

# Future Trends in Data Management

## “Prepared” text

By Michael Scofield

A lecture to the Medical Librarian Group of Southern California and Arizona Convention  
Long Beach, California  
Presented February 3, 2005  
Revised Feb.22, 2005

### Preface

This document is a reduction of the remarks and concepts presented on Feb. 3 to these medical librarians. Many of these topics are addressed in a very superficial manner—there simply was not time to go into more depth. So I have inevitably left out important concepts or sub-issues in each one. However, this does give, I hope, a suitable introduction to the important aspects of data management, and what is being done in the field.

### The difference between data and information

The first major difference between data and information is the amount of thought which goes into its capture or expression. Data consists mainly of observations, either recorded mechanically (as in a seismograph, traffic counter, voting machine, or ATM), or passing through the senses of a human (i.e. an observation) before the human records it (pencil, keyboard, etc.).

The second major difference is the utility of what is recorded and presented. Raw data is never as useful as information, which is full-defined data and often data in context of other data or other information.

The difference in utility can be described a number of ways.

### *Data plus context yields information*

Data is useless with definition. And definition of any item of data (an observation of reality) is basically composed of two parts. Static definition, and dynamic definition. They are both essential to understanding the meaning of a fact or observation.

To illustrate the importance of definition in making a piece of data meaningful, and thus useful, let us start with a simple piece of data--an observation.

85

That is not a particularly useful piece of data, in spite of the fact that it is quite accurate. Why? Because we don't know what it means. It has no definition. Let me, step by step, add definition to this piece of data employing words and symbols.

85 Degrees

That’s a start. But “degrees” is a very ambiguous term. It could mean temperature (on two scales) or angular dimension. So let me be more specific.

85 Degrees F

Now we know it is a temperature. But temperature of what?

85 Degrees F air temp

Now we know we are describing air temperature. We can relate to that somewhat. It is fairly warm. We can imagine how it feels on our skin. Certainly warmer than room temperature. But we don’t know where or when. So we need to add that definition of this data.

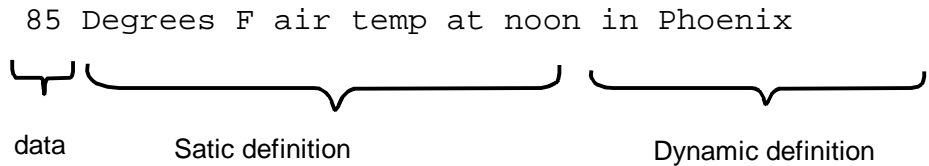
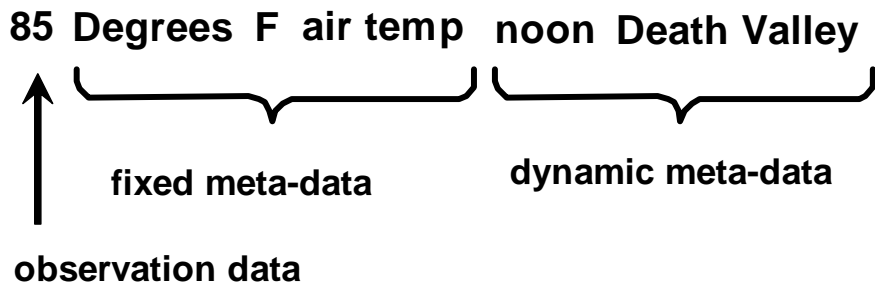
85 Degrees F air temp at noon

Now we have a time dimension to the data. It is a noon measurement of air temperature (we can make many assumptions here, including that it is the “air temperature in the shade” which is a common convention for weather measurements). But where was it taken?

85 Degrees F air temp at noon in Death Valley

Now, we added a spatial context (dynamic definition) to put around that observation, and then it becomes meaningful. Bringing our cultural experience of U.S. geography, and climate to this, we realize that this is not a very hot day. Death Valley can get to be very hot.

Again, let us parse out those words into what function they have.



The data itself is a small part of this text string. The static definition does not change over time or place for an “air temperature in the shade” observation. The dynamic definition tells us where (as distinguished from other noon air temp observations), this observation was made.

So not all data elements are created equal, or have the same function or role. Therefore, tests of data quality must consider the role the data element plays.

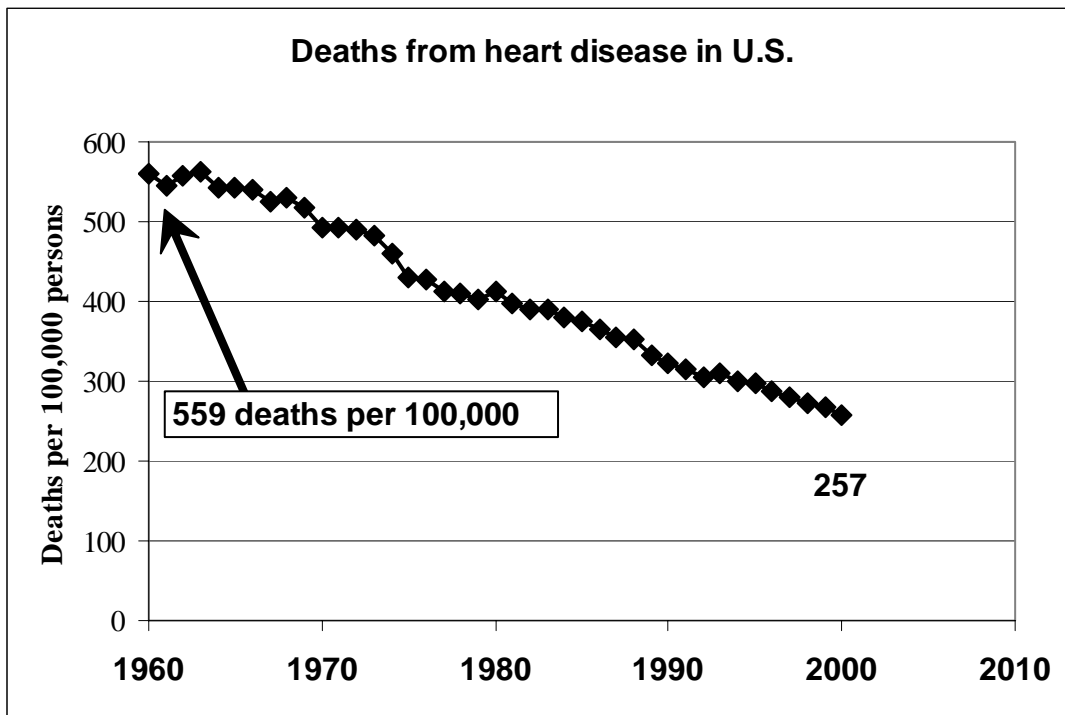
In many means of storing and communicating data, it is not fully-defined. The speaker (or writer) assumes that the listener (or reader) shares a somewhat common cultural context, and thus many aspects of the definition are left out. When we hear “It is the bottom of the eighth, and the batter is two and two”, we know we are listening to a baseball game, and 8 refers to the inning, and 2 and 2 refer to two balls and two strikes. We share a common cultural context with the speaker, and the clue words in the sentence help us understand what is being described.

But when data is recorded, and transferred in some way to a very different context, or read by someone not familiar with the culture, it may be misunderstood.

***Context can be provided, in part, by other data.***

My favorite is the statistic 257. Is that good or bad? Well, first I must define it. It is the number of deaths per 100,000 persons in the U.S. resulting from heart disease in CY-2000. But even when we know the statistic, we have no context to evaluate it if it is good or bad.

An excellent way to give context to any statistic is to show what it was previously, or depict it in a time series. I do that with the graph below.



Now, the statistic takes on more meaning. It is a part of a trend, and the trend is an improvement in that fewer people are dead due to heart disease. The euphoria is mitigated somewhat by our understanding that this trend may be because other things are killing them first. This statistic, to be truly meaningful, must also be combined with other statistics such as life expectancy (which is rising in the U.S. and elsewhere).

***Meta-data is absolutely essential to understanding the meaning of any fact of data.***

Meta-data is “data about data” or perhaps more broadly expressed “information about data”. It is far more than the database directory. It includes business definitions about both tables (what in the business world the table represents) and columns (fields) in the table (what attribute or characteristic is described).

Dynamic meta-data is the on-going behavior of data as live systems update the database with “observations” about the real world on a “go-forward” basis. Dynamic meta-data includes current usage of fields, and the current quality of the data. Usage, behavior, and quality of data cannot be determined prior to implementation.

***Context of data has a cultural component.***

Many statistics are uttered in a cultural context, without the speaker making much effort to fully and rigorously define the fact, data, or statistic. The speaker can get away with that because everyone in the culture (except for the newcomers) knows (almost intuitively) what the speaker means.

In a hospital situation, “patient count” may be assumed to exclude victims laying on gurneys in the E.D. The assumption is “patients formally admitted to the hospital and occupying a bed, and expected to stay overnight”. That excludes any kind of walk-in, walk-out clinical patients, coming either for diagnosis or treatment. That probably excludes gurneys rolling from area to area in the out-patient surgery center.

Does all this have to be spelled out? Well, to an outsider it might be. Organizations often have to report to accrediting bodies and government agencies various statistics which demand precise definition, and whose definition is often different from “the culture” of the institution.

Industries, too, have cultures, as well as do professional disciplines. Two endocrinologists talking about patients probably don’t define all the terms they use because they are common to the discipline.

## **A brief history of data**

## ***Early written record-keeping***

Early record-keeping (i.e. “writing things down”) is one of the first characteristics of civilization. Recordkeeping was an essential element to commerce, trade, the development of complex societies (and governments) and the division of labor which enabled them.

There were generally two kinds of written records: text and data. Perhaps a third, being graphics (illustrations, maps, etc.).

Early texts included history, sacred writings, philosophy, scientific theory, and creative writing (drama and fiction). On the other hand, raw data served science --first astronomy and climatology (when to plant the crops), and then commerce (property and land records).

## ***Babylon***

Ancient Babylon was the first society and government to make extensive use of formal record-keeping. Tablets have been found in archeological digs which reflect a wide variety of contracts, loans, mortgages, employment, wages, etc.

The “enabling” factors that moved from a primitive, agrarian society to a complex, semi-urban society (with labor specialization) include a strong government or authority, a monetary system and medium of exchange, and fundamental concepts of property and rules (the Hammurabi code). Without those factors, there was no division of labor, specialization, and hence the need to enforce more complex relationships among people and economic units.

So here, the recording of tabular data for commercial purposes first emerged.

## ***Melitus***

The Greek town of Melitus on the western coast of Turkey was a vibrant center of culture and learning. In the sixth century B.C. an astronomer in the town named Thales became the first astronomer to accurately predict an eclipse. The only way he could have done it was with meticulous and voluminous record-keeping, tracing the trajectories of the sun and moon with sufficient accuracy. This reflects the infancy of the tabular recording of scientific data.

## ***Early hospital records***

The first hospital in the United States (or, I should say, “the American colonies”) was in Philadelphia, founded in 1752. Ben Franklin was one of the trustees. In this hospital, medical recordkeeping began as a simple ledger. In each row were the following facts:

1. Patient name
2. Address
3. Disorder
4. Admission date
5. Discharge date
6. Patient’s security

That was it. No regular monitoring of vital signs. No diagnoses codes. No record of treatments given.

### ***Early accounting records***

The first double-entry bookkeeping financial statement dates back to 1399 in Italy. Entries were made for assets and liabilities, income, and expenses. This example happened to be expressed in three levels of Barcelonese currency (our U.S. system has just two levels—dollars and cents).

### ***Tabular records***

The parish records in the English country church include records of births in the community. These records are maintained in rows and columns, so that the facts (names, date of birth, etc.) gain their definition by their spatial position on the page, i.e. what column they are in.

So the definition of a column is written only once, at the top of the page, and does not need to be written again.

This actually forms the precursor of the relational model and relational database design. The simple concepts of tables with rows and columns (fields) has been powerfully useful in the most popular kind of database in use today—the relational database management system (RDBMS).

### ***Bletchley Park***

Bletchley Park was a country estate about an hour's train ride north of London (going out of Euston Station) where during World War Two, Alan Turing and a bunch of vary bright mathematicians and linguists worked to break various enemy codes. The grounds are now a museum, and a pilgrimage which every computer enthusiast should make. (Check their web site to ensure they are open on the weekend you wish to visit.)

It was here that a number of electro-mechanical computers were devised to speed the trial-and-error process of cracking German codes and ciphers. The total effort grew into a rather extensive bureaucracy as temporary buildings scattered around the grounds housed over 2,000 employees. Much of the effort was clerical in nature, taking “intercepts” and gleaning from them facts about persons, places, and operations of the German military.

The computers developed here (one known as Colossus) were used not for data management, but for number crunching. They didn't have much memory. The real management of a mountain of data (facts gleaned from observations) had to be done on cards and paper files, stored in baskets and drawers. But their efforts were successful, and they made a significant difference in the progress of the war.

## **Communications, an essential component of IT**

Information technology has a spatial component, and spatial challenges. Not all the employees of an enterprise live and work in the computer room (which would be awfully cold). Hence, the movement of data from where it resides to where it is needed is an essential leveraging component to IT.

### ***Speed and capacity***

The movement of data within a bureaucracy and between companies and enterprises (be they for profit, or non-profit) was first accomplished on paper documents, which required human intervention to interpret it. Then came cards, and later magnetic tape. Tapes were shipped between computers even into the 1980's until data communications networks were established. This required, in part, the expansion of capacity of networks (both voice and data, mostly run by various telephone companies).

The internet has greatly increased the flexibility of data communications. And the physical infrastructure has expanded its capacity by many powers of 10 as we moved from twisted pair, to coax, to fiber optics. Ironically, cellular telephone, in spite of all its flexibility, is confined to a very narrow bandwidth, and is susceptible to many kinds of interruption or failure. While the addition of cameras to cell phones may boost sales, any picture communication is necessarily very slow because of the narrow bandwidth of cellular communication.

### ***Semaphore of the British Admiralty***

The mighty British fleet was commanded by the Admiralty in London. The primary naval yard was down in Portsmouth, and from here many ships sailed with their final orders. Once out at sea, there was no way to amend those orders or give them new orders.

The Admiralty sought a way to speed final communication between London and Portsmouth, and in 1796 built a semaphore system—a network of buildings on hilltops, in sight of each other. Atop each tower was a structure with six large circular panels—large enough to be seen through a telescope from the adjacent hilltops in either direction. Men in these buildings would manipulate the panels to conform to the adjacent hilltop from which the message was coming.

These six round panels could be turned in an “on” or “off position; each panel thus was a “bit” in a six-bit

This was the first time that a message traveled further than the horizon and faster than the speed of the medium (i.e. a document, in a pouch, carried by a messenger on a horse). A test “ping” from London to Portsmouth (a distance of 64 miles) and back again required only two minutes.

But this mode of communication was very labor intensive (at least two men on each hilltop). For it to work, everyone had to be on duty when a message came along, and all do their jobs without error.

## ***Analog vs. digital means.***

The first forms of communication (telegraph using Morse code) was actually a digital means, although it required a human to interpret the clicks. While voice telephone was an analogue means of communication, digital forms expanded using teletype and punched paper tape. Television also employed an analog signal. Only in recent years with the advance of signal processing chips, is communication technology moving back to digital modes (even for carrying voice), which has numerous advantages.

## ***Symbiosis between IT and communications.***

There is now a profound symbiosis between computers and communication. First, without computer-driven telephone exchanges, we could not have the phone system we enjoy today. Before switches, human (usually female) telephone operators did all the switching manually. But this, too, was labor intensive, and as the system grew, digitally-operated switches and exchanges took over much of the work. That was fortunate, because it was estimated that to accommodate the growth (or market penetration) of the modern telephone in America, it would have eventually taken all the women in the country just to run the telephone exchanges.

Rising to a much higher level of sophistication the modern cellular telephone system relies on automation for a variety of functions, such as establishing the optimal power levels for both the phone and the tower (or cell site), and evaluating the relative signal strength from the portable phone as it moves out of range of one tower, and into the range of another. That “hand-off” is performed by a number of computers working together, and doing it so fast that the user cannot even hear the hand-off taking place.

## **History of automating tabular data**

### **Cards**

Hollerith cards (punched cards) were the most common form of data storage in the 50's, and well into the 60's and 70's for some organizations. Each card has 80 columns of punches, and 12 rows of punches. Once punched, the hole could never be restored again. So these were essentially “write-once-read-often” media. If your data (such as a long text field) required more than 80 characters (or 70 characters with the identifier is included on the card) you were up the creek.

### **Tapes**

Electronic magnetic tapes had greater capacity than the cards. More data could be stored in a smaller volume, and could be read and written faster. Most business applications used tapes emulating the card systems which went ahead of them. The problem with tape is to find one record, you had to read the entire tape.

### **DASD**

DASD (direct access storage device) is also known as disk, and allowed access to an individual record much more quickly than tape. Indexing systems, and eventually database management systems (DMBS) allowed the organization of data increasingly independent from the application itself.

### **Personal computer – major consequences in data management**

Until the introduction of the personal computer, all electronic data storage and processing was intensely institutional. It had to be run on mainframes, and shepherded by the programmers and computer operators—a kind of priesthood.

The personal computer changed all that, and with the spreadsheet, people were able to create their own data files (albeit usually keying in the data by hand) and manipulate the data into reports and graphs without waiting for help from the geeks down in the basement. This was a good thing.

The downside to the personal computer was data was created redundantly, and without any central control or coordination of definition, format, and rules. As a result, it was common for the VP of sales and the VP of finance to come up with totally different numbers when attempting to answer the same executive question.

### **Killer application—the spreadsheet**

The personal computer was only a toy for savvy experimenters until applications were developed to allow useful personal activities. The first real “killer application” was the spreadsheet. There were three different products in this area. First was Visi-Calc, which was superseded by Lotus-123, which in turn was eclipsed by Microsoft’s Excel.

Other heavily used applications which made the PC indispensable for many Americans were e-mail, text editors (or word processors), and finally the browser which allowed graphical access to the World Wide Web. But even web browsing became much more interesting and easy with search engines (first AltaVista, and now Google), which allow one to find a desired combination of text or words among billions of pages. This says nothing to the quality of what is found; only that it can be found quickly.

### **Tremendous improvement in performance of IT**

We continue to see staggering improvement in the price-performance of information technology, from large computer systems down to the personal computer. This performance improvement includes....

- The speed of the central processor
- Memory capacity and cost of memory
- Disk capacity and the cost of that capacity
- Speed of writing to disks.
- Clarity, brilliance, and granularity of visual displays

## **Modern IT environments**

A very useful model for understanding the total IT environment includes five basic elements, or layers. They are...

- The business
- The data which describes the business
- The applications which capture and store the data which describes the business
- The operating system which provides an environment for the application to operate
- The hardware infrastructure on which the operating system sits.

There are some models which are somewhat similar but with more layers

### **The business**

The business is what the data describes. In the broader sense, the business includes the environment of the controlled business. Both the business and its environment have logical data architectures. They are abstract, but can be described by a data model or E-R diagram. The model of many modern, complex businesses in intensely competitive industries and market is under pressure to morph, often to greater complexity.

### **Data**

Data is what describes the business and its environment. Data is an asset which has a unique characteristic, of the capability of being copied and transferred without being consumed. But much data loses its value over time, as it becomes stale, and potentially fails to describe current reality.

### **Applications**

Software programs (known as business “applications”) are written to capture data from normal, routine business processes, store that data, combine it, report it, and make it available for the “micro-decisions” of the enterprise. Applications can be home-grown, or purchased packages designed in a general manner by software vendors. All applications have some kind of data storage strategy, usually sitting atop a relational database which is designed to support the particular application.

Some applications (home grown) can share data from a common database, sometimes called an Operational Data Store.

### **Operating systems**

Operating systems (such as OS, MVS, Unix, and even Windows) provide an environment where the business or software applications can operate. In a sense, the operating system buffers the application from the hardware infrastructure, and provides common “utility” service to the application (such as file reads and writes).

### **Infrastructure**

This includes things like servers, networks, routers, DASD (direct access storage devices) and the protocols between them. Designing, installing, and maintaining these components requires some advanced skill sets, but they have little to do with the business. Indeed, this layer is about as far from the business as you can get.

## Data management essentials

In large, modern enterprises, data management (when it exists) includes a number of inter-related disciplines and roles.

**Data architecture** Building logical data models of the enterprise and its data and systems.  
**Meta-data management** Maintaining dictionaries and directories of data and its meaning.  
**Data quality** (assessment and improvement)  
**Data warehousing & business intelligence** Integrating data to support decision-making. Includes data mining, etc.  
**Data stewardship**  
**Data acquisition** (finding valuable data sources outside the enterprise)

Data management is a discipline which is often lost in an IT culture more oriented towards attention to process, programming, and hardware infrastructure.

## Data quality

Data quality is an emerging field as large, complex enterprises discover that their data asset may have lapses in quality. Detecting data quality problems can be either by anecdote (accidental discovery of erroneous data, often found by customers) or a deliberate effort to survey the latent data in the organization, as well as data coming in from external sources.

High quality data accurately describes the reality beyond it. Data quality has a set of sub-definitions, shown as a series of questions, with ever greater rigor and specificity in the box below.

Instance (row) present? (issue of scope of entire file)  
 Cell populated? (need to recognize null condition)  
 Is value in cell valid? (compare against rules)  
 Is value in cell reasonable? (requires context)  
 Is value in cell accurate? (requires definition)  
 How precise is the data in the cell?  
 Is value in cell current? (time dimension of definition)  
 Is the definition consistent over all dimensions?

A vigorous program of data quality improvement involves finding the bad data (through data profiling, data analysis), and a wide range of small, separate quality improvement projects conducted by those units of the business closest to the data capture events.

## Data warehousing and data integration

Data warehousing is a significant movement in information technology (IT) in that it makes data and information easier for the non-technologist to get to. Data warehousing is sometimes also known as business intelligence (BI) and decisions support systems (DSS). The boundary between these three concepts is subjective and somewhat blurred.

Data warehousing is data-centric, rather than process-centric of business applications systems, or hardware centric. And hence, successful data warehousing converts raw data into information which gives high-level insights into the behavior of the overall business and its environment.

### ***Descriptive elements of data warehousing.***

A data warehouse is a database which nearly always contains a *copy* of business data held elsewhere. Data is copied from original sources on a predictable schedule. Between the copy or update events the data is stable so that the same query posed twice (perhaps an hour apart during the business day) will yield the same results. It can be terribly frustrating to pose the same query and get different results (and hence the image of instability or unpredictability).

One of the most valuable roles of a data warehouse is integrating data from multiple (and diverse) sources into a common database. When an executive wishes to understand the cause-effect relationship in a process or phenomenon (such as the effect of advertising upon patronage), the data about cause and the data about effect generally come from separate sources—separate application systems and the databases beneath them. To bring the data together in a useful way requires that it is semantically integrated as well as physically integrated—and thus the “dimensions” around the data—the codes and time periods are compatible.

### ***Four-stage model of data warehousing.***

#### **1. Source databases**

The original applications in a business capture and hold data for the micro-decisions and business processes that they are designed to support. The databases underneath these applications are not necessarily designed for ad hoc queries and long scans of the entire range of data. Hence, it is best to copy such data once into a separate environment so that queries and reports against this data do not degrade the transaction processing performance of the application.

#### **2. Audit & Archive Database (staging)**

The audit and archive database is an exact replica of the original source databases, only containing more history.

#### **3. Integrated (“wholesale”) data warehouse**

The integrated “wholesale” data warehouse is a large, complex database containing granular data from a variety of sources. Those sources can have different logical data architectures, and hence the design of this database may be quite difficult. The transfer of data into this database may require translation, and filtering.

#### **4. Topic-oriented data marts**

Data marts are small, agile databases, often with star-schema structures to enable ad hoc queries by business users (rather than skilled programmers). BI (business intelligence) tools provide the access and flexible reporting capabilities in this area.

### ***Challenges of semantic data integration***

Bringing data together from structurally disparate source systems requires semantic data integration (not just making the data tables co-exist on the same platform) which can be quite difficult. For example, merging patient data from two different systems requires that the patient identifiers be compatible, both in format, and in meaning and behavior.

If, for example, Mary Smith has a patient number of 1045 in System-A, and she has a patient number of 8849102 in System-B, it becomes more problematic to match the data records from those systems, together, to create a more comprehensive understanding of Mary Smith.

Semantic data integration is a complex design process with many potential points of failure.

### **Converting data into information**

Information must be delivered in the language and paradigm of the intended user. And that translation is part of what moves raw data to useful information. Summary reports, exception reports, and graphical techniques are most useful in converting (and aggregating) raw data into more meaningful information for the decision-maker.

### **The Data-ization of knowledge**

A common phenomenon observed in the growth of large, complex enterprises their use of IT is the inevitable movement from integrated knowledge to fragmented data. This is actually a degradation of information into data.

### ***CRM***

The history of business (of any kind) includes growth of successful enterprises. Most begin as a sole proprietorship, and grow through the addition of more workers. As more people are involved, the enterprise fragments into either specialization, or parallel functions. Where

there is functional specialization, there must be exchange of information. When verbal interactions are not enough, information is reduced to data written on forms (which give structure and consistency to the information). In a digital age, those forms are the seed of computer application and systems design.

Large, complex bureaucracies today often contain much data about key objects of interest (particularly customers) in a variety of databases, and that fragmentation of the data makes it difficult to get a complete picture of the customer (or patient).

CRM (customer relationship management) is a technology which attempts to re-integrate all this disparate data to form a more intuitive-like picture of the customer, by which management and customer-contact employees can have, quickly, a more complete picture of the customer, his unique characteristics, and his value and relationship to the bureaucracy.

## Conclusion

Data is a significant asset of any enterprise, as is knowledge and information. The emergence of new information technologies have given us much more data, and many more options how to use it. Yet managing that flood of data, and making it useful to decision-makers has been a major organizational challenge. The management of data requires a skill set and understanding far beyond mere programming.

Progressive organizations of all kinds are beginning to recognize the importance of managing data and information as an asset.

Comments are welcome and should be addressed to the author at [NMScofield@aol.com](mailto:NMScofield@aol.com).

**Michael Scofield** is a popular speaker and consultant in data quality and data management. He is an Assistant Professor in Health Information Management at Loma Linda University. He has held numerous posts in data architecture, data quality, and data management. His energetic delivery, laced with generous humor and vivid graphical examples, keep audiences engaged and learning.

His articles on data architecture and data quality techniques have been published in Information Week, IBI System Journal, Data Management Review, the Cutter IT Journal, and the Database Newsletter. His speaking engagements include DAMA-International conferences, Meta-data Conferences in London and the U.S., over 12 DAMA chapters, 4 Oracle User groups, DB2 user groups, and various CASE user group conferences. He also writes humor, published in the Los Angeles Times and other journals.