Integrating methods and data: reflections on archaeological research in an IT environment

Torsten Madsen

In 1991 a Danish National Research Foundation was established with funding of US$330M. The strategy of the foundation is to fund only a few projects, but then to give these (literally) a golden future. Although the money from the foundation was expected to go mainly to the natural sciences, there would also be some funding within the humanities, and it was understood that archaeology, in particular, would stand a good chance.

The prospects of huge funding encouraged many groups of archaeologists and archaeological institutions to apply. In collaboration with colleagues from the Moesgård Institute, the Moesgård Museum, Ribe Museum and the Documentation Department of the National Museum I submitted a project entitled *Archaeology in the Information Age*. This project was different from the other proposals, in that it was not a project aimed at the production of archaeological data and knowledge from these data. Rather, it was a project aimed at developing ways of handling data and creating knowledge from these data within an information technology (IT) environment. In other words, it was a project aimed at developing archaeological methodology, given a framework of modern IT.

Our proposal was not funded, but the work that went into its formulation was not wasted. In putting it together, we had to take a new and coherent view of the archaeological research process, and we had to look at archaeological research from the point of view of information science, leading to new perspectives on a number of issues. There are good reasons to take a closer look at the proposal and some of its ideas, as we are dealing with issues which I believe will become important for archaeology in years to come. Indeed, archaeology must begin to work seriously on its methodology if it is to survive the current information explosion resulting from its increasing numbers of practitioners. All too easily archaeology can veer towards building castles in Spain, without regard or care for the data that should have been the building blocks for more solid constructions.

The core of the proposed project was to develop archaeological methodology within an IT framework

Methods in the Mountains: Proc. UISPP Comm. IV Mtg., Mt. Victoria, Australia, Aug. 1993
I.Johnson (ed) 1994 Sydney University Archaeological Methods Series # 2

and with full integration of methods and data. The first part of this paper is concerned with the research proposal itself. The problem addressed and the background to the proposition will be outlined, and a summary of its content will be provided.

In the second part, some issues of relevance to the idea of integration of data and methods in one dynamic system are presented. The most important issue is, of course, how we conceive data and methods in archaeology. Beyond this, however, lies the central problem of how we conceive the archaeological research process. I will start with the latter issue, and then turn to the nature of archaeological data and archaeological methodology. Finally, in the concluding section, I will state the basic requirements for an integrated IT based system as it was conceived in the research proposal.

The research proposal[1]

Problems addressed and background

A fundamental problem in archaeology today is the rapidly growing mass of primary data. It is increasingly difficult for archaeologists to cope with the information accumulating at all levels of research. The problem is aggravated by a segmented methodology rooted in traditions of the last century, and tuned to the use of pen and paper. Every type of task has its own procedures aimed at summing up information in written statements. Others may take up these statements as a starting point for further research, or they may leave them unused in the archives.

A good example is excavation. Traditionally, archaeology views excavation as a technical activity not directly associated with the formulation of culture-historical syntheses. Even if this view is outdated today, an excavation is still considered as an activity which is an end in itself. It is formally closed when all finds and documentation are in the museum archives. The excavator is only obliged to deliver the record in a proper form. (S)he is not required to analyse the finds, to append culture-historical syntheses of the excavation to the files, or to publish the excavation, however important it may be.

To see the problems facing us we may divide the archaeological domain into three areas, traditionally viewed as separate entities: *Data Procurement*, *Data Analysis* and *Data Dissemination*

Data Procurement: There are several areas of data procurement within archaeology. The most important and most troublesome of these is excavation. Data deriving from modern excavations are often too comprehensive and complicated for researchers to record and treat efficiently with current 'notepad' based methods. Attempts to use computer recording have so far been of little help in solving the problem of complexity. This is because the systems established tend to mimic the 'notepad' approach to data, rather than dealing with complexity from a multidimensional, relational point of view.

There is a major need for a computer based data recording and management system for archaeological excavations. The system must be based on an in-depth archaeological analysis of the structure of excavation data; it must be flexible to cope with different levels of particularity in field recordings; and it must have a global well-defined file structure that secures portability and re-use of data from applications other than those defined within the recording system itself.

There is a further need to improve many of the recording methods used in the field. This is not just a question of applying new field technology - many excavation Units do so anyway to save money. Rather it is a question of reassessing traditional methods of recording to evaluate if they optimally document the complexity of data. We must continually remind ourselves that the 'tyranny of pen and paper' has a decisive role in the way we do things.

Data Analysis: We normally encounter a high level of complexity in the processing and analysis of data, whether excavation data, data from regional studies, or studies of artefact groups. The versatility of the data is not easy to represent in the analyses, and the results often become a number of static representations that are far from satisfactory, considering the variability of the data in multiple dimensions.

There is an obvious need in archaeology to improve analytical methods, and this applies to all areas of the discipline. It is true that many sophisticated analytical methods exist in archaeology, for example multivariate statistics which have undergone amazing developments in recent years. Yet the multitude and diversity of analytical methods in archaeology, many of which are borrowed from other disciplines, often lead to poor applications. There is a clear need for better understanding, implementation and use of many methods.

During analysis of large datasets from excavations we must tackle various issues: stratigraphic analysis; distribution of artefact categories across features and/or stratigraphic levels; comparison of artefact distributions; phasing based on stratigraphy and artefact content of features; combination and incorporation of available information into a site model; etc. Although the principles are quite simple, the magnitude of the task when it comes to 'real world' data from large excavations can be insurmountable. Many excavations remain unpublished due to this problem.

Within regional studies we find major problems associated with the handling of map data. There are technical problems in acquiring relevant base maps, but above all there are problems with formal analyses of artefact and site distributions. Although regional studies are a major topic in Danish archaeology surprisingly few enhance or formalise analysis of the geographical component. Despite articles and books addressing this issue, especially in the Anglo-Saxon literature, there is currently no formal set of methods in use for comparing and evaluating artefact and site distributions on a regional level. Nor are there any established methods for comparing archaeological distributions with various types of mapped 'background' information (environment, topography, etc.). One reason is that many of the potential methods have had no chance to become established within the traditional methodology because they are too complicated and time consuming to carry out with pen and paper.

Analyses of artefacts employ a wide variety of methods. Typology is just one example, though a notable one, in serious need of renewal. For years now we have maintained that the typological method is an analytical decomposition of the items at hand, and a subsequent definition of types based on selected attributes from this decomposition. Increasingly, however, researchers have begun to avoid types or to talk about polythetic types, which are not types in the accepted formal sense. The reason is that typologies are rigid, monothetic and one-dimensional, whereas the reality to which they are applied is variable, polythetic and multidimensional. The problems have increasingly become clear, but no acceptable alternatives have so far been established, although the issue is much debated and many interesting papers have been written

Data Dissemination: Traditional printed publication often presents data in a way that makes them difficult to access for further study. Increasingly we find extensive information condensed into tables and lists. This information is valuable, in the sense that it displays the data used by the author. The problem, however, is that the data are not directly informative when read. Researchers wishing to use the data must generally convert them to electronic form from the publication, or go back to the primary sources, in order to use them in their own analyses. Resources are wasted on re-recording. Often publications take the form of 'black boxes' - the data are presented, but testing of conclusions or development of alternative interpretations based on these data often require a prohibitive amount of work. The consequence is generally a 'take-it-or-leave-it' attitude towards the publication's conclusions. A further consequence of this trend is that research increasingly becomes based on discussion of what others have said about data, rather than on the data themselves.

There is a further problem with the traditional printed publication. A book is a physical entity in itself, and any search system for information (e.g. indexes) is limited by its boundaries. To seek information on a specific issue means searching a number of books - there is nothing new to this. However, as the number of books grows, the search for information becomes increasingly tedious, and researchers become more and more restricted and specialised in their topics and in their outlook.

The problems we currently face with traditional publication create a pressing need for investigation into, and development of, new forms of inexpensive and easily searchable publication of data, based on electronic media. Expensive printed books could then be reserved for argued discussions and syntheses.

Summary of content

The purpose of the project was to improve the way archaeologists carry out research, in other words to improve archaeological methodology. This can be summarised in a number of statements:

- replace the current segmented methodology with an integrated methodology;
- improve specific archaeological methods within various research areas;
- improve the efficiency with which archaeological data are analysed and disseminated;
- make archaeological information available to more people faster, more easily and more cheaply;
- create a framework that encourages pluralism;
- establish an integrated IT based work environment for archaeology;

- create the necessary knowledge and tools to manage archaeological information as digitally stored data in an integrated environment;
- lead archaeology into the information age.

The proposal singled out four topics of research using the following headings: *handling of excavation data*; *regional studies*; *integrated archaeological workbench*; *dissemination of archaeological information*.

Handling of excavation data: The archaeological issues to be addressed here were: the structure of excavation data; data recording principles; methods of post excavation analysis of excavated data. Within the IT sphere this implies: creation of an object-oriented relational database system reflecting the structure of excavation data; establishing a recording system directly from the field to the database; creating applications that will perform the specified post excavation analyses.

Regional studies: The archaeological issues were: the structure of regional data; relating regional (archaeological) data to other types of geographically defined data (environmental, topographical, historical, etc.); comparing and interpreting distribution patterns. Within the IT sphere this implies: creation of a relational database system reflecting the structure of regional data; implementation of a geographical information system; development of archaeologically appropriate analytical applications in the geographical information system.

Integrated archaeological workbench: The variety of analytical methods that may be applied to archaeological data on the one hand, and the physical dispersion of data on the other, calls for integration. In the archaeological sphere we need to deal with basic methods for an analytical approach to data. Within the IT sphere we need, apart from implementing various methods, to create a basic common (communication) structure for all implemented applications to ensure that data can be implicitly passed between applications. Furthermore we need to concern ourselves with the introduction of distributed data and distributed processing solutions.

Dissemination of archaeological information: This topic addresses the dissemination of archaeological information in general. However, the project would have concentrated on the publication of archaeological excavations. Within the IT sphere this implies: the structuring and implementation of publication data on electronic media and experiments with a distributed knowledge-base system.

Theoretical issues

A model for archaeological research

The archaeological record is a static contemporary phenomenon. This is of course a trivial statement. Less trivial is the assertion that we cannot draw conclusions from archaeological data about past reality. It may appear that we do so, and we may indeed believe that this is what we do. I will argue, however, that there is no way we can infer, applying rules of logic, the past from the archaeological record. Except in trivial cases of physical determinants (technology, etc.) there is no set of rules as suggested by Binford (1977, 1983) that allows us to conclude that a given structure A observed in the data implies that a certain process B operated in the past.

The reason is simple. Any structure in the archaeological record can be the result of numerous relationships in the past. We have no means of knowing what specific combinations caused the pattern we observe. There is a one way logic that goes from the past to the present. If we knew the past reality, and if we knew all the successive transformations creating the archaeological record, then we could predict this record in great detail, but we cannot go the other way.

The same objection applies to Hodder's claim (1986) that we can read the archaeological record in the same way as we read a text. It is acceptable, I agree, to view material culture as meaningfully constituted, and readable as a text. We cannot, however, automatically transfer this property to material culture as preserved in the archaeological record. It is one thing to read the material culture when actively displayed in its live context. It is quite another matter to read the fragments of material culture preserved in the archaeological record, not knowing the transformations that occurred between the live and the dead context. Some elements of meaning may of course be better preserved and more 'readable' than others, but even if, in a specific case, we believe that we are able to read the data correctly, we still have a problem of making our readings plausible to others. Indeed, it makes little difference if you read or dream - the problem remains the same: how do you verify that you have correctly retraced the transformation?

Accepting that the logic goes from past to present, we may claim that, given a model of what the past

was like and a fair idea of the type of transformations applying, then we will know something of what this implies for the archaeological record. Starting from our model of the past, we can look for discrepancies between what we expect and what we find. In a dialectic process we can continuously change our perceptions of what the past was like, and of the data we create from the archaeological record, to maintain a concordance between expectations and observations. Although we will never know if we are correct in our expectations of the past, we will have to believe that gradually we can build up an increasingly correct picture of the past. The results we obtain are, of course, very much dependant on the paradigm from which we work. However, given a paradigm relevant to ourselves and our society, we will learn something valuable and 'true' about the past.

Within current reality archaeologists work at two levels. One is the matter of fact level of archaeological data. The other is the theoretical level of the mental construction we call the past. Archaeologists have a tendency to concern themselves with either one or the other of these levels. The essence of archaeology, however, is to create a concordance between them to minimise the conflict between the structure of our data and the structure of our conceptions of the past.

A very similar view of the archaeological research process has been stated by Read (1990). He proposed that archaeological research progresses through a comparison of two sets of models. One set - the theoretical models ($Model_T$) based on our theoretical standing - specifies what past reality was like. The other set - the data models ($Model_D$) - outlines structures of archaeological data. We compare the structure and consequences of our $Model_T$ with the structure and content of our $Model_D$.

Read presented this schema related to strict formalised mathematical modelling, and a processual theoretical view. The schema, however, is valid for informal modelling as well, and the theoretical view need not be processual either. Indeed, I would argue that archaeological research follows this twin modelling pattern in general, with a dialectic two-way interaction. At the low level end of the scale, theoretical models are often of a rather axiomatic nature. At the high end interaction is much more bilateral. Not only do we try to shape a $Model_T$ to fit a $Model_D$, we often try to manipulate data as well, to create a $Model_D$ that either fits or does not fit a specific $Model_T$, depending on whether we seek to promote this model or bring it into discredit.

How we arrive at our $Model_T$ is immaterial, but it is important that we are explicit when we formulate it, and explicit about its assumed consequences for a $Model_D$. The concordance check between the two sets of models is the only way in which we can exercise control and gain some confidence in our reconstructions of the past.

Archaeological data

There is a distinct difference between the archaeological record and archaeological data. The archaeological record is an unbounded and hence undefined part of current reality. This record is real in the sense that all reality is real, but it is certainly not objective, as there are an indefinite number of ways to view it. Only by imposing our personal views on this record can we make it usable for research.

Archaeological data may be viewed as the outcome of imposing our personal views on the archaeological record. The data are created through a categorisation carried out by individual archaeologists. They cannot be considered real, since they are defined abstractions based on preconceived notions of what is culturally relevant. However, despite the subjective act of creation, archaeological data can in some sense be considered objective. Once defined they become facts that may be rejected with reference to the 'history' of their creation, but not changed or redefined. They attain an existence as 'objective' data, where their original subjective birth is of little consequence. This is especially true with those data that becomes historical as a result of the destructive nature of excavations.

The archaeological record is, of course, a conceptual unity. However, when archaeologists work this record, they have to split the data they create into what they bring home with labels attached, and what they destroy on site. The latter data mostly consist of entities we usually call layers, and of information on relations between layers and relations between objects and layers. As these data become historical on recording, we have no means of changing them later. We may re-evaluate them against what we bring home from the excavation, and against what we learn from other excavations. As a result of our re-evaluation we may distrust and discard the data, but that is all we can do.

Movable objects can be anything we bring home, including complete 'chunks' of the archaeological record (and not just artefacts in the classical sense). Movable objects constitute a continuously available part of the archaeological record that allows extrac-

tion of new sets of archaeological data at any time. However, the exponentially growing number of objects, and their curation in dispersed locations, makes this availability a myth. Most of the time we will have to accept data categorised from the objects by other people, limiting the practical difference from data recorded in the field. Andresen & Madsen (1992) provide a detailed discussion of the various types of data from archaeological excavations.

It might be tempting to view archaeological data as being only what we categorise from context information and movable objects. This, however, would not cover what we actually use as data in our studies. If we return to the $Model_T$ and $Model_D$ concepts, we find that, what on one level is a $Model_T$, may on the next level be used to categorise information, and thus create new sets of data. These will be conditioned by the $Model_T$, but at the same time they constitute part of a $Model_D$. To take an example: the $Model_T$ - a posthole - may, when applied to a set of described layers, create a set of data which are considered postholes. These may enter a $Model_D$ describing the regularity with which they are positioned, leading to a comparison with another $Model_T$ - a house - and so forth.

We may thus distinguish between primary and secondary data. Primary data are those that we create by categorising the archaeological record directly, while secondary data are created by applying a $Model_T$ to existing data, primary or secondary. It should also be noted that the categorisation of primary data involves $Model_T$s of what constitute archaeological data, and not least of what constitute relevant archaeological data. In this respect there is no difference between primary and secondary data.

The difference between primary and secondary data is rarely heeded in archaeology, but it is a most important distinction if we are to attain more formal ways of storing and working with archaeological data in an IT environment. Thus primary data are categorised directly from the archaeological record. They are data we cannot change (context information) or can change only with difficulty (object information). Secondary data, on the other hand, are derived data that depend upon an explicit $Model_T$ within our explanatory framework. They can be discarded or changed by discrediting the relevance of this model in the actual context.

In a study dealing with data structures for excavation recording, Andresen & Madsen (1992) operated with three basic entities of archaeological information - layers, objects and constructs. Layers and objects constitute the primary data and, together with tables holding information on relations of layers with layers, objects with objects, and objects with layers, we could claim to represent all context information (the complex problem of object description was not considered). We viewed the construct entity as a means of interpretation of layers and objects. We argued that it is possible, through tables linking constructs with layers, constructs with objects and constructs with constructs, to build up a complete interpretative model of any excavation. The constructs and their linking tables are equivalent to secondary data, and they may clearly be viewed as entities that represent the consequences of $Model_T$s applied to the data from an excavation. This scheme was used as the basis for the integrated system outlined in the project description.

Archaeological methodology

In the same way that we understand our data through the theoretical dimension, we view and sense our data through the methodological dimension. I consider a *method* to be a systematic procedure through which we create, represent, transform or compare data. This also includes the comparison between the consequences of our $Model_T$, the 'expected' data, and the structure of our $Model_D$, the actual data. Archaeological methodology thus consists of a set of acknowledged systematic procedures for working with data. Acknowledged means that there should be a consensus within a community of archaeologists that the procedures are applicable. Systematic means that the procedures must be explicit and reproducible.

It is of course not possible to delimit precisely what constitutes archaeological methodology, nor can we expect that archaeological methodology in Scandinavia is exactly the same as in America. Within any one particular tradition, however, there is a fairly consistent assessment of what constitutes a valid methodology.

On the other hand, the demand that the methodology should be explicit and reproducible can and, I will argue, must be set up as a universal criterion. We cannot accept a methodology that is not made explicit and reproducible.

There are three areas where archaeological methodology comes into play. One is where data are initially categorised from the archaeological record. Although this categorisation is a subjective act of creation, it is far from unsystematic. On the contrary, much effort is placed in the definition of systematic procedures for the optimal acquisition of data. These procedures create standardisation and

quality control of the data acquired, but at the same time they also act as a conservative element in research, making it difficult for new types of data to enter the scene and become accepted as valid and relevant.

A second major methodological area is the creation of data models. This is the area where we look for pattern and structure in our data to test for concordance with our theoretical models. It is the general area of analytical methods in archaeology ranging from typologies to complex quantitative methods. Many of the more complex methods available are beyond the average archaeologist - they either ignore them, use them in a helpless fashion, or misuse them. This is unfortunate, as work with data models is essential for good results, and it is certainly one of the areas where marked improvements are needed.

The third area is concerned with control of concordance between theoretical models and data models. This is an essential methodological area, including a range of comparisons for concordance from very simple to very complex. As a low level example we may take the rule of stratigraphic succession. This is a strong rule, resulting in a clear and simple theoretical model that few archaeologists feel tempted to reject. As a well defined formal method to check for concordance with a data model we might use a directed graph (Ryan 1988:329 ff.). In this case, lack of concordance will not normally reject the theoretical model, but will make us look at the data model for misrepresentations, observational errors etc.

As an example from the 'high end' of the scale we may take Hodder's 'Domus/Agrios' model (1990). How do we perform a concordance check between this type of theoretical model and the data? We are certainly beyond a simple mechanical check in this case, and one may be pessimistic about our ability to formalise in this area. However, we must maintain as a minimal demand that it should be possible to build up arguments for concordance in a clear language. What are the consequences of the model for the data, and do the data we find fit our expectations? If we cannot do this, we have to discard the theoretical model as beyond the scope of our research.

An integrated IT based system for archaeological research

There is a marked need for an integrated IT-based system for archaeological research. Archaeological data are multiplying everywhere, and it is increasingly difficult for archaeologists to cope. Many respond to this by specialisation - they concern themselves with an ever diminishing part of the archaeological record and their outlook becomes increasingly narrow, while their $Model_T$s become increasingly monomaniac and stereotyped. Others let go of the connection to data - their $Model_T$s become increasingly vivid and sweeping, while their $Model_D$s rest on a thinner and thinner foundation, often being no more than a pretence of real data. If we are to stop this development we must concentrate much more on explicit and controlled linking of $Model_T$ and $Model_D$. Furthermore we must extend the factual basis of our $Model_D$s.

The main aim of the proposed research project was to create an integrated IT-based system that would have made it possible to counter these problems and shortcomings in archaeological research. I will argue that such a system would work as intended, provided the following requirements are met:

- The system must be of a general nature, able to cope with any type of excavation, yet tailorable to meet any specific recording demands;
- The system must hold not only the primary data acquired, but also the complete interpretative structure, including whatever secondary data is created along the way;
- The primary data must be clearly separable from the interpretative structure, including the secondary data, yet the two sets of information should be fully interlinked;
- Methods for pattern searching and concordance checking should be fully integrated with the database, so that actions can be invoked from the data side rather than from the application side. Further, they should be applicable across primary and secondary data in any combination;
- The system should be aimed at utilising future possibilities of distributed data and distributed data processing.

The first three points deal with the structure of the system. The first ensures that individual recording practices can be incorporated in the same system using the construct part, while the primary data are not influenced. The second ensures that the system will be fully dynamic, as everything created during interpretation and analysis becomes included and available for further analyses. The key feature is the separation of primary data from the interpretative structure while incorporating the complete interpretative structure and all derived data in the system.

The integration of methods that can be invoked from the data side means that a number of standard analytical and checking routines can be run automatically, whenever they apply to the data being handled. A system where methods are integrated with data in this way has enormous potential, offering improved analytical features and a much stricter control of concordance between $Model_T$ and $Model_D$ structures.

Future use in a distributed data and distributed data processing environment is perhaps the one feature that may begin to break down increasing specialisation and/or insulation of researchers from data. The growing availability of data at ones fingertips, and the possibility of linking various data sources using a common basic structure, will eventually create much sounder $Model_D$s to compare with our $Model_T$s.

---

1 An English version of the research proposal *Archaeology in the Information Age* is available from the author on request.