ARCHIVING DIGITAL DATA

doc. Ing. Milan Javůrek, CSc.

Department of Process Control, Faculty of Electrical Engineering and Informatics, University of Pardubice, Studentská 95, 532 10 Pardubice, e-mail:Milan.Javurek@upce.cz

ABSTRACT

At present, computer technology has penetrated into most branches of human activities. In many cases its implementation even is uncritical, without taking into account the impacts in the course of later development. The article tries to draw attention to potential problems and collisions which can occur later and should be considered today.

KEY WORDS

Digitizing, archiving, compatibility of systems, data formats.

INTRODUCTION

It has been stated [1] that in 2007 every person gathered, on an average, 50 GB of data, and this amount is expected to increase up to 300 GB in 2011. According to journal Think [2] which is published by IBM Company, the worldwide available amount of information increases by 15 petabytes per day, which represents an octuple of information contained in all libraries of the USA. Of course, this trend is untenable, which has to be solved by restriction of complexity of individual software systems, by more focusing on sharing and dynamism of utilisation of data, in order to prevent useless duplicity in storage of identical data. It is also obvious that these gigantic amounts of data must be stored – which is relatively easily possible with the present development of technological means.

But another emerging problem, which is almost not dealt with at all, is the durability of stored data in time. Naturally, the importance different data differs and their storage periods can differ diametrically. It is necessary to differentiate between two types of activities:

- archiving, i.e. long-term storage of selected pieces of information, individual files and/or directories
- backing-up, i.e. rather short term storage of contents of whole media, when the focus is on functioning of the whole system in the cases of any breakdown of HW and/or SW.

The text below predominantly deals with the archiving of data which have lasting value and are created first of all for future use, e.g., texts, books, pictures, photographs etc.

1 THE PRESENT ARCHIVING METHODS

The present existing technological means offer quite a varied selection of possibilities for data archiving. However, one of serious problems lies in the developmental changes of the data storage media themselves as well as changes of the instruments that enable the work with these media. If we remember the epoch of mainframe computers (each one occupied a whole big room), their usual backup medium was eight-track magnetic tapes. At present, there exist no instruments for reading them, and even if they existed, these media would be illegible after those more than twenty years. If we need, e.g., programs that were used in that epoch, we must have either a hard copy listing or, as a last resort, punch cards, which had description in their upper line. Let us move to a more recent epoch, after the arrival of PC-type computers: also here existed a number of media that are now unusable, e.g., 5.25“ floppy disks, Bernoulli’s disks, hard disks MFM, magneto-optic disks, ZIP disks, JAZ disks etc. Usually, these media **are** legible, but the instruments for their use **are not** available any more. Their legibility also very much depends upon the way of data storage: at one time there existed favoured methods of data compression (“packing”) by means of PKZIP, ARJ etc. programs, but they suffer from a serious drawback: a set of files is compressed into a single file, and if there is merely a single error, then the whole medium is lost; in the classic way of back-up, such error would cause the loss of (sometimes) only just one file of the whole set of files.

The present state of technological development, roughly speaking, offers four applicable ways of archiving from the “time” standpoint [1].

a) Archiving on DVD

This method is appropriate for continuous storing of smaller amounts of less important data. In the case of application of special archiving media the reported durability is 15 years, safety 60 %, and the expenses are only represented by the purchase cost of the medium. The amount of storable data depends on individual capacities of the media. The durability of record significantly depends upon the rate of recording – it is recommended to record data at the lowest possible speed.

b) Archiving on hard disk

In this context we mean a hard disk that is reserved and exclusively used for storing data. This method allows daily storing of data in any amount. The reported safety is 60 %, too. The expenses consist of the purchase cost of hard disk, and the amount of storable data is given by its capacity. The reported durability is up to ten years

c) Archiving on network disk via Internet

According to the present possibilities the data storage facilities on internet guarantee the durability of twenty years. However, the estimated expenses are quite high: 15,000 Czech crowns for storage of 1 TG of data, the safety being up to 90 %. The amount of data is not limited.

d) Distributed archiving on Internet

A higher level of data saving is deposition of data at several network points which are continuously checked and administrated. Here the reported durability is over fifty years, safety is 100 %, and again the amount of data is not limited. However, at present the costs of storage of 1 TB of data are ca 30,000 Czech crowns.

Apart from the above-mentioned possibilities there also exist other media, but a responsible guarantee of record durability in time is questionable – e.g., the favourite media Flash, whose reported presumed durability of data records is ca five years. Of course, due to commercial and advertising reasons various durability periods are claimed, particularly in infotainment literature. For example, in the popular journal "Květy" the author encountered a time period of three hundred years (sic!) for the data recording durability on CD media. Hard disks of SSD type begin to appear nowadays too: although they do not involve faulty mechanic components, the number of recordings is limited. However, it can be expected that this handicap will be remedied.

1.1 Formats of data backup

As already mentioned above, it is necessary to save files in their original form without any modifications. However, another problem emerges in this context, *viz.* what kind of format should be adopted. The files saved as ASCII will perhaps not cause any troubles; the ASCII table will most probably be valid still in the future for a long time. Another situation is encountered in the case of graphics, i.e. texts produced by modern editors, figures, tables etc. Here the situation is also complicated: e.g., the texts produced by means of the formerly favoured editors Chi-writer or AmiPro are illegible nowadays. The fate of Microsoft Word, which is so favoured at present, can be the same. In addition, we have to take into account the development versions, where the backward compatibility is not fully guaranteed. Another problem (also quite serious) lies in general incompatibility of various editors in the area of tables and graphics.

In this respect relatively promising appears to be the PDF or PDF/A format which was designed just for long-term data archiving. This format has the advantage of containing complete information inclusive of fonts; therefore, it is not dependent upon the PC configuration. Of course, possibility of its additional editing are restricted, at least with the present function ability of the Adobe Acrobat program.

A similar situation is encountered in the case of pictures and photographs, which concerns general public. Digital photography has become mass-spread, but it is a great question whether our descendents will be able to look at photographs of their ascendants. The graphical format that is most common at present, namely JPG, is little appropriate due to its maximum extent of compression; the formats PNG and TIFF can rather be recommended from a long-term standpoint. Their structure has been designed as an open standard and they are not so much compressed, hence they are less prone to damage.

The area which is rather neglected in this respect at present is the databases and their files. In this case we deal with pieces of information that are really precious and even irretrievable. Of course, there is a vast amount of data of temporary nature, such as bank account statements of a particular day and hour, SMS messages, phone conversations etc., but a quite different category is estate register, personal data of citizens, judicial records, birth/death records. And it gives you willies when you see the speed at which such databases are digitized in countrywide extent. Even such minor items as registers of university studies have to be archived for 30 years in order to enable issuing of diploma copy if somebody loses his/her one. And these databases usually do not belong among unified ones; mostly they are programs "made to measure" to fit a particular structure of information, they are products of their time and conform to the extent of corresponding information prescribed by law in the particular epoch. It can be objected that these pieces of information having countrywide significance are continuously updated; however, who can guarantee that all the data have been safely transferred in unchanged form to a higher version of program? This is usually found out only after confrontation of owner's documents with the data in database: then it can turn out that tract numbers or acreage of landed estates are discordant. Not to mention the hackers' interventions in such databases, which are quite common nowadays!

In this area it is more than troublesome to try to specify some format of data archiving, particularly in the case of interrelated databases. It is quite generally known that one of the biggest

problems in developing such systems lies in the data transfer – i.e. in the possibility to make use the whole existing sum of data in newly compiled program systems.

1.2 Program platforms

Apart from the above-mentioned problems concerning durability of records and methods of archiving, still another problem cannot be neglected *viz.* the hardware and system platform. All of us know what a substantial difference there is between 32-bit and 64-bit computers and/or the corresponding operation systems, even if we only consider the most common PC system in this country and the systems of Windows. Not to mention the Apple computers and the operation systems Linux, OS/2 etc.

It can be anticipated that the development will not stop in this field either; and will it be possible, e.g., to start Word 2007 in a PC after 20 years? Maybe, in this area the backward compatibility will be dealt with, but in the case of the databases mentioned in previous paragraph such possibility will not exist: it will be impossible to process particular historical facts (e.g., the wage records with ever changing law!). It will be necessary to archive also the corresponding program, but who will guarantee that it will operate correctly? The author even encountered a case in which the program stopped to operate after mere changing from Service Pack 2 to Version 3 in Windows XP.

Of course, also in this area various solutions are offered; e.g., the Windows program offers to enable starting of a given program in a regime compatible with the lower version (properties of starting icon on the monitor), but such an attempt usually remains in the domain of wishful thinking, as perhaps everybody who wanted to use this function found out.

Moreover, there exist emulation media such as, e.g., DOSBOX for the DOS system, Virtual PC (Microsoft) or VirtualBox (Sun) which supply environment for installation of any system of lower version as an application within the framework of the existing system. If we ignore the fact that it is almost impossible to install Virtual PC on some types of computers due to inability to boot from the installation media (no such problem is present in VirtualBox), we must not forget that lower versions lack the drivers for newer peripheral devices, and the communication with environment is considerably restricted. As a rule, functional communication exists with hard disks, optical devices and network. If new equipment will be introduced in future, e.g., the classic hard disks will be replaced by SSD disks, the situation will be still worse. However, not even the emulation of simple medium such as MS DOS is full-valued, and it fails in more demanding applications such as the development medium of Pascal; not to mention the server application of the type of Informix, Oracle etc.

CONCLUSION

It is perhaps impossible to find univocal solution to the above-described problems. It is impossible and also undesirable to stop the development of knowledge, technologies, means, materials etc. No other way exists either, with regard to the present amount of information, its flow and rate of processing.

The biggest problem of the present and—most probably—also future epoch lies in overestimating of economical aspect of the problem. Only in order to save money, we neglect, e.g., the form of hard copy documents and only archive our data in digitized form, although we know that books, journals, paintings... will last hundreds of years if properly handled, of course. However, in this respect, too, we encounter a modern phenomenon – whoever left his/her hard copies printed by means of a laser printer in a plastic cover for a longer time period surely knows what the author is speaking about.

First of all, the most serious item is databases of permanent character: in this area the "paperwork" processing and archiving should in no case be abandoned. This is one of the areas where insufficient competence of representatives of government, state and local authorities create conditions that can lead to collapse in public information.

In this context the author sometimes tells his students:

"If old chronicles of people have written on a computer, then most probably we would not have the slightest idea about it."

References

[1] LITTSCHWANGER, T.: Data For Ever. Prague: *Chip 01/2009*, pp. 46-49, ISSN 1210-0684.

[2] Dynamic Infrastructures for 21st Century. Prague: *Think 2/2009*, ISSN 1803-4527.