Archiving Archaeological Data

Harrison Eiteljorg, II

Abstract

Archiving archaeological data is a crucial task for the next decade and beyond. As we are all obliged to rely more upon studying materials already found and less on the excavation of new sites, we will find ourselves dependent as well on the excavation records that should accompany the artifacts. Those records must be available to us in useful forms if we are to continue working with the artifacts. The Archaeological Data Archive Project is concerned with preserving those records that are in computer form, and making them available as required. There is an erroneous assumption that computer records from field projects are properly looked after, but this is not necessarily so; archival storage of machine-readable documents is more complex than normally appreciated. The requirements are not onerous, but they are very different from the requirements for archiving paper-based information. The misunderstanding of the depth of the problem and the casual treatment of machine-readable data should concern us all. Personnel from the Archaeological Data Archive Project advise those preparing data for archival storage, archive data when required or help others to do so, provide information about data in the archive, and make those data available to scholars in useful forms. Problems with this seemingly simple process have included misunderstanding of the perils involved, reluctance to share information, and institutional pressures against public access. These and other problems will not disappear, but progress is being made.

The Archaeological Data Archive Project was conceived at a meeting of the computer committee of the Archaeological Institute of America in December of 1992. Over the following months of gestation the idea was discussed and refined until the final form of the project emerged - an attempt to provide a safe, secure home for archaeological data in machine readable form. Directing the project fell to the Center for the Study of Architecture and to me personally; the Archaeological Institute of America did not consider this to be an appropriate project for the Institute despite its involvement in the development of the archive.

We on the committee assumed the value of archiving archaeological data without argument or discussion, but we have learned that the importance of archival storage must be made explicit. It is, we believe, a clear professional responsibility, not simply from some pious sense of preserving all information, but because, as archaeologists are so fond of saying, we destroy much of our evidence as we work. Despite this, few of us would claim to have understood so well what we dug as to preclude the possibility of someone else learning more - if and only if that person has access to the excavation information that should be available in our data files. So we need to preserve the information about excavations as well as the artifacts so that others may learn from both.

I would also argue that archaeologists of the future will find fewer and fewer chances to excavate and, as a consequence, need to mine our data for new clues to the questions of the day. As we dig less in the ground and more in museum storerooms, we will need the data from old excavations as well as the artifacts, the excavation records that should accompany the artifacts and make them meaningful. Of course, those of us who are working in the field will be equally beholden to the records of past excavations for our raw data, since those records are crucial for dating and *comparanda*. Thus, the records from old excavations must be available to us in useful forms if we are to continue working with the artifacts we find in the field or the ones we find in museum cases and storage rooms.

"Useful forms" has several meanings. At its simplest, it means that the data files should be in the forms used by the creators and susceptible to the same kinds of analyses in the distant future as when they were created. Thus, database files should be able to generate the same tables, respond identically to the same queries, and create the same statistics for a user in the future as for their creator yesterday.

That seems to be a simple requirement, but we must recognize that it is more difficult than it may seem, because computer data files are linked inextricably to the programs designed to use them and the operating systems of the computers on which they live. A database file or a CAD file requires a program that can use the file if it is to be useful and a platform on which that program can run. What is obvious to some, as a corollary, is that data files we use today must be different if they are to be equally useful tomorrow. That is, the data in the files must be put into new and different formats if those data are to accessed and manipulated in the future as they can be today.

An archive for computer data, then, requires that data files be changed to retain their utility, not kept the same. Of course, we are talking about careful, controlled change here. We must change the format of the computer files so that the data remain as nearly unchanged as possible. This is, of course, commonly called data migration.

Data migration is not technically demanding in most cases. It does require familiarity with both the donor and the recipient data formats; it also requires familiarity with the data to be migrated. In some cases, when the data are very complex and include important multiple relationships,

familiarity with the data must be intimate. The best example I've encountered of the complexity we're likely to find is the conversion of a set of data tables from the database management system called Advanced Revelation to Microsoft's database management system *du jour*, Access. As many scholars doubtless know, Advanced Revelation permits the use of repeating fields - fields that allow multiple entries - and a single record may have several such repeating fields, each related to the other so that the first entry in one field goes with the first entry in another. It is possible to build an extremely rich and complex data set with such a system, but, of course, the files are far from standard. This is about as far from a normalized database as one can get, but it does, for instance, provide an easy way to record all figures on a pot without regard for how many there may be and without creating separate data tables.

Translating the Advanced Revelation data into an Access system has been extremely time-consuming. Whereas one might hope that data could simply be dumped from one system to another, that could not be done here. Not only were there intermediate steps, there was an entirely new file structure required. Many individual data tables were required where there might have been only one before. The job is still in progress, and, while the outcome is in sight, the cost will have been substantial.

Advanced Revelation is not alone in providing such potential problems for data transfer. Repeating fields are also permitted in File Maker Pro, and a similar feature is available in Fourth Dimension; so seemingly mainstream choices may leave an excavator with data requiring substantial work if they are to be useful to anyone beyond the excavation team.

One more example, this one from the world of CAD, computer-aided design. AutoCAD is so dominant in the PC world that its native file structure is often used as the standard. Autodesk, the parent company, even sells software to read and write the AutoCAD format - DWG files - so that other software developers can add that feature to their programs. What Autodesk does not tell consumers is rather startling, though. The specifications for DWG files are not simply a specific and restricted set of drawings entities (lines, circles, surfaces, and so on). Instead, the specifications provide for a set of supported entities and any new entity the programmer chooses to add, provided only that the new entity be identified according to certain rules. So long as the new entities are added correctly, the file is considered a proper DWG file. Note, however, that a new entity though included in a DWG file need not be known to Autodesk or to any standards body. Indeed, only the programmer needs to know of the existence of such an entity. As a result, no program can certainly read a file fully just because it is a DWG file. Indeed, even AutoCAD itself will not be able to read a proper DWG file if that file contains new entities unknown to the programmers or of no concern to them. (I should be clear about this. AutoCAD - or another CAD program that accepts DWG files - will read the file, but the unknown entities may be ignored, in which case the user will have no way to know of their presence. In other cases, the presence of certain entities may be known, but the system will be unable to use them, and the user will not be able to edit or query them.)

The entities in question are not arcane ones. It is true that many are design entities that permit parametric modeling, but some are simply better, more complex surface representations such as those we might use to define a column or a complex molding.

Imagine an archive with one of these non-standard DWG files. (An interesting bit of language there - a non-standard standard! - sometimes the electronic world seems even stranger than that of Hal). How does the archive treat such a file so that the user in the future can see and understand all entities in the file? Obviously, the archive must have information and personnel to a) understand the files received and b) understand how to translate the items into some more universal and useful form. Such facilities are neither easily obtained nor inexpensive.

I have sketched out a rather bleak picture; databases that must be completely reorganized to be moved from one system to another; CAD files that require substantial expertise to be useful even today, much less years in the future. Any archive, whether it be the Archaeology Data Service at York University or the Archaeological Data Archive Project I represent, must be concerned with these problems.

Up to this time, we at the Archaeological Data Archive have actually experienced few migration problems. AutoCAD files have not been migrated to Release 13 format, but release 12 files can be used by either Release 12 or Release 13; so there is no urgency. (As I write this, Release 14 is arriving; I do not know what that will bring.) We have moved some files to release 13 for internal use, and the process was automatic. An interesting and unexpected problem has been encountered, though. If we were to move the CAD files in the archive to Release 13, users of Release 12 would be unable to use the files; they are not backward-compatible. Since there is not a good generic format for CAD files (equivalent to DBF for data tables), this is a significant problem, and it may be necessary for the archive to maintain CAD files in multiple iterations. This is a good reason to seek generic, non-proprietary formats for data.

We have also been obliged to convert files that were provided in inappropriate or less useful formats. Both spreadsheet and ASCII files have been converted to DBF format and comma-delimited ASCII. Neither process was difficult, but I would not have wanted to entrust either job to someone not conversant with archaeological material. Furthermore, only when I tried to use one of the data sets myself did I realize that the simple translation had been inadequate and some reorganizing of the data was required.

If the reasons for preservation and migration of data are clear, is it also clear why there should be external archives? After all, we have for decades kept data from old excavations in the offices or archives of sponsoring institutions. Why do we now need an external organization to take care of excavation records?

The first reason is simple. Universities and other research organizations often do an excellent job of archiving records,

but some do no job. The archival care is especially weak when materials stay in the hands of excavators for years after the completion of the project and are, by the time of his/her death or retirement, more or less out of sight and out of mind. The need for maintaining data in perpetuity also stresses the good intentions of research organizations.

A more important problem exists for electronic data that are likely to be rather rarely used. Most university computing centers are prepared to backup data but not really to archive them - at least not in a routinized, scheduled-migration fashion. If files are in use and a user takes responsibility, those files may be properly cared for, but there is unlikely to be anyone with the specific responsibility to care for data files from long-forgotten excavations. Furthermore, if general technical personnel, not people trained in archaeology, are charged with the responsibility of caring for archaeological files, how well will those files be treated when they must be migrated? Will such personnel understand the requirements of the translation process for the data under consideration? In order to understand, they will need specific information about the files (current format, relationships, data types, and so on); will that information have been collected at the outset?

If, on the other hand, computer files are to be archived by traditional archival personnel, those experts must attend to an entirely new set of problems, since the standard archival processes for objects do not apply.

A colleague recently suggested, in a meeting at which we both spoke, that data stability should not be a problem with standards such as SGML. That is certainly possible for the text world, though it would require us to use SGML for material where that is not particularly suitable - or to migrate the files at the very beginning of the process. In any case, standards like that do not exist for databases or CAD models, and even good standards in the computer world, such as SGML, have a short, brutish life.

Data archives should deal with these problems with relative ease. Monitoring of data file types and needs for migration will be one of the prime jobs of the archive. Another will be maintaining accurate and complete information about the processes required for a wide variety of file types. (Note that these archives will probably be virtual ones - not located in a single place on a single machine. As a result, the data migration services, for instance, may be developed in different places for different data types but made available to all parts of the archive.) None of this is, as the current saying has it, rocket science. In fact, much is quite simple intellectually. But it must be done in a self-conscious, routine way.

An archaeological data archive will also have the advantage that its personnel will understand the nature and meaning of the data. That will be necessary for proper migration, as has been brought home to us already.

The one problem for an archive that may be more severe than for an excavation sponsor is money. Large institutions with long lists of supporters have an easier time raising money than do small, specialized ones. Hence, the ADS in Britain is a consortium, with funding widely spread among its members, which are themselves institutions. The Archaeological Data Archive Project expects to put together a consortium in the U.S. along roughly similar lines.

I have said things here that, by and large, have been said before - by myself and others, but the difficulties here are not well understood in the academic community at large. They have often been downplayed, lest the fear of impermanence frighten away those who are already uneasy about computers. So these things must be said again and again, but with a realistic appraisal of the remedies.

One other problem should be mentioned. Many of our colleagues equate putting material on the Web or on a CD with archiving that material. When pushed, they understand the difference, but they do not consider archival storage to be a need separate and distinct from putting material into electronic form. This distinction must be made clear to all, since archiving represents a crucial step beyond putting material into digital form.

Having concluded that archives are necessary and possible, I must turn to what they are not yet - real. Archives without data are not valuable. An empty archive is not, in fact, an archive. It may be an archive-in-waiting or an archive of the future, but it is not yet an archive.

We now have, in the Archaeological Data Archive a handful of files; this service has not been recognized by our colleagues as either necessary or desirable enough to encourage many to take advantage of it. Furthermore, we archaeologists have, generally speaking, a less than stellar record when it comes to handling our information. We often fail to systematize the material for others; we don't generally expect others to use our raw data but only the predigested parts that we put into publications. Yet we, of all scholars, should be among those most scrupulous about making our raw information available to others. We all know that, our own individual genius aside, little that we say about the ancient world today will be incontrovertible tomorrow. We all understand that we are obliged to destroy much of our primary evidence to obtain more.

Yet we sometimes treat our data as if they were our private possessions, to be shown to selected others when, if, and as we please. When we do make material available, we rarely consider it part of our responsibility to provide easy, indexed access.

Our institutions are often helpful only after the fact. They will archive material, but they don't regularly help scholars prepare their own records for archival storage. Certainly they do not give credit for the time and effort required to do that until it's too late and someone else must prepare material for archival preservation after the creator has gone.

I am not concerned here about the reasons for our sometimes unscientific handling of basic information; nor do I want to seem to be straining for a pulpit from which to preach. Rather, I want simply to make it clear that archival storage of our records requires that the creators of the records take the time and the responsibility to prepare them for the archive and then actually turn them over. Institutions must also recognize the importance of this work.

Such time and effort can, of course, be mandated. Sooner or later, it surely will be, as is already happening in Britain and

the U.S. However, I would like to see us as scholars, as the liberators and guardians of the most basic information about the material culture of humankind, take pains to treat our knowledge with the respect it requires before we are forced to by governments that will, in many cases, muddle the process beyond redemption. When the responsible bureaucrats are not archaeologists, they will rarely understand our problems or our needs. They will just make rules that are tidy but may not be appropriate. We need to preempt them, to create good, sensible rules about data archiving, and we need to start actually archiving our files to show our good intentions.

Permit me, please, to suggest a few rules we should use to head off the ones governments might impose.

First, we need to stop playing the waiting game. I'll publish when I'm ready, and the data will remain mine until then. I am not suggesting that data be made public immediately upon recording, but I am suggesting that data files be archived immediately. Those files can be kept private, but, were there a time limit on the privacy, there would be a simple spur to more prompt publication as well as the proper care of the data in the meantime.

Second, we need specifications about how data will be transmitted. That is, we need to know that certain file types are acceptable and what information (metadata) must accompany the files (see Miller, this volume). If we wait too long to agree about these kinds of things, we will, I fear, be presented with fixed systems for recording our information by host governments - our own or others - and the results will be catastrophic. Can you imagine a commissioner looking over your shoulder to tell you how to structure your data file? How to make your CAD model?

Third, we need to work on some of the common issues of vocabulary and authority lists that will help us all to work more efficiently and to make our data more easily understood. The people at the Getty have given us examples of this, some of which have been well received.

Fourth, scholars need to spend some of their time - as they now spend time on other professional duties such as reviewing articles and manuscripts - helping archives evaluate data files that have been offered for storage. An archive cannot and should not accept all things offered and must be sure that material stored meets certain standards. Scholars must be willing to treat reviewing potential archival material with the same seriousness they treat reviewing articles.

Fifth, we have to decide how to handle new information that is clearly related to information already in the archive. We need to be sure that two-way links are maintained to connect new information to old - and old to new. The individual data files must retain their integrity, but they must also be receptive to inclusion into a larger whole.

Sixth, it is very difficult now to obtain information from a variety of sources and to make all the information somehow compatible so that all can be seen and analyzed together. There is a strong temptation to try to fix that problem by creating data coordination schemes that will allow disparate data to be combined. Here I believe that we must resist temptation, because this problem will be solved by the commercial marketplace and I think, as a result, we are best served by spending our time on the other pieces of our puzzle. Microsoft and Oracle and the other players in this arena will worry about data incompatibility. (Indeed, this is already happening.) They must find a solution, and we will have to use it, whether we've already developed our own or not, because what they develop will be the standard. I do not believe that we belong in the software business.

If, as I have said here, archives are essential, and if, as I have also argued, independent archaeological archives are required for archaeological data, and if, as I have just pointed out, we will soon be required to provide archival data files to host governments, then, I say again, let us not wait and risk being forced into poor systems. Instead, let us begin to archive our records now, and let us cooperate in such ways that the archives can provide to us all the full and complete access to critical information that we need for our work.

Contact details

Harrison Eiteljorg, II
Archaeological Data Archive Project
Bryn Mawr University
PO Box 60
Bryn Mawr, PA 19010
USA
email: neiteljo@brynmawr.edu