## In this report

As the art of data-base organization develops over the next decade, corporate executives will be faced with a wide range of decisions which will determine whether their data-base system is an asset or a liability to their firm. These decisions will be complicated by a number of issues which range from technological inadequacies—the lack of appropriate hardware and software, for example—to basic misunderstandings of the limitations of data-base systems. Most important among these issues will be: (1) the conversion of data into information; (2) the development of hybrid systems; (3) the validation of data; (4) the control of data-base "pollution"; (5) the structuring of records; and (6) the preservation of privacy. Both users and manufacturers can work to resolve these issues in the coming decade. As networking replaces the single, in-house computer, however, the basic economics of computer usage will be the major force shaping the data-base systems of 1985.



# The Corporate Data Base: Asset or Liability for the Future?

The large corporate data base can be a valuable asset to a firm, with the potential to provide timely and valuable information to executives. Unfortunately, it also can cause incredible headaches.

The positive side of data bases—their potential benefits to management—has been stressed so many times that it hardly begs repeating. A modern corporation clearly could not operate without powerful information systems. However, the claims for the future development of such systems are too often exaggerated. The rosy projections painted by hardware and software suppliers alike range from "total" management information systems which surround the company executive with space-age gadgetry to complicated schemes for the computer professional that are couched in the language of graph theory or boolean algebra. But the problems which are likely to arise in the day-to-day application of data bases to the work of an organization have received little comment. These problems—and the prospects for their solution as the art of data base organization develops over the next decade—deserve some careful thought.

* * * * *

The issues which surround the creation and maintenance of a data-base system range from technological inadequacies to policy problems to basic misunderstandings of the limitations of data-base systems. Of these issues, six seem particularly important for the corporate executive:

### Issue 1. Understanding Information

Perhaps the greatest fallacy concerning data bases is the belief that data retrieved from a computer necessarily constitute information. There is no way to retrieve **information** from a computer, for the simple reason that information cannot be stored there in the first place. The only thing the EDP department can store into the company computer is **data**. Data are converted to information only for a particular user at a particular time. Unfortunately, the process by which this conversion occurs is very mysterious indeed.

In a modern hospital, for example, we may find a single record for Mr. Apple, who underwent an operation in May. The data in this record will be used by three different kinds of users: The administration will want to know how many beds, on the average, are empty in June. A nurse will want to know if Mr. Apple's X-rays have come back. And the medical researcher will be looking for possible correlations between symptom and complication patterns among all patients who have undergone the same treatment. In this example, it

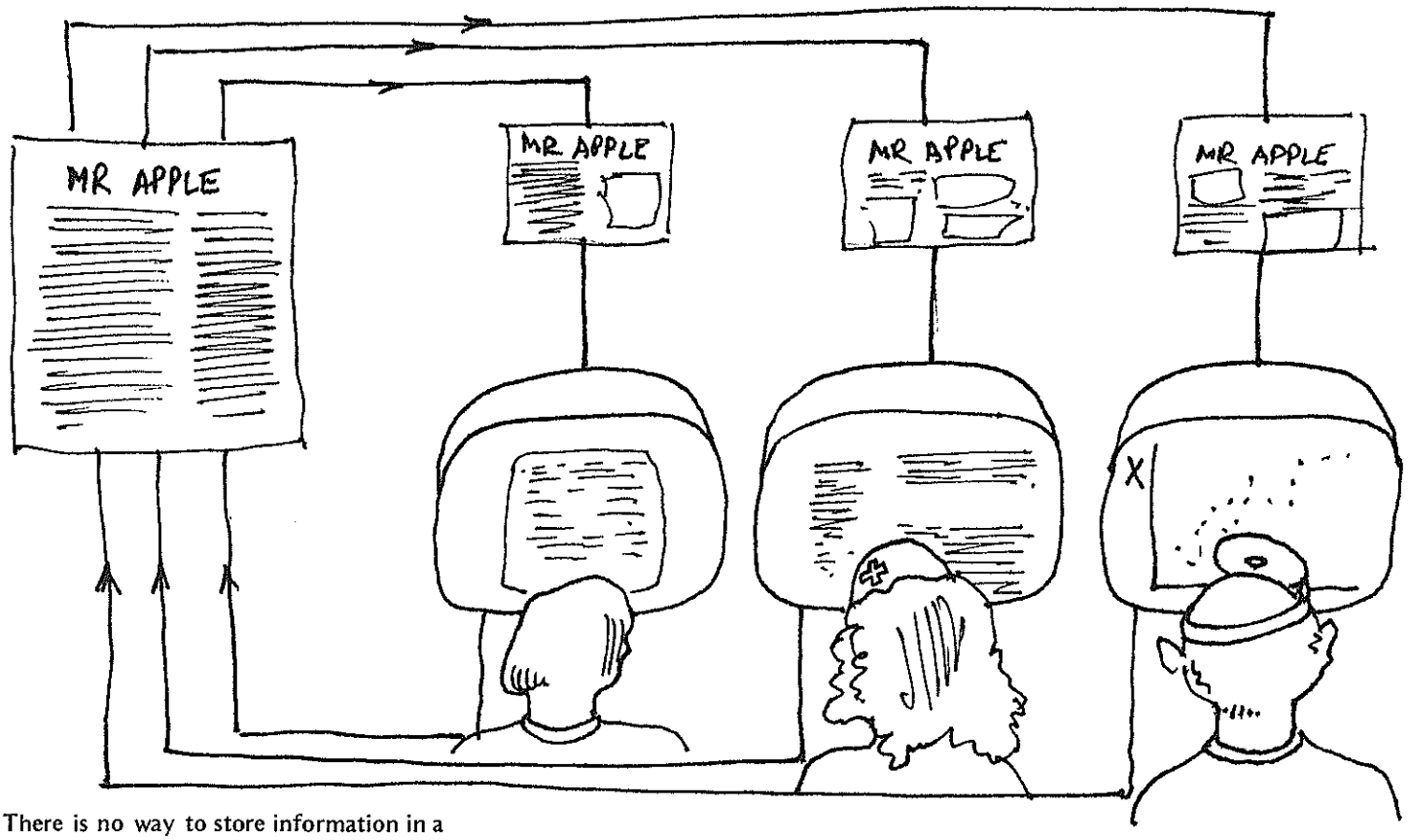
is easy to see that what constitutes information for one person may mean nothing to another—perhaps because it is couched in the wrong language, expressed in the wrong units, displayed in the wrong format, or presented at the wrong time. The need to support creative decision-making in this sense has not yet been taken into account by software designers.

In current systems, there is no provision, for instance, for storing the "retrieval profile" of a frequent user (the administrator or the nurse) and for applying it against the data base. Yet such a profile would itself contain useful information. In the absence of such tools, the task of finding an optimum structure for the data in storage falls upon the user's shoulders, while it rightly belongs to the computer. If you have been wondering why your company needs so many clerks to fill out data entry forms and record layouts, the answer is that someone has to anticipate the possible retrieval of every bit of information. This work has to be done by intensive human effort because the computer has not yet been endowed with the level of software sophistication that would enable it to take over this function. And unfortunately, there is no system on the horizon that will practically fulfill this need.

### Issue 2. Developing Hybrid Systems

In the coming decade, practitioners of the data-base field—as opposed to academic theoreticians or system implementers—will be placing increasing emphasis on the linking of disparate elements into hybrid systems. It is quite conceivable, for instance, that the manager of a corporate data center in 1980 will respond to some user demands by providing a text-editor to remote locations through a network; he will satisfy another group by maintaining a programming language facility (such as COBOL) at a central site; and he may offer a third group access to an information retrieval package to search engineering and patent files, personnel records, and customer data. No single system can meet all three needs in practice, and the manager in this example may have to live with a hybrid system, "patching" the various components together. And in spite of the emphasis on data integration and minimal duplication, the era of the grandiose, single-system approach seems to be over.

Some planners suggest that data bases should be maintained locally, under control of the people who originate the data, and that they should be linked in network fashion. Although this approach




There is no way to store information in a computer; only data can be stored. Data are converted into information for a particular user at a particular time.

would result in some duplication, they argue that, in many cases, such duplication is desirable. Smaller, more disseminated data bases, they say, will be more reliable than large centralized ones.

### Issue 3. Validating the Data

Another difficult problem is that of data validation. When a user obtains a record from the computer, how does he know that the information is accurate? Everyone has, by now, encountered a "computer goof." Typically, the problem may be traced to human error. The data entry clerk may have punched the wrong digit, for example, sending three planeloads of widgets to Minneapolis instead of three boxes to Atlanta. Some human error of this type is always possible under the most elaborate validation scheme. The level of error can be reduced in various ways, however, and in this respect, progress is fast being made. New computer terminals now have built-in controls that allow information to be entered only in certain fields and in a certain sequence. "Smart" software such as this can be developed to insure consistency.

### Issue 4. Controlling Data-Base Pollution

The greatest cause of "data-base pollution" is not poorly validated data, but a deliberate or unconscious bias. In this case, the data is entered correctly by the clerk, but it does not correspond to anything in the real world because the originators of the data were motivated to provide something other than the truth. During the Vietnam War, for instance, many high-level decisions were made on the basis of statistical information that gave a completely erroneous picture of events in the field. Local officers either had to provide information they did not have or were under pressure to "tell the people in Washington what they wanted to hear." The potential for such a bias exists in every corporation and in all branches of government, from welfare and environmental protection departments to crime statistics bureaus.

In estimating mineral and energy reserves, for example, the user of a data base must take into account the fact that the available information will be grossly affected by human bias, depending upon the nature of the commodity and the way it is taxed. If a depletion allowance is in effect, the quantity reported will be affected in a different way than if the owner is taxed on the reserve. Much of the confusion regarding national projections for minerals and fossil fuels undoubtedly comes from this level of bias in the data bases.

An illuminating account told by the French writer Colette demonstrates the problems of information pollution, even in the most prestigious data bases. She wrote in her "Reminiscenses" of spending a vacation in Brittany with a family friend who was a cataloguer at Paris's impressive Bibliothèque Nationale. The weather was often inclement, and he would then stay inside filling out little cards. When questioned about it, he stated flatly that he was working on the Master Catalogue.

"You know all these titles by heart?" asked Colette, very impressed.

"Not at all" he replied. "You see, I have noticed that we were sadly lacking in German manuscripts of the 13th and 14th century and in certain autographs, so I am making entries for imaginary works that could well exist."

"But how can you do such a thing?" she insisted indignantly. "The actual books haven't ever been written."

"Ah, you can't expect me to do everything!" answered the catalogue specialist.

The problem of bias—deliberate or unconscious—can be controlled. Astronomers, for example, have learned to correct their catalogues for "instrumental error" and "personal equation," but their practices have not yet been applied in the business world. Data-base pollution can only be detected by complex statistical examinations—by checking all of one source's entries against the distribution of all others, for instance. The expense represented by such processes should be included in any cost projections for new data-base systems and their use.

### Issue 5. Structuring the Records

There are several competing design philosophies which can underlie a data-base system. Data-base experts disagree as violently about alternative representations of records (as trees, graphs, or relations, for example) as automobile experts disagree on the various alternatives to the internal combustion engine. Their concepts of what makes the system "tick" are so abstruse that most users are discouraged from attempting to familiarize themselves with them. In selecting a system, however, it is crucial to understand its theoretical limitations and to project its possible obsolescence.

The current differences in conceptualization between "relational" data bases and other systems will not be resolved soon. Some experts state privately that they expect no breakthrough in this field within the next five to ten years. And even when basic differences in the representation of records are resolved, the user may still face difficult choices. Many human engineering questions (such as levels of access to the data base, computer assistance, and simplified languages for "interrogating" the data base) remain undefined. This paradoxical situation will continue to prove frustrating to users, bewildering to corporate executives, and disappointing to those software designers who thought ten years ago that a simple mathematical representation of information could soon be found and adopted by everyone.

### Issue 6. Preserving Privacy

The whole issue of privacy is adding a new variable to an already complex problem. It will make design problems more delicate in the future, forcing drastic changes of plans in corporations dealing with individual information or with public records. Internal security will also become a major factor as the issue of computer-related crime is better understood.

The computer industry is about to introduce new protection mechanisms in hardware and software to make data more secure. What impact greater security and privacy will have on various social groups (including social "deviants") remains to be seen, but the technology will soon be here to make data bases practically indestructible and, to a very high degree, private.

* * * * *

The resolution of these issues in the coming decade will depend largely on the priorities of users and manufacturers alike. Users, for example, could greatly influence the area of standardization. In particular, they could make their needs more clear through the various groups associated with the Conference on Data Systems Languages (CODASYL) and the American National Standards Institute (ANSI), which are currently dominated by com-

puter manufacturers. A common data description language could save much difficulty and special purpose programming as users move from one system to another. Also, a basic language for "interrogation" with simple conventions would save many training dollars and make access to company files much more understandable.

Manufacturers, too, can change the outlook for data-base development in the future through an investment in new software or new hardware. The implementation questions associated with data-base systems center on a single, simple paradox: computers are designed to perform calculations and are extremely effective in such operations as adding numbers or testing a numerical value. However, these operations are rarely required in data-base work. Instead, information systems need to locate tables, to move and format text, to match strings of variable length, and to organize displays. Very few computer languages reflect these operations in their command structure, and programmers are forced to "simulate" them by some combination of arithmetic and logical operations. A computer manufacturer could respond to this need with a careful study of such primitive functions, designing a language—or even a machine—around them.

In the absence of new efforts by users and manufacturers, the current trend to extend the COBOL language will likely continue well into the 1980s. Immediate practicality will take precedence over innovation until some crisis—possibly caused by the sheer volume of the information to be stored—forces corporations to consider new concepts. At the same time, however, the basic economics of computer usage (online costs and storage costs) as more and more companies switch to networking instead of relying on a single, in-house computer will shape the 1985 world of data bases.

As the marketplace evolves, with interactive systems increasingly available through networks, decision-makers should be alerted to several pitfalls (see page 5). Careful attention to these may prevent the transformation of that unique asset—the operational information a company needs—into a source of headaches or even a serious liability.

**In spite of the emphasis on data integration and minimal duplication, the era of the grandiose, single-system approach seems to be over; networking may well replace the single, in-house computer.**


# Ten Pitfalls in Data-Base Management

The most common management problems in the selection of a data-base system and the creation and operation of a corporate data base arise from misconceptions in the following areas:

• **Evaluating the performance of a data-base system by the speed with which it *retrieves* data.** Good response time to an interrogation is certainly important in data-base systems: Who owns account # 17581? What is today's exchange rate for the Japanese yen? The answers to such questions, given current computer technology, can be given arbitrarily fast. Performance in *update* speed, however, is the best criterion for evaluating a system. You pay for retrieval speed whenever you *update* the data base. So carefully test, time, and check the update phase.

• **Assuming that existing records can be simply converted to run under a new system.** Conversion is rarely a one-time operation. Will you lose some of the capabilities you were enjoying under a previous system of files and programs? If you move all your records under the new format, what will happen to all the COBOL programs your staff has already developed, using data in a different form to produce management reports? Be sure to take these questions into account when you make cost projections.

• **Ignoring the need for a processing capability during retrieval or update.** Suppose you want the names of all employees whose salary is above the average salary of their department. Suppose you want to know how many salesmen account for half of the orders received last year? Are you sure your data-base language can answer such questions directly? Look again. You will likely be told that someone will have to write a "small" special program. Very few systems have such a built-in capability.

• **Misunderstanding the complexity of the task of changing the records once a basic structure has been defined.** What happens if you have allocated 29 characters for each customer name and this turns out to be too short? Suppose you didn't realize that you would need to store the employee's birthdate as well as his age at hiring: are such mistakes irreversible? Will you have to reprocess the entire data base to make room for the new data? Examine carefully the system's ability to revise its structure dynamically.

• **Minimizing the importance of recovery and restart.** If the computer crashes during an "interrogation", the search can be resubmitted. But if you lose access to part of the system during an update, what happens? Can the designers of your data-base system guarantee that they will be able to recover lost transactions and restart the process exactly where it left off?

• **Delegating the task of data coding and data entry to clerks who are not familiar with the subject matter.** The difference between 63.5 and 6.35 is small—unless you deal with the composition of a new drug or the exchange rate of currency! Only a coder who understands the nature of the information can spot such errors at once.

• **Assuming that data in the computer file is necessarily valid.** There will be errors in the data base, some of them intentional, others resulting from subtle social or bureaucratic biases. If the system is well-designed, it should be possible to check for internal consistency of the data base, but it's more an art than a science! When you plan the data base, prepare a sizeable benchmark of basic questions to check for overall validity: How many of your employees are over 99 years old? You might be surprised at the answers.

• **Assuming that the system will replace the specialists who are currently handling the information.** The specialists are rarely replaceable. A well-designed system should not try to eliminate them, but to redefine their roles and to support them by making it easier for them to spot trends, control data quality, or to perform similar important functions. Do not be misled by the claim that "stepping up" to a data-base system will automatically result in reduced staff. There are some functions that a corporation simply cannot eliminate.

• **Ignoring the role of the data base as a social network.** Not only are different groups within the company originating (and claiming ownership of) various groups of records merged into the data base, but the many remote users accessing it from terminals will soon form a new type of community within the company—with its own rules, language, and rituals. In a company where various departments access the same system, you may find that there is a stronger bond among these users across departments than among any given structure in the company. This new entity will have its own way to resist change, to define new needs, and perhaps to introduce innovation; it might also end up being the tail that wags the dog.

• **Failing to recognize the need for new job descriptions for such positions as data-base administrator and user training specialist.** A well-designed data base may be poorly used or may fall into disorganization simply because the management assumes that it will "take care of itself." A data base changes from day to day as new records are entered, old ones are taken out, various "generations" of information are achieved, and the profile of the users changes. These facts define the need for new human roles and new job descriptions.