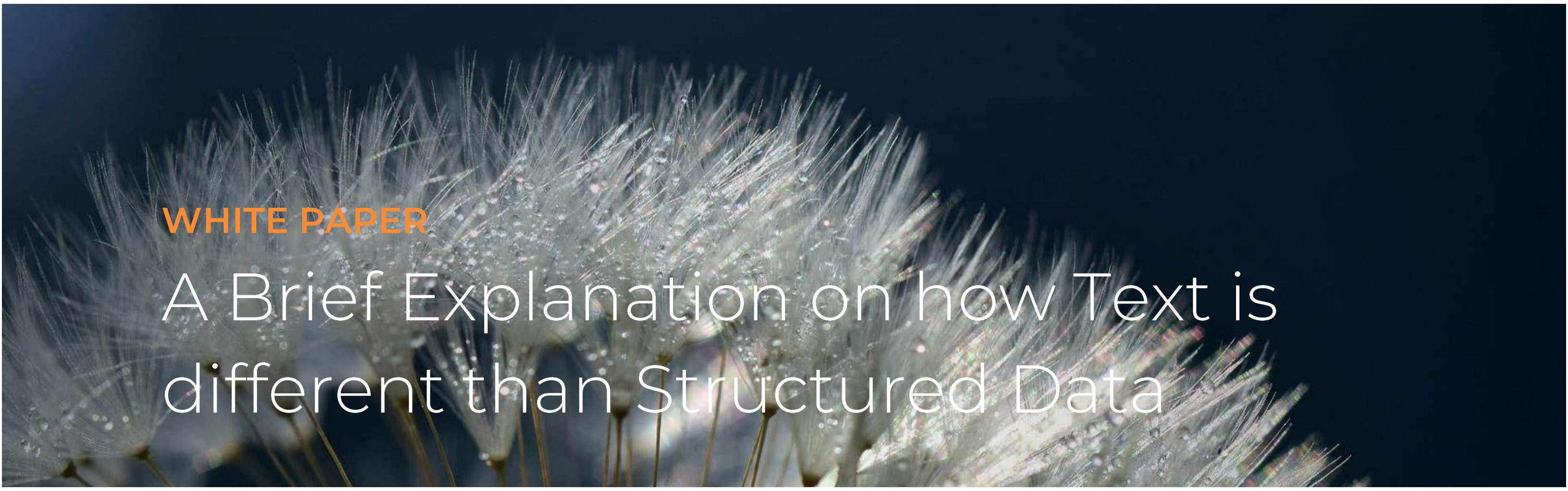
FORESTRIM TECHNOLOGY

LEARN MORE AT:
forestrimtech.com/whitepapers



WHITE PAPER

A Brief Explanation on how Text is different than Structured Data

by W H Inmon, Forest Rim Technology

So what are some of the ways that text is different than structured data? This question is sort of like asking – why don't you use snakes to pull a carriage? It is almost a non sequitur.

Here are MANY, MANY fundamental differences between text and classical structured data. Some of the many differences are –

Text is a self contained structure. When you say a sentence to another person the sentence needs to be entirely self explanatory. If you don't talk this way you don't make any sense. However in structured data, data is part of something larger. For example, when a purchase is recorded at a retailer, the data relating to the purchase is recorded entirely within the context of the purchase being made. For the most part structured data is NOT wholly self contained.

Text usually needs to be edited. There are all sorts of edits that are needed. Stop words can be removed. Spelling can be corrected. Date formats can be standardized, and so forth. It is entirely appropriate to edit text when preparing it for inclusion into a database. However, the same is not true for structured text. If a merchant receives a payment for $549.31 it is against the law for the merchant to round up the payment to $550.00. Structured data operates on the basis of precision. Textual data does not operate on the basis of precision.

In text, patterns of text become very important. The proximity of words affects the meanings of words. The context of words affects the meaning of text. The positioning of words within the document affects the meaning of words. There is a whole art to reading text and finding and interpreting the patterns of text that affect the meaning. There is no analogical equivalent of finding patterns in structured data.

Text can come in multiple languages. Text can be in English. Text can be in Spanish. Text can be in German, French or Mandarin. Yet the meaning of text remains the same, regardless of what language it is said in. For the most part, structured data is language independent.

Categorization of text can become important. It is through categorization that much of textual processing is done. Categorization of text is the means by which much of meaning is derived. There is no analogical equivalent of the categorization of structured data.

As important and as useful as text is, the context of text is even more important. Text without context is about as useless as a car without wheels. A car without wheels may be fun to camp in or even to get out of the rain in, but you certainly are not going on a trip in your car if it doesn’t have any wheels. Structured data too needs context. But the context of structured data is derived and handled in a completely different manner.

In most cases text is non repetitive. Consider emails. A person can write whatever he/she wants in an email. Therefore you cannot predict what text or the pattern of text that will be found in an email. But structured data is highly repetitive. Consider an ATM activity. Each ATM activity looks – structurally – like each other ATM activity. The only difference between ATM activities is the data contained in the record of the activity.

Text is subject to inference processing. When the doctor says – “you do NOT have cancer” that is a very different message than when the doctor says – “you do have cancer”. The entire meaning of what is being said is often reversed by a single word. Being able to recognize and interpret these words through inference processing is vitally important.

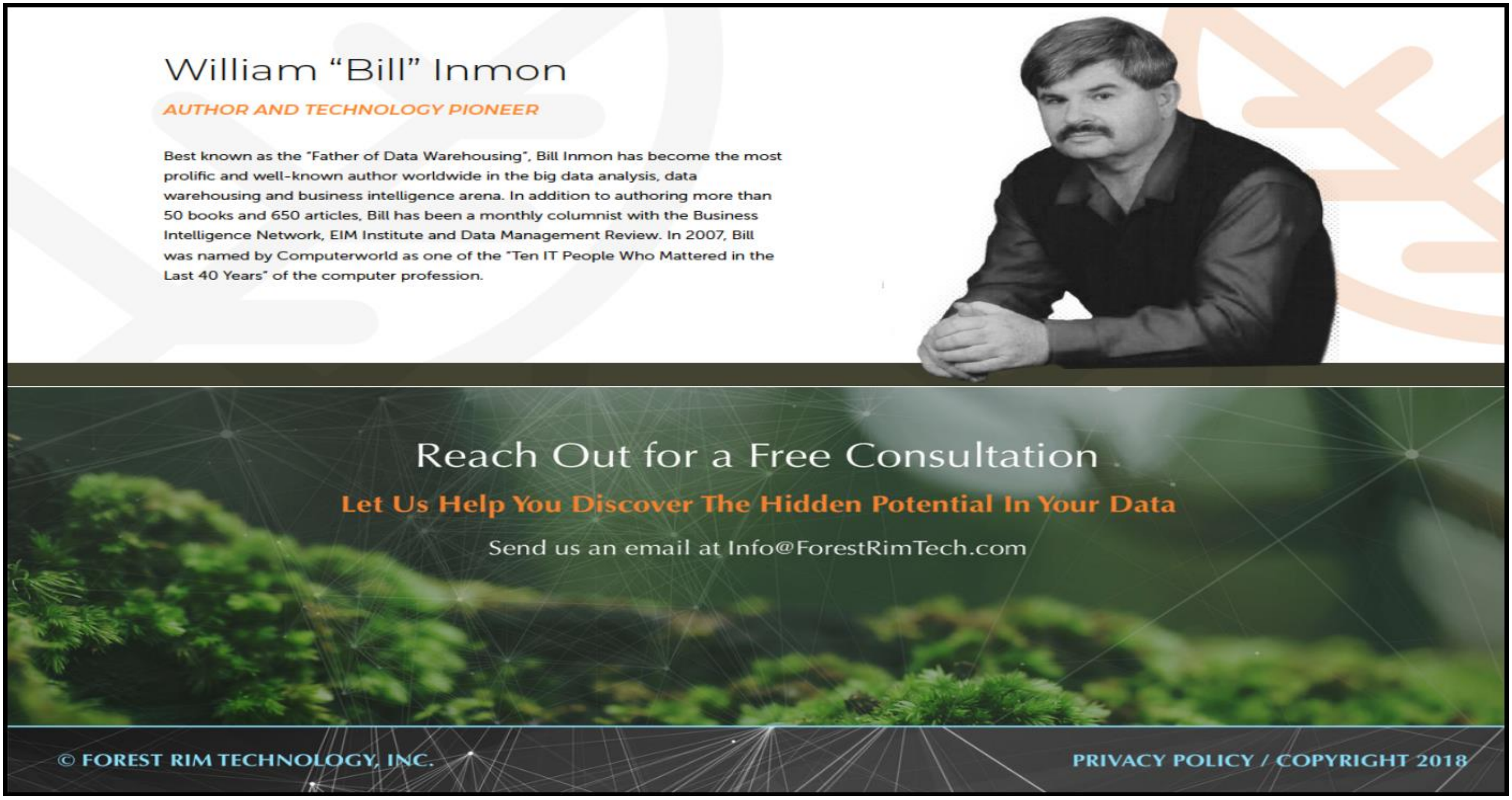
From a processing standpoint there are MANY significant differences between text and classical structured data. Whereas structured data is content to merely contain a value of data associated with a field, in textual data it is necessary to have a field of data, the type of field of data, AND the context of the field of data. If there is a single dividing line between textual data and structured data it is that it is not sufficient to merely identify and process text. Text MUST have context in order for the text to be useful inside a data base.

These reasons and more explain why handling text inside a data base is a complicated process and is quite different from handling and processing structured data.

Bill Inmon – the “father of data warehouse” – has written 57 books published in nine languages. Bill was named by ComputerWorld as one of the ten most influential people in the history of the computer profession. Bill lives in Castle Rock, Colorado.

Bill’s book on TURNING TEXT INTO GOLD, Technics Publications, a book that shows how text can be turned into business value. TURNING TEXT INTO GOLD is available on Amazon.com.

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

© FOREST RIM TECHNOLOGY, INC.

PRIVACY POLICY / COPYRIGHT 2018