The Importance of Data Structuring for Extracting Relevant Information

By Gregory J. Brown
Department of Archaeological Research
Colonial Williamsburg Foundation

Presented at the Digital Archaeological Archive of Chesapeake's Slavery Steering Committee Workshop. The International Center for Jefferson Studies, Charlottesville, Virginia. October 6, 2000.

This talk was intended to raise several issues that will have to be considered in developing the structure of the archive, as well as to suggest a few ways in which the archive will be a real innovator in the design of so-called "digital archives" in other disciplines. It is intended simply to promote thought about the "big picture" in terms of computer applications such as this, which are becoming very popular in many disciplines as institutions such as the Mellon Foundation provide seed money to develop digital collections.

The Digital Archaeological Archive of Chesapeake Slavery is, I think, proceeding in the right way by trying to encourage standardization among scholars from several institutions and disciplines. My remarks come from the viewpoint of someone who has worked with databases at one institution, Colonial Williamsburg, for over 15 years, but also from one who has struggled with a variety of computer software products and with the integration of historical as well as archaeological data from a variety of sites. Though Colonial Williamsburg's involvement with Dominic Powlesland of the Heslerton Parish Project in north Yorkshire, England, we have also seen some of the promise of careful organization of large amounts of archaeological data and in particular Powlesland's concept of "geographic information management."

Data Standardization

While there is general agreement of the rather obvious concept that the data must be fairly uniform (as highlighted by Jillian's very useful charts of the coding differences even among those institutions that use Re:discovery), it is important to realize that absolute consensus is not necessary. The data in the archive will be "massaged" to fit the categories agreed upon by the steering committee, but this should not imply that later researchers would necessarily be constrained from searching the archive using their own analytic categories. It should be possible in the archive to build a "relational thesaurus" of some kind, simply maps the querier's categories (e.g., "REFINED WARES") to the appropriate fields and field values in the tables.

That said, it is clearly important to resolve the completeness issue-that is , it is important to try to create the appropriate fields and authority tables so that cataloguing can proceed fairly easily and most, if not all, the significant attributes can be recorded while the fragment is laid out. The most time-consuming part of this process would potentially be

the omission of an important category, in which case each fragment would need to be re-located, re-sorted, and laid out a second time.

Primary vs. Composite Data

To understand the possible uses of the archive, it is important to understand that there are essentially two kinds of data in the collection. data consists of the most elemental units: artifact fragments and contexts. Composite data are larger units, mostly analytical, that include master contexts, phases, sites, "objects," and assemblages.

Composite data are often, but not necessarily, organized into hierarchies: Powlesland's so-called "master contexts" are defined as sets of related contexts, "phases" are defined in Colonial Williamsburg's scheme as sets of master contexts (e.g., the related postholes in a fence line), etc. Fraser Neiman has suggested the following structure: contexts are collected into sets of master contexts, which are collected into sets of "meta-master contexts," which are collected into sets of phases, which are collected into individual sites, and so forth. This is accomplished by creating fields in the Context Table for master context, meta-master context, phase, etc.

It is important to realize, however, that these groupings-whether master context, meta-master context, phase, or even site-are analytical constructions of someone, most often the site's excavator. To make the archive as flexible as it can possibly be, it is necessary to consider other possible analytic groupings. Perhaps an analyst will be interested in a set consisting of only some of the features contained in a master context or meta-master context, or, worse yet, some of the features in one meta-master context joined with some of the features in another one. Perhaps some question will require the consideration of only the lower, undisturbed levels of a series of sub-floor pits, or even only the lower, undisturbed levels of sub-floor pits that have been screened through fine mesh.

How do we create such an "assemblage"? It is possible, in fact the strength of a computer-based archive, to extract this data using Structured Query Language, usually known as SQL. A SQL (often pronounced "sequel") string can be generated of truly enormous length that will extract exactly the right data. But it is my argument that the process can be understood fairly simply by remembering the difference between primary and composite data.

Making a Query Set

The goal of the process of "assemblage creation" is to make a set of primary data (that is, contexts) that can be used to filter out the irrelevant data from the archive (of course, the "irrelevant" data is only hidden from view and analysis, not physically removed!). In the simplest case, it should be possible to make a query set by just asking for a single master context or meta-master context (e.g., I want only master context M12 from the Rich Neck Slave Quarter site). The program would then go and build a small table with all of the context numbers belonging to that master context, a table that it will be able to use later

when the querier asks to filter out all the data (i.e., artifact records) that do not have an associated context number that matches the filter table.

If you wanted a more complex query (context numbers from two or more master contexts, two or more sites, etc.), the process would be similar. By listing entire sets of master contexts, individually listing context numbers, or writing an SQL statement, the querier should be able to create another table containing all (and only) those context numbers that he or she wants. Ideally the querier could then name and save this table for later queries, and even make it available for other queriers (as "Greg's Undisturbed Sub-Floor Pit Levels," for instance).

To permit this sort of flexibility, it is essential during the data-entry stage to keep the distinction between primary data and composite data. Ideally each context should be the smallest possible unit (the depositional fill of a pit, the arbitrary excavation square of a selected layer). Even fills screened separately should be segregated and given separate context numbers-that is, the wet-screened portion of a pit fill should receive a number separate from the quarter-inch screened portion. Hierarchical categories such as "feature" should be assigned in the Context Table (probably as "master context") to ease later aggregation, but it must be possible to create ad-hoc groupings as well.

Composite Data on the Artifact Level

There was many types of composite data on the artifact level, ranging from the creation of cross-mended "objects" through minimum vessel counts and, in the case of faunal remains, minimum number of individual (MNI) determinations. This data will need to be stored and managed within the archive, even if it is not generated as part of the data entry process.

It is important to note that the creation of this composite data is entirely dependent on the units of aggregation that are used (site, master context, etc.). Therefore it is essential to figure out a way to store the "filter lists" described above with the minimum vessel count, object catalog, etc. It is all too easy to create a set of analytic data, whether it be a minimum vessel count or a statistical measure of variability, based on a certain grouping of associated contexts, only to later lose or forget exactly which contexts were used in the grouping! I speak from frustratingly long experience.

To the extent that some significant attribute data is currently entered on the "composite" level, as it is at Colonial Williamsburg, it will also be important to transfer this data "down" to the primary level. For example, at Colonial Williamsburg we traditionally do not identify vessel form at the "inventory" or sherd level-instead this data is entered into our "object catalog" at the time that the sherds are mended. It will be necessary to transfer the final object form determination, therefore, back to the individual sherds in order to maintain data standardization and comparability with other collections.

Logistical Flow

Another significant issue-in some ways the most important of all in terms of establishing the DAACS data structure as a regional standard-is the issue of practicality. Can an institution actually catalog an on-going site using this format? It remains to be seen, of course, but in large measure this is a function of the ease and seamlessness of entry screens.

It is important to establish data entry screens that correspond with the logistical workflow. For example, it may be best to enter a "rough inventory" first, containing all the "major" attributes such as count and weight, and then put an individual artifact class aside to do a more detailed data entry at a later time, when all of the objects (buttons, for instance) are laid out in front of the cataloguer.

Likewise, it is relatively simple to program MS Access to, for example, automatically enter "dependent" information to save the cataloguer extra work. For example, a menu selection of "wrought nail" should be able to enter "NAIL" in the form field, "WROUGHT" in the technology field, and "IRON" in the material field. It is also possible to "filter" authority tables to, for example, only present a subset of the technology codes once the ware type has been entered.

It is in the decisions about how many attributes to record, what to measure, etc. that the potential use of this archive will be tested. The more streamlining that can be done on a pure data entry level, the likelier it is that the practicality of taking individual measurements on individual fragments will be generally accepted. Everyone pretty much agrees that we can learn a lot by looking closer at artifacts-the real question is whether it can be done on a site's typical limited budget.

Problems that Can Be Addressed

Finally, I want to suggest a couple of issues that, it seems to me, the archive will have to address, and in so doing will make a real contribution to the technology of developing digital archives in other disciplines. These are the questions of time and of "fuzzy" data.

Time. Time (or more precisely stratigraphic sequencing) is difficult to manage in a traditional GIS (geographic information system), the technology that many digital archives are turning to in order to manage spatial (geographic) data. In my limited experience, the problem appears to me to involve dealing with different "layers" of time. Most GIS applications deal with overlapping, single-point-in-time regions of space, and "buffer zone" analysis and the like depend on each point in space having one, and only one, value. How does one deal with a typical site, where features and layers are overlying each other? How do you permit a querier to select a single point in time, or is it possible to allow him or her to deal with several different points in time (several different components or phases, for example) simultaneously?

Fuzzy data. Related is the issue of fuzzy data. It is a real problem for querying efficiency if an attribute cannot be assigned a definite value, but in archaeology this is often the case. For example, a building footprint (a "master context") may be dated in a relative

sense by finding a reference to the purchase of bricks in 1725. Other evidence may indicate that the building was up by 1740 for sure, but that the 1725 date is somewhat questionable. So how can this be handled in the database? Do we need a field for "certainty," and if so how do we assign it (and who does the assigning)? Do we show some types of data in grey to indicate that we are not sure about them?

To be sure, other disciplines must face this problem. But it seems to be, knowing what we all know about archaeological data, that we face it in an extreme sense. If we can figure out a way to allow users to ask flexible but complex questions of data that is ambiguous, difficult to pin down in time if not in space, and fraught with the kinds of natural problems that result from being excavated by different institutions under different conditions, we will be making a contribution that the Mellon Foundation should find to be a real contribution to a number of fields.