THE PROBLEMS OF PREPARING PRE-RECORDED DATA FOR COMPUTER INFORMATION RETRIEVAL AND ANALYSIS: A CASE STUDY

by

Naomi Iliff

Research Centre for Computer Archaeology
North Staffordshire Polytechnic

Many archaeological sites have now been fed to the computer for storage and statistical analysis. Some sites were originally recorded with computing in mind; some of the sites were used to illustrate some statistician's or programmer's theories and only the information relevant to his studies were used; some sites have been forced into formats of packages which were not written with thir particular problem in mind. As computing in archaeology becomes more commonplace there is more and more demand for older data to be looked at, and a new problem arises of preparing pre-recorded data and adapting it for computing analysis. The old adage of rubbish in and rubbish out is often produced by the furious programmer in reply to sarcastic archaeologists. But there is no excuse for rubbish, in or out, if problems of adapting data recorded for human consumption to data compatible with the computer are discussed.

The basic problem with preparing data to feed to computers is that humans and computers are simply incompatible. Some sort of efficient compromise must be worked out. A human uses strings of letters as a primary means of communication. The computer would like best to deal in binary numbers which is a rather difficult means of communication to a human. However high level languages for programming have done much to bridge that gap. Theoretically it is possible to write all information to the computer in long hand and program the computer to interpret it. Although the computer would be able to do this, the necessary programming would be very laborious. Humans, on the other hand, are quite efficient code-producers of a high level form. A human is also efficient in seeing mistakes and discrepencies in the data and has the initiative to cope with them.

Unfortunately it is difficult for the human to be rigorously consistent. The computer is quite capable of dealing with inconsistencies. Apart from human introduced inconsistencies, there are inconsistencies caused by differing standards of the records. Sometimes the data is the result of a specialised analysis; sometimes it might be the data of a workman who happened to notice some details. All information is important but it must be possible to differentiate between good and bad data. The human is trained to do this easily but the computer must be told the difference, especially the difference between negative evidence and "don't know". However it is crucial not to lose sight of the fact that data are of different standards.

When should coding be used? Is coding necessary? Yes, if you are going to use the individual characteristics for any numerical or correlative analysis. If you are just storing information it is not so necessary unless space is at a premium. There is little point in trying to code up information if you have more than 15 possible variations. For instance there is little point in coding up the length and width of houses in a site if they are all different. But if they were of four or five different lengths and widths would it be worth encoding them? The answer to that is, it depends on the sort of analyses you are intending to do. Sometimes within a culture settlement types fall into obvious groups with minor variations. How does one encode those data in order to maintain their family similarity but to point out their differences? It would be best to code up all the different attributes separately and then it would be possible to compare each characteristic, or use all the codes together as a string.

The more the original data is broken up into small groups the more difficult it is to get any coherent groupings after analysis. Coding up data produces groups in the data. They are implicit in discrete evidence. For instance in cemetery records the deposition of the body can only be in one of four ways, i.e. on the back, on the front, left or right sides. But depending on the sort of analyses that are to be done groups can be formed in continuous data. Perhaps the individual sizes of the objects under study are no longer important in themselves and it would create unnecessary background noise to study them by their actual measurements. Then it would make sense to use groupings. However, if the criteria of grouping are purely numerical it would be better to program the computer to group the evidence and for the full information on the object to be recorded. If the groupings are subjective, based on the archaeologist's specialised knowledge, or qualitative it is necessary to follow his code. It is important then that someone with specialised knowledge makes these crucial decisions.

How much data need be recorded? Leaving aside the question of whether it is possible to over record, it really depends on the sort of programs you intend to run. If you just intend to run a program once for a single analysis clearly only what is relevant need be recorded. If the data are to be permanently recorded and may be used in lots of different programs then as much data as possible should be recorded. Some items of data may be implicit in the others already recorded. For instance the volume of earth removed from an Iron Age pit need not be recorded; the volume is not actually measured but an approximation is worked out to some sort of formula. It is much simpler to get the computer to work out the volume each time than to record the actual volume.

What sort of codes should be used? Humans use and remember strings of letters best. High level programming languages make this viable. The results of analyses are only as good as the data that goes into them Therefore it is important to make the preparation of data as easy as possible. Therefore since letters are easier for humans it makes sense to use them. However the occasional digit helps to break up long strings of apparently meaningless letters.. This perhaps is of more help to the punch-girl but the more the mechanical accuracy of the data is improved the better. It also helps you to check your own data sheets and punched cards more easily.

As simple a coding as possible should be used. The sort of coding used by libraries which consist of smaller and smaller breakdowns of a subject is unsuitable. It is amazing how rarely someone else's ideas of categorisation correspond with one's own. This tendency is fatal to a computer retrieval system. In a library, in desperation, one can browse and find what one wants but with a computer system, short of asking for a full printout, some areas of data might never be found again. The best sort of coding is brief and has meaning in itself. For instance for the sex of a skeleton M for male and F for female cannot go far wrong. Instead of 1.1 meaning "carnivore, canis" it would be better to put "DOG". It would be easier to program the codes for the animals back into carnivores etc later. Also "DOG" conveys a meaning if you lose your coding translations. It makes it much easier to program analyses, without loss of human efficiency, if the codes for "unknown" or "unnoted" are well away from the main body of the code. For example when using letters to put X or U, and when using digits to use O. If all the unknowns are coded as the same symbol it also saves programming time.

I have now discussed in general the problems behind converting pre-recorded data and put forward some of my ideas. I now hope to illustrate them by reference to a case study.

My particular job recently has been the computer analysis of a small Anglo-Saxon cemetery. First I wanted to store the information. Then I wanted to derive statistics from the data with the hope of reconstructing and explaining that area of the past.

The site I am working on is in Yorkshire and is called Sewerby. The cemetery was discovered in 1959 when the farmer had an extension built on to his farm. Several burials were found in this way by the workmen. It was partly excavated by Philip Rahtz in 1959. More extensions were planned in 1975 and Sue Hirst directed a small dig in two sites where a petrol tank was to be put and where a chicken house was to be built. Now Sue Hirst is writing up the report as an MA thesis.

The site is right on the coast, near Bridlington. The cliffs are gradually being eaten away and the site may quite well have been more inland in Anglo-Saxon times. The site is on a small ridge. It is Sue Hirst's opinionthat the cemetery does not extend much beyond the ridge. The site may only consist of 100-150 graves. However as yet no actual limits of the cemetery have been found.

The site is on a gravel and sandy soil. The gravel is very acid and many of the graves had no bones in them and sometimes only a set of teeth would remain. On the other hand the sand preserves the bones well and wherever the graves are cut into the sand the bones are all there. About 60 burials have been found so far. Unfortunately several were found before the archaeologists arrived, although some of the bones were kept by the workmen as well as some of the finds and the workmen and farmer remembered the positions and orientation of the bodies. However, of the remaining burials that were excavated only about 20-25 have all the attributes recorded.

Ms. Hirst wanted some correlations worked out by the computer but was content for me to use the data for my own purposes. As I have said I wanted to store the information. I also wanted to make the system suitable for the storage of Dark-Age burials. I decided originally that I would want to access the grave using a grave number but put in enough information to access it by other attributes later with the use of inverted files.

Since I wanted a direct access system I had to create a unique number for each grave. I decided to add the county and site name so that it could later be retrieved by county or site. Thus the number consisted of YK for Yorkshire, SW for Sewerby and a four digit number which was the grave number within the cemetery. I used the numbering already given to each grave by the excavators. This however created its problems. Sometimes a number had been forgotten, and in one place a feature listed as a post hole was thought later to be a grave related to another, so they became 35 and 35a. Double graves also create a problem as there are two skeletons to list. I gave these odd graves a number at the series with a note in both their "comment" sections on their relationship.

Again in the hope of making my system suitable for general use I added the ordnance survey reference number. I hope later to do some work on the relationship of the cemetery to other geographical features, manmade as well as natural.

I have included two pieces of code which may seem unnecessary. They are the date excavated and the disturbance code. I included them for retrieval purposes. It is now easy to suppress any data before print-out time that may be too badly excavated or disturbed. I have used a code for whether the site was found or excavated, and when. The disturbance code makes a distinction between ancient and modern disturbance as well as human and animal. Thus you can assess how useful the information retrieved will be.

Although at Sewerby there are no cremations or barrow burials I thought it necessary to add all possible variations found in Anglo-Saxon burial customs. For inhumation I have used "I" as the code and for cremation "C" in accordance with my belief in using brief, meaningful codes. There are also codes for barrows (and whether they are Bronze Age or contemporary), secondary and primary burial, multiple, double or mass burials.

In site reports it is normally assumed that you can see a plan but that is rather awkward for a computer. Therefore I have put in coordinates. When I do work on more than one cemetery site I shall have to decide on some absolute point from which to measure coordinates rather than a vague point which happens to be the left hand corner of the plan. The coordinates are all measured to the position of the skull within the grave.

I have measured the dimensions of the grave from the plan. I took the widest point across the skeleton and for the length took the longest part from head to foot of the grave. I thought it was important to record the depth of the grave in order to get some sort of measure of the effort each grave took to dig. Unfortunately, although there is a column for depth it was rarely filled in and when it was it was in reference to some vague point such as modern ground surface or to the undisturbed natural. Therefore I have left spaces for depth but will not fill them in until I have decided on a site datum and worked out the depths from that.

After a short discussion on whether to record the orientations as a discrete group , it was decided that the orientations should be put in in full. It was originally thought that when a grave was orientated at perhaps 9 degrees it meant that the grave was orientated to the north and that that was as accurate as it was possible for the Anglo-Saxons to be. However many people now believe that slight changes in orientation may be affected by the seasonal movement of the sun and therefore precise orientations were necessary. If later the individual measurements of the orientation make it impossible to form any cohesive groups then I shall program the computer to make up groups. The problem with randomly imposing a modern idea of conventional compass points is that it may break up any real grouping in the data.

The relationship of one grave to another or to a feature is very important for dating in a cemetery where there is little vertical stratigraphy. I have made no alterations to the associations that are on the data sheets. but I have coded them up. All the features in the cemetery have been given a number, so that G25 means grave 25, F9 means feature 9 and P6 means posthole 6. Therefore I have put codes such as CBG35 - cut by grave 35 or NEAP6 which means near posthole 6. I have put NONE if there is no relationship apparent.

After the details of actual position and size come the close description of the skeleton itself. Here I was greatly helped by the excavator who had worked out her own codes for many of the attributes. She has used an alphabetic code throughout. To each I have added an unknown code of X. The only problem arose when she put 'deposition B of the upper torso and C for the rest.'. I solved this by going back to the grave plan and seeing which predominated. Only three columns were not coded. They were number of bones present, sex and age.

The "bones present" slot on the sheet was filled with a list of the actual bones present. In fact I think it is the sort of information that does not go well into a code. When I had thought why the information was present I decided that although it was essential information its use was to warn you how much reliance you could place on the following pieces of data. Therefore I looked through all the possible variations there could be and put them into groups to which I gave a code number.

The main problem with the "sex" column was that the evidence had been recorded to different standards. Some of the skeletons had been sent for a pathological study but some had not. Therefore I have used a different letter to show the difference between the two types of evidence. I have used M and F for male and female among the definitely sexed skeletons. I have used A and B for male and female in the skeletons sexed by their grave goods only. In this group are included the graves where the skeletal evidence had gone and we will never know better. There are two skeletons that don't fit into either category. Two skeletons have been sexed as definitely male but have female grave goods. The excavator says the finds are probably right. Therefore I have put B and added a comment later.

The "age" column also suffers from differential recording for the same reason as the sexing. Those burials which were analysed have very accurate ageings and some of course only being fragmentary could not be aged at all. Also there is some information from the farmer about the age of the burials. In fact his evidence on one burial was corroborated. He said the burial looked as if it were old and later when the bones were examined it was found the person had suffered from acute arthritis. So the age column ranges from accuracy such as - 39 + or - 5 years - to just 'adult'. But each is valid on its own though not suitable for direct comparison. Yet it is important that no information be lost. I thought for a long time about the role of the age in the analysis of burial customs. Age is crucial to the understanding of social systems as illustrated by burial customs, but is each individual year important? I therefore decided to make age groups. By making two overlapping series of groups it was possible to combine the accurate and the rather more general groups. So 1 - 6 cover from 0 - 50 years in decade gaps, except for the first 10 years which was divided into 5 years each. 7 was 50 plus or "old". 8 covered 15-35 or young adult; 9 covered 20-50 or adult. Again one can make one's own decision on how reliable the evidence is by the group it belongs in.

The study of any structural elaboration of the grave has become important. It is evidence that does not lend itself to efficient coding. Therefore I have just used a presence/absence code and details are in the comment section. I have included all the postholes quoted as being in association with graves as evidence of grave markers. Rather than use the category of 'coffin' I have used a category called 'container' which would cover cremations in a pot or bag. As with the structures it is just presence/absence with a fuller description below. Until there is a great deal more known about grave structures it would be pointless to try to compare grave structures as they simply do not occur commonly enough.

The full information about the finds has not been filled in yet. Therefore I have just put in presence/absence of various categories. Later I hope to create another file which you can access from the main file. This may later become a source of errors due to the categorisation.

As I have emphasised it is important to make the recording of the data as easy as possible and therefore to make it as accurate as possible. It is easier to program the computer to cope with codes that are clear to humans, than it is to get the computer to correct mistakes in the data due to complicated or difficult codes. Make the coding as simple as possible. Let the archaeologist bring in the complicated specialised knowledge to make sensible codes or groups. When the archaeologist has explained his needs and the programmer has considered his hardware and software and adapts the data to fit, then mistake-free data and sensible results will be produced.