DATA MODELING AND RELATIONAL DATABASE DESIGN IN ARCHAEOLOGY

by
Manuella Kadar

Abstract. Data from archaeological excavation is suitable for computerization although they bring challenges typical of working in non-scientific subjective areas. We present some issues with regard data modeling in the specific field of archaeology.

INTRODUCTION

Archaeological projects provides large quantities of written documents, notes, and forms, as well as drawings (plans, sections, and sketches) and photographic images of the ancient sites, architecture, and artifacts that are recovered during survey and excavation. Effectively utilizing all this material in the pursuit of research goals has always been a major challenge. The sheer amount of data which must be processed and evaluated by project members quite often necessitates the adoption of new tools and strategies for interpretation and analysis.

An archaeological site comprises a complex three dimensional matrix of deposits, cuts and interfaces which can combine great physical and chronological depth, continuity, discontinuity and stasis. Huge variations in deposit formation and deformation processes, deposit make-up and deposit visibility make the recording and interpretation of these deposits and their interrelationship both with one another and with the materials they contain, a complex and complicate process.

The emerging technologies of computer-aided drafting (CAD) and surveying instrumentation, remote sensing/satellite imaging, digital scanning, Global Positioning Systems (GPS), photogrammetric mapping, and digital/video (multimedia) imaging are of critical importance for modeling archaeological data, and Geographic Information Systems (GIS) and Exploratory Data Analysis (EDA) software are of critical importance for archaeological analysis.

Data from archaeological excavation is suitable for computerization although they bring challenges typical of working in non-scientific subjective areas. Meaning and significance within data are established on-site and afterwards by a heuristic process of discussion and contestation, a process at odds with the rigorous demands of database design.

MODELING ARCHAEOLOGICAL DATA

A common and powerful method for organizing data for computerization is the relational data model. Relational databases have a very well-known and proven underlying mathematical theory, a simple one (the set theory) that makes possible automatic query optimization, schema generation from high-level models and many other features that are now vital for mission-critical Information Systems development and operations.

When designing database one has two consider two categories: the first one is **data warehouse** that contains data gathered to perform a relatively limited role only in a particular project. The database is intended to provide the researcher with a particular set of data, but have no particular function or role at the conclusion of the project.

The second category comprises general purpose, resource databases or **operational databases**. A good example of a resource database are county archaeological sites and monuments records, or national monuments records. These databases are not project specific but are intended to be of use to a wide variety of users. Resource databases usually attempt to be comprehensive within their "domain of discourse", are maintained and updated, and are made available to interested parties. As these databases attempt to be comprehensive in order to accommodate unpredicted enquiries and research, they include a wide variety of data which in turn requires a complex "data structure", or way of storing the information. This type of "data structure", however, provides great power and flexibility both for the retrieval and for the handling of the data, but also for future expansion of the database to include other information and materials.

Conceptual data modeling is a very important phase of an archaeological application. This is the process of formally describing the application's structural and behavioral properties for purposes of understanding what is required from the application and communication between users and developers. Whether one is designing a data warehouse or an operational database for multiple uses, the basic principles of data modeling and database design prevail.

Entity-relationship (ER) diagrams are the essential part of the overall planning and maintenance of the information resources. Recently, new techniques of dimensional data modeling (star schema) and normalized dimension tables (snowflake schemas) have been introduced to support data warehouse design (Owen, 1998).

We will discuss here the traditional modeling theory which involves three steps: **analysis, design, and implementation.**

Analysis is the process of creating a conceptual data model independent of the target database technology. The typical result is an ER model.

Design is the process of creating a logical data model. This step is already dependent on the target technology (relational, hierarchical, or network), but not on the specific database implementation.

Implementation is the process of creating a physical model or schema for one specific database system, the result is an optimized physical design.

Data modeling tools that maintain a separation between conceptual and logical models must provide logic to translate ER constructs into relational tables and generate those tables into a relational model. The results of the forward transformation can be a pure logical model that does not include tables for a specific relational database management system (RDBMS).

Physical database design starts from a given relational model which is the definition of a set of tables and their respective columns. The objective of physical database design is to fulfill the performance requirements of a set of applications by optimizing the use of the Database Management Systems (Năstase et al., 1999).

Key areas include: optimizing the index configuration, data placement, and storage allocation.

One of the greatest impacts on the application's performance is the selection of indexes. For operational databases an index that improves data retrieval performance may degrade performance for all kinds of updates, because the maintenance cost for the index has to be paid for each update. Since there is not a lot of update activity in a data warehouse, the number of indexes is not as much of a concern. Selecting the optimal indexes for a given set of tables (an index configuration) is a non-trivial problem that requires trade-offs between the different kinds of database operations (retrieval operations, update transactions, and utilities). Given an index, one must determine its properties, such as the column ordering for multi-column indexes, and whether or not the index should be clustering, partitioning, or ordering.

While index selection is the most important and complex problem in physical database design, data placement and space optimization are important as well. One must allocate the optimal table space to a given table, and choose whether a simple, segmented, or partitioned table space is most appropriate. Allocate enough space to maintain the clustering properties, without wasting too much space. Index selection, data placement, and space optimization assume that the structure of the relational schema is stable. However, physical database design can also change the relational schema; you may need to split tables into different partitions to take advantage of concurrent operations, or merge tables in cases where different tables are frequently joined, so that performance improves if one merged table represents the stored join. Decisions can also be made on the tuning of the DBMS run-time parameters.

Most applications in archaeology use traditional ER modeling to create and maintain operational databases. These databases are typically very specialized and are designed to support very specific application requirements. Recent trend is to build data warehouses. These new mega-databases enable end users to access information based on data that was previously unavailable to them in a single place. For the data professional, the newest challenge is to design an optimized relational database that satisfies a much different set of requirements.

We will discuss some aspects regarding modeling archaeological sites and archaeological collections.

When modeling an archaeological site the basic feature is that many artifacts can be found in one excavated unit (context). Other details are also important, such as the size and appearance of an excavated unit, and its stratigraphic relationship with other contexts.

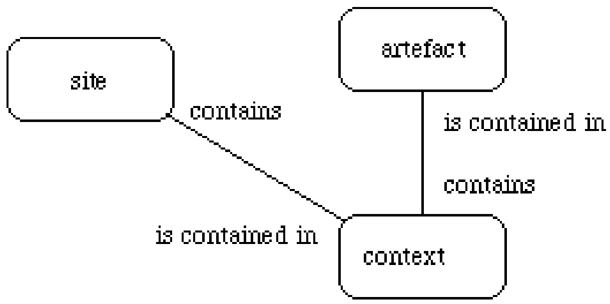

Figure1. Relationships

The main interest of the database is the relationship between excavated places - the contexts - and finds.

The main subjects are the excavated places - the site and the contexts - with the artifacts found there.

The name of the entity is given in the singular.

One context has many artifacts; each artifact is defined as belonging to one Context

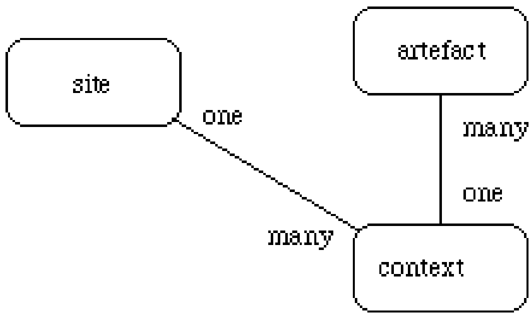

Figure 2. Degree of relationships

These diagrams do not give full expression to the complexities of the location of finds, the position of finds and the stratigraphy. This is to simplify the situation of

the one-to-many relationships. In a full implementation it is necessary to take account of the stratigraphical position of the contexts - and thus the finds.
The context and the stratigraphical position of the context are taken as entities, and account can thus be taken of the relationships between the artifacts and the features with which they were associated. The stratigraphical position itself is defined by the context and the feature.

Contexts contain both remains of features and artifacts. But the remains of features can be found in more than one context – so there is a many-to-many relationship between context and features. The stratigraphical position enables the many to many relationship between context and remains to be handled (any context can have more than one feature in it; any feature can be found in more than one context, for contexts occur at different levels). The attributes are shown for the entities shown in figure 3.

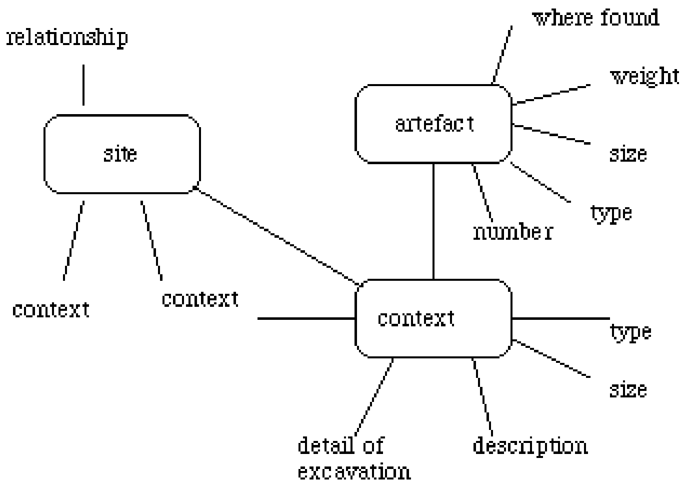

Figure3. Attributes of the entities

Attributes have to present all the details about the entity. They become important later in that the values of attributes are the data that is to be recorded in the database. Attributes represent the data that is to be kept about the entities.

The collection data which has to be modeled is essentially made up of three sets of related data:

- the collection details which are present for each collection;
- the provenance (collection site) data also present for each collection;
- the published details present for published collections.

A pottery collection may be made up of material from more than one site. A site may yield more than one collection, the same site could be re-excavated producing a second collection. Similarly a collection could be published in more than one source and a publication could include more than one collection. This results in there being many-to-many relationships between collections and sites and also between collections and publications. However, although a collection may be mentioned in more than one reference the published details for that collection must be constant because these are qualities intrinsic to the collection. There is therefore an optional one-to-one relationship between collections and their published details plus an optional many-to-many relationship between collections and references.

Within the three categories of data there are also multiple relationships between some categories their attributes. For example a collection may have been amassed as a result of a number of different methods of recovery, casual pick up, systematic field walking and excavation perhaps. Thus there is a one-to-many relationship between collections and recovery methods.

Within the data there are therefore a number of different relationships between the elements:

- optional relationships where data may or may not be present (published details, parish, etc.)
- one-to-many relationships
- many-to-many relationships

The relationships between the data items describing a collection can be formalized as a data model detailing these relationships.

Usually the model creation is followed by implementation of the data structure, interface design and testing. This flow chart emphasizes early planning, methodical completion of stages and testing as an integral part of the process. However, this strict linear process rarely occurs in practice. A more realistic representation of the development of an archaeological application is given in Fig.4.

Archaeological database design has to fulfill the following requirements:

- A high degree of data validation using referential integrity
- Control by the user of data validation
- A thorough representation of all possible relationships within the recording system

Deciding which changes to implement is not an easy task in order to create a robust design before implementation. However in some cases, changes to the data structure have to be made after implementation. Such changes are costly in developers time due to the high degree of referential integrity. A simple change to a key field can reverberate throughout the database structure many times.

Once the core of the database is created, a friendly interface is needed to give users access and the ability to input data as seen in figure 5. A central concern in creating the interface is to create a simple clean set of forms which give access to a complex set of tables. Data entry could only be achieved through a ”high level” knowledge of the database system and a more in depth knowledge of archaeological recording systems. This restricts the casual user but means that data entry becomes more of an intelligent process ideal for teaching basic skills required in creating archaeological archives.

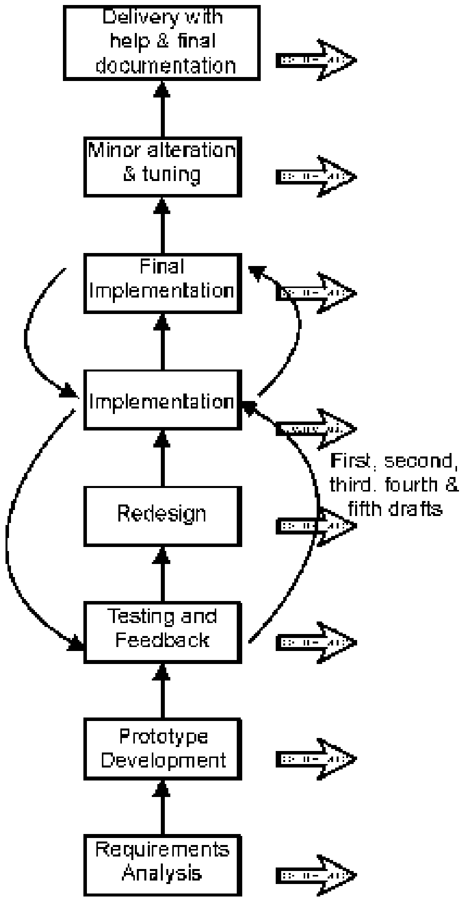

Figure 4. Flow chart for the development of an archaeological application

CONCLUSIONS

The design of relational databases raises many issues about the "real world" situation that is being modeled. Any small variations in the paper based recording system are soon exposed in a database that enforces data integrity. However, the data integrity is clearly what makes the database powerful as a research tool.

In the process of archaeological database design it is essential to have a complete understanding of the existing archaeological system before any attempt to create a more sophisticated model. First step is to understand the archaeological recording system and then to translate it into the structure of a relational database. Difficulties arise where the recording systems does not have the same level of cross-referencing of data as is required by the database system. The solution is either to design a database which is based on current practices and so which reflects some inconsistencies, thus making the database less powerful and potentially confusing, and introducing problems like redundancy of data, or to re-organize the present system, using the database as a catalyst for change in working practices in the field - a solution which is by no means simple in the "real world" of archaeologists.

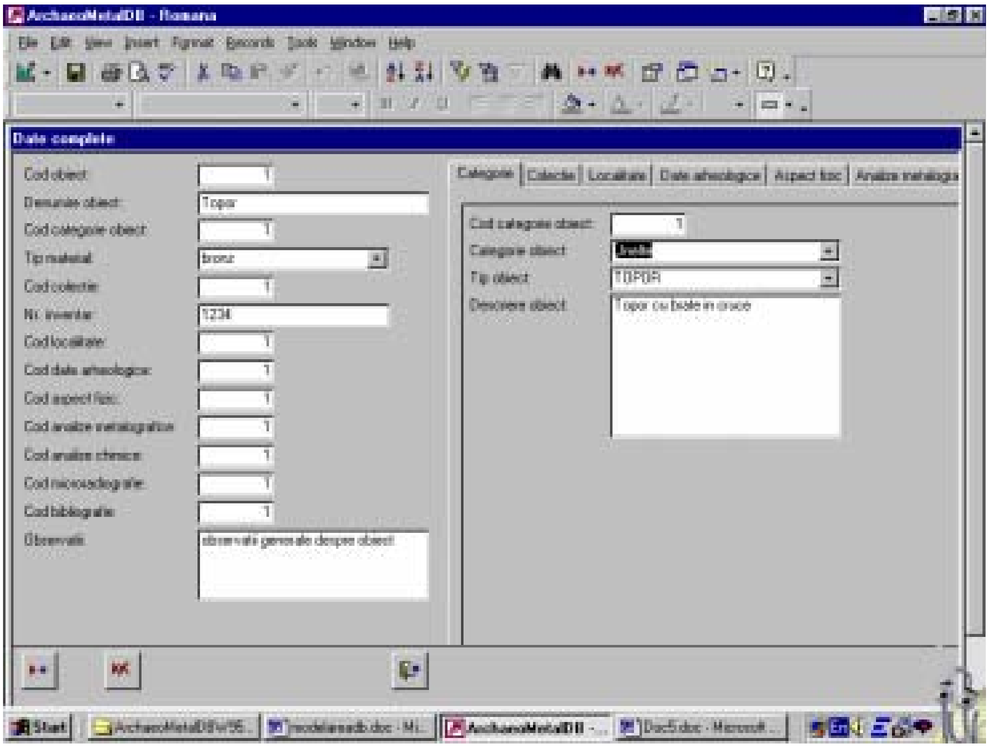

Figure 6. Data form for archeaometallurgical records

Bibliography

1. Năstase Pavel, Florin Mihai, Luana Cosăcescu, Bogdan Bărbulescu, Andrei Stanciu, Robert Aurelian Şova, Liana Covrig, *Baze de date Access 2000*, Ed. Teora, Bucureşti, 1999.
2. Owen Cathy, *Data Modeling and Relational Database Design* , The IDUG Solution Journal, march 1998-Vol.5, No.1.

Author:
Manuella Kadar- " 1 Decembrie 1918" University of Alba Iulia, Romania-
Department of Computer Sciences, e-mail: manuka@cristalsoft.ro