More Is Not Always Better

Tineke Manford

UNITEC
Auckland, New Zealand

tmanford@unitec.ac.nz

ABSTRACT

The data resource in many companies today is in a state of disarray. Hidden data resources may exist, if the organisation at large is not aware of all the data that is created within the organisation. There are many problems associated with this bad practice; low data quality is one of them. Since data forms the basis for decision-making, the quality of decisions can only be as good as that of the underlying data.

If Information Technology professionals lack understanding or do not take into account user needs when they design and administer databases, redundant data resources will be created by the users. It is important that computing students are not only taught the technical procedures for creating databases, but are also taught the consequences of ignoring user requirements.

This paper uses the story of a large company in Auckland that found itself in a situation of having thousands of redundant data sources, to illustrate the problem.

Keywords: Data quality, redundant data, data administration, computing education

1. INTRODUCTION

November 1999. The Y2K project team is quietly pleased with itself. It has been working hard for 18 months collaborating with programmers, analysts and vendors, to make sure the start of the new century won't cause problems on the company's computers. All systems have been assessed and fixed where necessary. The organization is an international billion-dollar company that has many big enterprise databases including human resources, payroll, financial, supply management, customer loyalty, marketing and booking systems. Now that all these have been checked, the only thing remaining is to see if there are any user-created spreadsheets or databases on the company's 400 or so desktop PCs that pose a Y2K risk. The Information Technology (IT) department has been charged with identifying these MS Excel and Access files and they are expected to report back soon. There is surprise and horror when the list is finally produced. Nobody has anticipated over two thousand files. Why did end-users create this many files when surely the organisational databases meet all the data and information requirement of the company? Is there enough time to make the necessary corrections in these files? An investigation gets under way: spreadsheets and databases are inspected and file

owners surveyed and interviewed. It turns out some of the files are no longer used and can be discarded.

The remaining are patched up. Everything gets finished on time and when the century changes, no Y2K problems are experienced with any of the systems or data. Thus we have a happy ending to this story. Or do we?

This paper explains what causes the creation of unnecessary data resources and the reasons why redundant data resources are a problem. It argues that IT professionals are partly to blame and that the problem can best be overcome with better communications between them and their users.

2.   WHAT CAUSES UNNECESSARY DATA RESOURCES?

The Y2K team in the case study found that the Excel and Access files, created by the users, contained a variety of information that ranged from not-so-important departmental birthday lists to very important operational data on which spending decisions involving millions of dollars were being made. Why did the users create these data resources? The enterprise databases already contained most of the data and there should not have been a need to duplicate this. Interviews with the users discovered many reasons; the main ones were:

♦ The user thought the enterprise reports were inadequate.
♦ The user considered the enterprise databases to be user-unfriendly and not meeting their needs.
♦ The user did not trust the data in the enterprise databases.
♦ The user thought the data did not exist in the enterprise databases.
♦ The enterprise databases only contained some of the data the user wanted to work with.
♦ The enterprise databases were slow and the user wanted faster access to information.
♦ The user did not have access to the enterprise databases or reports.

Since several of the enterprise systems were legacy systems with character-based user-interfaces, it is understandable that users found it hard to use these systems. Standard reports in many cases did not reflect user needs and user-defined reports were hard to create with most users having to wait for the IT department to do the job for them. It is not surprising therefore that users tried to retype or export data to Excel where information can be easily manipulated and graphed. Documentation was poor and users did not fully understand the data stored in the company's databases causing them to mistrust the data or think it was not stored in the enterprise databases. Some of the databases were very big and query performance could be very slow during certain times of the week and month. Consequently, by keeping MS Access databases, users found it gave them fast access to information, even if it meant they had to re-type a lot of data into these systems. There were also users, with no or limited access to the enterprise databases, who were keen to, or were asked to, carry out tasks involving 'getting answers to questions'. These users had no idea of the data already available to the company and collected their own data and stored this in Access databases.

Was this company unique in having multiple unnecessary data resources? Bracket (2000) suggests this is not the case. He found that most organizations are in the same situation and for the same reasons: users cannot find data, users don't trust data or cannot access data in existing systems and therefore create their own data resources.

3.   WHY ARE REDUNDANT DATA RESOURCES A PROBLEM?

There are several problems related to end-users creating their own data resources. When the same data is stored in multiple locations, there is a risk that the different copies get out-of-synch. It may become unclear which of the data resources has the most up-to-date or accurate information and therefore it becomes impossible to determine which is the preferred data resource. The problem may not be restricted to one or two copies of data items; in fact Bracket (2000) believes that for companies that have been in existence for a number of years, ten copies of each data item are stored.

Another issue related to multiple copies of data is the cost of maintaining them. For each copy, the company is paying an employee to keep it up to date and this is an unnecessary expense where redundant data is concerned.

In sophisticated databases, such as Financial or Supply Management systems, much effort has gone into designing a system that ensures the best possible data integrity. PC databases, such as MS Access are easy to use and with little training people are able to create quite large databases. Most end-users, however, are unaware of issues surrounding data integrity and the methods of ensuring it (Perry, 2001) and as a result databases created by end-users often pose serious data quality risks. When information from these databases is used as the basis for multi-million dollar decisions, the business could have a major problem. The Data Warehousing Institute estimates that data quality problems cost U.S. businesses more than $600 billion a year (TDWI, 2002).

Computers can fail for a variety of reasons but enterprise databases are, to some extent, protected from disasters by backup and recovery procedures. Data resources stored on an end-user's hard disk, however, often have no such protection and loss of data is a severe risk in these cases.

The right to use data in an organisational database is usually determined by strict security rules and enforced with logins and passwords to control access. When end-users take data and store it in personal databases or spreadsheets on their hard drives, no such restrictions are applied, allowing anybody to read or change the data and posing a significant risk to security and privacy.

Anecdotal evidence suggests that when users maintain their own data resources they become less interested in the corporate data stored in the enterprise databases. This often leads to an increased lack of understanding and awareness of the corporate data and consequently the need to create more personal data resources (Bracket, 2000). As a result, unless appropriate action is taken, the problem gets worse and organizations may find redundant data resources perpetuating themselves by encouraging the continued creation of further redundant data (Bracket, 2000).

4. DOES IT HAVE TO BE LIKE THIS?

Database administration involves implementing, maintaining and optimising the physical database environment. The database administrator is responsible for the day-to-day operations of the database. Data administration, on the other hand, is concerned with integration and integrity of organisational data. Its general goals are to improve data quality and accessibility (Little, 2001). Data administrators have close contact with end-users and listen to their requirements and complaints and their activities should be incorporated into all information systems planning, development and maintenance activities. Most companies realise the importance of database administration, but frequently the functions of data administration are not taken seriously enough and users are not consulted and listened to. These organizations often have computer systems that are unchanging, even though the business needs may have changed (Smith and Paredes, 2002). By not providing information that meets users' needs, IT departments are almost forcing users to create their own Excel and Access databases.

Anecdotal evidence shows there are many IT professionals who deem business requirements less important than technical requirements and consider end-users the cause of most problems. On the other hand there are users who believe computing is easy and things can be worked out without involvement of the IT department, or worse, the IT department has nothing to do with their job (Smith and Paredes, 2002). These users tend to 'do their own thing', regardless of the threat this poses to the business's infrastructure. The real issue is that there is a communication gap between IT professionals and end-users (Smith and Paredes, 2002). There is not much hope of improving the situation while each party refuses to listen to the other and both have little respect for each other's knowledge.

5. IMPLICATIONS FOR EDUCATION

Data administration is often not covered well by computer science courses. Even in business computing courses where a topic such as data integrity is included, integrity issues are usually limited to the physical structure of the database. Courses in infrastructure management normally also only consider the data from a technical perspective, for example in terms of the need for data protection (e.g. backup and recovery). It is important that the topic of data administration is taught to both technical and business students. Courses that address systems implementation need to include data administration issues in training for new users.

Only by tackling the problem from both 'sides' will the lack of communications be addressed.

6. CONCLUSIONS

There are many factors that contribute to poor data quality and redundant data resource is one of them. This can have adverse impact on the profitability of the business and the quality of its decision-making. The solution is more attention to data administration (Little, 2001). The quality of organisational data, however, cannot be the responsibility of one person or department. Every person in the company must be accountability for their role with data and information, regardless of whether they are IT professionals or business users. Managers can help by fostering a culture of co-operation and communication between the IT department and end-users. In education we can help by making sure we do not only teach technical skills in our software development courses. Computing students must understand the place of IT in the wider business and must recognise the importance of working closely with end-users in relation to their requirements, because by failing to do so, the problems related to data quality and security will continue.

REFERENCES

Bracket, M.H. (2000). 'Data Resource Quality: Turning bad habits into good practices', Addison-Wesley, London.

Little, J. (2001). 'Assessing a Data Administrative Program', in Purba, S. (ed), 'High-Performance Web Databases', Auerbach Publications, New York.

Perry, W.E. (2001). 'Ensuring the Integrity of the Database', in Purba, S. (ed), 'High-Performance Web Databases', Auerbach Publications, New York.

Smith, P. and Paredes, D., (2002). 'Managing your idiot users', MIS magazine, New Zealand, April 2002, pp 29-33.

TDWI (2002). 'Data Quality Report'. Accessed May 10, 2002, http://www.dw-institute.com/