Information Management in the Higher Education -- The Role and Importance of the Different Technologies

András Keszthelyi
Keleti Károly Faculty of Economics, Budapest Tech
Népszínház u. 8, H-1081 Budapest, Hungary
keszthelyi.andras@kgk.bmf.hu

Abstract: Knowledge is power. Information is value. I could be considered to talk in clichés, but on closer examination we can get interesting results.

1 Definitions

First of all let me allowed to make some definitions. Data is nothing else than any kind of facts (a more precise definition could be made but this pure one is good enough for our purposes). Information is the meaning of the pure data, the meaning assigned to the data by thinking persons. For example: the string "sK_17.W#SJ" is pure data for all of us now. At the moment when somebody can suppose that this is the main password to the treasury of a given bank, it becomes (a very important:) information.

Large amounts of pure data can be handled in some structured ways by the help of computers. Try to imagine tables of data of different sizes. Each table has at least one named column and zero or (usually) more rows. Column names are attributes of the different items represented by the rows. Each table can (or better to say: must) have so-called a key column with unique values in its each row. These key values serve as pointers to different rows (items) of other tables. This method is the basis of relational databases.

2 An Example

The most common example for this is an old, well-known paper based library catalogue. Data elements written on a card represent the columns (e.g. author,
title), a catalogue card itself represents the row (e.g. Shakespeare, Hamlet) of a database table. Data element on such a card called shelfmark has a unique value on each card. This value is a pointer to a given location of a given book-shelf in the storehouse.

A paper based library catalogue can be considered a (paper based:) database as well as its computerised version. In both cases we can search for a specific book by author, by title or by keywords. Above this similarity there are many differences between them. The paper based solution needs as many set of cards as the number of the search types (in this example: 3), and extended searches can only be performed very slowly, by reading up all the cards one after one: e.g. if one searches for a book which was printed in 1963 and has the word "monkey" in its title.

The computer based solution needs only one set of (virtual) cards, extended searches can be made as well as simple ones. Extended searches could be a bit slower than elementary ones, but doesn't need so much extra time than in the case of a paper based system.

3 Ways of Getting Knowledge

There are two different kinds of getting some knowledge by the help of a database. In the first case we use the database just like an encyclopaedia, finding 'Hamlet' we can see that the author is 'Shakespeare'. Another example: let us imagine a database to manage the students' data of their school achievement. We can search for 'Test, Eve (Miss)', and we can see, e.g., that she has got the best mark for her latest paper or the date of her birth. We can name this method as browsing. The difference between the time demands of browsing in the case of a paper based and a computer based solution is not too significant.

The other way, named selective data handling, is a bit different. Let us use the above example of the school achievements. We know that about half of the students cannot finish their studies. We would like to know more about this problem. How can we answer such questions as e.g. "Are there any relationships between this unsuccessfulness and the data of school achievements or personal data -- in the last five years?" Practically these kinds of questions cannot be answered without using a computerized database. These are the question types which shows not only the advantage of computerised databases but that of the selective data handling as well, opposite to the paper based data registering and simple data browsing.
4 Databases as Resources

To plan and build up a database needs not little cost in living labour and in money. If one has a well-planned database loaded with error free data he/she can answer such extended questions as the above mentioned by which he/she can save money, time or other resources, can better plan the future activities, can better adjust the activities of the enterprise to the different and quickly changing conditions of (business) life.

Databases must be considered as resources according to the above. As well as other resources databases also must be managed in their complete lifetime from the planning to deciding the rules of the everyday use.

Comparing a computer based information system with a paper based one we expect and should get the following: computer based systems should be much quicker and should need significantly less living labour than paper-based ones.

5 How to Get a Good Information System

To reach this result one has many important rules to comply with. The first step is data modelling. Better data modelling can (only can!) result in more useful database. Of course, even the best data model can result in a useless implementation if the programming is not cautious enough including the selection of the applied programming techniques, methods and software elements. After the most cautious programming based on the best data model users can put the database even out of order during the everyday use when if they enter improper or incomplete data.

One important circumstance is the selection of the applied software tools and technologies which, among the others, determines the goodness and effectiveness of the resulting system.

6 The Field of Interest

I try to measure the influence of this selection in a special field. The most important administrative task in the higher education is the management of all the educational data of the students. In Hungary nowadays one of two systems can be chosen, most of the institutes uses one of them and the rest uses the other.

In our polytechnic which has about 10.000 students. I was a faculty administrator of the first mentioned system for a whole year. I have used it as a teacher for
three years. Among these circumstances I could (and can) think much about the low effectiveness of our system. I cannot believe that the effectiveness of such a system might not be much better than that of the system we use. The system is not only slow but gives interesting results sometimes as well. The result of a query was the most remarkable, when we tried to determine the number of lessons per week our colleagues had. (One of us was told by the system to have 96 lessons per week which is simply unimaginable. The reason originated from a data modeling mistake.)

I cannot have any access to the data model nor to the source code of the two mentioned systems, so I cannot study them. Of course the hardware elements also have an effect on the result, but only to increase the hardware performance is not the best way of increasing the effectiveness of the whole system. There is only one parameter which can be tested among these special circumstances: the selection of the applied software techniques and protocols.

7 System Specification

The officially used system consists of three main parts. The database itself runs on a Windows 2000 Server under Oracle. The access to the database is granted by Windows 2000 terminal servers. Users can connect to the terminal servers by Windows remote desktop. The bottle-neck is the Oracle with the Win2k server. It can handle only about 200 database threads which means that less then about 250 users can connect simultaneously. I thought that this was very pure.

8 The Purpose of the Measurement

My purpose was to try to test whether I can reach (much) better performance by choosing the platform-independent http protocol. The main measured parameters are the response times and the number of simultaneously handled connections. These two parameters determine the effectiveness felt by the users.

I expected the result of the measurement should show significantly more effectiveness in the test environment opposite to the official one. I wanted to determine by this first and the possible simplest measurement that my theory is well established and a more sophisticated measurement could be planned on its basis. I also wanted to prove that better performance can be reached in a free software environment.
9 The Test Environment

I could not use as good hardware as the official system. I used a pure PC with a slow ATA hard disk (24 MByte/sec buffered disk reads), a P-IV processor of nearly 3 GHz and 512 MB RAM. The software environment consists of only free software. The operating system is Linux, kernel 2.4.24, Apache web server 1.3.31, PHP 4.3.7 and MySQL 12.22 (distrib 4.0.20) with factory default settings.

In the real life the two most critical fields which cause the highest load of such a system are the following: (a) application for an examination date; (b) application for courses of different subjects. Studying these most critical fields and the possible database structures which could describe these fields I found that average activities are done among about three database tables with two relationships.

10 The Measuring Method

I generated the first table with 10,000 rows of random data. The second table had also 10,000 rows. The third table between the first two had 30,000 rows. The relationships between the first and the third tables (it would be better to say: between the two relations) are one-many and mandatory-mandatory. The relationship between the second and the third tables were the same.

I generated scripts of different number of queries. These queries were done by the help of the offline browser wget. Wget browsers started to run in the background, which caused that these queries could run nearly simultaneously.

I performed the test tasks on a single machine. Http access to the local web server was made via localhost. I logged the response times and the load of the system. The results of the queries were redirected to local files.

11 Measured Values

I ran different tests with increased number of queries in each step. The last one I ran had 900 queries. 790 of them was successful, in 110 cases I got a timeout error. Time between the first and the last answer was 85 sec. The max. load of the system was 47.93 (load average of the previous one minute).
12 The Possible Error of the Measurement

There may be (and I think that there are) errors in this measurement. The most relevant possible errors are the following: the client side was the same as the server side, so client activities could decrease the measurable effectiveness; missed fine-tunings also decreased the effectiveness (see the 110 timeouts). On the other hand there are two circumstances that could increase the effectiveness: the calculation of the average activity, the circumstance that the test database had only read activities. I think that these errors of the measurement sums up to approximately zero.

Conclusions

The first and possible simplest test shows that the number of the successfully answered queries were 3.16 times bigger than that of the official system. This is good enough to say that a significant difference can be seen between the two environments. This difference is good enough to state that the selection of the operating system and the applied software tools is also an important step if we want to increase the effectiveness of an information system. This difference is not good enough to state that the software environment based on only free software is much better in any circumstances than Microsoft based ones, but is enough to plan further and more sophisticated tests which I would like to do.

References