DATA WAREHOUSING IN THE HEALTHCARE ENVIRONMENT

By W H Inmon
For years organizations had unintegrated data. With unintegrated data there was a lot of pain. No one could look across the information of the corporation and see information from the corporate perspective. Instead there were many unintegrated parochial views of data. The pain of the lack of corporate information was widespread. Finance, marketing, sales, management and others needed to see a corporate perspective of information but the many local parochial systems that held that information simply did not yield the integrated corporate perspective that was needed.

Then the concept of data warehousing appeared. Data warehousing required that the many disparate systems that held unintegrated data become integrated. In many circles the older unintegrated systems were called “legacy systems.” The process of integration for legacy systems was never clean and easy. But organizations found that the ability to look at corporate data in an integrated manner was worth the pain and trouble of integration of the data. In a short while, data warehousing began to be a corporate necessity. Today data warehousing is conventional wisdom in information processing.

HEALTHCARE INFORMATION SYSTEMS

But data warehousing did not find its way easily and readily into medicine and healthcare. While retailers, financial institutions, government agencies, manufacturers and a host of other industries were reaping a bonanza of corporate information, healthcare and medicine were stagnant when it came to data warehousing. There still were the needs for data warehousing in healthcare and medicine, but the information needs of medicine and healthcare were fundamentally different than those of other institutions.

The first difference between the information needs of healthcare and medicine and other institutions was in the form that information takes. As an example of the fundamental difference between the information found in healthcare versus the information found in other institutions, consider the events that cause the gathering of information. As an example of the way corporations gather information, consider an airline. In airline commerce, a transaction occurs in many places. A reservation is made. A flight takes off. Cargo is loaded. A seat is selected. Baggage is delivered, and so forth. Practically every interaction between the consumer and the airline results in a transaction. Once the transaction is executed, information is gathered.

In the business world a transaction is repetitive and is dominated by numerical data. The same activity occurs again and again, with different consumers and different numbers. And the same pattern of executing transactions emerges for banks, insurance companies, manufacturers, and so forth.

But in healthcare there is no such set of repetitive transaction activity (or at least not nearly as many transactions as are found elsewhere in business). In healthcare there is emergency room care. There is outpatient care. There are visits to the doctor’s office. There are procedures, and so forth. There simply is not the same transactional repetition that occurs in institutions outside of healthcare. Each transaction or encounter in healthcare is fairly unique, as opposed to business where each transaction is very repetitive.
TEXT, NOT NUMBERS
Another factor that separates healthcare data from other institutional data is the very nature of the recording of the interactions that happen between patient and care giver. In almost every case, the result (in terms of information) of the encounter between patient and care giver is verbiage – text. The doctor writes notes about the office visit. The hospital writes notes about the stay. The nurse writes notes about the procedure, and so forth. Instead of repetitive numbers that are generated by a business transaction, the care giver generates text – descriptive information that depicts an incidence of healthcare.

Data warehouses have traditionally lent themselves to transactions and numbers. The institutions that have been successful with data warehousing – banks, retailers, manufacturers, and so forth – have files full of transactions and numbers. Healthcare has neither the transactions nor the numbers that are found in business. Instead, healthcare has textual descriptions of the different medical encounters.

It is because of these fundamental differences in information gathering and storage that healthcare and medicine have not been successful with data warehousing, regardless of the value that data warehousing could bring to the healthcare industry.

The technology that houses a traditional data warehouse is designed to manage transactional data that is highly dominated by numerical information. When textual, non-transactional information is encountered, the classical data warehouse technology of today is simply at a loss to handle healthcare information.

If data warehousing is to be a success in healthcare, a fundamentally different approach to the technology of information handling is needed. In short, what is needed is a “textual”, non-transaction oriented data warehouse.

A TEXTUAL DATA WAREHOUSE
Certainly there is some small amount of data in the healthcare environment that is transaction oriented and structured. Usually this data relates to payments and insurance coverage and claims. While this data is certainly important, for many purposes in the healthcare data warehouse environment there is other data that is far more interesting than billing and payments data.

The first step to the creation of the more interesting textual data in a healthcare data warehouse begins with the integration of text (just as the first step to a traditional data warehouse begins with the integration of numbers and transactions from the legacy transaction environment.) The need for the integration of text can be explained by – suppose that a data warehouse is to contain information from a lot of different sources – hospitals, outpatient clinics, doctor’s offices, emergency rooms, and so forth. In addition, the doctors that feed the data warehouse represent different disciplines – pediatrics, cardiology, epidemiology, orthopedics, gynecology, and so forth. Each of these disciplines has their own terminology. Furthermore the information going into the data warehouse is written by different levels of people – physicians, nurses, technicians, accountants and so forth.
THE TOWER OF BABEL
By ingesting all this diverse textual information, it is discovered that what is happening is a recreation of the Tower of Babel. Everyone is speaking a different language and no one really understands what anyone else is saying. What is needed is a common vocabulary. The first part of building a data warehouse for healthcare then is to read in all the different dialects and then to reduce the dialects to a common vocabulary. Once this fundamental ingestion and integration of textual information occurs, then there is textual agreement as to what is being recorded. Without this fundamental textual ingestion and integration, the data warehouse contains gibberish – many dialects that do not recognize other dialects where there is massive overlap that is unrecognized. The Tower of Babel has been recreated in healthcare when there is no integration of textual information before the information is placed inside the data warehouse.

TEXTUAL INTEGRATION
The term “textual integration” is not a commonly used term. Specifically what is meant by textual integration? Textual integration entails at least the following activities:

- Removal of stop words such as "a", "an", "and", "the", "which", "what", and so forth
- Word reduction to the Latin (or other) common stem – move, moved, moving, mover all recognized as coming from the stem “mov”
- Synonym resolution by replacement – there are at least 20 ways to say a bone is broken. Recognition of these terms all being synonyms for the same thing is one of most important first steps in the integration of text
- Synonym resolution by concatenation. In some cases synonyms are replaced; in other cases synonyms are recognized and are concatenated with the verbiage for the common term
- Homograph resolution by replacement. The term “ha” means heart attack to a cardiologist and “hepatitis A” to an endocrinologist. These homographs need to be recognized and resolved if the text is to make sense.
- Alternate word spelling resolution. Names and procedures all can have alternate spellings and acronyms. The system must recognize the terms and prepare a resolution
- Negativity exclusion. When “no” or “not” precedes a term, the term needs to be excluded from indexing or further processing
- External categorization of words. Healthcare text must be exposed to and analyzed versus different external categories of data to determine its relevancy
- Treatment of both words and phrases. It is not sufficient to look just at words. Phrases must be able to be processed as well
- Creation of internal themes form the text that is being read
- And other activities

In a word, there are many different editing and text modification exercises which must be done in order to achieve integration of healthcare data before that data can be placed inside a data warehouse. Once the text has been edited and integrated, then it is fit for analysis and decision making.

A COMMON VOCABULARY
In order to execute these activities, it is necessary to have a common healthcare and medical vocabulary. Sometimes the vocabulary is in the form of an ontology. In other cases
the vocabulary is in the form of a simple glossary. In any case, an extensive and exhaustive healthcare and medical vocabulary is needed. The healthcare vocabulary is needed in order to “normalize” the unedited healthcare text coming from the many different sources.

In prior days, the process of integration of healthcare text and terminology was an onerous task. The process of normalization and integration of healthcare text coming from many diverse sources had to be done manually and because there was so much data and so many synonyms the task was tedious, to say the least. Today there are tools that can do textual integration and healthcare vocabularies that are available for the purpose of automating the process of integrating healthcare and medical text. It no longer is necessary to manually attempt a large healthcare textual integration effort.

In the classical data warehouse development, there is a similar step to the achievement of integration of data inside the data warehouse. In the classical data warehouse, data is run through what is termed “ETL” technology. ETL technology stands for “extract/transform/load”. ETL software is designed for integration of transaction based, numeric based legacy systems data. While the software for the integration of unstructured textual healthcare data does the same functionality as ETL processing, the difference is that in the unstructured medical world it is text that is being integrated, not transaction based numerical data.

**THE PHYSICAL FORMS OF DATA**
Assuming that the ingestion and integration of healthcare text can take place, the next step is to place the textual data into the data warehouse in a meaningful manner. In actuality, the integrated text takes on several forms as it is placed in the data warehouse. Some of the forms that the healthcare text assumes as it passes into the data warehouse, is that of:

- Simple strings of text
- Text organized by internal themes
- Text organized by external themes
- Keyword indexes to selected groups of text

In addition, while actual text can be placed in the healthcare data warehouse, pointers to other text outside the data warehouse may be included as well. In other words, it may be useful to point to text not included in the healthcare data warehouse when the text that is pointed to is very voluminous. Or in other cases it may be desirable to point to text that has not been edited and integrated.

The net effect of the building of a healthcare data warehouse that contains a large amount of text is to keep the text that is the most important in the data warehouse and to keep the text that is of less importance or that has a lower probability of access in the external devices where the text can be accessed if needed.

**INMONCIF.COM**
For a much more complete description of how unstructured text should be organized and structured inside a data warehouse, refer to the web site inmoncif.com and look for the section on DW 2.0™. The treatise on DW 2.0™ fully describes how unstructured data should be integrated into a modern data warehouse.
Once the text is organized as described, it can be accessed by standard query tools or by textual visualization tools. The integrated text in the data warehouse can be accessed and analyzed by textual content – that is – a query can be made that looks only and exclusively at integrated healthcare text. Or the integrated text found in the healthcare data warehouse can be linked to parts of the structured portion of the data warehouse.

**PROBABILISTIC MATCH**

Linkage between text and structured data is made by means of what can be termed a “probabilistic match”. A probabilistic match is a match that may be accurate but may also point to a false positive. In other words, with a probabilistic match, the match is associated with a probability of accuracy. If a match is found where there is a strong likelihood of validity, then the probability of a match is high - .9 or higher. On the other hand if a match is found that is unlikely to be real, the probability of such a match is .1 or less.

As an example of the usage of a probabilistic match between the structured environment and the unstructured textual environment, consider several matches. One match is between two individuals – one in the structured environment and another in the unstructured textual environment – named “Pat Smith”. The name Pat Smith is a very common name, so the match may or may not be valid. But other information is found, and the two individuals have the same birth date, the same birth place, the same current address, and the same passport number. Under these circumstances it is a high probability that the Pat Smith in one place is the same Pat Smith in another place. The probability of a match would be around .95.

Now suppose there are two individuals in the structured environment and the unstructured textual environment. One individual is named William Inmon and the other individual is named Bill Inmon. Are these the same person? Even though the last name is not common, the first name is not the same, but is never the less related. So the probability of a match is probably .30 or thereabouts. If the names had been William Jones and Bill Jones, where the last name is very common, the probability of a match would be .10 or less.

By using a probabilistic match, the linkage between the structured data in the data warehouse environment and the unstructured textual data in the data warehouse can be used in synch with each other.

**TIME VALUE OF INFORMATION**

There is one major difference between data in the institutionalized world and data in the healthcare world, especially the research portion of the healthcare world. That difference is the time value of information. In the institutionalized portion of the world, data has a a short amount of time in which the data is useful. Once an airline flight takes off, reservation information about that flight does not have much value. In a bank, once the statement cycle is complete, the record of transactions has limited value. In an insurance environment, once a person has died and a life insurance claim has been paid off, the information about the policy is not worth much (at least to the bookkeeper. Perhaps the actuary finds the information interesting, but the accountant doesn’t have much use for it.).

But in the healthcare environment – especially the healthcare research environment – data has a long life indeed. For the purpose of studying a disease, medical records that are 50
years old may be extremely valuable, even where many of the people whose records are represented are departed.

For these reasons then, the time value of information inside a healthcare data warehouse is quite different than the data found outside the healthcare milieu.

EXTERNAL CATEGORIZATION
In a traditional data warehouse, data is typically organized by subject area. The organization by subject area is a convenient way to structure data for easy access and for the assurance that the data is integrated. There is a similar concept for the textual data, although strictly speaking the healthcare textual data is not really organized by subject area. Healthcare textual data inside the data warehouse can be organized by what is termed “external categorization.”

Almost anything can be an external category – Sarbanes Oxley, knowledge about cancer or specific kinds of cancer, ethics, profitability, improper language, and so forth. In order to organize data by external categories, the unstructured medical text is passed by one or more external categories and the number of “hits” made is noted. Some medical text may relate strongly to orthopedics, other medical text may relate to liver cancer, and yet other medical text may relate to HIPAA or billing.

IDENTIFIERS
One of the major challenges of management of unstructured medical data is that of finding or creating key or identifier information. There are usually two basic kinds of identifiers – document identifiers and content identifiers. A document identifier is something that identifies a document, such as a report title and date, an email address and date, a document registration number and so forth. A content identifier is information in the content of the document that identifies what the information is about. Simple content identifiers may include a patient number, a hospital entry number and date, a social security number, and so forth.

One of the challenges is the fact that a given patient may have different identifiers in different places. In one document the patient has a name, in another document the patient has a patient number, in another place the patient has social security number, and so forth. In order to integrate the unstructured data into a data warehouse, the identifiers must be consolidated and normalized.

SOME CONSIDERATIONS
These then are some of the considerations of a data warehouse for the healthcare community. It has been seen that an effective data warehouse for the healthcare community is very different from a data warehouse for other environments. The healthcare data warehouse consists primarily of text. The text must be integrated before being placed into the data warehouse in order for the data warehouse to make sense and be usable.