Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, CAA2010
F. Contreras, M. Farjas and F.J. Melero (eds.)

Sustaining Database Semantics

Kintigh, K.W.
School of Human Evolution and Social Change, Arizona State University, Tempe, Arizona, United States
kintigh@asu.edu

This paper argues that the semantic content of digital databases is rarely adequately documented. It proposes a specification for what is necessary to document the semantics of a database and therefore sustain their analytical utility. It outlines an approach to documenting database semantics utilized by Digital Antiquity's repository, tDAR. It concludes with a discussion of how the metadata documentation used by tDAR can be used to facilitate the integration of data across databases employing different recording protocols.

Keywords: Databases, Metadata, Sustainability.

1. The Problem

Formal databases have for the last several decades served as the central mechanism for storing systematically collected observations about the archaeological record. With databases I include everything from sophisticated, integrated multi-table databases to rather less regimented spreadsheets that are often used in recording a single class of observations. These databases are sometimes used for direct data entry, are frequently used for assembling simple tabulations, and provide the data stores from which targeted sets of analytical data are extracted for more sophisticated quantitative analysis. Because of their systematic nature, these databases very often serve as the primary data stores that are expected to remain useable and to carry much of the burden of documenting an archaeological project into the future.

It is a simple fact that achieving sustainability imposes a strong set of necessary conditions on both the infrastructure—the financial, organizational, and technical components of the repository holding the data—and the metadata documenting the databases. Generally, failure in any of these conditions leads to the partial or complete loss of irreplaceable data (MICHENER *et al*., 1997). While all these considerations require serious attention, this paper focuses on the last, and perhaps most difficult of these challenges, that of maintaining the semantics—the meaning—of the observations repre-sented in the database. This includes documentation—table by table, column by column, and nominal value by nominal value—of the data contained in the database.

In the US, museums and other formal repositories of physical collections resulting from archaeological field work frequently take a rather passive approach to the collection of metadata concerning these critical information resources. These organizations are ordinarily rigorous in ensuring that they have the key elements of information needed to accession the artifact collections associated with a project (project name, investigator name, project sponsor, date, etc.). When it comes to the technical and semantic information needed to sustain a digital object, however, they are typically less thorough, at least in part due to a lack of on-staff technical expertise. It is my sense that, overwhelmingly, these institutions have not assumed the full burden (or taken on the challenge) of making sure that the necessary semantic metadata are obtained and maintained for the digital collections that they curate. My suspicion is that few of these repositories of physical collections go much farther than holding the database documentation (in paper or digital) that was provided them.

There are, however, a number of serious international digital publication and preservation efforts directed to archaeology (ADS, DANS, OpenContext), cultural heritage (DARIAH, Europeana), and the arts and humanities more broadly that have undertaken or are engaged in addressing this challenge. The University of York's Archaeology Data Service (ADS), for example, requests of their depositors information regarding the name, description, and data type of each field of each table of each database, along with maps of the relationships among the database entities (ARCHAEOLOGY DATA SERVICE, 2008).

In this paper, beyond outlining the problem, I will explore the question of what constitutes adequate semantic metadata for a database. I will then describe a project in

which we have attempted to prototype a framework that can approach that standard.

2. What Constitutes Adequate Semantic Metadata for a Database?

Basically, I argue that what is needed to truly document the semantic content of a database is sufficient information for an archaeologist not familiar with the specifics of a project to make sensible analytical use of the data. While this is not a clear-cut standard, in nearly all real world cases it will be more comprehensive than what most analysts, most of the time, would think to document and submit with a database. Indeed, we all know that projects sometimes don't make sensible analytical use of their own data, for a number of reasons including a lack of internal communication of key semantic information.

I am also aware that some might argue that it is either not possible or not realistically feasible to provide truly comprehensive metadata. I take the point, but it seems to me that this is tantamount to saying that no one can use anyone else's data. However, if each of us is not prepared to recreate archaeology from the ground up (i.e., if the discipline is to be at all cumulative, which is not to say perfectly cumulative), we have to be able to critically evaluate others' evidence and accept or reject their conclusions. We must accept that we can, at some level, share data.

So, having asserted that adequate semantic metadata is the documentation needed to make sensible analytical use of the data, how are we to proceed? The most obvious way to begin is to take each database table in turn and document the individual columns. Minimally, this might require associating the column label with a longer more descriptive label. For columns that describe metric variables, necessary metadata would include the measurement units and a textual description of how the measurement was taken. For nominal variables, we need both to decode values represented by numeric or abbreviated textual codes—that is to associate with the codes a descriptive label and textual discussion of how the associated value is defined and distinguished from other values.

Less obviously, the nominal values should also have an indication of whether the value was systematically recorded by the relevant analyst. It is not uncommon for analysts, in any one instance, to use only portions of a master coding sheet. For example, faunal analysts might use a master species list, but in a given analysis might only identify some classes of remains at a higher classificatory level. While the coding key may have a long list of bird species, it may be that in a particular case the analyst would simply identify specimens as undifferentiated birds. Someone else using the database and seeing bald eagles on the species list, might conclude that the absence of bald eagle bones in the database would indicate an absence of eagles, when it in fact it indicates that the analysis did not distinguish among bird species.

The *critical* reader will observe that this specification seems a bit loose about how much would be enough and that any description is going to depend on reasonable agreement on the meaning of what constitute primitive terms for purposes of the documentation. These points are well taken. Of course, without the minimal documentation described above, I think there is no hope of sensibly using the data. I am sure that I am not the only one who has had to give up on using an important legacy dataset because there was no key that provided an interpretation of numeric codes. More subtly, documentation that would be adequate for someone with a similar background to the original analyst, might not be adequate for someone who comes out of a very different academic tradition. Some documentation will be inadequate for any use, but the degree of adequacy will, to an extent, be a relative.

The *practical* reader will observe that for an analyst to provide this documentation effectively would both require both a high degree of circumspection and a great deal of time. These observations are indeed warranted. On the first point, the development of coding keys/or metadata descriptions needs to be done not from the standpoint of what the analyst might need a few years hence to be reminded of exactly what was done, but instead taking into account what a relatively naive user would need to know. This will be time consuming. However, honoring our ethical obligations to the archaeological record demands this investment in data documentation—at least as much as field documentation that we'd never think of skipping.

In the end, tradeoffs will need to be made, but they should be made with the long-term issues of data sustainability very much in mind.

3. Documenting Database Semantics in the Digital Archaeological Record

I have argued that adequate semantic metadata is the documentation needed to make sensible analytical use of the data. I also suggest a corollary that one cannot really know whether the semantic metadata is adequate until a naive archaeological analyst has tried and succeeded in analyzing the data contained in the database.

Over the last three years, a group of scholars in archaeology, informatics, and computer science, have approached the problem of database semantics from this general direction. We initially attacked this problem not because of our interest in metadata or digital libraries, but because of our interest in understanding socio-ecological dynamics in the archaeological record at temporal and spatial scales that greatly exceed those of a single, self-contained project. That is, we were driven by a compelling need to integrate data from multiple archaeological projects directed by different investigators employing incommensurate coding schemes.

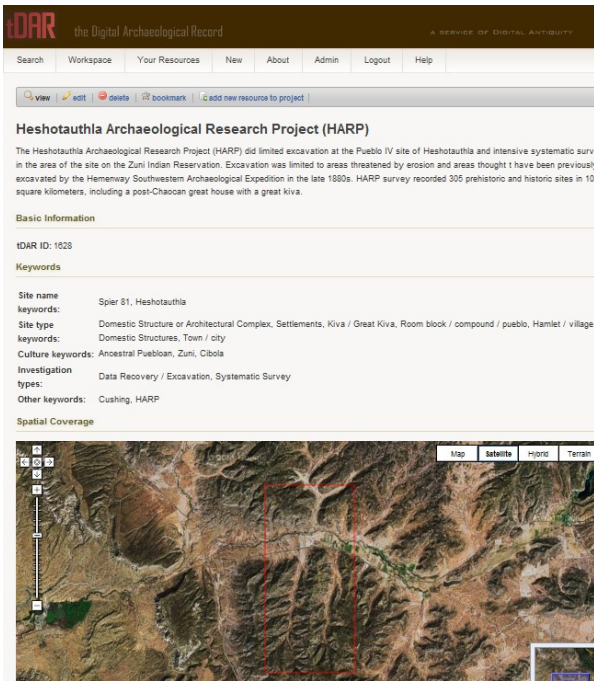

Figure 1: *tDAR project metadata.*

Driven both by a substantive need for integration of data across databases, and also by concerns for preservation of and access to archaeological data, this multidisciplinary and multi-institutional team has been working for the last several years to establish a trusted digital repository for archaeological data and documents in the United States, embedded in a sustainable organization, Digital Antiquity (MCMANAMON *et al.,* 2010. Digital Antiquity's repository, known as tDAR (for "the Digital Archaeological Record") targets documents and data derived from ongoing research as well as legacy data collected through more than a century of archaeological research in the Americas.

Because of the scale of archaeological work in the US—more than 50,000 field projects annually are mandated or authorized by the federal government alone (DEPARTMENTAL CONSULTING ARCHEOLOGIST, 2009)—and to minimize out-of-pocket costs associated with the entry of data into the repository, tDAR offers a Web interface that allows archaeologists to up-load their data and documents and to provide the associated metadata documentation through a series of interactive Web forms.

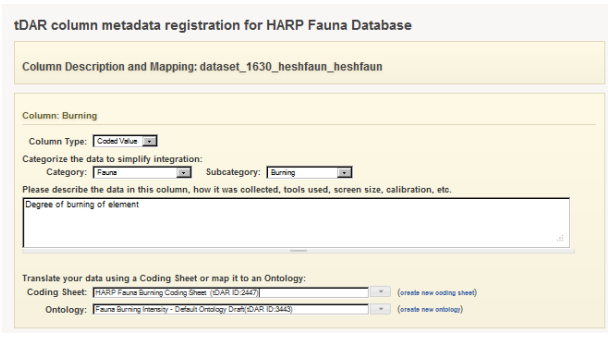

Figure 2: *Database column documentation.*

At the highest level, tDAR collects metadata for an archaeological project (or intervention; Figure 1) that is shared across the project's component information resources, such as documents, databases, and images. Next, technical, bibliographic, and substantive metadata are collected separately for each information resource. For databases, the process of metadata documentation continues along the lines suggested in the previous section. tDAR reads the database and identifies the component tables and columns. Then, for each column it prompts for information about the nature of the value being represented (Figure 2). If it is a metric measurement, it asks for the measurement unit and a description of the value being measured.

If a column represents a nominal (categorical) variable, tDAR then seeks the entry or upload of a database-specific coding key information for that column (called, in tDAR, a "coding sheet"). In general, that consists of a set of triples for every different value in the classification being recorded: a code, a textual label, and a description of that value. This coding sheet (which is reusable) is then permanently associated with that column in that table in that database (Figure 3).

Figure 3: *Coding sheet.*

Users who go to tDAR (http://tdar.org) are able to search the metadata (and, if they wish, the data) to identify relevant information resources (Figure 4). Once they agree to properly cite any data they download, they are able to download the dataset and associated coding keys, or they can download—for their own analysis—a "decoded" database in which the nominal values are replaced by the textual labels from the appropriate coding key.

4. Beyond Discovery and Access

While archaeologists frequently complain about a lack of data, most of us do not do a very good job of making use of data that others have are already collected. There are, of course some good reasons (e.g., the difficulty of discovering and accessing relevant data) and not so good reasons that this is the case.

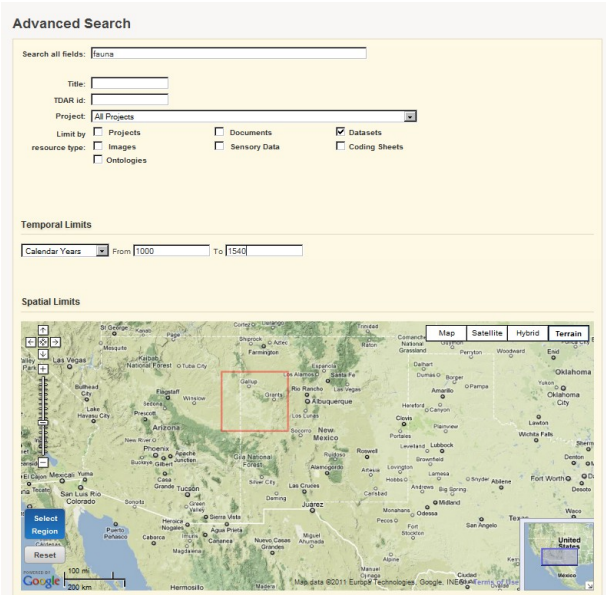

Figure 4: *Search screen.*

Substantial attention, appropriately, has been devoted to the problems of discovery and access. However, if we want to synthesize data from multiple sources using different recording standards, then discovery and access are not sufficient, even with comprehensive metadata for each dataset. We must provide tools to facilitate the integration of archaeological databases (KINTIGH, 2006; SNOW *et al*., 2006). By database integration we mean the process of transforming a set of input datasets that were recorded using different protocols into a form in which the observations are comparable across them.

Here lies a key payoff for tDAR's approach of having the semantic metadata online and tied directly into the databases. With these metadata in place, it becomes possible to automate tools that enable users to perform data integration.

tDAR has implemented a key data integration component that resolves conflicts in the recording of nominal variables. As noted above, in tDAR one step in the metadata documentation of a database is the user's association of a database-specific coding sheet with each database column representing a nominal variable. Integration of these variables requires a further step in which the values in each coding sheet are mapped to nodes of a shared, hierarchically organized ontology for that variable (Figure 5). With this accomplished it becomes possible to integrate database columns that are mapped to the same set of ontologies (without, of course, having altered the content of the original databases).

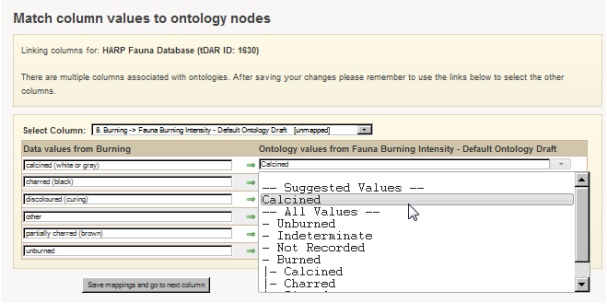

Figure 5: *Mapping coding sheet to ontology.*

The integration process proceeds with the user identifying the databases of interest though a search or other means. The user then selects the column from each selected database that is associated with each variable to integrated (e.g., for the variable, burning, pairing the "BURNED" column of one database with the "BURN" column of another). For each variable, the user then prunes the ontology tree associated with each integration variable based on the nodes actually represented by observations in each of the selected datasets. Thus, if one analyst recorded bones as simply burned or unburned and another recorded different intensities of burning, such as scorched, charred, and calcined, the integration of the two datasets would be sensible only at the level of burned vs. unburned.

Once the pruning is accomplished, the integration can proceed. It yields an output dataset in which the integrated variables in the original datasets are all reported in a common classificatory scheme derived from the pruned ontology. This unified database can then be subjected to statistical analysis.

This approach to the data integration problem highlights the need for domain-specific ontologies that not only make integration possible, but that can also greatly strengthen resource discovery capabilities. The effective deployment of these ontologies, however, will require investments by many user communities in building mutually acceptable ontologies. It is our expectation that user communities will make this investment because it enables new archaeological syntheses of data that are of interest to them.

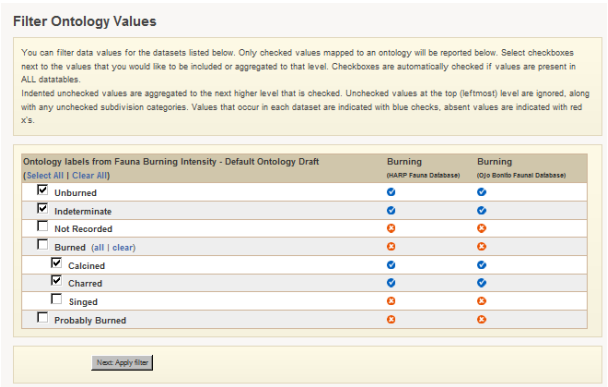

Figure 6: *Pruning the ontology tree.*

For the same reason, data contributors will recognize a payoff (individual and collective) for the time they invest in documenting the semantic content of their databases, thus increasing their willingness to provide these critical metadata in the first place. It is, of course, exactly this effort that is an essential step in making the semantics of their database sustainable in the long term.

Conclusions

The premise of this article has been that semantic metadata are adequate to the extent that they enable an archaeologist not familiar with the specifics of a project to make sensible analytical use of its data. We have certainly not addressed all of the issues of data comparability (which, for example, also has much to do with assessing the comparability of field data collection procedures). Nonetheless, the data integration framework that we have implemented in tDAR is important because it prompts data contributors to more fully encode in formal metadata their knowledge about the data and it empowers users to pursue synthetic analyses on a scale that would heretofore have been unthinkable.

Acknowledgments

The work presented here depends upon contributions of many individuals at Arizona State University with whom I have had the privilege of collaborating: Chitta Baral, K. Selçuk Candan, Huiping Cao, Tiffany Clark, Matthew Cordial, Hasan Davulcu, Subbarao Kambhampati, Allen Lee, Shelby Manney, Ben Nelson, Margaret Nelson, Yan Qi, Karen Schollmeyer, Katherine Spielmann, and Joshua Watts.

Digital Antiquity collaborators at other institutions include Jeffrey Altschul at the SRI Foundation/Statistical Research, Inc., John Howard at University College-Dublin, Timothy Kohler at Washington State University, W. Fredrick Limp at the University of Arkansas, Julian Richards at the University of York/Archaeology Data Service, and Dean Snow at the Pennsylvania State University.

This material is based upon work supported by the National Science Foundation under Grant Nos. 0433959 and 0624341 and generous grants from the Andrew W. Mellon Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or the Andrew W. Mellon Foundation.

References

ARCHAEOLOGY DATA SERVICE 2008. Guidelines for Depositors, Version 1.3, March 2008. http://ads.ahds.ac.uk/project/userinfo/deposit_guidelines/deposit_create2.cfm?datatype=database, accessed June 15, 2010.

DEPARTAMENTAL CONSULTING ARCHEOLOGIST 2009. *The Secretary of the Interior's Report to Congress on the Federal Archeological Program, 1998-2003*. Archeology Program, National Park Service, Washington, DC. http://www.nps.gov/archeology/SRC/src.htm.

KINTIGH, K.W. (Ed.) 2006. The Promise and Challenge of Archaeological Data Integration. *American Antiquity* 71(3) pp. 567-578.

MCMANAMON, F.P; KINTIGH, K.W. 2010. Digital Antiquity: Transforming Archaeological Data into Knowledge. *The SAA Archaeological Record* 10(2): 37-40.

MICHENER, W.K.; Brunt, J.W.; Helly, J.J.; Kirchner, T.B.; and Stafford, S.J. 1997. Nongeospatial Metadata for the Ecological Sciences. *Ecological Applications* 7(1) pp. 330-342.

SNOW, D.R.; GAHEGAN, M.; GILES, C.L.; HIRTH, K.G.; MILNER, G.R.; MITRA, P; WANG, J.Z. 2006. Cybertools and Archaeology. *Science* 311, p. 958-959.