11

Practical considerations for long term data conservation and analysis

Daniel Arroyo-Bishop

M. T. Lantada Zarzosa

(GDR 880, 'Terrains et Théories en Archéologie' du CNRS, Université de Paris 1, Institut d'Art et Archéologie, 3, rue Michelet, 75006 — Paris, France)

11.1 Introduction

The research being undertaken for the ArchéoDATA System has always had to take into account the fundamental problem of the free flow of data between computers and programs and bringing it all together for analysis and conservation. This has been one of our principal concerns, and one for which we are dedicating a substantial effort into understanding. For the present we have been concentrating especially on micro-computers and their programs as these are by far the most widely used in archaeology today and will present the most problems in the future. We would like then to present some ideas on this subject.

Long term use and conservation of data generated by micro-computers usage has never really been considered, by the casual user, as being a particular problem. He believes that by just buying another box of diskettes or a larger hard disk, and making a few backup copies, everything will be taken care of. It is only later when he will want to go back to some data, or transfer it to a new machine or program, that the surprises will come. This is of course at a minimalist level, because the data that the archaeologist conserves on his machine is not only his, but it is also part of our heritage, as much as any other archaeological document or find, and it should be treated and preserved as such.

11.2 Problems

If we take into account some of the physical dangers that await our data, apart from pure accident or incompetence,[1] we can see that we are far from secure.

Data held on floppy disks should be transferred as soon as possible to more durable media as their storage life, due to progressive demagnetization, does not seem to be perfectly assured beyond three years. The inherent fragility of this media also makes it particularly vulnerable to physical and environmental damage arriving from high temperatures, creasing, liquids, *etc.*

Even though the mean time between breakdowns are in the 30,000 to 60,000 (some even go as high as 150,000) hour bracket, hard disks seem to suffer substantially from progressive wear not reflected in the aforementioned figures. The greatest problem seems to be that the sectors containing data become unreadable after that data was recorded and it becomes partly irrecoverable. This is because the disk mechanism ages and the magnetic heads are no longer aligned in the same way as when the data was originally recorded. The sector tracks are consequently imperfectly read with the corresponding loss of data. The chances for conserving data integrity can be improved by re-recording all the data on the disk to partly compensate for the aging process and physically re-formatting the hard disk when this process has gone too far.

Progressive demagnetization is also a problem in hard disks, but this is much slower than with diskettes and should not be a primary problem. As the hard disks contents will be around longer than the computer,[2] this presents us with the problem of forced transfer, sooner or later.

The optical disk situation is very confusing at the moment as there are many techniques present and their development is rapid. If archaeology were a business where records are kept for only a few years it would be possible to invest in any system that would work for say five to ten years and the replace it with the latest equipment at that time. This of course is not the case in archaeology. A problem seems to have arisen with aging, which was supposed to be practically nil, of the disks aluminum recording surface. According to accelerated aging tests, we cannot, at the present time, be sure that they will last more than ten years without loss of data.[3] This could possibly be worse with the new optical-magnetic reusable disks. Even though the physical dimensions of the recording media is standardized, the formats and drivers are not. This makes it particularly difficult to read disks from different vendors, even though the prime manufacturers, Ricoh and Sony, make the drive mechanisms for practically everyone.

A new generation of optical disks should offer the possibility to read CD-ROMs and read/write optical-magnetic reusable disks on the same machine. Two other CD types that could be very interesting for archaeological material in that it can simultaneously store data and good quality graphics and photographs, are CD-I and CVI optical disks. The technology is there, the hard and software seems to be there, but the product is not really there yet. The only large scale use in France are the parts catalogs of the car manufacturer Renault and there seems to be no will to go further at the present time.

In the case that the optical disk route is taken it would be advisable to base the system on some large public project which has similar views on data conservation. We in France

[1] These can be anything from an electrical fire or water damage from a burst pipe to the inadvertent erasure of storage media by inexperienced excavation workers.

[2] We estimate, except for very undemanding chores, that the viable (performance/replacement cost ratio) life of a micro-computer to be less than five years.

[3] This widely cited number has been disputed by some manufacturers, but their own documentation on sec disks they estimate at thirty years their stable life-span, which is anyway substantially less than eternal.

are looking into what is going to be used for the new French National Library to be built in Paris and for which extensive use of this technology is foreseen.

We should end up somewhere in the near future with multiple use machines which can read, and maybe write, several types of discs. For the time being, for archaeology, we are not sure if optical disks are full of promise, or promises.

11.3 Basic data transfers

Data transfer and data conversion are fundamental to data conservation. Basic data transfers on the ASCII level do not pose any particular problem as there are many communications programs, when it is not the programs being used themselves, that can handle these. Things are considerably more difficult when these also involve the transfer of data with their field types and formatting, as would be the case of a full Paradox to dBase transfer. The inverse would be automatic with Paradox as it is able to read dBase files directly, although this is not the case for most. Several specialized programs, MacLink Plus (Mac) from DataViz, Lap-Link (Mac and PC) from Traveling Software, are available, among others, to facilitate these particular cases.

In the case of very complex or rare file types it will probably be necessary to go to one of the specialized companies for help. They exist in practically every country and they can also do many types of media conversions. However it is best not to find oneself in need of these services, particularly through thorough planning, as they are very expensive. This cost factor should be considered one of the main reasons why it will may not be possible to recover independent data at a later stage.

11.4 Graphic data formats

The extensive use of graphics in the ArchéoDATA System's development has made convertibility a fundamental problem. We have found more than thirty odd graphic file types used in the Macintosh alone. The problem is compounded if we take into account the basic incompatibility between the Macintosh and the IBM type PCs, our two main working environments.

Some people work around the transfer problem by using the same program on both the Macintosh and on the PC. This, albeit expensive solution, works quite well for most things, but some surprises are to be had, especially concerning graphics. Practically everything in the two versions of PageMaker transfer well except that it only sends bit-mapped images from the Macintosh, and not MacDraw type vectorized drawings. Databases such as Blith Software's Omnis 5 could be interesting for those institutions who want to run a homogeneous program in an heterogeneous environment.

One of our first observations for the conservation of computer graphics is that if special propriety drawing formats are used it would be best that copies be made in one of the more common or universal file types, which are sure to be around in the future. Even though this might entail the loss of some quality in that the new format might not be able to store all the information contained in the former's data files. It is always better to loose a little quality now, than everything later on. Several programs that can handle graphics conversions quite well are: Arts and Letters Graphics Editor (PC) from Computer Support, The Curator (Mac) and Glue (Mac) from Solutions International and Hijaak (PC) from Inset Systems.

An example of how we are trying to work around some of these drawing format problems is to be seen in one of the schemes being worked out for excavation drawings. The DXF file is a professionally recognized data format for Computer Assisted Drawing or CAD and is used by the micro-computer industry standard drawing package AutoCAD, as well as for other micro, mini and mainframe packages. Practically all other programs use or can store in the DXF format. If all archaeologists used professional CAD packages there would be little or no compatibility or conservation problems. Even though a package such as AutoCAD is now available on the Macintosh, it is a high end program meant for a dedicated user, certainly not the typical archaeologist who has neither the organization to have a dedicated specialist nor the time to become one himself. He is probably doing good Mac Draw type drawings on his Macintosh and the problem becomes how to get these two worlds, to their mutual gain, to communicate and exchange their data. Two main benefits are to be noted, the first is that eventually the archaeologist will want to take advantage of the more sophisticated possibilities of higher level CAD packages, but to do so he will not want to have to redo the work already done. Secondly, and most important for future, the DXF. standard will offer greater assurance for long term conservation of the document.

To illustrate this we offer the following practical examples of vectorized drawing packages in four levels of use:

Level one: Canvas 2.1 from Denoba is a base level vectorized Macintosh drawing program with its own propriety format. A very good program, it offers several other storage formats including Mac Paint (bit-map), EPS (Encapsulated Postscript), and the PICT Macintosh standard for vectorized files.

Level two: Claris CAD 2.0 from Claris is a medium level CAD program which can import and export the main Macintosh formats. This quite powerful program is very user friendly and will appeal to the non specialist archaeologist. It not only imports and exports PICT files, it also does so with the DXF format. This program can then be used directly as the main CAD (2D) package for excavation work or be used as an interface for data transfers.

Level three: AutoCAD 10, a not user friendly, but a powerful specialist oriented CAD package, which demands substantial investment of time and money. The DXF files it creates are the industry norm and fully compatible with other programs. The Macintosh version is but the last of a very long list of platforms on which it runs.

Level four: SPANS and ARC/INFO, powerful GIS packages, can import and convert to their own propriety formats, among several others, the DFX. files. An archaeologist can in this way work with easy to use

Macintoshes and be able to recover a good deal of his drawings and data on much more powerful programs for which there are no Macintosh equivalents at present (albeit a very slow A/UX 1.1 version of GRASS).

This system of data exchange has the added advantage in that the archaeologist can have his drawings enhanced by the CAD specialists of a regional authority and still be able in many cases to use them on his micro through a 'downward' transfer of the data file.

Cost must also be taken into consideration as this will be, in many cases, the determining factor for archaeologists. If we take the different levels presented in our example and give an equivalent cost factor of one-hundred to level one, level two will be four-hundred, level three, sixteen-hundred and level four, five to ten thousand. This of course does not take into account the further expense in hardware necessary to run the more demanding software.

11.5 Conclusion: centralized conservation

Centralized conservation is the only possible long term solution for preserving our archaeological computer data as any other solution would lead either to its probable loss or prove prohibitively costly and complicated to implement and maintain. We cannot expect that any organization could keep abreast of the work necessary to conserving thousands originals, with all their different types:

- Dozens of different operating systems
- Hundreds of different file types and formats
- Thousands of different programs

Even though the problem is being studied, the truth is that the situation for archaeology is quite bleak as there is no central authority ready to handle the problem and we are condemned, before things get better, to loosing much data in the coming years. This last, and crucial stage, must be in the hands of specialists whose jobs it will be the conservation of this part of our cultural heritage.

Of course simple centralized conservation of ASCII data could be very much improved upon if the data were structured in a similar way, so that it might be related with that of other sites. This would make it possible to engage in extensive correlation and analysis of archaeological data and not have these possibilities reserved to the few who can invest a great deal of time and energy to re-format data. The standardization of data formats, especially those used for graphics, will be a very important factor in the recovery and conservation of the data. This would also makes it possible to use sophisticated programs, such as Geographical Information Systems, where it is necessary to have data unity.

This of course brings us back to one the main reasons for the ArchéoDATA project.