38
The Archaeological Data Archive Project

Harrison Eiteljorg II
Center for the Study of Architecture, Bryn Mawr, USA

38.1 Introduction

The use of computers to record archaeological field data and to assist with individual scholarship has grown exponentially in the last decade. However, careful attention to problems with the storage and preservation of data files has not accompanied this growth. Too many files lie unattended and ignored on university mainframes, on hard disks on desktops, or on floppy disks in drawers. After they have served their original purposes, the files are, for all practical purposes, forgotten.

No less than the notebooks, plans, and catalogues, though, computer files from excavations are important records. Their preservation is crucial to scholarship, and access to them is no less crucial. In the case of data sets gathered by individual scholars, the importance of the files to other scholars varies widely. Nonetheless, the labour which was spent to create the data sets should not be wasted through neglect.

Unfortunately, neither individual scholars nor universities that sponsored their work are well prepared and equipped to deal with the problems of data storage. Therefore, an archive to house and care for data sets of value to archaeological research should be created. That is the goal of the Archaeological Data Archive Project

38.2 The Archaeological Data Archive Project

The idea of the Archaeological Data Archive Project (ADAP) grew out of a discussion at a meeting of the computer committee of the Archaeological Institute of America. Although the participants in that first discussion were uniformly excited about the potential of network access to computer data, on reflection we realised that there were immediate problems with the storage and preservation of such data. In particular, we were concerned about files from excavations and other data files that form the core of our knowledge base.

A long and careful process of examining goals and exploring possibilities followed. During that time it was decided that the archival concerns deserved first priority and that the Archaeological Data Archive Project should be independent of academic or professional groups. The ADAP is directed by Harrison Eiteljorg, II, and operates as a unit of the Center for the Study of Architecture, which was already building an archive of CAD models.

Initially, the appeal of an archive had more to do with access to computer data than with storage and safety. The idea of providing Internet access to huge quantities of data is so appealing to us all that it is seductive. However, it has become clear that the more crucial issue for the moment is the preservation of data files that are at risk and the building of an archive to preserve the files that are being created even as we speak.

Although scholars understand that paper records are subject to many kinds of damage and decay, we have often assumed that computer files are far more stable than they actually are. In reality, data storage media are subject to decay. Unfortunately, the damage is generally recognised only when access to the data shows problems; then it is usually too late to rectify the situation. Less obvious, but equally devastating, the data in computer files have often been compiled to assist with analysis for the person who created the files and no one else. As a natural result, the utility of the files is severely limited if they are used by anyone other than their creator. Taken together, these problems with existing computer data make urgent the task of proper storage and preservation. We must begin now so that we do not lose more information, and, in fact, the ADAP has actually begun to accept data files.

Early uses of computers for archaeology necessarily involved data cards, tapes, and mainframe disks. As time passed, of course, the development of the microcomputer led to the use of floppy disks and desktop hard disks. How much data resides on cards that are now be unreadable, tapes that may now be without appropriate tape drives, or floppies made for machines long since antiquated we do not know. Nor do we know how much data may have been left on a hard disk, unused and unrefreshed, for years. That is one of the reasons the task is urgent. An appropriate archive must be prepared to deal with all those storage forms – not to mention file formats – simply to preserve the information. To be realistic, however, we know we cannot deal with all media; so we must make some difficult choices.

The first choice has to do with the physical media on which data lie. The ADAP can deal with media from PCs, Macs, and Sun workstations without difficulty. Of course, files can also be sent over the Internet. What about the other media? Here there is no simple answer. We have, for instance, been assured of the co-operation of a colleague should we have KayPro disks. He can access the information and transfer it to DOS-formatted disks; other CP/M disks may also be accessible with this system. One scholar has asked about dealing with disks from Apple IIs, and we have found what must be done to accept them. Those are relatively easy problems.

Dealing with data cards, tapes, and other such media is another matter. We cannot predict what problems we will face until someone brings us a specific request. The ADAP will certainly not own card readers or tape drives which have become obsolete, but we can expect to have help from those who have the data and from others who are experiencing similar problems with old media. After all, we are not alone in this. We have, for instance, heard from one IBM employee who spent years dealing with old files at IBM, and the US government has been obliged to give up on some important data files because they were on tapes no longer supported. It is doubtless too late for some files, but we can prevent the unnecessary loss of more by starting this process now.

Similarly, we cannot predict the file format problems we will face, but we can be sure that there will be some we have not expected. Fortunately, ASCII provides a fall-back choice that, generally speaking, will preserve the data adequately. (Sooner or later, Unicode or the 32-bit ISO standard will probably supersede ASCII.) But many files should be preserved in far more complex formats so that they can be used to their fullest. It would be a pity to settle for ASCII files as the lowest common denominator if we start with files in sophisticated database formats. On the other hand, ASCII files permit anyone to use the information; access to a specific database management system is not required.

But how many different formats can we store? How many should we keep? Here there is a difference between the archival duty and the question of effective access. It may be argued that the archival purpose is satisfied with ASCII files, but those files can be far less effectively used than the data files which preserve the complexity of a good database management system. Indeed, some of the data complexity would surely be lost if the ASCII files were not accompanied by thorough descriptions of the data files, the relationships between and among fields, the authority lists used, and so on.

I have not answered the question posed about what formats should be accepted, because, at least for the moment, we will again emphasise our archival responsibility and accept any format. That may seem ridiculous; some file formats will be all but useless. But we and our colleagues are better served if we have the files well preserved and properly maintained in an obsolete format; such files can be translated if they are important. Once left to decay, however, they are unrecoverable.

Whenever possible, we will ask that files be supplied in their native format and ASCII. If the database management system will do so automatically, we will also ask that the files be supplied in .dbf format, since that seems to be a widely-used effective standard.

More important than the file format, however, will be the documentation that accompanies the files. That documentation should make it possible for a user who has access to nothing more than the published material about a site to utilise the files. The descriptions of files, fields, relationships, and so on must be complete and accurate; they must make absolutely clear the ways the data can be fitted together. Without such documentation, the data files can never be fully understood.

But we cannot refuse to accept files if there is no documentation. To do so would be to deny our archival function. Therefore, the ADAP will, indeed, accept files that have no accompanying documentation if and only if the scholars who created the files are no longer able to supply such documentation. Of course, we hope that will not be necessary, because undocumented files are of so little value to others until someone has spent the time required to document them.

One of the functions of the ADAP will be to make certain that, once a part of the archive, files will be maintained in the most current and useful formats possible. No matter the format received, the files will be maintained and migrated into new formats as old ones fall out of use. This, of course, will be one of the most valuable services performed by the archive. Individual scholars will not be required to transfer their own data from old to new formats, though they will obviously do so with data files still in use. Over the very long term, this will be a crucial service, as files pass from generation to generation, with each new generation providing more effective data access.

That brings us back to the original interest of the participants in the first discussion of the ADAP – access to the archive. The ADAP is not and will not be a closed archive, accessible only to a select few. It will be a networked archive, open to all who have access to the Internet and its successors, as well as anyone with a computer. The files stored in the archive will be available either over the networks or on disk. No files will be accepted unless they may be made public, though we will not require that they be made public at the moment they are contributed to the archive. (Access to some data items may be restricted, however, to keep potential looters from learning the exact location of a site, for example.)

Files to be archived include, first and foremost, any file from an excavation. Other files, however, will also be archived – catalogue files, CAD models, GIS files, images, authority lists, and so on – in sum, any data file; we do not expect to archive text files other than those text files that are required to describe data files. The archive will accept files from any cultural, geographic, or chronological area.

Many individual scholars and institutions will wish to discharge their archival responsibilities directly and will not, therefore, want their records archived elsewhere. The ADAP will co-operate with them, assisting in those ways possible, and will use the facilities of the network to include their files in the system without duplicating them on site, when possible, or to provide information about where and how to find them.

An archive such as this raises numerous thorny but important problems about standards, particularly because easy access to information – not simply data files – is severely limited by an absence of standards, In fact, it is tempting to begin working on the standards questions first, waiting to work on the archive when the standards have been established. We have decided, however, that the urgency of the archival needs – coupled with the likelihood that creating standards will be both a protracted process and, at best, influence future computer files only – requires us to move forward with the archive. Here again, the need for preservation supersedes the need for information.

We will work on issues involving standards, and have begun to do so already. But we will not delay the important archival work, nor will the ADAP attempt to apply standards which are not fully accepted by the archaeological community.

In conclusion, I should say a few words about the progress of the ADAP since its announcement last November. I have been delighted by the response to the announcement, indeed, nearly overwhelmed. We have had offers of data, and, in fact, some files are on their way to the ADAP now. We have had numerous offers of support and questions about process – from people in the computer world as well as scholars. We have also learned about relevant projects and are actively co-ordinating our efforts with those of other scholars. But as we stand at the very beginning of the computer era, so we stand at the start of the process of archiving archaeological computer information. Indeed, the direction is so uncharted that I hope all of you will help to guide the process with your interest, your suggestions, and your expertise. We have a long way to go, but the journey should be exciting, rewarding, and full of surprises.