Data Integration in the Service of Synthetic Research

Keith W. Kintigh, Katherine A. Spielmann, Adam Brin, K. Selçuk Candan, Tiffany C. Clark, and Matthew Peeples

Anthropological archaeologists are committed to achieving scientific understandings of complex social processes that operate on centennial or millennial scales, notably including segments of societies that are absent from or underreported in recorded history. Many, including us, also believe that this research on the past should also have the potential to inform social action in the present for the future.

Recent articles in the *Proceedings of the National Academy of Sciences* and *American Antiquity* (Kintigh et al. 2014a, 2014b) propose a set of 25 grand challenges for archaeology intended to represent the most compelling questions facing our discipline. The challenges include, for example, "Why and how do social inequalities emerge, grow, persist, and diminish, and with what consequences?" and "How do humans perceive and react to changes in climate and the natural environment over short and long terms?" The challenges do not focus on new discoveries, nor are they peculiarly archaeological; rather, they address major issues in the social sciences. Answers to these challenges will

ABSTRACT

Addressing archaeology's most compelling substantive challenges requires synthetic research that exploits the large and rapidly expanding corpus of systematically collected archaeological data. That, in turn, requires a means of combining datasets that employ different systematics in their recording while at the same time preserving the semantics of the data. To that end, we have developed a general procedure that we call query-driven, on-the-fly data integration that is deployed within the Digital Archaeological Record digital repository. The integration procedure employs ontologies that are mapped to the original datasets. Integration of the ontology-based dataset representations is done at the time the query is executed, based on the specific content of the query. In this way, the original data are preserved, and data are aggregated only to the extent necessary to obtain semantic comparability. Our presentation draws examples from the largest application to date: an effort by a research community of Southwest US faunal analysts. Using 24 ontologies developed to cover a broad range of observed faunal variables, we integrate faunal data from 33 sites across the late prehistoric northern Southwest, including about 300,000 individually recorded faunal specimens.

Abordar los retos sustantivos más convincentes de la arqueología requiere una investigación sintética que explote el corpus grande y rápidamente en expansión de datos arqueológicos recopilados sistemáticamente. Esto, a su vez, requiere un medio de combinar conjuntos de datos que empleen sistemática diferente en su grabación mientras que al mismo tiempo preserva la semántica de los datos. Para ello, hemos desarrollado un procedimiento general que denominamos integración de datos en tiempo real basada en consultas, que se despliega dentro del repositorio digital el Digital Archaeological Record. El procedimiento de integración emplea ontologías que se asignan a los conjuntos de datos originales. La integración de las representaciones de conjuntos de datos basados en ontología se realiza en el momento en que se ejecuta la consulta, en función del contenido específico de la consulta. De esta manera, los datos originales se conservan y los datos se agregan sólo en la medida necesaria para obtener comparabilidad semántica. Nuestra presentación dibuja ejemplos de la aplicación más grande hasta la fecha: un esfuerzo de una comunidad de investigadores de analistas faunísticos del suroeste de Estados Unidos. Utilizando 24 ontologías desarrolladas para cubrir una amplia gama de variables faunísticas observadas, integramos datos faunísticos de 33 conjuntos de datos que investigan el suroeste septentrional prehistórico tardío, incluyendo más de 300.000 muestras de fauna registradas individualmente.

Advances in Archaeological Practice 6(1), 2018, pp. 30–41

DOI:10.1017/aap.2017.33

not and cannot emerge through intensive study of individual cases. Instead, they require research that synthesizes data and information—from a region, a hemisphere, or even the globe—to achieve knowledge that includes novel understandings of fundamentally important social processes (Kintigh et al. 2014a:5). Other scholars working at regional and macroregional scales have similarly recognized the need to integrate diverse sources of data (e.g., Arbuckle et al. 2014; McKechnie et al. 2014; Manning et al. 2016; Mills et al. 2013).

WHAT SYNTHESIS REQUIRES

Achieving the kinds of synthesis envisioned here requires the resolution of both technical and social problems (explored initially in Kintigh 2006 and Kintigh et al. 2015 and in more detail in Altschul et al. 2017, nd). This article focuses on one particular problem, achieving effective integration of data across multiple datasets. By data integration, we mean the process of transforming datasets that were recorded in different ways into a single, unified dataset with analytically comparable observations.

The Need to Integrate Primary Data

Synthetic research on the scale that grand challenges require entails deriving and comparing data-driven interpretations of primary data recovered by other archaeological projects. Today, syntheses are too often based on the data summaries and conclusions drawn by the original researchers. While this mode of synthesis is efficient and has undeniably been important, it also has liabilities. Conclusions that are erroneous or based on inconsistent premises become entrenched in the literature as "facts" that persist as faulty premises in subsequent scientific arguments. For example, for several decades archaeologists cited DiPeso's (1974) dating of the cultural chronology for Casas Grandes—a major center of the late prehistoric American Southwest/northwest Mexico region. The errors in dating were not corrected until Dean and Ravesloot (1993) reinterpreted the primary data—the tree ring dates from Paquimé (see also Whalen and Minnis 2001). Equally important, the potential to explore multiple, large sets of primary data provides the opportunity to discover important cross-dataset patterns that could never be seen when comparing higher-level interpretations.

The Need for Discovery and Access to Data

The explosion in the quantity and complexity of archaeological data has led to large databases, obtained at great expense, with immense potential to contribute to science. Nonetheless, datasets that could be extremely useful for synthesis are often unknown or not readily accessible to scientists. Fortunately, the needed technical infrastructure is now available through digital repositories that provide effective discovery, access, and long-term preservation of datasets, notably the Archaeology Data Service (Richards 2017) in the United Kingdom and the Digital Archaeological Record (tDAR; McManamon et al. 2017) in the United States. While that preservation and access infrastructure is now well established, only a tiny fraction of the potentially useful datasets developed in recent decades have been deposited in these repositories or are otherwise accessible.

The Need to Integrate Data across Projects and Areas

Comprehensive, regional-scale data are never collected by a single research team; data must be compiled from many projects. Integrating data across projects is essential to archaeologists' efforts to recognize phenomena operating on large spatiotemporal scales and to conduct crucial comparative studies.

The Need for Comparable Observations

Although large-scale and synthetic research demands the integration of data across projects, recorded observations from different projects are often not directly comparable. This issue may be due to the variables chosen, inconsistent measurement techniques, evolving or conflicting taxonomies, or differing collection intensities. In the absence of tools to resolve these discrepancies systematically, researchers rely on text descriptions or verbal communication with the original investigators; or (too often) they proceed with analyses unaware of the implicit difficulties, thereby inviting spurious results.

The Need for Adequate Metadata

Adequate metadata for each variable in a dataset are essential to assess the comparability of observations in different datasets and to the task of aligning those observations to make them comparable (Kansa and Kansa 2013; Kansa et al. 2014). Metadata include technical information, such as file formats and character sets used. They also include semantic documentation of individual tables, columns, and nominal values in a relational database or spreadsheet. Is a variable a count, a measurement, or a nominal value? If it is a measurement, what are the units, and how were they measured? If it is a code, what does each different value of the code represent, and how were the values distinguished?

Kintigh (2013) has elsewhere argued that to be considered adequate, metadata for databases must include sufficient information for an archaeologist not familiar with the project to make meaningful scientific use of the data. While meeting this standard demands considerable effort, it is necessary for datasets to be used in data integration. In addition to documenting the variables represented in the dataset, it is also important to provide key contextual metadata that typically do *not* appear anywhere in the dataset itself, such as dates, location, sampling intensity, or recovery technique.

The Need for General-Purpose Data Integration Tools

For decades, archaeologists, including us, have integrated multiple datasets. This process typically involved examining the representation of each variable under consideration in every one of the subject datasets and recoding all of these variables to an ad hoc standard in all of the datasets. Most who have done this would agree that it is an often frustrating and nearly always time-consuming endeavor. Furthermore, these efforts are typically tailored to the specific datasets involved and not readily generalized or extended.

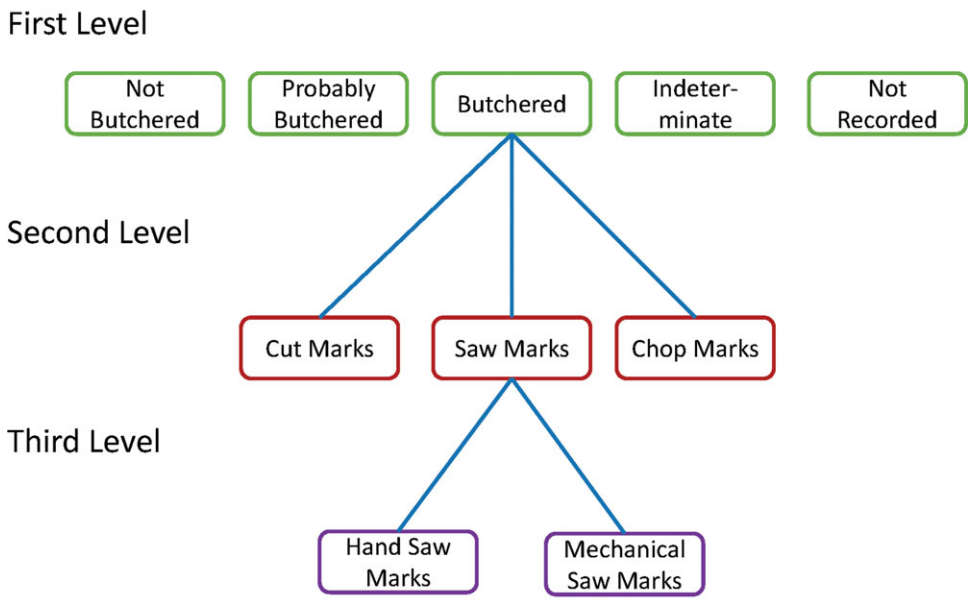

FIGURE 1. Butchering ontology.

QUERY-DRIVEN, ON-THE-FLY DATA INTEGRATION

Starting almost 20 years ago, archaeologists at Arizona State University were frustrated in their efforts to move beyond their individual areas of study to examine regional patterns. We envisioned a cyberinfrastructure tool that would facilitate the broad integration of data across our separate regional cases. Beginning in 1999, we sought National Science Foundation (NSF) funding to this end and in 2004 received our first award, whose goal was to assess disciplinary needs for cyberinfrastructure (reported on in Kintigh 2006).

Subsequent awards from the NSF and the Andrew W. Mellon Foundation funded the development of tDAR. It was, in fact, the research need for data integration and the associated demands for discovery and access that drove tDAR's initial development. The preservation component that is now integral to tDAR was soon added as a natural and important complement.

The ontology-based approach to data integration described here was developed over the last 15 years as a product of close collaboration among archaeologists, computer and information scientists, and software engineers.[1] In most cases, the refinements that we have implemented were direct responses to researcher requests. Thus far, the most intensive use has been by the community of archaeological faunal analyses, and our examples are drawn from that experience (Spielmann and Kintigh 2011).

As noted above, the standard approach to the integration of extant data sources is to do an ex post facto normalization of the subject datasets to a project-specific standard. In this approach, datasets that do not meet minimal data standards are rejected. When the best datasets have a precision that exceeds the set standard, that resolution is effectively discarded.

We chose instead to reconcile data source observations *with the data requirements of the query under consideration* rather than attempt global reconciliation of data sources. Because nominal variables (e.g., ceramic type, floral or faunal taxon, lithic tool type) are central to most archaeological analyses, reconciling nominal variables recorded using different classification systems is a central challenge for data integration. In this framework of query-driven data integration, each classification system used in the original recording of a nominal variable (e.g., butchering or faunal taxon) is represented by a set of values, each of which is explicitly linked to a node in a concept hierarchy, that is, an ontology (see below). In responding to a query, datasets using different coding systems can be used together *as long as each separate classification system is linked to a shared ontology*.

The source datasets (e.g., in Microsoft Excel or Access) are always maintained in tDAR in their original form (as well as in open-standard preservation formats). This policy is important because we do not want to lose the ability to see the data as they were originally recorded.

Ontologies

The integration software depends on agreed-upon ontologies for the database variables (columns) that are to be integrated. For our purposes, an ontology is a treelike hierarchy of concepts of increasing specificity. Figure 1 shows an ontology for the faunal variable butchering.

Ideally, ontologies are designed by a user community (in this case, archaeological faunal analysts) to capture the diversity of concepts used within the specialist community. Ontologies arrange concepts hierarchically, enabling more and less specific assignments (and the human lumpers and splitters) to peacefully coexist.

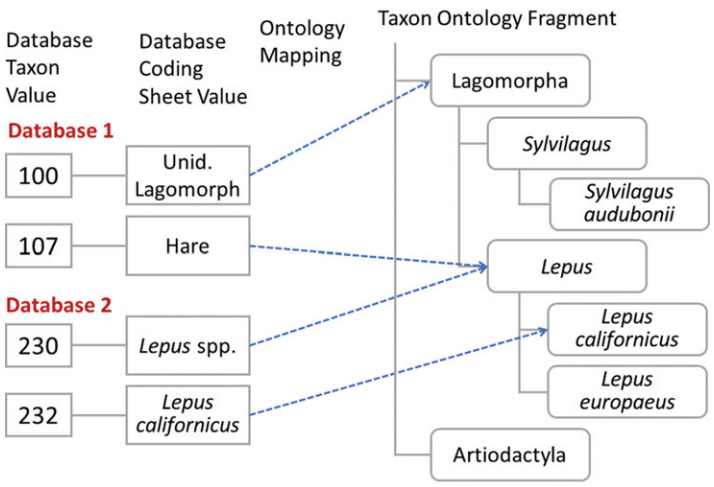

FIGURE 2. Faunal taxon variable in two datasets mapped to the taxon ontology (partial).

Coding Sheets

In many cases, the original datasets employ arbitrary numeric or textual codes to represent the individual values for a column. To document the meaning of these codes, tDAR allows the contributor of a dataset to enter a "coding sheet" that provides a translation of the codes to meaningful values (along with optional documentation of each value). Thus, the coding sheet might associate the code "100" in the taxon column of a particular database with the meaningful value "unidentified lagomorph." These coding sheets can be unique to a specific dataset or shared across multiple datasets.

Mapping Coding Sheet Values to Ontologies

The analyst contributing the database and coding sheet then maps the individual coding sheet or database values to specific nodes in the ontology associated with that variable, as illustrated in Figure 2. In this way, any number of datasets recorded in different ways can be represented within the unified framework provided by the ontology.

The tDAR ontologies accommodate synonyms that can assist in the mapping process. For example, coding sheets that employ common names for taxa can be easily mapped to the taxon ontology because common names are maintained as synonyms to the names in the Linnaean taxonomy.

Data Integration

Using datasets whose columns for the relevant variables have been mapped to the relevant ontologies, the stage is set for data integration. Assume that we want to compare artiodactyl indices (an indicator of reliance on large game) across sites in Southwest US datasets. The artiodactyl index is defined as the number of identifiable specimens (NISP) of artiodactyls (e.g., deer and antelope) divided by the total NISP of artiodactyls plus lagomorphs (rabbits and hares).

In that case, the integration process selects the subset of cases (rows) from all observations in all of the source (mapped) datasets in which the taxon variable is recorded as artiodactyl, or lagomorph, *or any subcategory of either of those higher-level values*. It returns a single dataset with the source dataset taxon values from all source datasets transformed into "Artiodactyla" or "Lagomorpha" as appropriate. The individual posing the original query can read the unified output database in Excel and use a pivot table or count occurrences of each value by site to easily calculate the artiodactyl index.

Now, let us say that same analyst wants to calculate the lagomorph indices for the same sites. The lagomorph index (indicating aspects of the local environment) is the ratio of the *Sylvilagus* (rabbit) NISP to the sum of the *Sylvilagus* and *Lepus* (hare or jackrabbit) NISPs. In this integration, we select from all the rows in all the datasets only the rows that are mapped to the genus (*Sylvilagus* or *Lepus*) level or below and produce a combined dataset as described above.

In this lagomorph index–directed integration, datasets that only classify bones by taxonomic order (Lagomorpha, Artiodactlya, Rodentia, etc.) would be ignored altogether because the information they contain does not address this integration query. However, those datasets *would* be used in calculations for the artiodactyl index integration query discussed above because they encode all the taxonomic specificity that is needed. The lagomorph index integration will also ignore (fail to select) any rows in the source datasets mapped directly to the Lagomorpha node (corresponding to an original dataset value of "unidentified lagomorph") because these rows do not inform this particular query, which requires distinguishing the genus.

By retaining all the specificity contained in the original datasets and performing the integration on the fly *at query time*, we are able to take advantage of those datasets using less specific classifications where they are relevant to the specific query while retaining the ability to use the finely classified data to satisfy those queries that demand such refinement.

A convenient data management and data exploration by-product of this integration process is that in a single operation it permits the specification of complex selections and hierarchical

TABLE 1. Artiodactyl Index Computation for Cíbola-Area Projects.

Project	Artiodactyla	Lagomorpha	Artiodactyl Index	Project Reference
Cibola Archaeological Research Project	3,008	7,823	0.28	Watson, LeBlanc, and Redman 1980
El Morro Valley Prehistory Project	84	94	0.47	Schachner 2012
Heshotauthla Archaeological Research Project	46	880	0.05	Kintigh, Glowacki, and Huntley 2004
Ojo Bonito Archaeological Project	87	3,367	0.03	Kintigh, Howell, and Duff 1996
Rudd Creek Archaeology Project	137	255	0.35	Clark et al. 2006
Upper Little Colorado Prehistory Project	4,864	14,303	0.25	Duff 2002
Total	8,226	26,722	0.24	

data aggregation across a great number of databases. Thus, if one wished to find all the macaws identified in these datasets, it would not be necessary to understand the coding and search 20 different datasets; one would simply look in an integrated dataset at the taxon column for the genus *Ara*.

Example: Artiodactyl Index

We illustrate key features of the data integration process with a simple example from the Cíbola (Zuni) area of the northern Southwest United States and then present a substantive result based on our ongoing research. Following the artiodactyl index discussion above, an integration of six faunal databases from Pueblo III and Pueblo IV sites (roughly AD 1200–1350) yielded a table with 31,700 rows representing 34,948 bones (in some cases a single row represented more than one bone with all the same characteristics). Table 1 summarizes these data by project area. The table shows considerable variation in the artiodactyl index, which we might expect to relate to the project's proximity to relevant habitats and perhaps to the kinds of sites investigated.

Table 1 shows an extremely high value (nearly as many artiodactyls as lagomorphs) for the El Morro Valley Prehistory Project, which primarily investigated a local post-Chacoan center close to the Zuni Mountains (a productive habitat for deer). The Ojo Bonito Archaeological Project, which primarily investigated another post-Chacoan center relatively distant from any mountains, had a very low proportion of artiodactyls (only 3% of the combined assemblage). The Cibola Archaeological Research Project and the Rudd Creek Archaeology Project, both close to major mountains, show moderate values, with artiodactyls representing about a third of both classes combined. However, a similar value is seen for the Upper Little Colorado Prehistory Project, which is somewhat farther from deer habitats. Seemingly anomalous is the Heshotauthla Archaeological Research Project, with a very low index value (0.05) despite being not much farther from the Zuni Mountains than the Cibola Archaeological Research Project and El Morro Valley Prehistory Project sites.

APPLICATION

Datasets

Spielmann led a large, NSF-supported collaborative synthetic effort through which 13 faunal analysts and other archaeologists contributed 42 datasets for faunal assemblages from 59 sites in the northern Southwest United States. Combined, these datasets contain more than 364,000 individually identified faunal specimens. The analysis presented here includes 297,839 specimens from 33 of these sites dating to the Pueblo III or Pueblo IV period (ca. AD 1150–1500). Datasets varied considerably with respect to which variables they recorded, though a core set was consistently recorded. All but seven of the 42 datasets (https://core.tdar.org/collection/16056) are now freely available in tDAR for anyone to use; the remainder are temporarily embargoed.

Ontologies

Faunal analysts mapped the dataset columns to a set of 24 ontologies (Table 2) for faunal variables developed by a series of working groups. (In most cases, this mapping was done by the original analysts.) With the exception of the taxon ontology, these are general-purpose ontologies devised to cover most nominal faunal variables recorded for prehistoric contexts in the United States and the United Kingdom. We used the Integrated Taxonomic Information System as a standard for the faunal taxa, including only taxa appearing in archaeological contexts in the US Southwest and adding indeterminate categories that were not encompassed by the taxonomic hierarchy (e.g., "large mammal"). All these ontologies are freely available for anyone to use in tDAR (https://core.tdar.org/collection/15376).

A broad-based working group of faunal analysts devised the initial drafts of the faunal ontologies. Ontologies were comparatively easy to develop for faunal variables that have well-defined categories and are generally recorded in similar ways (e.g., taxon, element, dorsal/ventral, and side). Other variables exhibit considerable diversity in how they are recorded (e.g., butchering, condition, completeness, and gnawing), but it was nonetheless possible to achieve agreements on ontologies.

The draft ontologies have been refined through interactions with working groups focused on the northern Southwest United States, on the United Kingdom, and on the Archaic period in the eastern United States. As analysts mapped their datasets to the draft ontologies and were unable to fit their categories, the draft ontologies were refined to the point that a large fraction of the analysts with whom we have interacted believe that their recordings can be reasonably represented within the system. This agreement was possible, in part, because analysts did not have

TABLE 2. Shared Faunal Variable Ontologies.

#	Variable Ontology
1	Age
2	Anterior-Posterior
3	Burning Extent
4	Burning Intensity
5	Butchering
6	Completeness
7	Condition
8	Confidence of Identification
9	Digestion
10	Dorsal/Ventral
11	Element
12	Erosion/Preservation
13	Fusion
14	Gnawing
15	Measurements
16	Origin of Fragmentation
17	Pathology
18	Proximal/Distal (long bones)
19	Recovery
20	Sex
21	Side
22	Taxon (by region)
23	Weathering
24	Worked Bone

to abandon their own coding schemes to use the ontologies. Mapping coded values to ontologies enables effective standardization without forcing analysts to accept them at the level of their individual analyses.

While working groups of users developed and uploaded the faunal ontologies used here, any data contributor can upload, use, and share an ontology. However, the more that an ontology represents shared agreement within a user community, the more it is possible to share and synthesize data across datasets recorded using different systematics.

If a research community shares a general approach to recording a variable, it may well be possible to reach a consensus on an ontology, using the hierarchical nature of the ontologies to accommodate disputes and differences among analysts. For example, there may be widespread agreement on the upper levels of a ceramic typology but variation in how "lumpers" and "splitters" deal with the finer points. In other cases, a community may be split with different factions recording certain variables in fundamentally incompatible ways, as is the case with some approaches to lithic typology. Even in these cases, there is value in employing ontologies to integrate data on the contested variables within each subgroup and sharing ontologies across the community on variables on which agreement is possible (e.g., lithic material).

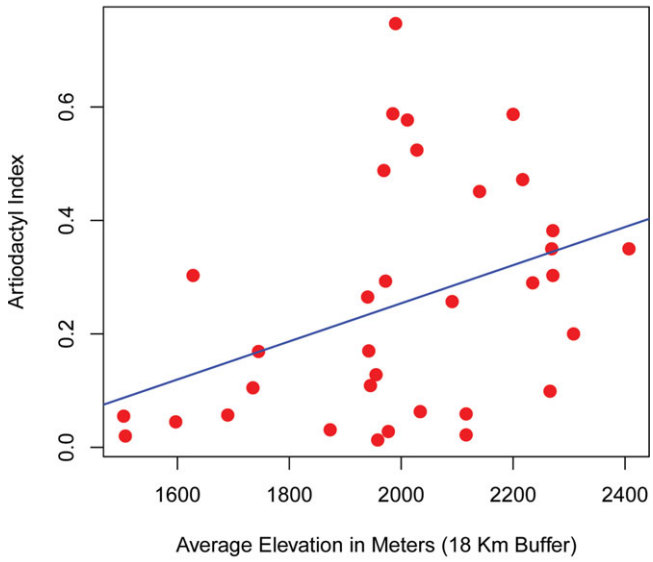

FIGURE 3. Scatterplot of site elevation (averaged over an 18-km buffer) vs. artiodactyl index. Regression $R^2 = 0.15$; $p = 0.02$.

Hypothesis

One substantive goal for this collaboration was to examine the hypothesis: human population *persistence* and *concentration* on the landscape result in large-mammal resource depression. As a part of this investigation, we needed to explore an alternative: Are environmental differences a significant factor in resource abundance (i.e., regardless of human demography, do more mesic environments favor larger game and drier environments, smaller game)? In the US Southwest, higher elevations tend to be more mesic, and elevation is a reasonable proxy for habitat productivity (e.g., Schollmeyer and Driver 2013). In this region, the large mammals used for food are overwhelmingly artiodactyls, with lagomorphs being the other major faunal food resource. Therefore, large-mammal resource depression is indicated by a decrease in the artiodactyl index.

Results

An initial view of the data (Figure 3) indicates a fairly strong positive relationship between elevation and project area, suggesting that differences in the artiodactyl index may be due more to elevation than to human predation. This finding, of course, is not unexpected, as higher elevations are favored habitats for deer. However, further investigation revealed a bimodal distribution of elevation of the sites and projects investigated, as shown in Figure 4.

If we plot only those project areas located above 1,900 m elevation (Figure 5), it is clear that there is *no relationship* ($R^2 < 0.01$) between elevation and artiodactyl index at these higher-elevation sites. Rejecting the idea that the artiodactyl index is simply a function of elevation for the higher-elevation sites, we are able to proceed with further analysis of the original hypothesis with the higher-elevation projects. The key point here is that by looking at

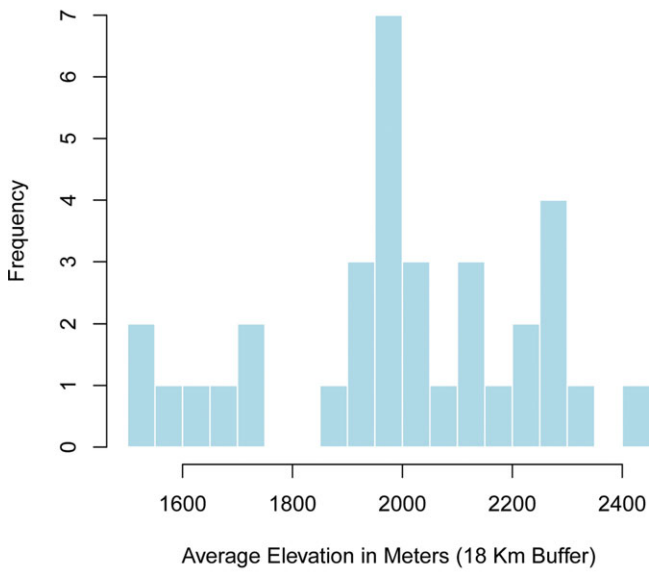

FIGURE 4. Histogram of site elevation (averaged over an 18-km buffer) showing bimodal elevation distribution.

only one or a few cases, we would *never* have been able to see this patterning.

Employing integrated data from these 34 sites, we are currently exploring taphonomic processes and other confounding factors. A full treatment of this synthetic research is in preparation and will appear separately. The data underlying Figures 3–5 are embargoed until that study appears but will be available in tDAR (at https://core.tdar.org/dataset/438729).

THE DATA INTEGRATION PROCESS IN tDAR

Dataset Ingest and Metadata Documentation

Data integration necessarily includes a number of steps. First, of course, the datasets to be integrated must be documented with appropriate column metadata, with coding sheets for nominal variables mapped to shared ontologies. tDAR provides software that interactively guides the user through each of these steps.

Thoroughly documenting a large dataset is a substantial undertaking. This process can be expected to go smoothly if the dataset is "clean" and was developed using good practices (e.g., Archaeology Data Service and Digital Antiquity 2013:73–84). It is further simplified to the extent that the same coding sheets (each of which is a separate tDAR resource) are used by multiple datasets because a coding sheet needs to be mapped to the corresponding ontology only once. tDAR also allows one documented dataset to serve as a template for others, so an analyst using a consistent coding scheme will find that while uploading and documenting the first dataset will take considerable time, uploads of subsequent datasets will go very quickly.

However, tDAR's data ingest and metadata documentation process can also reveal problems in the dataset design (e.g., having the interpretation of one column depend upon the value of a different column) or errors in coding that can be time-consuming to correct. For example, a dataset may contain numeric values that do not appear in the coding sheet for a particular variable. In that case, the analyst needs to determine whether this situation represents an omission in the coding sheet (in which case the coding sheet needs to be corrected) or whether the value was miscoded. If the value is miscoded, the analyst would attempt to ascertain whether this mistake was an error in transcription from a paper form (in which case the digital dataset can be corrected and reuploaded) or whether the value was initially coded incorrectly (in which case the value would be converted to the code for a missing value when it is impractical to reanalyze the specific specimen, as is usually the case).

Data Integration

Data integration proceeds by first selecting the datasets to be integrated. The user then chooses the variables to be integrated by selecting from a list of ontologies represented in the selected datasets (the process can only integrate variables that are mapped to shared ontologies). The user can then select one or more "display variables" for each dataset. These variables, for example, site identifier, provenience identifier, time period, or project name, are included in the output dataset but are not otherwise processed. The user also has the option of identifying a "count" variable (indicating that this row of the database represents that number of identical observations with respect to the variables recorded) used to statistically weight a case.

Finally, the user has the opportunity to control how the integration operates for each integration variable. For each variable in turn, the software displays all ontology values with check marks indicating which values are present in which datasets. The user then selects the ontology nodes to be output (Figure 6). Whenever a node is selected, tDAR automatically aggregates

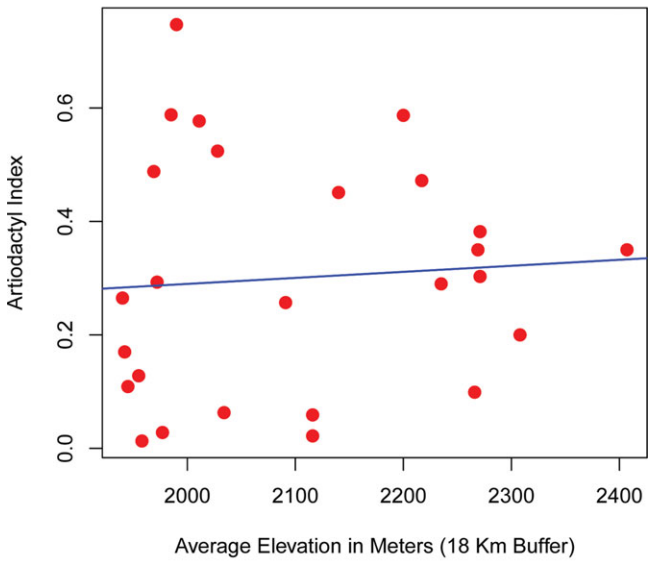

FIGURE 5. Scatterplot of average site/project elevation vs. artiodactyl index for sites above 1,900 m. Regression $R^2 < 0.01$; $p = 0.72$.

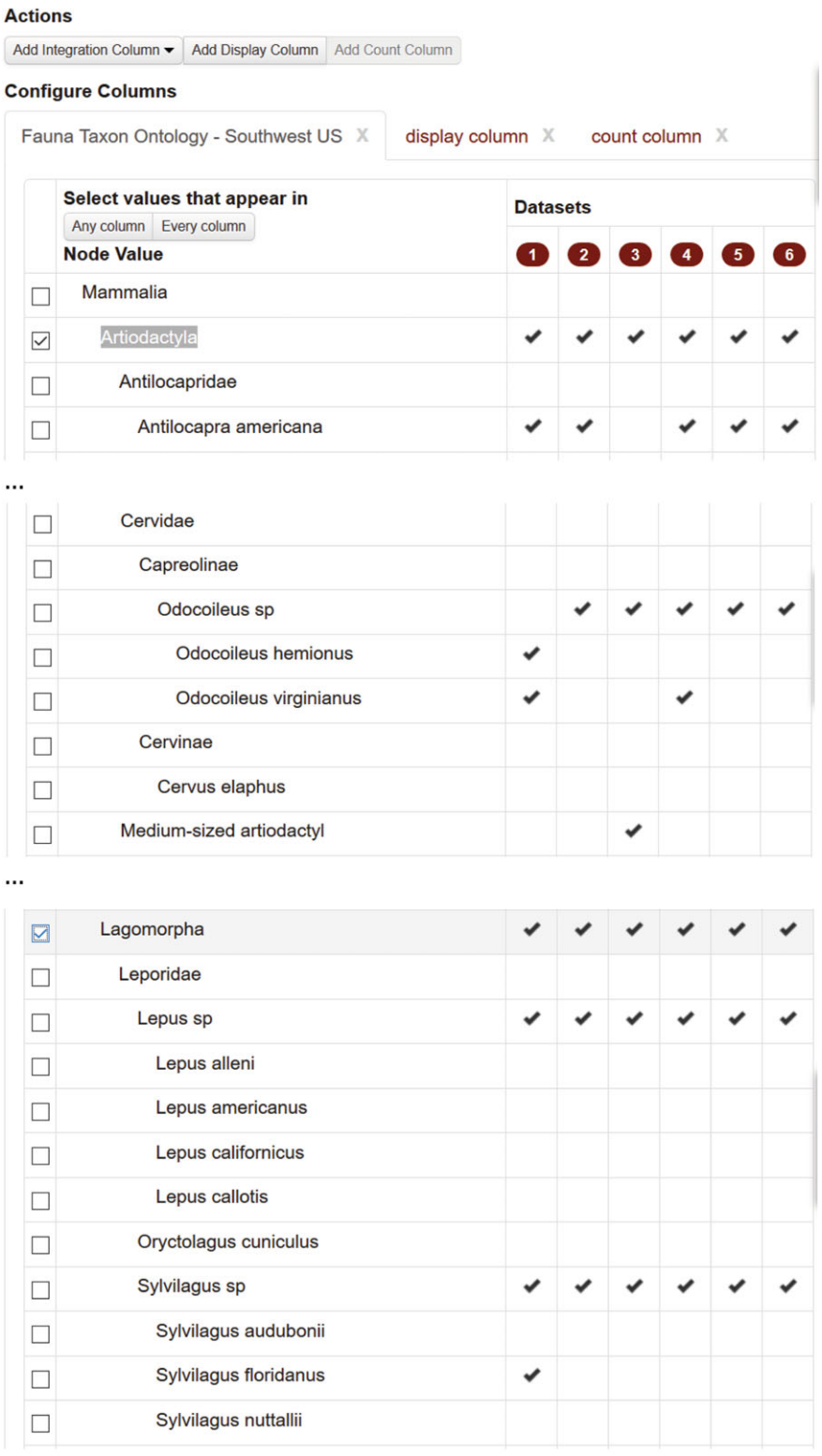

FIGURE 6. Screenshot of the Digital Archaeological Record data integration window (partial).

into the selected node all cases with unselected values that are hierarchically below it.

DATA REUSE, DATA INTEGRATION, AND ANALYTICAL COMPARABILITY

If, as a discipline, we are to accomplish syntheses that advance our scholarly and public objectives, we need to become more serious about sharing data in ways that promote data reuse (Kansa and Kansa 2013).

Data Reuse

The reuse of datasets depends on a number of factors:

- *Relevance*. The relevance of the dataset to the research questions, geographically, temporally, in terms of material and sample size, and on other substantive grounds.
- *Discoverability*. The ability of a potential user to discover the existence of a dataset and evaluate its relevance.
- *Accessibility*. Once a dataset is discovered, the ability to acquire a copy of it or to otherwise analyze it and the related costs.
- *Adequacy of Metadata*. Datasets can be rendered useless by inadequate metadata—the information that documents the content of the dataset, at the level of the dataset as a whole, the individual tables, the columns, and the nominal values that appear. For example, if the dataset uses arbitrary codes to represent nominal values, such as species or ceramic type, and the coding key is not documented in the dataset or its metadata, there is no meaningful access to the data.
- *Availability of Contextual Information*. The availability of key contextual information is important for establishing the analytical comparability of datasets and their constituent observations. This information includes spatial location, temporal assignment, depositional context, sampling intensity, and recovery technique. Too often, this contextual information is not a part of stand-alone analytical datasets.
- *Ease of Use*. Datasets may be difficult to use because they are stored in obsolete or obscure formats or because they are structured in a way that makes them difficult to employ in quantitative analyses without extensive data cleaning and reorganization. In contrast, datasets that are encoded in widely used formats and that employ or are linked to standards or other shared vocabularies or ontologies are more easily and effectively reused.

Data Integration and Data Reuse

Use of the data integration tools described here greatly facilitates data reuse. And because each data integration is responsive to the specific demands of the query, the datasets are exploited to their maximum potential. While there is a substantial one-time investment in developing the ontologies and mapping datasets to them, the payoffs can be enormous. Consider the examples presented above. If one simply had a copy of each original dataset in whatever form it was last used along with a pdf of the coding key, the example analyses described above that we completed in minutes would have taken literally months of effort. At any time, new or revised queries can quickly and easily be run on all the datasets mapped to the same ontologies. As new datasets are added to tDAR and mapped to the ontologies, the data integration queries can be saved and easily rerun to incorporate the new data. In this way, research communities can, over time, build ever more powerful integrated datasets.

User communities can range tremendously in scale. For example, faunal analysis lends itself to broad generalization because it deals mostly with biological characteristics, some of which (notably taxon and element) have established standards.

At the other end of the scale, Kintigh has worked with a number of students and close colleagues on several survey and excavation projects in the Cíbola area. At any particular time, the research teams shared a single coding sheet for recording ceramic type and form. However, over time, the forms evolved as they investigated new areas and refined some of their observations. To combine the results of the ceramic recordings for these projects, the easiest—and best—solution was to upload them separately to tDAR with their original coding sheets. In this way, tDAR preserves and maintains the data as originally recorded. It was then easy to develop type and form ontologies that captured the variation in the coding across the projects. Having mapped the project-specific coding sheets to the ontologies, the data integration tool made it easy to obtain a unified dataset with any desired aggregation of categories. In this case, the integration involved six projects and 11 datasets (survey and excavation were sometimes in different datasets), with about 240,000 individually recorded potsherds. This integrated database not only is easier for Kintigh (2016) and his immediate colleagues to use but is freely available for reuse.

Establishing Analytical Comparability

If the datasets have strong column- and value-level metadata and the mappings of dataset values to ontologies are reasonably consistent, data integration is highly effective at the variable level. However, data comparability also depends on contextual (including time, space, depositional context, sampling intensity) and taphonomic characteristics of the datasets as a whole—information that is often not directly documented in the datasets. For example, if an excavation project screened all deposits and a testing project recovered large numbers of artifacts from backhoe trenches without screening, then quantitative comparability is lost. More subtly, datasets may differ in their mix of contexts—one might have excavated largely room contexts, and another, mostly midden contexts. In this case, one would need to determine whether observed differences between the datasets are due to the different kinds of contexts investigated or to actual differences in the sites themselves.

Datasets, or contexts within them, may also differ in terms of the taphonomic processes that shaped the formation of the collections. Because of taphonomic differences, even consistently recorded datasets representing similar mixes of depositional contexts may not be comparable with respect to some kinds of questions. While these comparability problems can never be ignored, being able to integrate datasets in the way that we have proposed enormously facilitates their resolution. For our analysis of the collections from 33 Southwestern sites described above, we developed an analytical protocol that uses variables recorded in most or all datasets to evaluate statistically the

taphonomic comparability of the datasets for different purposes (Clark 2014).

What too often limits the assessment of analytical comparability in practice is that datasets—especially those derived from specialist analyses—often do not contain contextual data about the proveniences investigated, because the analysts did not have this information in the first place or it was never later integrated with the specialist data. The absence of contextual data makes the datasets less useful than they otherwise would be. Of course, this issue is not a technical limitation of the databases or of tDAR's data integration tool. It indicates a serious deficiency in the workflows that produce, analyze, and archive digital data (McManamon et al. 2017).

Costs and Responsibilities

Making datasets suitable for reuse entails planning, effort, and some direct costs. The direct cost of depositing a dataset in tDAR is low ($10 or less for a 10-MB database). The larger cost is in the effort devoted to properly preparing and documenting a dataset in a way that it can responsibly be used by others. Making datasets accessible and suitable for reuse is an ethical responsibility according to the "Society for American Archaeology Principles of Archaeological Ethics" (Society for American Archaeology 1996; discussed in Kintigh 2006). Similar ethical responsibilities are laid out by the Chartered Institute for Archaeologists, the Register of Professional Archaeologists, and the European Association of Archaeologists. In many cases, there is also a legal obligation to make publicly funded data accessible (Cultural Heritage Partners 2012).

Depositing a dataset in tDAR immediately provides for easy discovery, both through tDAR's Web interface and through Google and other search engines. While tDAR users need to register (at no cost), all use of data in tDAR, including downloads of datasets, is free. Although tDAR does not force every depositor to provide ideal metadata with a dataset, its interactive interface prompts the depositor to provide thorough metadata at the dataset, column, and value levels. Not only are those metadata available to any subsequent user; they are directly exploited by the data integration tools. Mapping data values to shared ontologies so they can be used in data integration constitutes another level of metadata. That is, each mapping constitutes an assertion that a given value of this particular variable is reasonably equivalent to the value in this node of the shared ontology. We are not aware of any other repository that provides such comprehensive tools to gather critical dataset metadata or the ease of use and analytical power of tDAR's data integration software.

CONCLUSIONS

To answer many of the most pressing questions of concern to archaeologists, to scientists more generally, to policy makers, and to the broader publics to which we are responsible, archaeology needs to conduct synthetic research. That synthetic research requires that we integrate primary data from multiple projects that do not typically collect data in completely consistent ways. As a result, we must have means of integrating observations across datasets in ways that maintain their semantic integrity. However data integration is accomplished, it places heavy demands on the metadata that document not only the tables, columns, and values but also collection procedures and other information often not contained in the datasets themselves.

As data integration has traditionally been done, it is a highly time-consuming and often frustrating endeavor. The efforts are typically one-off and are not readily built upon. Through this article we have sought to draw attention to data integration as an important component of our disciplinary analytical processes. We have also sought to highlight what we believe are unique tools in tDAR that make it possible, with reasonable, incremental efforts, to integrate very large numbers of datasets in ways that are directly expandable. tDAR's query-directed, on-the-fly data integration tools not only enable the kinds of synthetic research that we need; they facilitate many other kinds of analysis, and they greatly foster data reuse.

Acknowledgments

This article draws on the collective efforts of Arizona State University archaeology and computer science colleagues who contributed to our initial National Science Foundation proposals. In addition to the authors, these individuals include John Anderies, Chitta Baral, Huiping Cao, George Cowgill, Hasan Davulcu, James DeVos, Michelle Hegmon, John Howard, Subbarao Kambhampati, Allen Lee, Peter McCartney, Francis McManamon, Ben Nelson, Margaret Nelson, Charles Redman, Arleyn Simon, and Sander van der Leeuw. Recent conversations with Jeffrey Altschul have also helped frame this discussion. The article was substantially improved based on comments from the editor and three anonymous reviewers.

The members of the Southwest Faunal Working Group contributed data used in this article and helped refine the data integration workflow: Nancy Akins, Robin Cordero, Kathy Roler, Vincent LaMotta, Barnet Pavao-Zuckerman, Karen Gust Schollmeyer, and Christine Szuter. In addition, Linda Cordell, Jonathan Driver, Barbara Mills, Alison Rautman, Kari Schmidt, the Arizona State Museum, the Crow Canyon Archaeological Center, Los Alamos National Laboratory, and the School for Advanced Research provided datasets used in the analyses. We are very grateful to them and the members of the North American, Eastern US Archaic, and UK faunal working groups, which together developed the faunal ontologies used here.

This article is an extended version of a paper presented in the invited session "The Future of Big Data in Archaeology," organized by Erick Nolan Robinson at the 82nd Annual Meeting of the Society for American Archaeology, Vancouver, British Columbia, on March 31, 2017.

This material is based upon work supported by the National Science Foundation under grant numbers 0433959, 0624341, 1016921, and 1153115 awarded to Arizona State University and grant number 1353727 awarded to Indiana University of Pennsylvania. It is also based on work supported by a joint award, PX-50022-09, from the National Endowment for the Humanities and the Joint Information Systems Committee (UK). Permits were not required for this work.

Data Availability Statement

All data used in this article are available through tDAR: The Digital Archaeological Record. Links to specific data resource references are supplied at appropriate places within the article text.

REFERENCES CITED

Altschul, Jeffrey H., Keith W. Kintigh, Terry H. Klein, William H. Doelle, Kelley A. Hays-Gilpin, Sarah A. Herr, Timothy A. Kohler, Barbara J. Mills, Lindsay M. Montgomery, Margaret C. Nelson, Scott G. Ortman, John N. Parker, Matthew A. Peeples, and Jeremy A. Sabloff
nd Fostering Collaborative Synthetic Research in Archaeology. *Advances in Archaeological Practice*, in press. https://doi.org/10.1017/aap.2017.31.
2017 Fostering Synthetic Research in Archaeology to Advance Science and Benefit Society. *Proceedings of the National Academy of Sciences* 114:10999–11002. DOI:10.1073/pnas.1715950114.

Arbuckle, Benjamin S., Sarah Whitcher Kansa, Eric Kansa, David Orton, Canan Çakırlar, Lionel Gourichon, Levent Atici, Alfred Galik, Arkadiusz Marciniak, Jacqui Mulville, Hijlke Buitenhuis, Denise Carruthers, Bea De Cupere, Arzu Demirergi, Sheelagh Frame, Daniel Helmer, Louise Martin, Joris Peters, Nadja Pöllath, Kamilla Pawłowska, Nerissa Russell, Katheryn Twiss, and Doris Würtenberger
2014 Data Sharing Reveals Complexity in the Westward Spread of Domestic Animals across Neolithic Turkey. *PLoS ONE* 9:e99845. DOI:10.1371/journal.pone.0099845.

Archaeology Data Service and Digital Antiquity
2013 *Caring for Digital Data in Archaeology: A Guide to Good Practice.* Oxbow Books, Oxford. Electronic document, http://guides.archaeologydataservice.ac.uk/.

Clark Tiffany
2014 A Faunal Taphonomic Protocol for the Southwestern US. Electronic document, https://core.tdar.org/document/437822.

Clark, Tiffany C., Gregson Schachner, Suzanne L. Eckert, Todd L. Howell, and Deborah L. Huntley
2006 Rudd Creek Pueblo: A Late Tularosa Phase Village in East Central Arizona. *Kiva* 71:397–428.

Cultural Heritage Partners
2012 Federal Laws and Regulations Requiring Curation of Digital Archaeological Documents and Data. Report prepared for the Office of General Counsel, Arizona State University. Electronic document, http://www.digitalantiquity.org/wp-uploads/2013/05/2013-CHP-Legal-Analysis-of-Fed-Req-for-Curation-of-Dig-Arch-Docs-Data-.pdf, accessed October 17, 2017.

Dean, Jeffrey S., and John C. Ravesloot
1993 The Chronology of Cultural Interaction in the Gran Chichimeca. In *Culture and Contact: Charles C. Di Peso's Gran Chichimeca*, edited by Anne I. Woosley and John C. Ravesloot, pp. 83–103. Amerind Foundation, Dragoon, Arizona; and University of New Mexico Press, Albuquerque.

DiPeso, Charles C.
1974 *Casas Grandes: A Fallen Trading Center of the Gran Chichimeca*, Vols. 1–3. Amerind Foundation Series No. 9. Northland Press, Flagstaff, Arizona.

Duff, Andrew Ian
2002 *Western Pueblo Identities: Regional Interaction, Migration, and Transformation.* University of Arizona Press, Tucson.

Kansa, Eric C., and Sarah W. Kansa
2013 We All Know that a 14 Is a Sheep: Data Publication and Professionalism in Archaeological Communication. *Journal of Eastern Mediterranean Archaeology and Heritage Studies* 1:88–97.

Kansa, Eric C., Sarah W. Kansa, and Benjamin Arbuckle
2014 Publishing and Pushing: Mixing Models for Communicating Research Data in Archaeology. *International Journal of Digital Curation* 9(1):57–70. DOI:10.2218/ijdc.v9i1.301.

Kansa, Sarah W.
2015 Using Linked Open Data to Improve Data Reuse in Zooarchaeology. *Ethnobiology Letters* 6:224–231.

Kintigh, Keith W.
2013 Sustaining Database Semantics. In *CAA 2010: Fusion of Cultures. Proceedings of the 38th Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, Spain, April 2010*, edited by F. Contreras, M. Farjas, and F. J. Melero, pp. 585–589. BAR International Series 2494. British Archaeological Reports, Oxford.
2016 Cibola Prehistory Project Integrated Ceramic Data. DOI:10.6067/XCV8NS0XGV.

Kintigh, Keith W. (editor)
2006 The Promise and Challenge of Archaeological Data Integration. *American Antiquity* 71:567–578.

Kintigh, Keith W., Jeffrey H. Altschul, Mary C. Beaudry, Robert D. Drennan, Ann P. Kinzig, Timothy A. Kohler, W. Fredrick Limp, Herbert D. G. Maschner, William K. Michener, Timothy R. Pauketat, Peter Peregrine, Jeremy A. Sabloff, Tony J. Wilkinson, Henry T. Wright, and Melinda A. Zeder
2014a Grand Challenges for Archaeology. *American Antiquity* 79:5–24.
2014b Grand Challenges for Archaeology. *Proceedings of the National Academy of Sciences* 111:879–880. DOI:10.1073/pnas.1324000111.

Kintigh, Keith W., Jeffrey H. Altschul, Ann P. Kinzig, W. Fredrick Limp, William K. Michener, Jeremy A. Sabloff, Edward J. Hackett, Timothy A. Kohler, Bertram Ludäscher, and Clifford A. Lynch
2015 Cultural Dynamics, Deep Time, and Data: Planning Cyberinfrastructure Investments for Archaeology. *Advances in Archaeological Practice* 3:1–15. DOI:10.7183/2326-3768.3.1.1.

Kintigh, Keith W., Donna M. Glowacki, and Deborah L. Huntley
2004 Long-Term Settlement History and the Emergence of Towns in the Zuni Area. *American Antiquity* 69:432–456.

Kintigh, Keith W., Todd L. Howell, and Andrew Duff
1996 Post-Chacoan Social Integration at the Hinkson Site, New Mexico. *Kiva* 61:257–274.

McKechnie, Iain, Dana Lepofsky, Madonna L. Moss, Virginia L. Butler, Trevor J. Orchard, Gary Coupland, Fredrick Foster, Megan Caldwell, and Ken Lertzman
2014 Archaeological Data Provide Alternative Hypotheses on Pacific Herring (*Clupea pallasii*) Distribution, Abundance, and Variability. *Proceedings of the National Academy of Sciences* 111:E807–E816. DOI:10.1073/pnas.1316072111.

McManamon, Francis P., Keith W. Kintigh, Leigh Anne Ellison, and Adam Brin
2017 tDAR: A Cultural Heritage Archive for Twenty-First-Century Public Outreach, Research, and Resource Management. *Advances in Archaeological Practice* 1–12. DOI:10.1017/aap.2017.18.

Manning, Katie, Sue Colledge, Enrico Crema, Stephen Shennan, and Adrian Timpson
2016 The Cultural Evolution of Neolithic Europe. EUROEVOL Dataset 1: Sites, Phases and Radiocarbon Data. *Journal of Open Archaeology Data* 5:e2. DOI:10.5334/joad.40.

Mills, Barbara J., Jeffery J. Clark, Matthew A. Peeples, Wm. R. Haas, John M. Roberts, J. Brett Hill, Deborah L. Huntley, Lewis Borck, Ronald L. Breiger, Aaron Clauset, and M. Stephen Shackley
2013 Transformation of Social Networks in the Late Pre-Hispanic US Southwest. *Proceedings of the National Academy of Sciences* 110(15):5785–5790. DOI:10.1073/pnas.1219966110.

Richards, Julian D.
2017 Twenty Years Preserving Data: A View from the United Kingdom. *Advances in Archaeological Practice* 5:227–237. DOI:10.1017/aap.2017.11.

Schachner, Gregson
2012 *Population Circulation and the Transformation of Ancient Zuni Communities.* University of Arizona Press, Tucson.

Schollmeyer, Karen Gust, and Jonathan C. Driver
2013 Settlement Patterns, Source-Sink Dynamics, and Artiodactyl Hunting in the Prehistoric U.S. Southwest. *Journal of Archaeological Method and Theory* 20:448–478.

Society for American Archaeology
1996 Society for American Archaeology Principles of Archaeological Ethics. *American Antiquity* 61:451–452.
Spielmann, Katherine A., and Keith W. Kintigh
2011 The Digital Archaeological Record: The Potentials of Archaeozoological Data Integration through tDAR. *SAA Archaeological Record* 11:22–25. Electronic document, http://digitaleditions.sheridan.com/publication/?i=58423&p=24, accessed October 17, 2017.
Watson, Patty Jo, Steven A. LeBlanc, and Charles L. Redman
1980 Aspects of Zuni Prehistory: Preliminary Report on Excavations and Survey in the El Morro Valley of New Mexico. *Journal of Field Archaeology* 7:201–218.
Whalen, Michael E., and Paul E. Minnis
2001 *Casas Grandes and Its Hinterland: Prehistoric Regional Organization in Northwest Mexico*. University of Arizona Press, Tucson.

NOTE

1. The Open Context data publishing platform similarly recognizes the need to integrate databases across projects and has used archaeological fauna as a major focus of application (Arbuckle et al. 2014; Kansa 2015), employing a "linked open data" approach to data integration.

AUTHORS INFORMATION

Keith W. Kintigh ■ School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85287-2402, USA (kintigh@asu.edu, corresponding author)

Katherine A. Spielmann ■ School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85287-2402, USA (kate.spielmann@asu.edu)

Adam Brin ■ Center for Digital Antiquity, Arizona State University, Tempe, AZ 85287-2402, USA (abrin@digitalantiquity.org)

K. Selçuk Candan ■ School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287–8809, USA (candan@asu.edu)

Tiffany C. Clark ■ Applied EarthWorks, Inc., Pasadena, CA 91107–3414, USA (tclark@appliedearthworks.com)

Matthew Peeples ■ School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85287-2402, USA (matthew.peeples@asu.edu)