THE DIGITAL ARCHAEOLOGICAL RECORD

THE POTENTIALS OF ARCHAEOZOOLOGICAL DATA INTEGRATION THROUGH TDAR

Katherine A. Spielmann and Keith W. 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; McManamon and Kintigh 2010; 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 them 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 Southwestern archaeologists at Arizona State University to synthesize regional-scale archaeological data on socioeconomic 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 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 of general ontologies for archaeozoological variables, and (3) results of a pilot analysis of Southwestern faunal databases that is informing our software development and helping refine protocols for integrated analyses of faunal data.

Archiving Databases in tDAR

tDAR not only allows users to discover and download data and documents of interest, but also to contribute their own archaeological data. In contributing an information resource. one uploads a file and enters the archival and semantic information—metadata—that will enable it to be discovered by a search and permit its long-term preservation and scientific use. As we discuss below, detailed metadata make it possible for observations to be made comparable across databases. For databases (and spreadsheets), the metadata includes information describing the individual 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. 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 sets of hierarchically organized concepts. For example, "burning" might have been recorded by one analyst as "burned/unburned," while another may have effectively subdivided "burned" into charred, burned, and calcined. If communities of researchers agree upon a shared ontology, e.g., for the variable burning, "charred" and "calcined" are subcategories of "burned" (Figure 1), 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 recorded differently in different databases can be integrated because the original encodings are mapped to shared values.

Developing general ontologies involves a community of

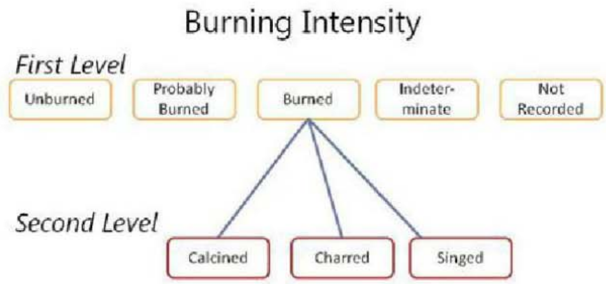

Figure 1. Burning ontology.

users moving toward a consensus on a framework. 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 U.S., and to meet with a British Faunal Working Group organized by Archaeological Data Services in York, England. One objective of these meetings was to develop general ontologies for the variables that archaeozoologists typically code. For most variables 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). The general ontologies developed by the faunal working groups are now publicly available in tDAR.

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 would perform these mappings; however, a tDAR user can create these mappings her or himself. Datasets to be integrated are then moved into the user's workspace. tDAR's integration tool allows the analyst to choose the variables 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 to more specific taxonomic levels under artiodactyl or lagomorph (e.g., *Antilocapra sp.* or *Lepus sp.*) aggregate up. The output from the integration can be exported as an Excel file that can be uploaded into a statistical package for further analysis.

Pilot Analysis

In preparation for a recent meeting of the North American FWG Spielmann undertook an integrated analysis of nine Southwestern faunal databases. tDAR now houses at least 17 Southwestern faunal databases representing over 220,000 faunal specimens in tDAR.

The intent of the pilot was to investigate patterns in faunal resource depression between A.D. 1200 and 1400, the period 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. 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

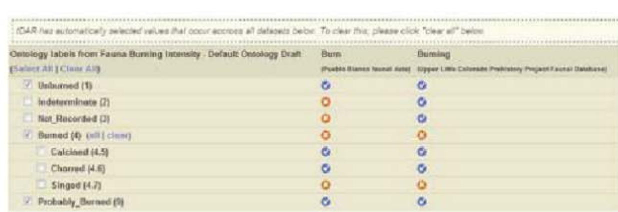

Figure 2. Burning ontology mapping

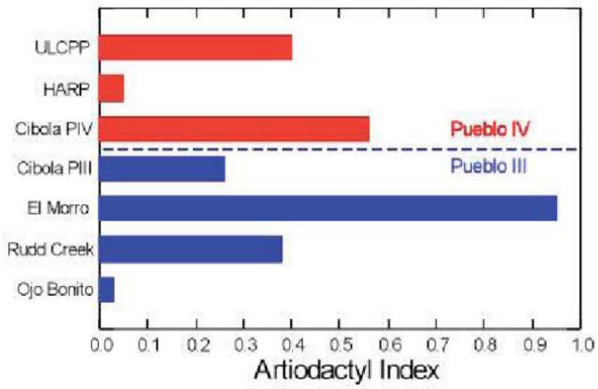

Figure 3. Zuni artiodactyl indices.

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 intersite 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 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 subregional 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. 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, Spielmann calculated the artiodactyl index (Artiodactyl NISP /Lagomorph NISP) for the Zuni sites in the sample. These data are provided in Table 1 and Figure 3.

As Table 1 and Figure 3 indicate, even within the Zuni area the artiodactyl index is quite 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 (3) 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.

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

Table 1. Artiodactyl indices for A.D. 1200–1400 Zuni datasets in tDAR.

Zuni Ojo Bonito	Rudd Creek	El Morro	Cibola PIII	Cibola PIV	HARP	UCLPP
1200-1300	1200-1300	1250-1300	1250-1300	1300-1350	1300-1350	1300-1400
.03	.38	.95	.26	.56	.05	.4

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

datasets and analytical protocols that allow us to assess the comparability of these datasets, and commitment to understanding the particular contexts from which these data are 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. We thank the Faunal Working Group members from North America and Britain for their insights concerning faunal integration. 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.

References Cited

McManamon, Francis P., and Keith W. Kintigh
2010 Digital Antiquity: Transforming Archaeological Data into Knowledge. *The SAA Archaeological Record* 10(2):37–40.