



ENTERPRISE CONTENT MANAGEMENT AND TEXTUAL ANALYTICS

**A WHITE PAPER BY
W H Inmon**

Over time it is normal for corporations to collect important information that is on paper. The only difference between one corporation and another is the rate of collection, the amount of papers collected, and importance of those papers. Some corporations collect a lot of paper and collect it quickly. Other corporations collect relatively small amounts of paper and at a slow rate. In any case, the collection of paper documents with important information on the documents is a fact of life in the corporation.



**Over time,
corporations collect
a lot of textual (or
unstructured) data**

Fig 1

An interesting question is – what happens to the papers over time? More importantly, what happens to the information on those papers over time? At least two nasty things happen when large amounts of information are collected and stored over time –

- the paper starts to disintegrate. When the paper disintegrates the information on the paper is lost forever,
- there starts to be some many papers that information on the paper becomes lost. Important information gets to be buried in a ton of lesser important information.

As an interesting case in point - what happens when a reader goes to a library, selects a book, then replaces the book in an incorrect location on an incorrect shelf? The book that is misplaced “hides” on the shelf until some librarian notices that the book is out of place and replaces it into its proper location. Until the book is replaced, it is effectively lost.

The same phenomenon occurs with corporate papers. Once there starts to be a sheer volume of papers, finding even the most basic document becomes a real issue. And as time passes and more paper aggregates, the problem becomes progressively worse, not better.



**There are many challenges
in trying to make sense of
textual data**

Fig 2

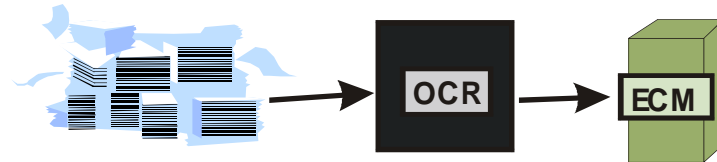
In recognition of the challenges associated with storing information on paper, corporations discover that for the long term, it is a much better proposition to store information that is originally found on paper in an electronic format. There are some really important reasons why storage of information found on paper in an electronic format is a good idea. Storing paper based information on an electronic format is –

- cheaper. The information can be stored in a condensed fashion. The information contained in a box of paper can easily be stored on a small disk,
- accessible. Instead of having to manually thumb through paper documents, a computer can be used. Computers are much faster and more accurate than any human when it comes to looking for data. In addition, humans can only look at so much data before their minds become numbed. Although computers are not as smart humans, computers do not suffer from having their minds numbed.
- able to be protected from corrosion. While computers and electronic storage media certainly do corrode over time, it is much easier and much less expensive to protect electronic data from corruption and erosion than it is to protect paper from long term corrosion.

These then are just a few of the important reasons why taking paper based information and placing it in an electronic format is simply the best solution, certainly for the long term.

A standard way of collecting paper based data is through the process known as OCR. OCR stands for optical character recognition. OCR technology has been around for a long time and is a proven technology. With OCR, paper data is read and is converted into an electronic based format. If the information on the paper is in a standard font, the paper based text can be read easily. If the font is non standard or if the text is light or essentially difficult to read, the transformation through OCR is more difficult. It is normal for there to be a certain amount of manual adjustment to be done as the OCR process is in operation. The challenge of OCR is to keep the manual adjustment down to a minimum and let the machine that does OCR do the vast majority of transformation. In any case, OCR technology is proven technology and has been around for a long time.

After the documents have gone through the OCR process, they are normally placed in what is termed as an ECM store. ECM stands for enterprise document management. The text is passed through OCR and the results are stored on a document by document basis. Now the results are in an electronic format. Now all of the benefits of storing data electronically can be enjoyed.



Often times the text is put into an electronic format by OCR then stored in ECM
Fig 3

But the organization that has just gone through an OCR and ECM process wakes up one day and starts to ask an interesting question. That question is – “I have all of this data stored electronically (and there are real benefits to having done that), but how do I now start to do analytical processing against my data that I have in ECM?”

The simple answer is that I take a computer and let the computer start to churn through all of the data. And you can do exactly that. But there are some challenges to just letting the computer churn through all of the data. Some of those challenges are –

- it takes a long time to churn through the data. I either need a big computer or I need a lot of time,
- looking through data is an imprecise art. One person writes data as 2008/03/27. Another person writes date at March 27, 2008. One person writes name as John W. Higgins. Another person writes name as J. W. Higgins. Making comparisons from one unit of written data to another is actually a very difficult thing to do (as you find out when you try to do it for the first time.)

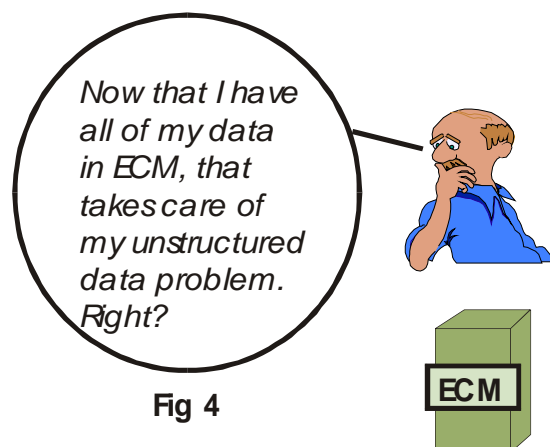
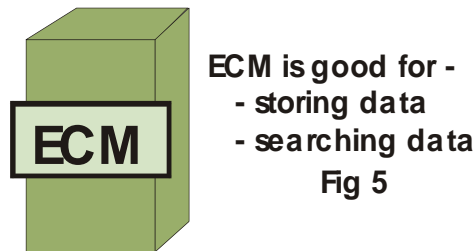


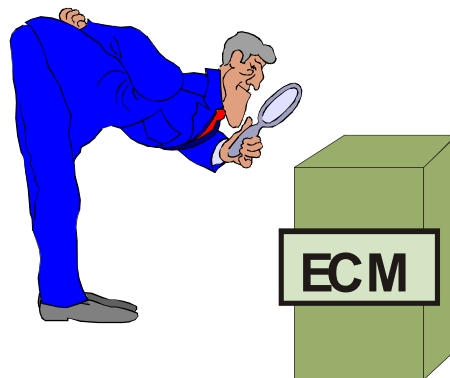
Fig 4

The organization quickly comes to the conclusion that – OCR and ECM are good for taking information from paper, and OCR and ECM are good for storage and simple searches of data, but when it comes time to do analysis of the data that is now electronically captured, that is another story entirely. In order to actually make use of the electronically captured data, it is optimal to put the electronically captured data in the form of a data base.



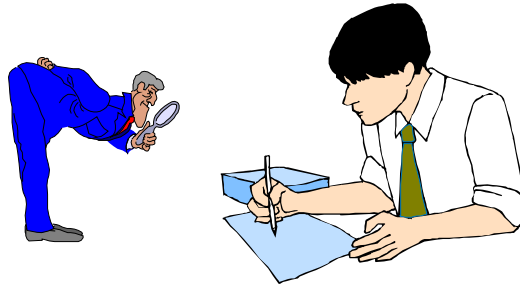
The organization finds that when people access and analyze electronic data that there is a need to look at the data in a very sophisticated format and structure. As powerful and as useful as it has been to take information off of paper and place the information on an electronic medium (and make no mistake, there are big benefits to doing just that), the problem of meaningfully searching and analyzing the data has just begun.

The data – as stored in ECM – is in a crude format and structure. In order to make the data useful to analysts, the data must be taken from its crude format and placed in a sophisticated data base format.



**So what's wrong with
searching data?
Fig 6**

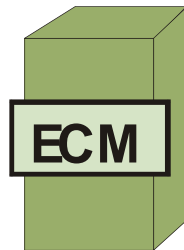
It is at this point that the analyst discovers the big difference between search and analysis. A search merely looks for raw pieces of data based on some parametric information. Analysis goes much deeper into the fabric of the data and allows the analyst to ask very sophisticated questions that cannot be handled by a mere search.



**There is a big and important
difference between search
and analysis**

Fig 7

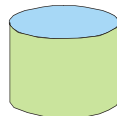
In a word, search is done on raw document based data found in ECM.



**Search is done
on raw data**

Fig 8

But analysis is done on integrated data, and that data is typically found in a data base, where the data is managed by a data base management system (dbms).



**Analysis is done
on integrated data**

Fig 9

Moving raw document based data into a dbms is no small effort. The major challenge faced by the organization that attempts to make such a transformation is that classically dbms structures have been used to handle what is referred to as structured data. In structured data there is repeatability of the same type of information. In a bank, people cash checks all day long. There is no difference from one person cashing a check for another other than the amount of the check and the account number. In an airline, reservations are made every day. There is no difference between one reservation and the next other than the person making the flight and the day and flight number. Other than that the same types of information appear over and over again and fit very nicely into a standard dbms.

But when it comes to textual data there is no such repeatability of information. When a person writes text in a document, there are no particular rules dictating how that document is written or how to say anything. Text is free form, and there just is no repeatability of information, as there is in banks, airlines and many other places.

Until now there has been no good way to take text and place text into a data base management system. But now there is textual ETL, by Forest Rim Technology. The patent pending technology by Forest Rim Technology reads textual information and structures the information so that it easily and comfortably fits into a standard relational data base management system.

Once the data is in ECM, it is actually easy to transform the textual data into a standard relational dbms using textual ETL by Forest Rim Technology.

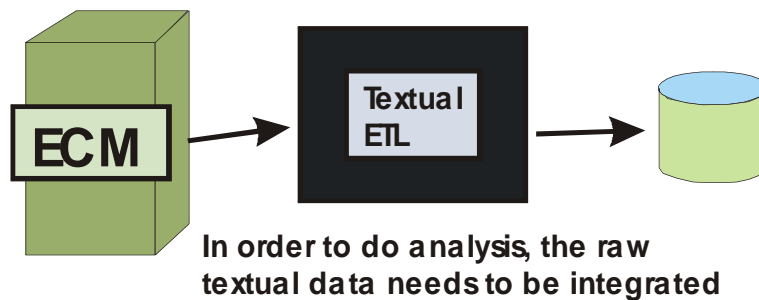


Fig 10

There are many facets to the transformation of textual data into the form and structure needed for standard dbms management. (Note: the process of textual transformation is patent pending by Forest Rim Technology.)

Some (but not all) of the aspects of textual ETL include -

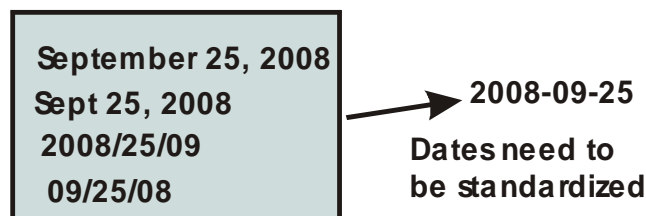


Fig 11

- standardizing dates. People write dates in many different formats. Forest Rim reads those dates and converts them into a standard format that can be managed and understood by a dbms,



Text needs to be converted to numerics

Fig 12

- converting textual numbers to numeric numbers. People may write numbers in prose. But dbms don't understand prose. Dbms understand numbers. So Forest Rim converts text into numbers, when appropriate,

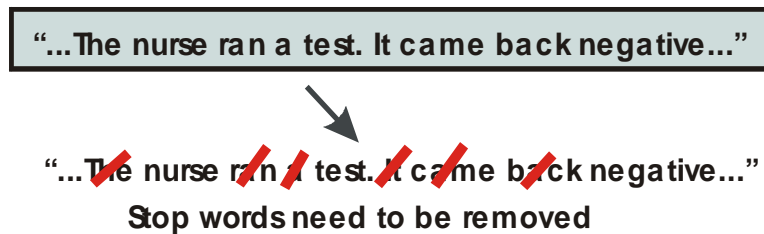


Fig 13

- removal of extraneous text. Most languages have words that are necessary for grammar but are not useful to the meaning of conversation. These are called stop words. Typical English stop words are “a”, “and”, “the”, “is”, “was”, “that”, “which”, “when”, and so forth. Forest Rim removes those words so that they don't get in the way,

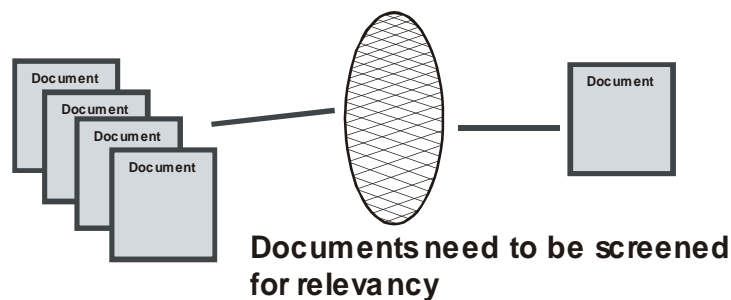
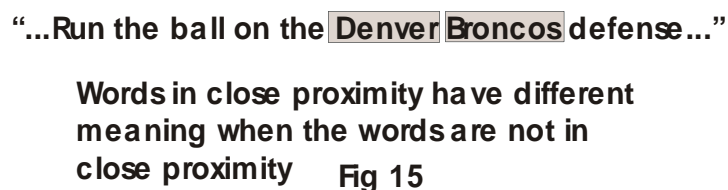


Fig 14

- many times an entire document is not useful or relevant to the business. For example an email says “Let's go out on Saturday night.” This message isn't relevant to the business of the corporation, and is best deleted or not entered into the data base at all. Forest Rim allows screening of document and the text contained therein to occur,



Proximity analysis. In text, on occasion when words are close together they have one meaning. When the words are separated they have another meaning. In a document there is a discussion about how the city of Denver was formed. At the end of the document there is another discussion of the early rodeos that were held and the broncos that were in those rodeos. These words are far apart, and have their own meaning. But when the words are put in close proximity to each other, they form an entirely different meaning.

When someone reads about the Denver Broncos, they don't think about rodeos or the early town of Denver. Forest Rim allows these proximity variables to be recognized and managed,

“...Broken bone...”
“ disarticulated tibia...”
“ fractured ulna...”
Terminology is different
even though meaning
is the same or similar
Fig 16

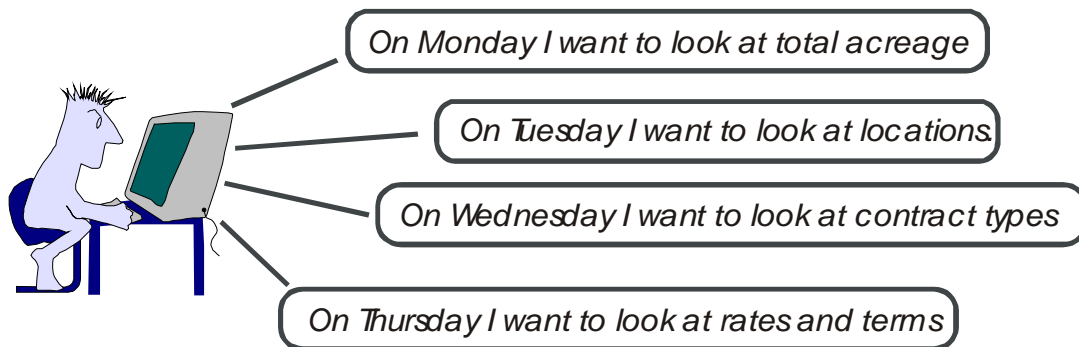
- different terminology is used by different people for the same thing. Doctors for example, have many terms for the human body. In order to do effective analytical processing there needs to be a resolution of different terms meaning the same or similar things. Forest Rim provides that capability,
-

“This is an agreement made the 14th day of March, 2008 between...”
Certain text is considered to be so important that
an index is created for the type of data
Fig 17

--- recognition and management of important words and phrases that need to have indexes created for them. In most documents there is the need for identifying and capturing certain information that is needed to separate one document from another. Take, for example, a collection of resumes. The analyst would like to know standard information about

- whose resume it is
- what education the person has had
- where the person lives
- the age of the person
- the salary of the person, and do forth.

Forest Rim provides the ability to look into a document and create common indexes on data that has been identified as important.



One advantage of automating text is the opportunity to change your mind as to what you want to look for

Fig 18

- once the data base has been created, one of its great advantages is the ability to support many different kinds of queries. One day the analyst wants to look for one thing. The next day the analyst wants to look for something else. Forest Rim supports the building of data bases in standard relational technology such as NT SQL Server, Oracle, DB2, Teradata, and so forth.

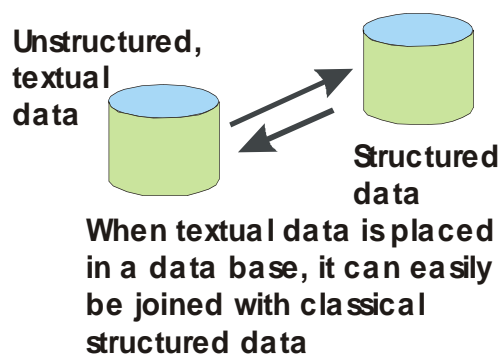
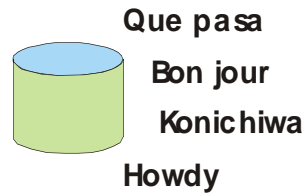


Fig 19

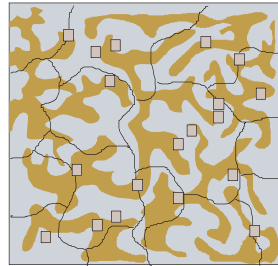
- once the data base has been created, one of the most powerful things that can be done is to combine the unstructured textual data base with classical structured data. Being able to tie textual data to financial data, for example, is a very important capability. By tying textual data to financial data, whole new classes of analysis can be done that simply were impossible to do otherwise. Forest Rim supports the linkage of one data base to another.



**Another advantage of
integrating text into a
data base is that
multiple languages can
be handled**

Fig 20

- text often comes in different languages. When the analyst does a query, in some cases the query operates on text that has been written in more than one language. Forest Rim supports multiple languages.



**Another advantage of
integrating text is that
visualizations can be
created**

Fig 21

- while query capability is important, often times it is useful to visualize data. Forest Rim Technology supports visualization of text.

These then are a few of the capabilities of textual ETL. Once the textual data has been captured in ECM, the next logical step is the movement of the document data into a standard data base. Forest Rim Technology's patent pending technology provides that important linkage.

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

Forest Rim Technology is happy to provide you with an actual demonstration of the techniques and tools described in this document.