Torsten Madsen

Computing in Archaeology - Integration or Disintegration

Paper presented at 16th Annual Meeting of Gesellschaft für Klassifikation - Dortmund 1-3 April 1992

Appear as read, with minor corrections to grammar and language.
Torsten Madsen 30-9-2003

Archaeology is flooded with data, and has been so for years now. Take Denmark for instance. There are well over 100 archaeologists employed in permanent positions, and in addition to their wages, a sum of approximately 5 million D-mark is used annually for excavations. The resulting amount of information is staggering.

Up until the Second World War only a handful of archaeologist were doing excavations, and it was easy for one man to keep track of all that happened. A good researcher could even keep track of much of what happened in other parts of Europe as well. The last person in Denmark, who was able to keep track of everything, was Professor C.J. Becker, but somewhere in the sixties, even he was overtaken by the development and clearly lost insight in several important areas.

To day, we find three types of archaeologists.

There is the field archaeologist, who excavate all and everything, but who actually knows very little of the cultural context of the material he excavates, and who has more than enough to do processing the data he uncover and write up his reports.

The second type is the specialist, who has buried himself in specific problems of either data or periods. Within his domain of expertise, he is unbeatable. He knows what is worth knowing, and he is very productive in terms of publications. The only problem is that very few people read what he is writing, either because they are field archaeologist, and don't have the time to read that sort of stuff, or because they are specialists, and only read publications within their own field of expertise.

The third type of archaeologist is what Flannery once called the great synthesiser (not to be confused with the noisy music instrument). He can write up overviews and syntheses on everything, but his knowledge of the data, supposed to be the bases of his writings, is shaky.

Archaeology in Denmark, and I suppose anywhere else, clearly suffers from information indigestion. Taking a closer look at the problem, however, you are liable to find that it is not created entirely by the escalating amount of data mentioned in the opening paragraph. There are structural problems as well.

One such problem has to do with the structure of Danish archaeology. Archaeological research is almost exclusively run by museums, and indeed most archaeologist are scattered across more than fifty museums. Each of these museums perform their own research, has their own archive, and their own collection. There is of course a central archive giving rudimentary overviews of what has been found, but in order to study in detail say a period, you have to visit all museums, and go through all their archives. Moreover, do not believe that a visit to the National Museum provides you with the major part of the material. To day the National Museum cover only some 5% or less of the total activity. The rest is spread evenly across the country.

Another problem relate to the structure of the data themselves. As the level of archaeological activity rises so does the level of complexity of the data to be processed. Excavations tend to become larger and the information stemming from them more difficult to overview.

Obviously, archaeology is now faced with serious problems of information management and processing. We see elements of information overload, information bottleneck problems and information fragmentation. Where we would like to see a pool of integrated archaeological information, we see disintegration grow.

The obvious solution to such information handling problems ought to be the computer, and personally, I do believe it is the solution. However, if you investigate what has come out of using computers in archaeology so far, there is no obvious trend towards integration. On the contrary, the computer seems to have furthered the disintegration.

In Denmark as in most other countries, I guess, the computer has been applied to four main areas within archaeology. These are Sites and Monuments records; excavation recordings; statistical analyses of data; and publication production.

There is one central Sites and Monuments record system in Denmark - called DKC. It is a direct continuation of the so-called Parish Inventories that has been kept at the National Museum since the last part of the preceding century. For approximately 10 years now, all new finds has been recorded in this database, while at the same time a retrospective recording of old records is carried out. To day a little more than half of the 140.000 known archaeological sites in Denmark have been recorded in the database.

The database itself has so far not been opened to interactive search performed by persons outside the DKC office. You can get listings on paper from the database - preferably complete parish-by-parish listings as these are "on stock". You may also order listings of specific types of monuments or sites, but requests for complex searches are for various reasons not well received.

What you get when you order a listing is a stack of papers 10 times as voluminous as the stack of Xeroxes you used to get in connection with the old parish inventories. The reason for this is a more spacious layout, an inclusion of redundant information on the listing of each site, and the inclusion of additional information, which quite laudably has been merged during recording from other files at the National Museum.

Excavation recording is increasingly done on computer, but till now on heterogeneous and differentiated levels. In some cases the computerization is limited to the use of a word-processor on a portable computer in the work shed at the excavation. The advantage is entirely one of speed, as the stage of the hand written notes is avoided. The excavation report, however, does not differ in any way from the one produced without the help of a computer.

In other cases, a regular database system is used for the processing of the data. The benefit of course is that some steps in the post-excavation processing of data become easier, faster and more systematic. The outcome, however, is a printed excavation report, and due to the use of the database, and the ease with which data can be printed in different combinations and constellations, the report tend to be a lot more voluminous than the traditional excavation report. Indeed excavation publications based on a database recording also tend to become very voluminous, due to the ease with which "documentary" listings can be produced and included.

The use of the computer for analyses is a slightly different matter. Any analysis takes a set of data from one state to another, and often the transformation is so complex that it could not have been performed without the use of a computer. Nobody could argue that the computer is not a helpful, indeed indispensable, tool for analyses of data. But it does not provide integration of information in this case either.

Analyses of data has increasingly become a work of "specialists" - not statisticians and the like - but archaeologists who have specialised themselves in various often complicated types of analyses that they regard as useful in connection with the kind archaeological investigation they perform. It could for instance be archaeologists who has specialised themselves in the use of correspondence analysis to mention a recent, very popular and indeed very useful tool.

The application of correspondence analysis has a first state of atomisation of data for the input matrices. The data remains atomised, but from the analyses emerges structures, which may or may not be intelligible in terms of the archaeologist’s knowledge of the data.

Apart from the new insight into the data that the analyses often offer to the specialist, they do produce many extra data in terms of input and output from the analyses. These data find their way into the publications, where they will appear as documentation. Yet to

the reader they will often be another set of unintelligible data due to a lack of familiarity with the subject matter of the analyses, with the analytical method itself, or with both.

Computers thus tend to create more data in archaeology, and these data tend to end up as all other data on paper, and quite often presented in proper publications. And for publication production the computer is helpful. Due to the highly developed word processing systems available to all archaeologists, the publication productivity has been raised considerably over the last few years. Further, the computer has helped to make the publication production more efficient and much cheaper. Thus, publications stuffed with information appear as never before. Quite apart from whatever contributions has appeared in journals etc., my own institution for instance has published 7 monographs containing approximately 23-2400 pages within the last year.

Thus, so far the computer has helped to speed up some of the processes of work, it has undoubtedly led to a higher level in data analysis, and it has contributed significantly to a marked growth in printed information, but it has not in any way helped to create an integrated information environment. On the contrary, it has been used to speed up the traditional pattern of research leading to increasing information disintegration and isolation.

The major problem facing archaeology right now as I see it, is that information storage, processing, presentation and dissemination is based on the concept of the printed document. This concept of course is borne of tradition, and has been the foundation of knowledge in research and society in general to our civilization. Right now, however, the foundation is shaking. By virtue of sheer magnitude, we cannot any longer overview the information. The individual researcher either ends up with a general, limited knowledge, or ends up as a highly knowledgeable person on increasingly more limited aspects, and with complete ignorance on all others.

We can all read syntheses, but the possibility to search deeper into the basic information underlying the syntheses diminish as the amount of information grow. Further, even the specialist is using more and more of his or her time to gather information, and increasingly has to narrow the scope of investigation.

When for instance I and some colleagues in Copenhagen posses somewhat differing views on what the earliest Neolithic in South Scandinavia is like, this may partly be attributed to the fact that I base myself mainly on the west Danish material that I know very well, while they base themselves mainly on east Danish material. Thus the information base underlying our syntheses disintegrate, and our ability to present syntheses based on a sound broad spectrum of the available information diminish.

Can anything be done about this situation? - is there a way out, and more specifically do the computer offer us a way out? I think so. At least theoretically, there is an obvious solution that may offer a better information management and a better information access. In practice, it may turn out that it will take long time to implement this solution, and it may take even longer time to convince the research establishment that it is a viable alternative, and an efficient solution.

The first step is to realise that **printed information is dead information**. By this, I do not intend to imply that printed information is bad or not useful. On the contrary, there is probably no better way to communicate thoughts and ideas, and I cannot imagine how we should fare without printed papers to present our hypotheses and conclusions based on the archaeological data.

What I intend to say is that when we deal with complex, highly structured information on a computer, then printing the information on paper transforms the information from a structure and a media, where even huge amounts of information can be searched,

operated and analysed through strict, well considered algorithms, to a media, from where it can be searched, operated and analysed by the human mind only. However brilliant the human mind may be to intuitively grasp structure, there is a very low limit as to how much information it can overview clearly.

Further, there are very severe limitations with respect to what kind of data structures can be presented on a piece of paper compared to what kind of data structures can be kept in a computer. Archaeologists have often talked about multidimensional reality etc. but the types of descriptive and documentary tabulations that fill up archaeological publications are not multidimensional in structure.

A relational database is a dynamic information storage facility that can hold very complex multidimensional structures, and its content can be made immediately available as a succession of different one-dimensional views defined by the individual user at the point of querying. In contrast, printed information can be seen as a static one-dimensional view of a multidimensional reality defined by the author once and for all.

Storing information in a relational database, however, is no solution by itself. We may store all the archaeological information in the world in relational databases, and be no better off than before. What we need is to build a computer based information universe in which to store, access and analyse our information in an integrated environment. In order to build an information universe we need to concern ourselves with:

1) The domain of the universe - that is the kind of information it should hold,
2) The structure of the information - both the logical and the physical model in Data Base Management Systems terms.
3) The communication structure within the universe - both the physical "networking" type of communication, and the logical "language" type of communication.

First, what should the domain of such an universe be? Should all archaeological information enter into one huge information universe, should we have different universes for different categories of information, or should we create a universe for only part of the information and not at all for other parts? There is a rich ground for theoretical considerations here, and I am not going to make any final suggestions. Let me however point out one obvious logical division pertaining to archaeological information.

We may distinguish between archaeological sources and archaeological information. An archaeological source is the potential site to be excavated, or an artefact to be described. Archaeological information on the other hand is the excavated site as described through drawings, notes etc. or the artefact as it appears through structured descriptions. Now within archaeological information, we can distinguish two types. There is information that is historically unique, and information that is reproducible. An archaeological excavation for instance is a one-time event. Through a complete destruction the source is altered into information and that information is historical as it can never be legitimately verified or altered. An artefact, on the other hand, can be examined repeatedly, and on each occasion a new set of information can be created, identical or different from the previous.

I am very much in doubt whether we should attempt to include non-historical information like artefact descriptions into a rigid information universes. I am certain however that we should build an information universe for excavation data.

Following the decision to create a universe for excavation data we have to decide a structure for the information. First, we have to concern ourselves with the logical model for a structure, which is a problem that archaeology has to deal with itself. There is a bewildering variety of ways of doing archaeological excavations, and if, what you see is what you get, then there is little hope to create a common structure that everyone could agree upon.

I shall not venture any deeper into the discussions of this point, but only point out that recently Jens Andresen and I have published an article, where we argue that it is indeed possible to find a common basic structure. Basically, most of the variety we see in excavation recordings is more apparent than real, and basically due to differences in terminology, and adherence to different operational strategies. We are convinced that it will be possible to create a universe for excavation data with a common basic structure, even if different users may wish to view the system differently through individually tailored user interfaces, and use it differently as a reflection of their habits.

Compared to the logical model of archaeological information from excavations, the physical model is certainly a minor problem, although it is something that has to be agreed upon to a certain extent if we are not going to make thing too complex for ourselves. It is not however an item of immediate concern.

A much more pertinent problem is the communication structure, and that includes both the more physical aspects of how communication is set up, and the logics of how it is performed. The reason why this is such an important point should be obvious. What we are seeking is a better way of information access and dissemination than the one we have. A better structuring of the data is only part of it. To succeed, it is imperative that we find a much better way to communicate information, and we should keep in mind that there can be no halfway solutions in this game. Communication is a unity concept, where all parts have to be tuned for the exact same wavelength.

For the physical aspects of communication, two important conditions must be fulfilled. The one is that it must not make any difference to communication where the user is positioned physically, though it is a precondition that it has to be within the same linked set of networks. The other condition is that it must make no difference where information to be communicated is stored physically, again if it is within the same set of linked networks.

What I am getting at of course is that communication of information should be based on the concept of a distributed database in a widely available physical network. I see no way that a model with a centralised database would stand a chance, at least not in Denmark, and I doubt very much in other countries as well. Thus, a database of excavation information in Denmark would physically be placed in up to say fifty different positions, and yet the user should feel that the data were all placed at the hard disk of his own machine.

A distributed database opens up for immense problems. Avoidance of sharing violations becomes a very complex and difficult matter, and things are not made easier by the necessity of distributed data processing as well. That is data base operations and data analyses must take place at the same machine, as the data is stored on. Otherwise, the response time on many data processing operations would be painstaking slow due to the excessive data transport that would have to take place across the network.

Computer scientists are working intensely, trying to solve the problems of distributed databases and distributed data processing. To day, it is more of a concept than a reality, but rather soon we will begin to find the technology of distributed databases commonly available, and not only in the high end of the computer market.

The physical structure of information access and dissemination however is only one part of the problem. An entirely different matter is the logics of information flow within the universe - the language to be spoken if you wish. This is not as much a question of how data from the database end up at the users machine, as it is a question of how the data is flowed and utilised at the users front end. That is how data are presented to the user, and how they are moved between different programs of analyses. In essence, it is what makes the difference between segmented information handling systems, and totally integrated information handling systems.