about. Not for the faint of heart.

While there are many facets to textual disambiguation, the most intriguing aspect of textual disambiguation is that of deriving the context of text while placing text into a structured format.

While textual disambiguation is interesting, the strategic value of textual disambiguation is that it enables text to be placed into a standard data base. In turn, once text is placed into a standard database, text can be used for corporate decision making. And strategically that is very important.

If you don't grasp the strategic importance of being able to make decisions based on text, think about this. It is estimated that 80% to 90% of the data in the corporation is based on text. But what data is being used as a basis for making decisions? Most corporate decisions are made on the basis of reading and analyzing 10% to 20% of the structured data in the corporation. Does this make sense?

It is like saying that only men over 65 who have college educations should make all the political decisions for the entire population. What about women? What about people younger than 65? What about people who do not have a college education?

We would never stand for a political system that was so misshapen and so elitist. But that is exactly what we do for the data found in our corporations.

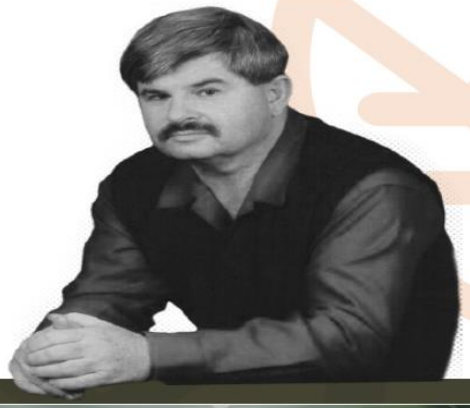
Workers of the world unite. Start making corporate and management decisions on your unstructured data.

Forest Rim Technology was formed by Bill Inmon in order to provide technology to bridge the gap between structured and unstructured data. Forest Rim Technology is located in Castle Rock, Colorado.

William "Bill" Inmon

AUTHOR AND TECHNOLOGY PIONEER

Best known as the "Father of Data Warehousing", Bill Inmon has become the most prolific and well-known author worldwide in the big data analysis, data warehousing and business intelligence arena. In addition to authoring more than 50 books and 650 articles, Bill has been a monthly columnist with the Business Intelligence Network, EIM Institute and Data Management Review. In 2007, Bill was named by Computerworld as one of the "Ten IT People Who Mattered in the Last 40 Years" of the computer profession.



Reach Out for a Free Consultation

Let Us Help You Discover The Hidden Potential In Your Data

Send us an email at Info@ForestRimTech.com