the Digital Archaeological Record: The Potentials of Archaeozoological Data Integration through tDAR

Katherine Spielmann and Keith Kintigh

Katherine Spielmann and Keith Kintigh are Professors in the School of Human Evolution and Social Change at Arizona State University.

For the past several years archaeologists, archaeozoologists, and computer scientists have collaborated on creating a Web-based, public-access cyberinfrastructure, tDAR (the Digital Archaeological Record; http://tdar.org). tDAR not only serves as a digital repository providing preservation and access to archaeological datasets uploaded by users, but also allows users to integrate digital datasets by combining datasets recorded using different protocols into a single dataset with analytically comparable observations.

Our original motivation to develop tDAR derived from a collaborative effort by a group of Southwestern archaeologists at Arizona State University to synthesize regional-scale archaeological data on socio-economic change. Rather than simply synthesizing the *conclusions* of many separate analyses, our objective was to create integrated datasets of original observations that could be subjected to new analyses focused on larger spatial and temporal-scale questions. Not surprisingly, our efforts at synthesis were frustrated by the practical difficulties both of acquiring the original datasets and then of integrating them, given that they were collected by different investigators from across the US Southwest. As we developed the integration capabilities of tDAR, archaeozoological data have been our specific focus.

In this paper we briefly discuss 1) data archiving in tDAR, 2) the development and use of general ontologies for archaeozoological variables such as taxon, completeness, butchering, and burning in tDAR, and 3) results of a pilot analysis of Southwestern faunal databases that is informing our software development and helping refine protocols to enable integrated analyses

of faunal data. The members of two Faunal Working Groups, one in North America and the other in Britain, have contributed significantly to the development of these protocols.

Archiving Databases in tDAR

tDAR accepts a wide range of digital archaeological data. In order to contribute digital data to tDAR, users must first register and agree to the terms of a user agreement. In tDAR, digital documents or data are registered either as *independent* digital resources (for example, a single methodological article) or as part of a suite of information resources associated with a single *project* such as a single excavation or a survey. The digital *information resources* resulting from a project might include any number of separate datasets, documents, and image files resulting from the field work.

In registering a project or information resource, the contributor provides the archival and semantic information—metadata—that will permit their long-term preservation and scientific use. To simplify metadata entry, the project-level metadata (e.g., sponsor, location, culture) applies to all of the project's component information resources (databases, documents, etc.). Within each project, individual information resources are described with additional metadata specific to that digital object to enhance its ability to be discovered by a search and to be properly preserved and used in the very long term.

As we discuss below, integration of multiple datasets is necessary to address larger-scale research, and detailed metadata make it possible for observations to be made comparable across databases. For databases (and spreadsheets), the metadata includes information on the individual tables, and columns along with the coding sheets that provide the semantic labels for encoded values. For example, a column labeled "Taxon" encodes information on a bone's taxonomic

assignment and in that column the database value 101 may represent "*Lepus*". A translation function in tDAR creates a dataset with both the value labels and the original numeric codes.

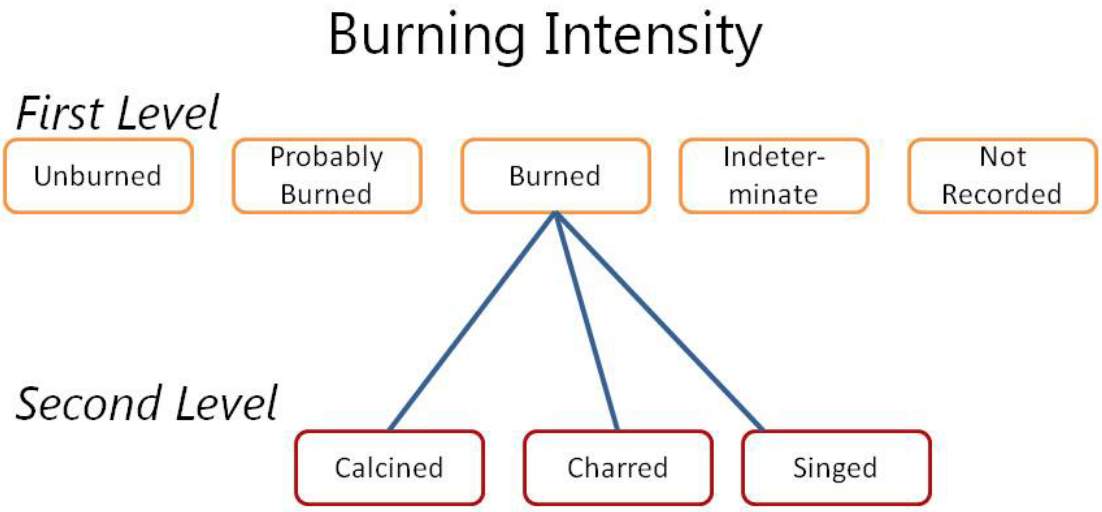
Developing Faunal Ontologies

We recognize that there is significant variation in how researchers code archaeological data, including fauna. Our goal is *not* to standardize what individual analysts do, but instead to make it possible to integrate their data with those of others using a shared conceptual framework for analysis. In tDAR we feel that it is essential to maintain the data as they were originally encoded, along with the associated coding keys. To accomplish that goal while enabling integration, we employ "ontologies." In tDAR, ontologies are hierarchically organized maps of concepts. For example, a variable "burning" might have been recorded by one analyst as present/absent, while another may have subdivided "present" into charred, burned, and calcined. If communities of researchers agree upon a shared ontology, for example for the variable burning such that charred and calcined are subcategories of burned, the data integration tools of tDAR allow an analyst to map the individual translated codes in their databases (e.g., *calcined*) to ontology values (in this case, *burned*) used by other analysts. The result is that variables that are recorded differently in different databases can be integrated because the original encodings are mapped to shared values.

The process of developing general ontologies involves a community of users moving toward a consensus on a framework that can be shared. Over the past year we have had two opportunities to convene the North American Faunal Working Group (FWG), comprised of faunal analysts working in the southwestern and eastern regions of the US, and to meet with a British Faunal Working Group organized by Archaeological Data Services in York, England. One of the objectives of these meetings was to develop general ontologies for the variables that

archaeozoologists typically code. For most variables (e.g., burning, discussed above) this proved to be a relatively straightforward process for both groups. There was general agreement that there should be a first-level option often at the level of presence/absence, as well as indeterminate, or unrecorded values. Beyond that, for most variables there was a second level of greater specificity regarding presence (e.g., charred, burned, or calcined; Figure 1). The general ontologies developed by the faunal working groups are now publicly available in tDAR, and members of the North American FWG have been working with them.

Figure 1. Burning ontology.

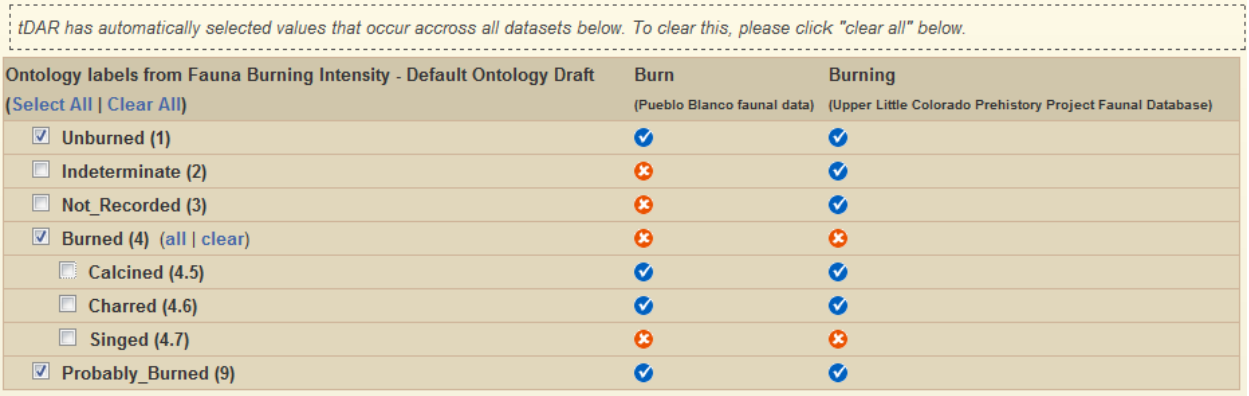


Mapping to Ontologies and Data Integration

Data integration in tDAR requires that the variables of interest have been mapped to the shared ontologies for those variables. Ideally, the original analyst or person uploading the dataset would perform these mappings; however a tDAR user can create these mappings her or himself. Datasets (ones' own or developed by others) are then moved into the user's workspace, and those to be integrated are identified.

The tDAR faunal integration tool allows the analyst to choose the variables that are to be integrated, as well as the level at which integration is to take place. For example, while two datasets may have specific degrees of burning intensity coded (e.g., charred, burned, calcined), the analyst may only be interested in the presence or absence of burning. In that case, as illustrated in Figure 2, selecting "burned" would include all those cases coded to a more specific "burned" value. Likewise, if an analyst were interested in comparing artiodactyls and lagomorphs, she could choose only those taxonomic values. Cases coded more to more specific taxonomic levels under artiodactyl or lagomorph (e.g., *Antilocapra sp.* or *Lepus sp.*) would aggregate up.

Figure 2. Burning ontology mapping



The output from the integration can be exported as an Excel file in which each dataset is a separate sheet. These spreadsheets, or a combined spreadsheet, can then be analyzed or uploaded into a statistical package for further analysis. We are currently working with our computer scientist partners to streamline ontology mapping and to make it possible to re-run previously used integration run streams.

Pilot Analysis

In preparation for an October meeting of the North American FWG Spielmann undertook an integrated analysis of nine Southwestern faunal databases. tDAR now houses at least

seventeen Southwestern faunal databases, which contain over 220,000 faunal specimens entered and associated with coding sheets in tDAR. To perform an integrated analysis at this scale without digital integration would be unthinkable.

The intent of the pilot was to investigate patterns in faunal resource depression between A.D. 1200 and 1400, the period of time represented in most of the current tDAR Southwestern faunal databases. Investigating regional-scale faunal resource depression has been a goal of the tDAR project since its inception. The results of the pilot, however, pertain more to determining the comparability of datasets, which is necessary for faunal data integration to be viable both practically and scientifically.

Temporal information. Integrating multiple archaeological datasets requires project-level metadata on the period of time the site or sites date to. In addition, if a single dataset contains multiple time periods, it is critical that temporal information be contained in the dataset so that observations pertaining to each period can be distinguished. Two large and interesting Southwestern faunal datasets currently in tDAR were not useable in the pilot, for example, because they contained data spanning a 200-year period and the file did not contain temporal information.

Ontology mapping. A review of how taxonomic categories were mapped by different analysts to the Southwestern general taxonomic ontology revealed some variation in mapping. For example, some analysts mapped codes to both the “large mammal” and “artiodactyl” categories, while others only used “artiodactyl.” And some mapped to “small mammal” while others mapped to “small unid. animal.” Prior to undertaking an integrated analysis, an analyst should be aware of different patterns in ontology mapping.

Taphonomy. Archaeozoologists routinely collect data relevant to taphonomic processes (e.g., fragment size, condition, weathering, and animal gnawing), but these data generally are not readily accessible or are only reported in a summary fashion. The integrated analysis of multiple datasets requires that we first evaluate the degree to which zooarchaeological remains from different sites have been subject to similar taphonomic processes. Controlling for taphonomic processes will allow us to identify patterning in the zooarchaeological record that is not due to biases, and perhaps to reveal patterning that we did not previously have the ability to detect. tDAR makes this possible by making available the full faunal datasets, complete with variables related to taphonomy, and an integrative tool that allows these variables to be analyzed by taxonomic category across multiple datasets.

When faced with information on fragment size, condition, weathering, and animal gnawing, however, it is not immediately obvious how to take this rich information and evaluate taphonomic comparability. We are thus preparing a proposal in part to fund the development of a protocol, which we will invite the archaeozoological community to evaluate, for determining the degree to which faunal datasets are taphonomically comparable.

Context. It is well-documented that people may choose to dispose of different animal taxa or different portions of taxa in different contexts. Thus context must be controlled in inter-site comparative or synthetic analyses. As with temporal information mentioned above, control for context requires that intelligible contextual information be included in the project metadata as it pertains to site type, and within the datasets themselves as it pertains to the excavation context of individual specimens.

As with taphonomy, controlling for context is not as straightforward as it might appear. In working with contextual information from across the Southwest, it is clear that some

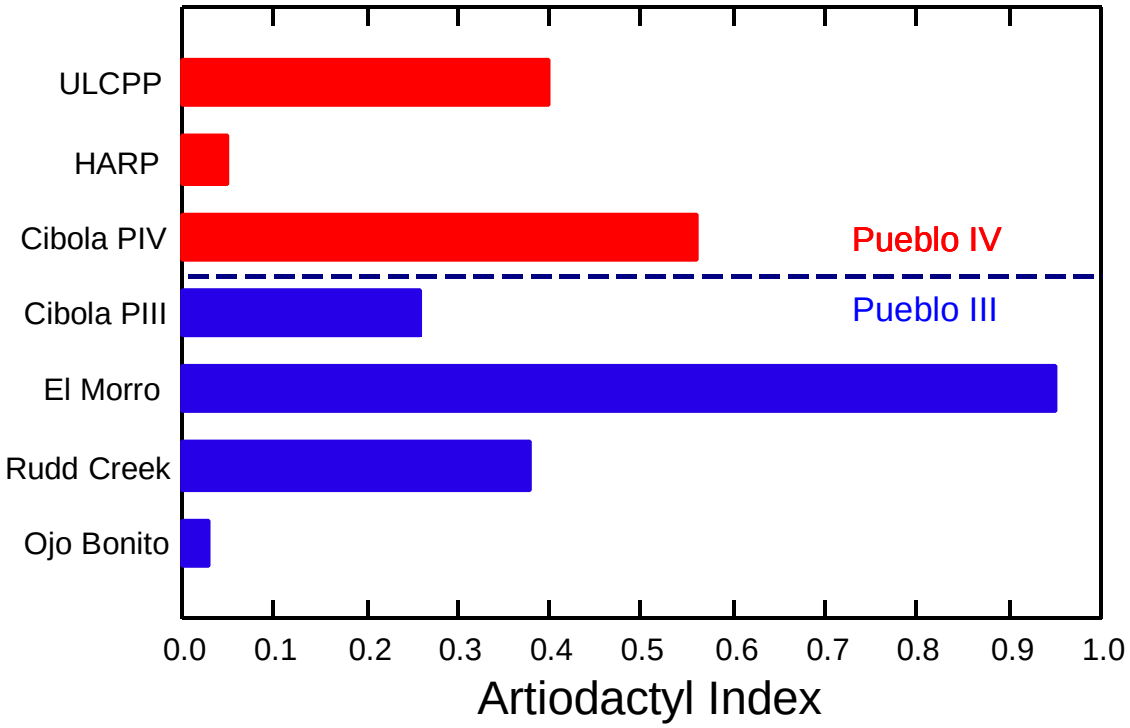
contextual coding schemes are far more detailed than others. An integrated analysis is thus likely to allow only broad control over context (e.g., intra-mural vs. extra-mural; midden vs. pit). Moreover, even where contextual information is similarly detailed across the sites of interest, sample size issues are likely to require the aggregation of multiple contexts. At this point we do not know whether controlling for broad contexts of faunal deposition is sufficient for integrated analysis. To our knowledge there has not been a systematic analysis of patterning in faunal disposal at the regional or sub-regional level, and thus this is a second area in which we are proposing to undertake research.

Results of the pilot. The datasets in tDAR that spanned the A.D. 1200-1400 period range were largely from the Zuni area. One Black Mountain Phase Mimbres dataset and two Salinas data sets also fell within that time period (Table 1). After exploring a few taphonomic variables and the ontology mapping differences discussed above, as a first evaluation of whether resource depression had occurred over time, particularly in the Zuni area, Spielmann calculated the artiodactyl index (Artiodactyl NISP /Lagopmorph NISP) for the sites in the sample. These data are provided in Table 1 and Figure 3 below.

Table 1. Artiodactyl indices for A.D. 1200-1400 Southwestern datasets in tDAR.

Mimbres	***Zuni***							***Salinas***	
Animas Village	**Ojo Bonito**	**Rudd Creek**	**El Morro**	**Cibola PIII**	**Cibola PIV**	**HARP**	**UCLPP**	**Pueblo Blanco**	**Pueblo Colorado**
1200-1300	1200-1300	1200-1300	1250-1300	1250-1300	1300-1350	1300-1350	1300-1400	1300-1400	1300-1400
0.35	0.03	0.38	.95	0.26	0.56	0.05	0.4	0.71	1.57

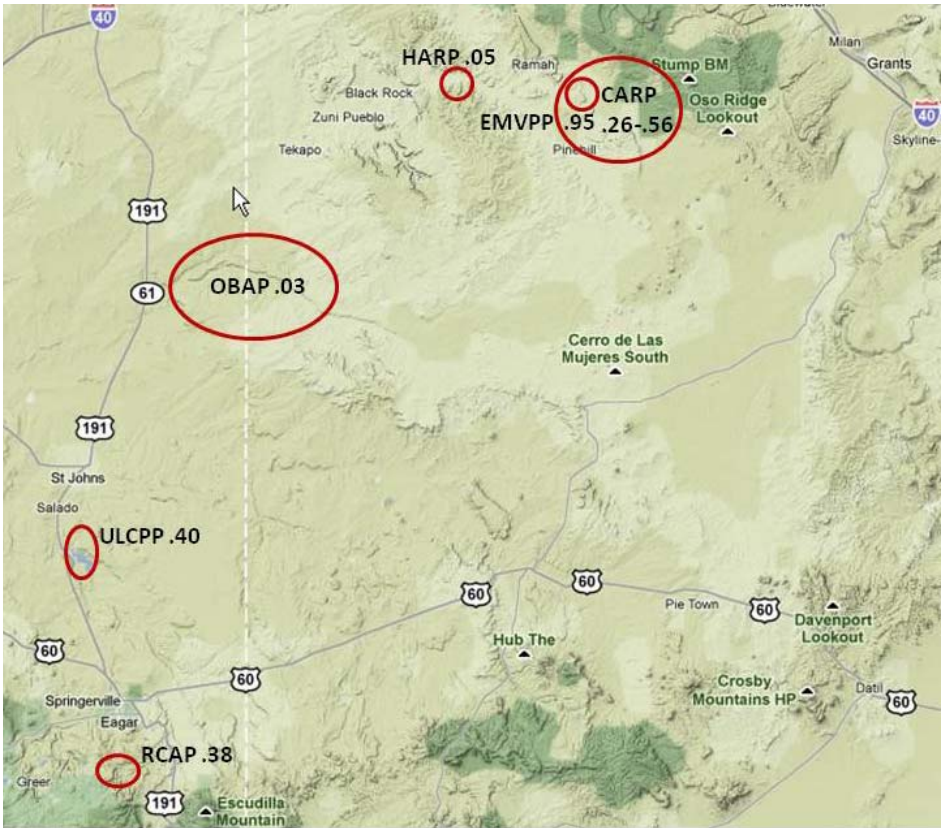
Figure 3. Zuni artiodactyl indices.



As Table 1 and Figure 3 indicate, even within the Zuni area the artiodactyl index is immensely variable and does not pattern temporally (from left to right in Table 1 and bottom to top in Figure 3). In discussion with Kintigh, whose datasets these were, it became clear that 1) intensity of long term settlement on the landscape, 2) site type (e.g., post-Chacoan great houses vs. village settlement), and proximity to higher elevation areas (Figure 4) all likely played a role in long-term artiodactyl availability on this landscape.

In moving forward we do not harbor any illusions as to data integration and analysis being a straightforward undertaking. Nonetheless, the rewards of being able to address regional-scale anthropological research questions at a depth and breadth that have not thus far been possible using zooarchaeological data are compelling.

Figure 4. Locations of Zuni archaeological sites with associated artiodactyl indices.



Conclusion

Inquiry on a regional scale requires changing archaeological practice to promote a new approach to data sharing that includes the adequate documentation of these data so that they are broadly useable in scientific analyses. tDAR provides a technical infrastructure for the preservation of and access to archaeological data, and has prototyped data integration tools that empower synthesis. What is necessary now is the accumulation of large numbers of well documented data sets and analytical protocols that allow us to assess the comparability of these data sets, and commitment to understanding the particular contexts from which these data derived. We invite SAA and ICAZ members worldwide to upload their projects into tDAR so that they may be shared and to experiment with the integration tool.

Acknowledgments

The work reported here has been supported by grants from the National Science Foundation (04-33959, 06-24341) and a grant jointly funded by the National Endowment for the Humanities (PX-50022-09) and the Higher Education Funding Council for England of the United Kingdom acting through the Joint Information Systems Committee (JISC). Development of the Digital Archaeological Record has also been funded by grants from the Andrew W. Mellon Foundation. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation, the National Endowment for the Humanities, the Joint Information Systems Committee or the Mellon Foundation.