Digital recording of excavations: do we need data standards and common strategies?

Torsten Madsen

This contribution will outline what we should understand by data standards and common strategies for recording digital information. Further, it will argue what to avoid and what to pursue in connection with these issues. My focus point is archaeological excavation, but the argument can easily be applied to other issues in connection with museums, and thus be of interest to others than those concerned with excavations.

Data standards

The concept of data standards refers to two different areas of standardisation. The one concerns the standardisation of description- and classification systems for data to be recorded. This I will refer to as *standards of content*. The other concerns the way data will be recorded and stored in digital format. This I will refer to as *standards of form*.

It is seldom to see a clear differentiation between these two areas. Normally they are considered to be closely associated in the sense that content determines form. Actually it is two entirely different issues without logical interdependence. It is therefore necessary that we treat and evaluate them independently. In the following I shall argue that we need to develop standards of form as a base for a mutual co-ordinated recording of excavation data. These standards must be of a nature that allows for widely differing description- and classification systems. This is a necessity because standards of content must be avoided at all costs, I will argue.

Let me start with standards of content. Why should they be avoided? This question brings us deep into a long-standing discussion concerning the nature and role of classifications in archaeology. Through more than 25 years of intensive discussions it has become increasingly clear that it is neither possible nor desirable to reach consent on standard classifications. The debate came into focus with a well-known paper by Hill and Evans from 1972 with the title "*A model for classification and typology*" (Hill & Evans 1972). In this paper it is demonstrated that classifications are always dependent on the questions posed. With an increasing number and variety of questions more and more diverse classifications become necessary to elucidate a given set of data.

The key issue is the role that classification plays in research. Is it a neutral objective tool that may be used by the researcher to depict structure and qualities of materials investigated, or is it a mean by which the researcher may assign qualities, structure and meaning to materials investigated? In other words is it a passive or an active tool?

All claims for standards of content are logically founded on the idea that classification is a passive, neutral tool. Only if so, it may be argued, that universal description and classification systems can be created that will produce independent, objective results when applied.

The prevailing attitude in archaeology today is that description and classification systems are the means of the researcher to categorise and to assign qualities to materials, and then use these actively to create meaning. Exactly what categories are created and what qualities are assigned will depend on the goals and attitudes of the researcher. Thus the choice of description and classification in many ways will determine the results obtained, and different choices may well lead to different results (Jensen & Nielsen 1997)

If archaeology, one way or the other, ends up with standards of content in its recording systems the research process is bound to fossilise. The sudden jump in knowledge that a completely new way of looking at data may lead to will be difficult to achieve. Reproduction rather than knowledge gain will become the rule. All experiences show that it is very difficult to make archaeologists accept standards of data content. Everyone thinks that their own ways of looking at materials are the best and most meaningful. Basically, the unwillingness of archaeologists to accept standards is very healthy. It shows the views of a researcher and not a technician.

With standards of form it is different. The databases we create for recording are basically just containers of the data we wish to record. The way we design such a recording system should ideally be independent of the content. Unfortunately this is never true in praxis. The rule is that the description and classification systems used here and now will determine the design of the database. As a consequence, different databases for the same type of material with designs based on different recording and classification systems turns out to be logically incompatible.

An important question in this connection is, if this incompatibility is determined by the existing database technology, or if it is determined by the practice of database design. 10 or 15 years ago the database technology was certainly a limiting factor. Today it is very much less so. We may not yet be able to do all that we want to do, but it is obviously more the practice of design than the technology that constitutes the major limitations.

The traditional approach to design of database systems is first to analyse the prevailing practice in the part of the world of realities that the database should depict, and then transform this practice into a formalised database structure. This approach works fine with administrative systems where procedures and standards of content by itself are the central issues of the systems. It is not, however, an acceptable approach in connection with research oriented applications, where the goal by itself is innovation. In connection with this type of application we should try to limit ourselves to a very general level, when we design the structure of the system.

The starting point for establishing a research oriented recording system should be a general conceptualisation of data at hand. The design should primarily focus on this conceptualisation, and not so much on the different ways that practice has developed in connection with the handling of these data. With this starting point the aim becomes to create a design, which is sufficiently general to accommodate different structures, classifications and descriptions of data. We should try to reach a database design that contains no restrictions on data beyond our basic conceptualisation of data.

Let us take a closer look at such an endeavour and try to isolate some of the conditions for such a design. My basis for doing this is work I have carried out together with Jens Andresen on a general recording system for archaeological excavations (Andresen & Madsen 1992, 1996a, 1996b).

Increasingly, it has become clear that the research process can be viewed as a dialectic process between a theoretical modelling on the one hand, where pictures of the past are painted, and a data modelling on the other, where observations are organised into meaningful structures (Madsen 1995). The core of the process is an interaction between these two kinds of modelling with a continuous dialectic inter-flow of information. Classifications and descriptions are an integrated part of the data modelling process and as such it is an active research component in its own right. It is built into this process that we vary and change our data models to be able to confront our theoretical models with new and different data structures. The moment we subside and accept a particular classification as everlasting the research process stops. We thus need classifications and descriptions to change continuously. Further we need alternative competing and supplementary classifications to be available for any particular set of data. This leads us to a first demand for a standard design:

- It should be possible to classify data through more classification systems at the same time. In principle the number of such classification systems applied simultaneously to any particular set of data should be unlimited.
- It should be possible to add and change classification systems "on the fly" while they are being used for recording. However, the integrity of already recorded data should be enforced (i.e. it should not be possible to delete description or classification elements, if they are already in use).

If we look at the structure of classifications it would be unacceptable if these were limited to having all classes at one single level. Most classifications are hierarchical, but also more complex structures are possible. The way classes are structured in relation to each other is by itself meaningful, and it is hence important that the classification structures become implemented in the database in such a way that they are fully operational. It should be possible to perform searches that are dependent on the classification structures (i.e. find all instances that are not recorded as subclasses of class x). This leads us to our third demand for a design.

- It should be possible to incorporate tree-structures in the definition of the classification systems and use these to operate on the classified data.

Archaeological data are to a high degree contextual. They not only relate to each other somehow, but the precise way in which they relate is meaningful and important. Hence we should not only be able to establish relationships between different records but also be able to qualify and quantify the nature of these relationships. A good example of what this implies can be found in a paper by Costis Dallas with the title *"Relational description, similarity and classification of complex archaeological entities"* (Dallas 1992). Our fourth demand for standard design then becomes:

- It should be possible to categorise and quantify all relationships between recorded data.

For each class in a classification system it should be possible to assign an unlimited number of variables. It should be possible to define the variables using a basic set of scales. Thus we should be able to define: nominal scale variables with or without multiple choice and with the possibility to separate between alternative and dichotomy structures; ordinal scale variables with rank order recorded; ratio scale variables with values recorded as points or intervals on a continuos scale.

In addition to this we need to be able to define how the variables associate with the structure of the classifications rather than just with individual classes. A variable assigned to a class should automatically be assigned to all subclasses of this class. What we should aim towards is heredity between classes in a classification system with respect to its descriptive variables. This lead us to a further two demands for a design standard.

- It should be possible to assign an unlimited number of description variables to each class.
- Inheritance between classes should be established so that all variables of a class automatically exist as variables of all its subclasses.

Creating a system, where the structure of the content is independent of the recording system, means that we will have no foreknowledge of what kind of data will be stored in the system. There is thus no reason to create "code books" with definitions of the content in the database, as these will change continuously. Instead it is important that definitions of content are kept together with the data in the database. In this way the database will become self explainable with respect to content. This leads us to the seventh and final demand for a design standard.

- All definitions of classification and description systems used must be kept in the database, so that all recorded data becomes understandable without the use of external information.

One of the strongest arguments for standards of content has always been that this makes it possible to compare data from different recordings. If such standards do not exist we will end in a Babylonian confusion, where nothing can be compared or analysed together, it is claimed. As already mentioned standards of content are not acceptable. The question then is if everything as a consequence will end in chaos. This need not be the case, I will argue, if we carefully create design standards that will allow comparisons regardless of content. Comparisons may not always be meaningful to perform, but that is another matter.

An outcome of the design standards outlined above will be that regardless of what content is entered into the database it will all have the same logical structure and be stored with the same physical structure. From an operational point of view everything is comparable on database level regardless of actual content. This leaves us with the problem of concordance in the structure of content.

If the structure of various parts of content is very different then comparisons cannot be made, and neither will it be meaningful to make comparisons since the point of departure and the aims for making the different structures obviously has not been the same. We should not try to force compatibility, if it was never intended. Where the differences are small and possibly only a question of different naming we can add a demand for thesauri to our design structure.

If the administrative powers, as they invariably will, demand standard classifications for certain parts of data, this will pose no major problem, as the design standard ensures that simultaneous recording by different parallel classification systems can take place. A little extra work at the keyboard is all it takes.

Common strategies

With common strategies we move into a completely different area with no research based constraints. We are here dealing with technical and not least political issues, the most important of which is how we store and secure our digital data for the future. More peripheral, but still important, is if we can further productivity and the use of digital recording through the establishment of a common infrastructure, and through the choice of hardware and software solutions.

Digital storage

When we speak of storing digital data we will first and foremost have to make clear the nature of the computer medium. Digital information in electronic components does not have a stable physical form, nor does it possess analogue qualities that may be seen or heard. It exists as patterns of presence and absence of electrical currents only. When the power is cut off the data is gone.

To store electronic information it is necessary to convert it into a more stable form, which can be quickly generated, and from which the electronic information can be quickly re-established when the power is turned on again. Up until now the dominating media for recording has been a surface that can be magnetised either on tape or on disks. Increasingly, today optical media takes over the storage role for digital information.

Both magnetic and optical media have the virtue of storing the digital information as a more or less direct copy of its electronic form, and hence it allows the writing and reading of this information with tremendous speed. Neither of the two types of media are however physically stable when measured with museum archival demands. Magnetic media can only be expected to last a few years, while optical media will last a good deal longer, but in all probability not as long as we would wish them to.

Confronted with this problem the reaction you normally meet is that all information should be kept in paper copy. If the digital data should disappear this would be unfortunate but at least we would have a paper copy. This, however, is neither expedient nor possible. First of all why use digital recording if the end product is a paper archive? Secondly, the analogue presentation of data on a piece of paper or on the screen for that matter is not equivalent to the digital data. It is an interpreted version. If we should make a direct presentation of digital data on paper it would have to be as a patterned occurrence of small black dots. Thirdly it is only possible to produce a reasonably correct interpreted version on paper of digital information, if the digital information itself is more or less a direct translation of some information originally kept on paper. It would be difficult for instance to give sound and interpretable presentation on a piece of paper. Nor would it be possible to give the content of a complex relational database an understandable presentation on paper, even if it is possible to write out all its constituent parts.

The conclusion of all this is that the digital format binds. If we wish to preserve the information that we have created in a digital format, then we can only do this by keeping it alive in that format. We must make sure that the representations of digital data kept on magnetic or optic media stay valid. We must periodically read and rewrite them to make sure of their validity, and we must keep more copies in separate locations to safeguard against accidental destruction.

Keep the data alive

The danger of physical destruction of information is however only one of the problems we have. Digital information is coded information and to decode it (take it from a digital to an analogue format) we need both hardware and software that can handle the digital format. The outdating that continuously takes place with both hardware and software due to the fast development of information technology gives us even greater problems. With hardware it is a question of choosing the newest *de facto* standards for data storage, whenever files are copied from one media to another to safeguard against physical destruction. This is a simple problem to solve. Worse, however, is the software side of the problem. Here it is not a question of copying, but a much more complex one of converting data from one format to another. If you record your data today using an off the shelf software product, and then do not touch the data any more, the likelihood is high that in 20-30 years time you will not be able to find a software product that can access the data.

Again the conclusion must be that there is only one safe approach, and that is to keep data alive. This means that every time a change in a data format of the software product used occurs data has to be converted to the new format. If changes occur in the market, making the software product used marginal or on its way out, data has to be moved to another and more viable product. This may be an expensive and time-consuming process. It is very likely that you have to go through a lot of work to recreate the database structure in the new product before data can be moved (Small 1997).

One of the purposes of museums is to record and keep cultural history information for the future. By changing to digital recording of this information you engage with a media that poses completely different and much sterner demands to resources, knowledge and organisation than do the traditional paper filing. The questions therefore become: 1) can we expect that cultural history museums with often only a few employees can fulfil these demands not only now but for the safety of the records for the future. 2) and is it rational that, for instance, the odd fifty museums in Denmark with an obligation to secure archaeological information must fulfil these demands, if it means that each one of them has to employ a trained person to do so?

The answers are self-evident and the conclusion must be that there is a need for a common strategy to secure cultural history information in digital format. In a country like Denmark only a national strategy is relevant. In this strategy it must be clearly defined who has the responsibility for the survival of data and how such a survival can be achieved, including the creation of a system that makes flexible recordings possible as argued in the first part of this paper.

Systems, software & storage

The situation in Denmark today is that digital recordings have begun, but there is no strategy whatsoever. There is a well-defined recording system for sites and monuments on a national level (DKC), a recording system for objects in the National Museum (GENREG), an administrative system for cultural history museums (DMI) that some use and others do not; experiments with recording of textual information from excavations (IDEA), and no qualified bid on how to handle digital plans from excavations. It is to be expected, not to say it is certain, that the DKC office will be made responsible for storing and maintaining digital information. It is less certain, if there is political will and courage to enforce a common strategy, and if so, how such a strategy will appear, and not least if it will respect the research demands for flexibility. If the latter is not the case it is foreseeable that the strategy will be undermined, as a lot of information recorded will never reach the central archive because it cannot be fitted into this.

The introduction of a centralised storage of digital data results in various practical problems, which should also be addressed under the heading of common strategy. One such problem is the actual communication between museums and the central authority responsible for the digital information. It is of course imperative that this communication is flexible and efficient. Data should continuously and automatically flow from the museums to the central authority and at the same time the museums should continuously and, just as automatically, be able to draw on the information stored centrally. Today there is in reality only one solution – the Internet. It cannot be too much to demand that the museums should be hooked up to the Internet in such a way that even large amounts of data can be transferred at speed. Nor can it be too much to demand that the central authority ensure an efficient way of collecting and redistributing data via the Internet. A common strategy for such data collection and data dissemination is a necessary but also trivial political decision.

It is more problematic to introduce a common strategy for the software to be used for recording, analysis and storage of data. At first glance it may appear to be a trivial technical decision. My experience over the years, however, shows that this is not the case. Many users of computers are very conservative. When they have learned to use a program they are in no mood to change to another. Often they have had difficulties learning to use a computer, and when they have become familiar with a particular program they do not feel like starting all over. Others take a completely different attitude. They willingly try out new solutions, and if they find better ones they are likely to change without hesitation, and will not accept to be blocked by decisions of common standards. The results are a high degree of heterogeneity of software used, and little willingness to convert to common solutions.

On the other side of the fence we find the technicians taking care of the computer systems. Naturally they are very interested in establishing and maintaining standards for software used. A high degree of diversity gives them problems and a lot of work, and as a consequence they are most often very conservative, and do not like to see the systems changed, which in computing of course is a futile attitude. At my own institution, for instance, several of the computer technicians have only recently and reluctantly acknowledged the need to change from DOS to Windows based programs, and they have indeed received heavy support from those users who are still sitting comfortably with their WordPerfect for DOS.

A common strategy for use of software calls for mild pressure on the one hand as to what software to use, and an open mind on the other hand towards the recognition and acceptance of new and better software solutions. At the same time the strategy has to have a tolerance for individual preferences among the users.

Conclusion

It is almost 20 years ago that I was first introduced to a computer in the shape of the Aarhus University main frame computer, which could be addressed through punch cards and later through screen terminals. It is less than 10 years ago that I began to grasp what the development in personal computing could lead to in archaeology. It is only now that I see a movement in archaeology towards a general utilisation of the computer.

The last 10 years have in many ways been frustrating. In the beginning I was optimistic. I thought that very quickly we would see a focus on the possibilities of the computer in archaeology, and a serious attempt from the discipline to utilise it. It did not happen. On the contrary I found indifference and a lack of will by individuals to try to understand the new technology, its potentials, problems and limitations.

Will it change now? Does the *de-facto* spread of computers into all corners of the field of archaeology, and the fact that the most important raw material of archaeology - its data – increasingly will be stored in digital format, mean that archaeologists now actively will engage themselves in a learning process? We can hope, but there are reasons for pessimism. We can change our education systems and make courses for the older generation as much as we like. Only if the individuals realise that it is a necessity for an archaeologist to have a well-founded insight into information technology, and personally seek this knowledge, will we see results. Progress has been slow and still is. Even among new generations of students many are negative and dismissive.

Now, as information technology becomes an active tool, and as our data shift from being presentations on paper to digital representations, it becomes a serious problem to archaeology that lack of knowledge and lack of established procedures invariably will result in loss of data. The next 10-20 years will show how serious the problem will be. However, when Murphy strikes remember not to blame it on the information technology.

References

Andresen, J. & Madsen, T. 1992. Data Structures for Excavation Recording. A Case of complex Information Management. C.U. Larsen (ed.*): Sites & Monuments. National Archaeological Records*. The National Museum of Denmark, pp. 49-67

Andresen, J. & Madsen, T. 1996a. IDEA – the Integrated Database for Excavation Analysis. H. Kamermans and K. Fennema (eds.): *Interfacing the Past. Computer Applications and Quantitative Methods in Archaeology CAA95.* Analecta Praehistorica Leidensia 28, Leiden, pp. 3-14.

Andresen, J. & Madsen, T. 1996b. Dynamic classification and description in the IDEA. III International Symposium on Computing and Archaeology. *Archeologia e Calcolatori* 7, pp. 591-602.

Dallas, C. 1992. Relational description, similarity and classification of complex archaeological entities. G. Lock & J. Moffet (eds.): *CAA91. Computer Applications and Quantitative Methods in Archaeology 1991.* BAR International Series S577, Oxford, pp. 167-178.

Hill, J. N. & Evans, R.K. 1972. A model for classification and typology. D.L. Clarke (ed.): *Models in Archaeology*. Methuen & Co. Ltd., London, pp. 231-274.

Jensen, C.K. & Nielsen, K.H. 1997. Burial Data and Correspondence Analysis. In C.K. Jensen & K.H. Nielsen (eds.): *Burial & Society. The Chrono-*

logical and Social Analysis of Archaeological Burial Data. Aarhus University Press, Aarhus, pp. 29-61.

Madsen, T. 1995. Archaeology between facts and fiction: the need for an explicit methodology. In M.Kuna & N. Venelová (eds.): *Wither Archaeology? Papers in Honour of Evzen Neustupný*, Praha, pp. 13-23.

Small, J.P. 1997. How to transfer your DOS database into a Windows 95 database in 659 easy steps. *CSA Newsletter*, Vol. X, No. 2.