The Mass of Data

BY MICHAEL NORTON

As if there isn't enough for any harried IS manager to worry about, I was recently reminded just how temporal archives really are. Sometimes the loss is due to breakage or the loss of magnetic properties; sometimes the application used to process the data has disappeared completely. Regardless of the cause, the reality is, despite the best laid plans of mice and IS managers, data is always at risk. The spectre of data loss is enough to strike terror in the heart of any computer professional who has been indoctrinated with the primacy of data. For this reason most of us become pack rats when it comes to data, archiving rather than deleting. Do you have duplicate directories on your system? What about those tapes, disks and cartridges stashed away in various niches in your office, many of which are at best redundant and at worst completely useless?

DEFUZZING LOGIC

Of course, one man's trash is another's treasure, and there will always be someone with the insight and incentive to mine data stores to glean another bit of useful information from the data. Indeed, an entire industry has arisen to provide the business and program logic for data mining. Their task is to find the proverbial needle in a haystack. The task of finding data can be daunting, and at times, even impossible. Search for just about anything on AltaVista or one of the other Internet search engines and you are likely to receive thousands, if not hundreds of thousands, of hits. The prolific number of matches is not due to poor programming. Indeed the complexity of data retrieval technologies such as fuzzy logic attracts some of the brightest minds in the industry. Nor would faster hardware eliminate the problem. It is somewhat startling, even frightening, to realize that even our best logic is completely overwhelmed by the sheer mass of data.

Regardless of the cause, the reality is, despite the best laid plans of mice and IS managers, data is always at risk.

DUPES, DEADS, AND OTHER BOGEYMEN OF RECORD

Last I heard — and this has been a couple of years — there was something like 50 million web sites (not pages, sites). And by far the vast majority of data isn't even on the 'net but in hundreds of thousands of corporate databases, some of which are reaching mind-boggling sizes. There is an increasing public awareness and alarm concerning the amount of data being amassed on each of us, including not only our employment, credit, and medical history but also what we purchase, who we contact, and what we do for fun. Indeed, the Orwellian vision of a society tightly controlled by the use and misuse of information has become ingrained in our popular culture to the point it is assumed to be true.

What prophets of cyber enslavement fail to consider are the difficulties locating, accessing, and organizing that mass of data. In the first place, there is no central repository of data — nor could there be, with today's hardware and software. Big iron is not quite that big yet. Instead, information is fragmented in myriad independent databases. Just locating where data on a subject is stored can be such a labor-intensive task that using the data becomes unfeasible. Second, contrary to popular myth, corporations do not often simply turn over their databases to just anyone who asks — or even the highest bidder, for that matter (In case you're wondering, these lists are often provided to a bonded mailing house, independent of either company — and even then, only the salient information is exchanged). Much of that information about you and me will remain behind the firewall, protected (as it were) by varying degrees of security. But even assuming access to data, organizing it into something useful and meaningful is no small task. If the value of a hundred million email addresses was so immense, why do companies sell them for a couple of hundred bucks? Every IS professional knows the fallacy here: By the time you've weeded out all the bogus, erroneous, irrelevant, and duplicate entries, a fraction of the original entries remain as prospects to compensate for all the weeding out of bogus, erroneous, irrelevant, and duplicate entries.

PUTTING AWAY THE CHISEL

Indeed in many instances acquiring or requiring data is actually the more prudent course of action. Sometimes it is easier to (gasp!) just re-key the data. There is a psychological barrier here, however. Raise your hand if you've ever spent hours attempting to recover data it would take less than half that time to re-enter. One of the promises of computers was that we would finally be liberated from the onerous burden of having to constantly re-key data. Before computers, someone had to enter a customer's

name on a customer record, and then enter that same name on the solicitation letter, the shipping label, and each month's invoice. Computers changed all of that, of course, and we have come to believe that data in the electronic format is fundamentally different from the other media (except possibly stone) on which we have used to store information. This is true in many respects, but the differences are often exaggerated. There are many more similarities, the most fundamental being the inherent problems of managing information.

MANAGING KNOWLEDGE

Is it information we're trying to manage or is it really knowledge? Or data? In this column I've more or less used information and data as synonyms. Are they? Although the two words are interchangeable here, there is a distinction. There is much more data than there is information, more information than knowledge, and more knowledge than wisdom. The meaning of these terms varies from age to age, culture to culture, and even among individuals in the same time and place. However, the progression itself seems almost universally accepted.

And it is certainly the direction the computer world is moving. It is no accident that Lotus announced that the Domino/Notes product was moving toward "knowledge" management. History has shown that Lotus is way ahead of the curve in grasping the nuances of human/computer interaction, first eliminating the cumbersome file paradigm with databases,then popularizing the concept of groupware. The Internet, which Lotus was relatively quick to embrace, has made clear that it is not enough to simply have access to massive amounts of data; that data must be organized into information and forged into knowledge. It is actually rather uplifting to realize that it is the real mission statement of each one of us in the IS/IT world to achieve that objective. ts

Michael Norton is the network administrator at SofTouch Systems, Oklahoma City, Okla., which provides both mainframe and PC software solutions. He has written mainframe manuals in addition to articles for a number of publications. Michael can be contacted at norton@softouch.com.