Torsten Madsen

The IDEA of data standards

Paper presented at the conference Computer Applications and Quantitative Methods in Archaeology CAA97 held at University of Birmingham, April 1997.

Appear as read with minor changes to grammar and language
Torsten Madsen 2-10-2003

Introduction

A couple of years ago a questionnaire was sent to the fifty or so archaeological institutions in Denmark. They were asked to describe how they recorded excavations. Half of the answers contained more or less detailed descriptions of the recording principles used, none of which were identical. The other half of the answers just noted that they followed the standard, forgetting to specify which standard. Judging from those who took the trouble to answer in detail, it must surely have been their own.

The word standard probably makes most of us cringe. We immediately think of something big, nasty and bureaucratic, whether the Royal Commission or even worse, *The* Commission in Brussels. Standards, however, are too dangerous and devastating to be left to commissions. It is something that should concern us all, and not least should the issues to be standardised concern us.

The following is as a small, but seriously meant contribution to the discussion of data standards.

Data Standards

With the term *data standards* in this paper, I refer to two different areas of standardisation. One is standards of structure, classification and description of data. This I will refer to as *standards of content*. The other is design standards for databases. This I will refer to as *standards of design*. It is seldom to see a clear distinction between these two types of standards, because a very direct relationship between them is normally assumed, where content determines design. They are, however, two completely different things. It is imperative that a distinction is made, and that we value them differently. While attempts to develop standards of design have to be applauded, attempts to create standards of content should be abolished.

Starting with standards of content, why are they to be abolished? This question brings us deep into a long-standing debate in archaeology on the nature and role of classifications. For more than 25 years, intensive discussions have increasingly lead to the conclusion that standard classifications are neither possible nor desirable. The debate came into focus with a well-known paper written by Hill and Evans in 1972 called “A model for classification and typology”. In this paper, it is pointed out that classifications always depend on questions asked. As more and more questions are posed, increasingly diverse classifications are needed to elucidate any given set of data.

With standards of design, it is very different. The recording systems we design are merely containers for the data we wish to record. Ideally, the design of a system should be independent of what it will actually accommodate. This is unfortunately never the case in practice. The rule is that the intended content of a database decides its design, and hence different databases become incompatible, due to the differences in design.

An important question is, whether this incompatibility of design is something forced by the database technology we use, or if it is forced by the practice of design. Ten or fifteen years ago the database technology was certainly a substantial limiting factor,

but this is no longer so. Today it is more the practice of design than limitations in database technology that leads to incompatibility between different databases.

The standard approach of design is first to analyse the working practice of the particular part of the reality that the database should cover, and subsequently implement this practice into the table structure. This approach of design may work fine for business applications, but it is not a satisfactory solution for research applications, and I consider all archaeological databases to be just that.

When we deal with research, we have to start with a conceptualisation of the nature of archaeological data as such. Our design should be focused on this conceptualisation, and not on a specific practice associated with the handling of data.

Starting from here the goal is then to create a design that is sufficiently generalised to accommodate widely differing structures, classifications and descriptions of data. We should seek a standard of database design, with no prior assumptions of content, beyond those set by our conceptualisation of the nature of data.

Let us take a closer look at this endeavour, and try to isolate what should be the demands for such a design. We start with the research process itself. Increasingly, it is realised that the research process can be viewed as a dialectic process between theoretical modelling on the one hand, where views of the past are created, and data modelling on the other hand, where observational data are formed into meaningful structures. The core of the process is the interaction between the two forms of modelling, with a continuous dialectic flow between them. Classification and description are an integrated part of the data modelling process, and is thus an active research tool in its own right. It is inherent to the process that we vary and change our data modelling to confront our theoretical models with new and different data structures. The moment we subside and accept a particular classification or description to be *the classification and description* the research process stop. Thus, we need classifications and descriptions to change continuously. Further, we need alternative, competing or supplementary classifications and descriptions to exist side by side, applied simultaneously to the same set of data.

These observations lead us to the first two demands for a standard design:

1. Any instance of recorded data should be attributable to an indefinite number of classes from different classification systems, but not, of course, to alternative classes within the same classification system.
2. It should be possible to add, delete and alter classes and classification systems continuously, respecting of course the integrity of data already recorded.

If we turn to classifications, it is obvious that it would be unacceptable to constrain these to form a flat structure. That is, we cannot have all classes parallel to each other, available at the same level. Most current classification systems are hierarchical, but structures that are even more complex may occur. The way that classes interrelate in these structures is part of the meaning assigned to the classes. It is thus important that not only should the classification structure be implemented in the database, it should also be fully operational (i.e. you should be able to make searches that are dependent

on a classification structure, say: give me all instances of class A including all of its subclasses). This leads us to our third demand for a standard design:

3. It should be possible to apply network structures to classes, and use the defined structures to operate on recorded data.

Archaeological data are to a high degree contextual. This means, that not only do data relate to each other, somehow. The way they relate is by itself informative and important. Thus not only should we be able to establish links between different instances of data, we should also be able to qualify and quantify the nature of the relationships that these links represent. If you want to study an eminent example of what this implies, I would advise you to take a look at the paper of Costis Dallas in CAA 91 termed "Relational description, similarity and classification of complex archaeological entities". Our fourth demand for a standard design then is:

4. It should be possible to categorise, qualify and quantify all relationships between instances of data.

For each class, it should be possible to attribute an indefinite number of variables. These variables should cover the standard range of scales with their different variations. Thus we should have: nominal scale variables with or without multiple choice, as well as differentiation between an alternative and a dichotomy structure; ordinal scale variables with rank order recorded; ratio scale variables with values recorded as points or as intervals on the measurement scale, as well as a recording of what the scale is named.

In addition to this, there should be some sort of control of how variables are associated with classes bound together in a classification system. More specifically, a variable associated with a class must always be associated with all subclasses of that class as well. What we are looking for is a sort of inheritance among classes of a classification system with respect to all associated variables. This leads us to two further demands for a design standard:

5. It should be possible to attribute any class with an indefinite number of standard type variables.
6. Inheritance of variables should be enforced to the degree, where it is ensured that all variables of a class are always available in all of its sub-classes.

Creating a system, where the structure of content is independent of the design means that there is no way we can know in advance what will be stored in the database. Thus there is no idea in having codebooks with definitions of contents. Instead the definitions of content must be stored in the database along with the data. In this way the database will always be "self explanatory" with respect to its content. This leads to our sevens and final demand for a design standard:

7. All definitions of classifications and descriptions should be stored in the database to make recorded data understandable without the use of external information.

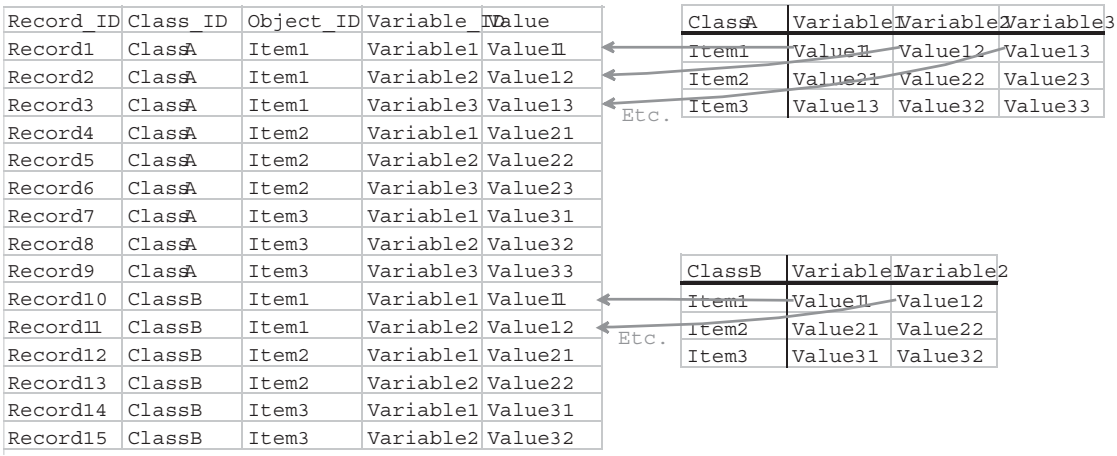
From idea to IDEA

Three years ago, in 1994, Jens Andresen and I began a joint development project called the "Integrated Database for Excavation Analysis" or IDEA for short. Initially our design was based on a straightforward analysis of excavation data, leading to the isolation of five basic entities of information from excavations. These became the core of the system, and to each of them, different areas of specific information were attached. One of these areas of information, naturally, covered classification and description systems. As work progressed, our efforts resulted in a design for classification and description, which fulfilled six of the seven above-mentioned demands. The only demand not met so far is the inheritance of variables.

Let us briefly take a look at the basic structure of our solution.

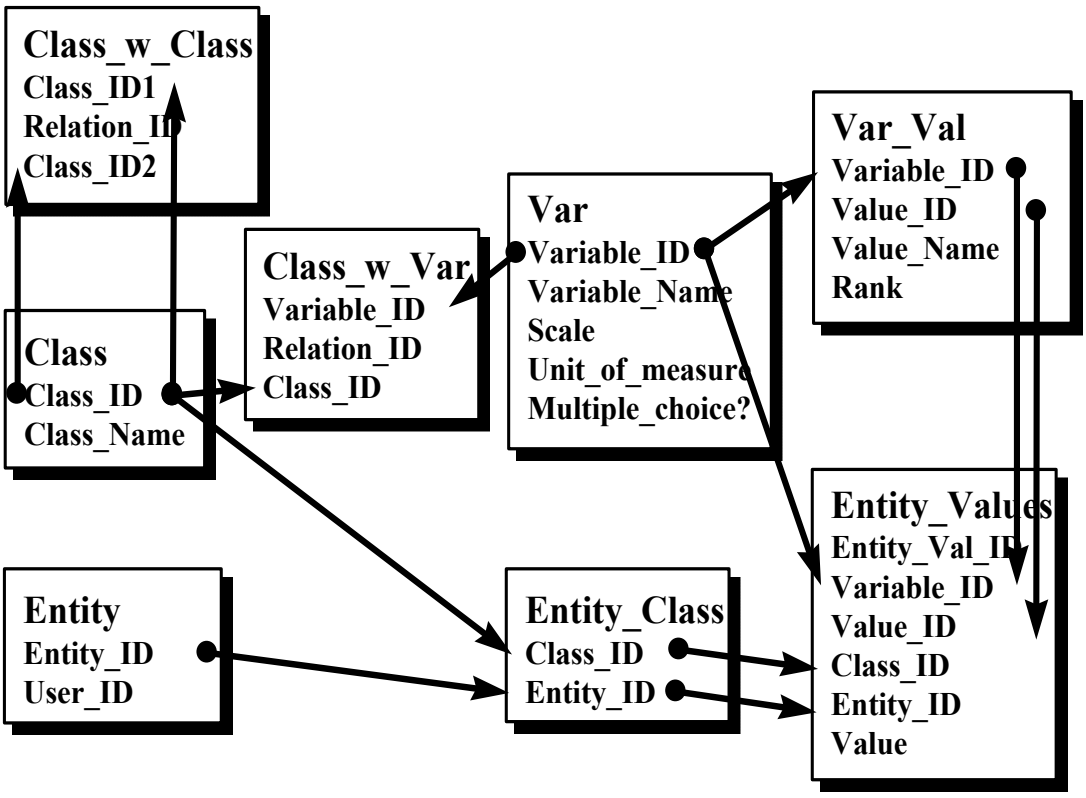
Usually, we record the description of a set of instances of a class in terms of a table with the instances in the rows, and the variables of the class in the columns. Consequently, instances of different classes, with different sets of describing variables cannot be placed in the same table. In order to make any instance fit the same structure no matter what class and kind, and what number of descriptive variables it posses, all instances must be described in one table and one table only.

To do this you may simply take every cell in the descriptive tables of the different classes, and transform them into a record of their own. This record holds all the information - *Class_ID*, *Object_ID*, *Variable_ID* - of the *Value* entry.



Record_ID	Class_ID	Object_ID	Variable_ID	Value
Record1	ClassA	Item1	Variable1	Value11
Record2	ClassA	Item1	Variable2	Value12
Record3	ClassA	Item1	Variable3	Value13
Record4	ClassA	Item2	Variable1	Value21
Record5	ClassA	Item2	Variable2	Value22
Record6	ClassA	Item2	Variable3	Value23
Record7	ClassA	Item3	Variable1	Value31
Record8	ClassA	Item3	Variable2	Value32
Record9	ClassA	Item3	Variable3	Value33
Record10	ClassB	Item1	Variable1	Value11
Record11	ClassB	Item1	Variable2	Value12
Record12	ClassB	Item2	Variable1	Value21
Record13	ClassB	Item2	Variable2	Value22
Record14	ClassB	Item3	Variable1	Value31
Record15	ClassB	Item3	Variable2	Value32

In terms of a relational design, our solution appears as shown in the illustration below. One of the demands for the design is that the complete classification and description system should be stored in the database. This is done through tables *Class*, *Var* and *Var_Val,* while link tables *Class_w_Class* and *Class_w_Var* establish the necessary structuring of the classification system. The actual data are stored in table *Entity_Values*, which is linked through table *Entity_Class* to table *Entity*, where the identification number of the instance recorded is stored.



This should make it obvious, why we can change content without changing physical structure. In contrast to traditional database design, all class entries relating to instances are kept in one field, all variable entries relating to classes are kept in one field, and all value entries relating to variables are kept in one field. The only "content", so to speak, stemming from our conceptualisation of data, is the assertion that an instance of data can be described in terms of a class, a variable and a value. For each value, of each variable of each class of each instance there will be a record in the database, and consequently for each instance there may be many records, and not just one.

During the development of the IDEA to its current version 1.1 we began to realise some intriguing facts about our design. Let me mention but two.

- Several of the areas of description attached to the five basic entities are conspicuously alike in design even if they deal with conceptually very different items of information.
- The five basic entities that we initially separated can be considered to be classes lying just one level above the top levels of the classifications that the user may implement, and indeed the five entities are just one level below an imaginary top level consisting of everything.

The question from a design point of view then is, why are we predefining basic entities, and why are we separating different areas of description to attach to the entities. Couldn't we combine everything into one single structure, where the user is responsible for the definition of all classes including what we analytically would consider basic entities, and where all description areas are user defined, kept within a shared common structure?

The answer is - *yes* - it is possible. It wont be easy with the tools we have right now, but it can be done, even within our current DBMS, which is Microsoft Access.

One day there may appear a version 2 of IDEA. If so, it will not be anything resembling version 1. Although it will have the same functionality as version 1 and probably a lot more, its user interface will appear very different, and its design will be very different. Version 1 has close to 100 tables, and is forever bound to our five predefined entities. Our first experiments with a new design have shown us that we will end up with about 20 tables, even with an extended functionality, and with the ability to handle an indefinite number of user defined entities. Further, it wont be predetermined to deal with excavation recordings. Although this area will still be in our minds during development, the system can be user customised into a recording system for something or perhaps anything else.

There is no time here to demonstrate the basics of the design we are currently experimenting with, but we will try to include it in the printed version of this paper. Instead, let me conclude on the question of data standards. Our work with IDEA has shown us that it is indeed possible to create design standards that will accommodate widely differing kinds and structures of data. We find that seeking generalised standards of design, at the same time will free us from the straitjacket of standards of content. Databases then can become important research tools in their own right, and not just containers for mass storage of dead information.

My only problem and worry here at the 25th anniversary of CAA, formed by my experiences with computing in archaeology over the last 15 years, is, how are we going to get the message across to a world of archaeology, that doesn't care.