Other People's Data: A Demonstration of the Imperative of Publishing Primary Data

Levent Atici[a*], Sarah Whitcher Kansa[b], Justin Lev-Tov[c], and Eric C. Kansa[d]

[a] Department of Anthropology, University of Nevada, Las Vegas, Las Vegas, NV 89074, U.S.A.

[b] The Alexandria Archive Institute, San Francisco, CA 94127, U.S.A.

[c] Statistical Research, Inc., Redlands, CA, 92374, U.S.A.

[d] School of Information, University of California Berkeley, Berkeley, CA 94720, U.S.A.

[*] Corresponding Author. Department of Anthropology, University of Nevada, Las Vegas. 4505 S. Maryland Pkwy Mail Stop 455003 Las Vegas, NV 89074, USA. Tel. +1-702-445-9149. E-mail address: Levent.Atici@unlv.edu (Levent Atici).

Abstract

Many scholars avoid employing "legacy" datasets, even if accessible, because of perceived unknowns in using data collected by others. This study explored issues in using data generated by other analysts. Three researchers independently analyzed a legacy, decades-old zooarchaeological dataset and then compared their analytical approaches and results. Although they took a similar initial approach to determine the dataset's suitability for analysis, the three researchers generated markedly different interpretive conclusions. In examining how researchers use legacy data, this paper highlights interpretive issues, data integrity concerns, and data documentation needs. In order to meet these needs, this paper proposes greater formalism and professional recognition for data dissemination, favoring models of "data publication" over "data sharing" or "data archiving".

Key Words

Data integrity, Blind test, Faunal analysis, Legacy data

1. Introduction

While archaeologists routinely manage complex and highly structured digital data, dissemination and communication objectives remain decidedly oriented toward print or digital analogs of printed documents (PDFs). The prevailing norms and expectations for print publication mean that researchers tend not to share the raw data they collect, thus precluding reuse and reexamination of these data. While data sharing is still rare, it is gaining traction as a key issue in scientific communications (Costello, 2009; Nature Editors, 2009). Scholars have discussed a multitude of semantic (Kintigh, 2006), technological (Snow et al., 2006), data preservation and longevity (Carraway, 2011; Richards, 2004), intellectual property (Kansa et al., 2005), and professional incentive concerns (Costello, 2009; Kansa, 2010) regarding data sharing. While most see data sharing as an important goal, much attention focuses on problems relating to supplying researchers with data, and less on how researchers can best consume and reuse data. Despite wide acknowledgement that approaches to data collection, recording, analysis, presentation, and interpretation vary among researchers, few studies have explored challenges researchers may face in the analysis of datasets produced by others.

Recent policy changes, such as the National Science Foundation (NSF) requiring "data management plans" of all grant-seekers (See NSF press release (May 10, 2010): http://www.nsf.gov/news/news_summ.jsp?cntn_id=116928), promise to raise the professional stakes in data sharing. In light of the increasing professional significance of data sharing, this paper highlights the perspective of "end users" who consume and seek to reuse data. We compare analytic results of three zooarchaeologists who conducted blind analysis on an "orphaned" faunal dataset from the 1960s. Our results highlight how publishing original datasets can be of value to future data consumers. Ideally data dissemination should be accompanied by

published syntheses, but our results also demonstrate that data sharing can be useful even in cases without accompanying reports, provided datasets have minimal documentation and demonstrate sufficient quality.

We show some of the analytical value of data sharing by demonstrating the diversity of interpretive outcomes when different researchers analyze the same dataset independently of each other. We find that no two zooarchaeologists will analyze the same dataset in exactly the same way. This point comes as no surprise, since research outcomes are heavily influenced by the research questions, and the analyst's background and choice of analytical methods. This alone provides a strong argument for primary data that contribute to a synthetic analysis to be shared in full. However, some researchers doubt whether "other people's data" can be used by others. Our study demonstrates that seeking for obvious indicators of quality and a level of integrity sufficient to permit analysis is a critical, yet far under-appreciated, aspect to using data produced by others. It is, in fact, the essential element that allows for intelligible reuse of a dataset whether to replicate or to explore new directions beyond the original analyst's work. Though this study involved zooarchaeological data, the implications of our results reach far beyond the scope of archaeology and our recommendations may apply to other disciplines.

2. Background

Like many field sciences, zooarchaeological research involves different stages, including sorting, identification, recording, quantification, analysis, and interpretation. These stages are usually interdependent and associated with primary and secondary data. Reitz and Wing (2008: 153) define primary data as "observations that cannot be replicated by subsequent investigators, such as element representation and taxonomic identification." Lyman, too, distinguishes between

primary and secondary data or fundamental and derived measurements (Lyman, 1994a, 2008). He defines the former as "an easily observed property of a phenomenon" (Lyman, 2008: 12) or as "observational units or empirical manifestations" (Lyman, 1994a). Primary data is associated with identification stage, whereas secondary data is associated with analysis and interpretation stages. Because secondary data are derivatives or abstractions of multiple primary data, they are more complex and sometimes difficult to directly observe, requiring additional levels of analytical manipulation and subsequent interpretation (Reitz and Wing, 2008).

The type of primary data collected from an assemblage will depend on the zooarchaeologist doing the work. The interpretive impact of various analytical decisions was demonstrated clearly by Gobalet (2001) in a blind analysis of an archaeological fish bone assemblage by researchers with different training and experience, as well as by the discussions between Turner (1989) and Klein (1989) concerning the Klasies River Mouth faunal assemblage. Rigorous and detailed recording of primary data using adequate samples as defined by Gamble (1978) will ensure the success of zooarchaeological research. In addition, budget and time constraints and inaccessibility of sites or archaeofaunal assemblages due to various factors can be avoided if analysts collect as much primary data as possible during the field or laboratory recording and identification stages.

Decisions regarding what and how to record vary depending upon the site (including its temporal period(s) and geographic location), recovery methods, bone sample sizes, experience of the analyst, and more importantly, the research design and the questions being asked (see discussions in Chaplin, 1971; Davis, 1987; Driver, 1991; Grigson, 1978; Hesse and Wapnish, 1985; Klein and Cruz-Uribe, 1984; Lyman, 1994a, b, 2008; Meadow, 1980; O'Connor, 2003; Reitz and Wing, 2008; Ringrose, 1993; Speth, 1983; Thomas, 1996; Uerpmann, 1973;

Uerpmann, 1978). Although a universal methodology that is employed by all zooarchaeologists and that is applicable to all faunal collections across time and space does not exist, many zooarchaeologists collect certain basic primary data such as those on skeletal part, taxon, symmetry, state of epiphyseal fusion, and nature of dental eruption/wear patterns. However, that these basic variables represent a minimal baseline, and zooarchaeologists may significantly differ in the complexity and amount of additional primary data they choose to record. Guidelines published over the past few decades have helped the zooarchaeological community work toward identifying priorities in data collection (see Clutton-Brock, 1975; Driver, 1991; Grigson, 1978; Meadow, 1978). The near-ubiquity of digital formats today calls for additional guidelines, taking into account the potential of the Web for communicating research and its implications for archiving and reuse of datasets (see Kansa et al., In Preparation).

Following the time-consuming and sometimes decades-long process of collecting and examining a zooarchaeological assemblage, researchers typically publish the synthetic results of the study. Print publication normally does not provide the space for listing full datasets that contribute toward the synthetic analysis. Thus, the primary dataset upon which the study was based is left behind. Readers are left with the interpretations and certain select (and usually summarized) tables based on the abstractions of raw data, but no means by which to return to the raw data to replicate the results or ask different questions not addressed by the original analyst. Lack of full disclosure of primary data results in poor scientific practices because others cannot reproduce the original analyst's results or cannot independently test a scientific conclusion (Costello, 2009).

The fact that even primary data carries implicit biases does not necessarily argue against the need for more and better data sharing. Despite the ubiquity of biases, researchers still find communication and sharing of results central to their specific goals, as well as to the broader

goal of incorporation of archaeological data into large-scale, multidisciplinary studies (see Amorosi et al., 1996). In this paper, we show that sharing the primary data allows us to better confront some of the biases in the data collection and analysis process, and do more informed research, rather than simply taking the interpretive publication at face value. Because vast quantities of primary data can be shared on the Web, the research community urgently needs strategies to best use other people's data, especially ways to document and describe primary data in ways that improve subsequent reuse by other investigators. This studies use of a "legacy dataset", once stored in an old and obsolete file format, then made available via the Web, highlights critical interpretive challenges and documentation needs for data dissemination and reuse.

3. Methods and Materials: The Blind Test

This project uses the publicly available dataset of over 30,000 animal bone specimens from excavations at Chogha Mish, Iran during the 1960s and 1970s. The specimens were identified by Jane Wheeler Pires-Ferreira in the 1960s and while she never analyzed the data or produced a report, her identifications were saved and later transferred to punch cards and eventually to Excel spreadsheets that we ultimately used. This "orphaned" dataset was made available on the web in 2008 by Abbas Alizadeh (University of Chicago) at the time of his publication of *Chogha Mish, Volume I I.* The full dataset was made available as a downloadable spreadsheet on the Oriental Institute Publications webpage for Chogha Mish Volume II (http://oi.uchicago.edu/research/pubs/catalog/oip/oip130.html). This original spreadsheet is also available in Open Context at the project page for Chogha Mish: http://opencontext.org/projects/497ADEAD-0C2A-4C62-FEEF-9079FB09B1A5.

Three researchers, each with over fifteen years of experience working with Near Eastern zooarchaeological assemblages, carried out a blind analysis on this dataset. Guidelines were minimal; researchers were told to use their own approach and carry out any analysis they deemed relevant, interesting, and feasible with the given dataset. The analysts documented the full process, from data cleaning to interpretation, and had no contact with each other or discussion of the project until concluding their independent analyses. The analysts then met in person to compare their methodological approaches, discuss their findings, and develop a collaborative analysis plan. Results of this project, including a link to the revised original dataset, and the data discussed in this analysis are available in their entirety in Open Context (Atici et al., 2010; Wheeler Pires Ferreira et al., 2010).

4. Results

The three zooarchaeologists all began their analysis by taking an inventory of the database to judge its overall "quality." This included checking for misspellings, mismatched taxon /element pairings, and errors that may have occurred in the translation to punch cards. While all three faunal analysts determined that the quality of the data was sufficient to move forward with analysis, they lamented that certain data were not present, specifically contextual and methodological information.

Verification of data integrity was a key step in this analysis. Data integrity refers to the internal consistency and structural coherence of a dataset, as well as data quality issues. If structural coherence was not maintained, then the values in different database fields could become muddled. Given the history of this dataset, where the information migrated from punch cards to Excel, structural coherence was a major concern. Similarly, if recording practices were not

sufficiently consistent, or fields showed too many errors (misplaced decimal points, too many instances where data values entered into obviously wrong fields), then the data would seem to lack integrity, making subsequent interpretations suspect. Since this dataset showed sufficient integrity, the three analysts decided to proceed and invest effort in its analysis. All three analysts independently determined that the only research question that could be addressed by the assemblage involved economic changes over time, as reflected in broad patterns in the relative proportions of taxa and ages in different periods. To address this broad question, each analyst manipulated the dataset, based on individual assumptions and operational decisions dictated by the lack of contextual and methodological background and information. Inter-analyst variation included: decisions about aggregation of phases and taxa; judgments about data reliability, consistency, and comparability; and the "threshold" at which the researcher decided s/he had made too many assumptions and could not conduct further analyses. Analytically, all three chose to look at relative proportions of taxa by period, then at the overall age at death of principal taxa, and finally at butchery and fragmentation patterns. Thus, all three researchers took the same approach to extract from the dataset some of the basic information they would seek in any dataset. Their results, however, varied considerably.

4.1 Data survey and cleaning

While many of the same inconsistencies in the original dataset were noted by all analysts, some found other problematic data where others did not. This led to different approaches to "cleaning" the dataset, where an inconsistency observed and omitted by one analyst was left in the dataset by another. Many of the checks centered on accuracy and consistency regarding fragmentation and portion designation. Below are some examples of data cleaning procedures used, in varying degrees, by the different analysts:

- All cranial elements and/or elements not bearing epiphyses such as carpals or astragali were assigned "not applicable" value with respect to fusion proximal and distal. Similarly, all astragali have been revised to eliminate incorrect directional terminology such as "proximal end with shaft."
- The Proximal/Distal field was revised and edited to assign each specimen to either proximal or distal categories to tally epiphyseal fusion stages.
- Specimens were assigned complete, nonidentified, not applicable, shaft, distal shaft, or proximal shaft as well. For example all cranial skeletal elements with assigned "proximal end" or "distal articulation with epiphysis" sort of entries have been corrected as Proximal/Distal not applicable. For scapulae, Proximal/Distal data have been reversed to correct the anatomical orientation and directional terminology with respect to epiphyseal fusion. Thus, all proximal portions have been reentered as distal to record distal epiphysis fusion stage of scapulae.
- All specimens that were not identified to a skeletal element, and that were assigned a fusion state were reassigned "not applicable" to their fusion state field.
- Seven specimens identified as *Equus equus* have been renamed as *Equus caballus*, as there is no such species as *Equus equus* in the Genus *Equus*. It seems that the analyst meant "horse" by not using other more generic categories.
- One hundred sixty specimens identified as gazelle have been renamed *Gazella* sp. to conform the Latin genus name; two hundred and four specimens have been renamed from goat to *Capra* sp; seventy five specimens have been renamed from sheep to *Ovis* sp.; and three horn core specimens identified to medium mammal have been renamed to *Ovis*/*Capra* as this is the most appropriate category for medium horn cores.

4.2 Quantification

Choices about data aggregation led to discrepancies even at the most fundamental level, as the analysts did not have any data regarding how the taxonomic and skeletal element/portion identifications were made at Chogha Mish. For example, in quantifying relative proportions of taxa by period, all three faunal analysts started with a different base dataset, one which they had "tidied up" before beginning their analysis. The Chogha Mish data suggest that the original analyst was very conservative and certain in her taxonomic identifications, since "Ovis/Capra/Gazelle," "large-size mammal," or "medium artiodactyl" account for large samples in many of the assemblages from almost all the periods. These are common "methodological categories" that zooarchaeologists employ when they lack confidence or certainty in identification. As discussed below, such categorization concerns need to be considered when crafting and applying ontologies (formalized conceptual systems) used in data sharing and data integration efforts.

As to basic quantification units, researcher assumptions and observations varied from the outset. The original analyst appears to have used fragment or specimen count as the basic unit of quantification, assigning a single line of data for every specimen. She did not designate a field to enter the number of specimens/fragments, as she did not group specimens for collective entry under a unique identification number.

4.3 Periodization

Designation in the dataset of some "periods" as "mixed" implies that there was a certain degree of mixing of the sediments and of their contents, either because of the insensitivity of the excavation method to the site's topography or because of site formation processes that resulted in

mixed deposits indistinguishable during excavation. The dataset offered only very limited descriptions of archaeological context. It presented contextual information only in vague terms, making it impossible to securely determine the compatibility of excavation units or the completeness and integrity of contexts. Moreover, the dataset lacked detailed description of recovery methods applied by the excavators. These gaps obliged all analysts to make some assumptions and guesses about data reliability and comparability with respect to context and stratigraphy.

All three analysts commented on the high potential for sampling bias and the discomfort they felt in working with data for which recovery techniques are unknown. All dealt with the potential effects of this bias by excluding certain periods from the analysis. However, the choices they made about which periods to exclude varied, thus impacting the interpretive results. For example, one analyst chose to subsume sub-periods into one broad period, while another left sub-periods as distinct units of analysis. These decisions relate to the comfort of the analyst when working with small sample sizes: the former researcher chose to avoid making assumptions based on small samples by working with lumped data, but thus took the risk of obscuring finer-grained patterns on a sub-period level; while the latter accepted the risks of working with smaller sample sizes in order to detect sub-period distinctions between the samples. The same was observed for choices the analysts made about how to deal with vague taxonomic identifications (such as "sheep/goat/gazelle" being left as is, lumped into "medium mammal," or omitted completely).

Figure 1 illustrates how each analyst made different decisions about how to aggregate periods and taxa in preparation for basic analytical tasks. One analyst included a wide variety of taxa (beyond genus-level determinations) and consolidated the cultural periods into only five groups

(Fig. 1a). The other two analysts similarly focused only on the predominant taxa (see Fig. 1b and 1c), but they took different approaches to combining the cultural periods. Thus, though they started out with the same taxonomic analytical units, by the time they added periodization, their results were different. This can lead to interpretive variability. For example, Figures 1b and 1c both show an increase in cattle in the Early Susiana. However, Figure 1a does not show a similar trend because of the analyst's more conservative approach to consolidating taxonomic categories by presenting a wide array of taxa. Since the analyst of Figure 1a did not lump "large mammal" and cattle together or did not proportionally allocate "large mammal" bones to various large taxa such as horse and camel, bones of medium and large mammals dominated the assemblages, masking the trends in species composition. Similarly, the analysts of Figures 1b and 1c document a steep decline in sheep-goat category from 90.2 percent during Archaic Susiana 2 to 56.8 percent during Early Susiana. This pattern can be accounted for by larger increases in cattle and pigs as documented by the analysts of those figures. Along the same line, Figure 1a compromises explanatory and interpretive power and resolution and does not account for the decline in sheep-goat category or increase in cattle and pig categories due to data aggregation decisions. Thus, these seemingly small choices can lead to vastly different interpretive results.

4.4 Finer-Grained Analyses

As each analyst had made so many different choices and assumptions in the early stages of analysis, the results of finer-grained analyses were incomparable. A useful example of the huge discrepancy in fine-grained analyses comes from estimations of age at death for sheep and goats. The original analyst used a category "DeathAge" where she listed the age of death in six-month increments, mainly for post-cranial elements but also for teeth in a few cases. One analyst relied on NISP counts, but noted that the original analyst recorded bone fusion status on many

specimens without assigning them an estimated "DeathAge." This analyst assigned an age of death to all the elements with fusion information, placed each in an age group (based on Silver 1969) and compared changes over time without collapsing sub-phases (Fig. 2a). Another analyst also relied on NISP counts and, like the first analyst, converted the original researcher's fusion data for ageable elements using Silver's (1969) work as a guide. This analyst also had five age groups for each period, but then chose to collapse sub-periods into broader periods (Fig. 2b). Finally, another analyst worked with the raw data from the original analysis and restructured and similarly inferred the epiphyseal fusion data by creating separate fields for proximal and distal epiphyses of long bones. However, in order to eliminate double-counting, this researcher used MNE values to estimate age-at-death and demographic profiles (Fig. 2c). The results were vastly different, not only because of the different approaches to quantifying the specimens, but also because of the different analysts' choices around aggregating taxa and periods and about which specimens to omit from the analysis (as described in the previous section). We must reiterate, however, that all three analysts consistently adopted a very conservative approach in order to minimize errors and biases, as the original analyst's methodologies were unknown.

5. Discussion: Recommendations and Critical Issues in Data Sharing

Blind analysis of the Chogha Mish faunal data has demonstrated what most researchers already assumed, that *any* dataset will see multiple interpretations depending on the analyst's research perspective and analytical decisions. Choices about data aggregation and splitting will depend on the research question(s) being asked. For example, regional syntheses differ from comparisons of faunal data with material culture at one site. Regional synthesis lend themselves to more "lumping" while more intrasite analysis / comparison with other classes of finds is often more fine-grained. Access to raw data is needed to make alternate aggregation options available.

Furthermore, published datasets should be well documented and analytical methods described in detail. For example, contextual information (time and place) *must* be provided with raw datasets in order to make them useful. In the case of Chogha Mish, the minimal contextual data available in the Chogha Mish faunal dataset could be supplemented through reference to related publications. Other critical information includes: the name of the original analyst, decoded data (or, at a minimum, use of a published code), and identification basics (taxon, element, portion, side, fusion, sex), as well as how identifications were derived (use of a physical reference collection, published studies used for determining age, etc.).

We would argue that, even in cases where analysis was never undertaken, it is essential to share raw datasets so that future generations of scholars can benefit from the work. However, broader archaeological/ anthropological questions require fine resolution data and adequate samples, so it is important to be aware of the potential uses and limitations of zooarchaeological data collected by other analysts. The current study is a case in point: laboratory identification of 30,000+ fragments reflects many months of work on the part of the original analyst. Analysis of this dataset forty years after the original identifications were made is not an ideal situation, but the information would have been lost entirely if project members had not taken the various steps along the way to preserve and, ultimately, post this unanalyzed dataset online. Fortunately, skilled analysts can find clues in any dataset that vouch for its quality. Even in the absence of detailed descriptive information about a dataset, some basic analysis can be conducted to inform archaeological interpretation. At the very least, the analysis of the Chogha Mish dataset has revealed some broad temporal and spatial trends that can be useful for understanding dietary and economic shifts in the region and at the site itself (see Lev-Tov et al., In Preparation). Rather than ignore "old" or unanalyzed datasets such as the one used in this study, researchers should

take more care to ensure that the dataset being used can sufficiently address the research questions being asked (Amorosi et al., 1996), and if need be, modify the research questions to work with the broad patterns that legacy datasets offer.

This exercise has highlighted some of the difficulties in using another researcher's dataset; however, the challenges are greatly compounded when we consider use of datasets from many different projects. One of the most commonly articulated goals advanced by advocates of data sharing is "data integration"— pooling disparate datasets to enable analyses across data sources. Most data integration methods require use of a common "ontology" or a formally described conceptual system shared by members of a disciplinary community (such as the Linnaean classification system commonly used to describe biological taxa).

Zooarchaeology is rather unique in archeological sub-disciplines with respect to ontologies. Unlike many other specializations in archaeology, zooarchaeologists already have many common recording conventions (biological taxa, skeletal elements, fusion data, etc.). These common conventions should make it easier to apply common ontologies. Somewhat ironically, zooarchaeology's common recording conventions also make ontologies somewhat less necessary, at least with respect to the interpretation and use of a single legacy dataset (ontologies are more useful for comparing across multiple datasets). Because zooarchaeological recording conventions are widely shared, the three zooarchaeologists in this study had few difficulties in understanding the vocabularies and terminologies expressed in the Chogha Mish faunal dataset. Tacit knowledge common to zooarchaeology proved sufficient to decode the semantics of this specific dataset.

While ontologies are generally useful and necessary for integrating different datasets, this study helps to demonstrate challenges in their application, even in zooarchaeology. In our study, each analyst chose to lump and split the dataset in different ways. Each analyst made different choices with regard to taxonomic identifications, age determinations, and chronological distinctions. Similar variability should be expected in mapping data to a common ontology. For example, Digital Antiquity's tDAR project (http://tdar.org) aims for data integration by relating datasets to common ontologies. Our study helps demonstrate that while ontologies can be useful (and essential), we should expect that different analysts will make different choices in mapping to a common ontology. Thus, data integration outcomes should not be taken for granted, since their methods require potentially contestable judgment calls made by informed analysts. This, again, reinforces the point that datasets should be documented in as much detail as possible by the original analyst to allow for informed reuse.

As demonstrated above, an analyst will approach an archaeological assemblage with research questions and analytical biases that will differ from those of another researcher. Inherent biases in analytical approaches from the outset of a study necessarily lead to interpretive differences down the line. Before the advent of the Web and increased capacity for sharing the vast amounts of data accumulated in the practice of archaeology, scholars shared only a small portion of their primary data, along with as much detail on methodology as permissible within the limits of print publication. No matter how well documented, however, choice of methods and research perspectives shaped the resulting printed analysis. Thus, up to now, the interpretive publication of a dataset usually stood as the official and authoritative "last say" on the assemblage and rarely was there the opportunity to return to the primary data with new questions.

Most datasets are now "born-digital," giving the researcher community new opportunities for sharing them via the Web. Despite hesitation and incentive concerns (see Harley et al., 2010) we are witnessing a change in scholarly culture (Kansa and Kansa, 2011). There is a growing expectation for access to primary datasets so that other scholars can reanalyze them with new questions and perspectives. As data sharing assumes greater primacy in professional communications, researchers need to develop methods and analytic techniques to most effectively use shared data.

In exploring these "end-user" concerns, this paper seeks to inform discussion of how to better document and describe shared datasets so that they can persist as quality scholarly resources. To recapitulate, we demonstrated that legacy datasets can be useful for analysis and reuse, provided that they are accessible, have sufficient quality, and come with at least some minimum level of documentation. To encourage analytically-meaningful practices in data sharing, we recommend the following:

5.1 Encourage Professional Rewards

Data such as those discussed in this study cannot be reexamined or reused without dissemination. Scholars need professional rewards for sharing their data, and these incentives must override fears of being "scooped" or that data are not "ready" for viewing by others. Increased contribution to science and consequent peer recognition, reputation, employment and promotion opportunities, collaboration opportunities, citation rates, access to a far wider audience in the professional field are justifiable motivations and some of the professional rewards for and benefits of online data publication (Carraway, 2011; Costello, 2009: 420-422).

5.2 Explicit Open Licensing

In some ways data sharing has stronger requirements for intellectual property "openness" than more conventional publication. A traditional paper is a more or less a stand-alone artifact, meant to be read and understood as a whole. It makes little sense to literally copy the abstract and discussion of a paper and combine those with the results and conclusion sections of another paper. In contrast, a shared dataset has the potential to be sampled and combined and recombined with many other datasets in new analyses. To unlock the analytic potential of datasets, permission for such sampling and reuse needs to be legally guaranteed.

This study obtained specific permission for reuse of these data from Abbas Alizadeh. Because copyright defaults to "all rights reserve" which prohibits duplication and adaption of content, a dataset made available on the Web without explicit permissions is left in a legal limbo ("you can look but not touch")(Kansa et al., 2005). If Alizadeh did not respond to requests for permissions, legal ambiguities would have inhibited this study. Thus, an important aspect of data publication is intellectual property licensing. Creative Commons licensing explicitly gives permissions for reuse without the need for data owners to grant permission for each and every request for reuse. In other words, Creative Commons licensing removes an important "transaction cost" (the negotiation of permissions) in data-sharing. Thus, open-licensing is a key requirement for efficient data publication.

5.3 Data Sharing as Publication

As described in this paper, data quality and integrity determinations play an important role in shaping subsequent reuse. If a dataset lacks sufficient quality, attempts at reuse may be fraught with problems. Data sharing venues should therefore find ways to encourage higher data quality, be it through editorial oversight or even through user-rating systems or other means. Some data

integrity issues may be more obvious, and others may require some specialized background to notice. For instance, only a trained zooarchaeologist may notice impossible combinations of descriptions of biological taxa and bone elements. Table I describes some data integrity issues, as well as methods to detect and prevent problems.

In our attempt to improve data integrity and quality, we need to recognize that quality requires effort. Similar effort and expertise may be required to provide adequate description of a dataset (see below). Therefore, we recommend that data dissemination should take on more of the formal (and, hopefully, rewarded) trappings of "publication," rather than informal "sharing" (see also Kansa, 2010) or even "archiving". While any effort to share or archive data should be applauded, we believe that data dissemination should be a more regular and integral part of professional practice. Informally "shared data", without many of the scientific conventions of outside review and description, may lack adequate documentation for many analytic purposes. The term "data archiving" has similar problems with "data sharing" respect to providing incentives to offer adequate data documentation. Though "data archiving" clearly indicates preservation goals, it can convey a sense of "file away and forget", giving data contributors little sense their data and data documentation efforts will be recognized and rewarded.

If datasets are to be recognized as first-class research outputs, they need to be properly documented, published, and archived in citable venues, like traditional print publications (see also Costello, 2009). In contrast with print however, datasets need digital publication in venues far more open and permissive with intellectual property rights than typical scholarly journals (see above). Thus, instead of advocating for the dissemination of "raw data" we should advocate for comprehensive *publication* of cleaned, properly documented, and usable data in editorially-managed venues backed by digital repositories. Professionally recognized, editorially-managed

data dissemination channels can better communicate expectations of quality and relevance to a specific disciplinary domain (in this case, zooarchaeology). Finally, the more immediate rewards and recognition that may come with publication (as opposed to the less immediate benefits of "archiving") can, hopefully, help motivate researchers to work under the guidance of "data-editors" to adequately describe their datasets.

5.4 Adequate Documentation

Adequate documentation helps ensure informed reuse, and the right documentation can improve the confidence of future reuse of data. Given that resources and human effort is scarce, certain forms of documentation should be prioritized. For example, published data should include some "fundamentals," including discussion of methods, research aims, and data collection practices. Baseline contextual information (geographic, stratigraphic, chronological) also needs to be provided. Shared data needs to be decoded (or coding systems need detailed documentation) to facilitate informed reuse and comparison with other datasets. In this case, decoding happened well before our study, probably sometime in the transition from punch-cards to Excel. Without such decoding, the dataset may have been useless.

This study attempted analysis and reuse of a legacy dataset accompanied by only minimal documentation. This situation is not ideal, since the lack of documentation created too much uncertainty for our analysts to feel confident about pursing certain kinds of questions, particularly questions requiring fine-grained understanding of context. However, because the Chogha Mish dataset contained some (coarse) contextual information that could be supplemented with available publications, the dataset could be used to address a number of more general archaeological questions regarding both the site and the region. Similarly, Amorosi et al.

(1996) show how broad patterns can still be discerned across datasets, even if collected by different researchers under different or even obsolete methodologies. Recovery methods also lacked directly documentation, but the analysts were able to glean information about the data from secondary sources (published reports on the site). "Forensic" analysis of the dataset itself, with relative propotions of certain size-ranges of elements, seems to indicate that the dataset came from a largely "hand-collected" (as opposed to screened) excavation assemblage.

Taken together, these findings suggest that while data documentation is important, we should not discount assemblages lacking detailed documentation. As demonstrated in this study, the background and tacit knowledge of experienced zooarchaeologists can be invaluable in understanding an old dataset, even without detailed documentation. In addition, our call for at least base-line data documentation should not be taken as a call for rigid standardization of recording methodologies. Some aspects of recording probably can and should see more standardization, particularly in taxonomic identification, bone element identification, fusion, and the like. However, researchers also need freedom to innovate and tailor recording methods to particular questions. Thus, we focus on the need to document and describe datasets, no matter what their recording methodologies, to inform future reuse. Since the point of documentation is communication, editorial review of datasets and data documentation can be invaluable. If data documentation can successfully communicate meaning to editors (with the appropriate subject expertise), such documentation has a better chance of informing a wider research community as well. Thus, data publication models can help make data documentation an effective aspect of data dissemination.

5.5 Implications beyond Zooarchaeology

This blind study, though focused on a zooarchaeological dataset, can serve as a guide to archaeologists and practitioners of small sciences more broadly. "Small science" is research undertaken by individuals or small teams. Small science domains typically generate many, relatively small but often complex datasets and often require data from diverse sources, sometimes beyond disciplinary boundaries (Onsrud and Campbell, 2007). Lessons learned here can inform not only scholars who wish to share their datasets and prepare them for reuse, but also evaluators of published datasets, for whom there are no current guidelines to help determine what a "quality" or properly documented dataset might look like. Many zooarchaeologists record their analyses on spreadsheets which are not linked in any way to the overall project. Thus, information about the site, context descriptions, and recovery methods are often disassociated from the zooarchaeological data. These other data provide essential contextual information needed by zooarchaeologists and future investigators wanting to reuse legacy data. Table II shows the information that is imperative to provide with a dataset if data are to be intelligible and reusable by another researcher. In some cases, this information will be part of the project description (metadata); in other cases, this information will be included with every item in the database. Archaeologists who intend to share their work should ensure that their datasets meet all of the essential criteria and as many of the less critical criteria as possible. Criteria in Table III relate specifically to zooarchaeological analysis, but can be adapted to other material types and fields.

6. Conclusion

This study explored some of the challenges and opportunities in using decades-old primary data collected by a prior investigator. After independently establishing that the dataset had enough integrity to merit further use, this study's three zooarchaeologists proceeded with blind analysis of the data. These blind analyses produced diverse analytical results. Though this is to be expected in any research, these differences highlight the fact that interpretations are contingent on many analytic choices and judgment calls. Access to primary data is needed for scientific replicability, so that others can evaluate such judgment calls or explore new questions. But in order for primary data to be reused, datasets must demonstrate some level of quality and intelligibility. In order to use legacy data with confidence, researchers need some assurances of data integrity and need to find sufficient documentation to guide their analyses.

Recognizing that data integrity and intelligibility is paramount to reuse, data dissemination efforts need processes to promote greater data quality and promote more comprehensive data documentation. As data sharing assumes greater professional acceptance, multiple systems are emerging to meet widely varying needs of researchers. While standards can help make data dissemination and reuse much more efficient, they can also constrain interpretive choices inhibit innovation in methods. Given archaeology's (and zooarchaeology's) widely varying research questions, theoretical perspectives, and methological needs, we encourage a diversity of distributed approaches and not "monolithic" centralization or overly-rigid standardization. Where researchers need to custom tailor their recording methods, they need to take extra pains to provide adequate data documentation. In many cases, editorially supervised "data publishing" models can help make data dissemination reach the levels of quality and documentation needed to enable confidence in reuse.

We hope this study encourages similar methodological innovation in zooarchaeology and beyond. While this study focused on concerns relating to the reuse of a single dataset, many research programs will require use of multiple datasets. Such integrative research will require the development of methods to guide data selection, assessments of data quality, data compatibility, and semantics. Future research should explore methodologies and approaches to improve the rigor of such "meta-analyses" that span multiple datasets. Expanding studies like this can help place integrative research on a firmer analytic foundation while informing best practices for archaeological data dissemination.

Acknowledgments

We thank Abbas Alizadeh (University of Chicago) for making this dataset publicly available and encouraging our use of these data. We also note that this study would not have been possible without Jane Wheeler's original analysis, and her contribution is recognized in Open Context, where a copy of these data is published and archived. This study is part of a broader endeavor exploring user needs in archaeological data sharing, carried out by the Alexandria Archive Institute and funded by a grant from the National Endowment for the Humanities' *Advancing Knowledge: The IMLS/NEH Digital Partnership* program.

References

Amorosi, T., Woollett, J., Perdikaris, S., McGovern, T. (1996). Regional Zooarchaeology and Global Change: Problems and Potentials. World Archaeology 28, 126-157.

Atici, L., Lev-Tov, J., Kansa, S.W. (2010). Chogha Mish Fauna (Overview). (Released 2010-08-24), in: Atici, L., Lev-Tov, J., Kansa, S.W. (Eds.), Open Context. <http://opencontext.org/projects/497ADEAD-0C2A-4C62-FEEF-9079FB09B1A5>

Carraway, L.N. (2011). On Preserving Knowledge. American Midland Naturalist 166, 1-12.

Chaplin, R.E. (1971). The Study of Animal Bones from Archaeological Sites. London: Seminar Press.

Clutton-Brock, J. (1975). A system for the retrieval of data relating to animal remains from archaeological sites. In: Clason, A.T. (Ed.), Archaeozoological Studies (pp. 21-34). Amsterdam: Elsevier.

Costello, M.J. (2009). Motivating Online Publication of data. BioScience 59, 418-427.

Davis, S. (1987). The Archaeology of Animal Bones. London: Yale University Press.

Driver, J.C. (1991). Identification, classification and zooarchaeology. Circaea 9, 35-47.

Editors. (2009). Data's shameful neglect. Nature 461, 145.

Gamble, C. (1978). Optimising information from studies of faunal remains. In: Cherry, J.F., Gamble, C., Shennan, S. (Eds.), Sampling in Contemporary British Archaeology (pp. 321-353). Oxford: Archaeopress.

Gobalet, K.W. (2001). A critique of faunal analysis; inconsistency among experts in blind tests. Journal of Archaeological Science 28, 377-386.

Grigson, C. (1978). Towards a blueprint for animal bone reports in archaeology. In: Brothwell, D., Thomas, K.D., Clutton-Brock, J. (Eds.), Research Problems in Zooarchaeology (pp. 121-128). London: University of London.

Harley, D., Acord, S.K., Earl-Novell, S., Lawrence, S., King, C.J. (2010). Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines. http://escholarship.org/uc/item/15x7385g (Accessed October 7, 2010).

Hesse, B., Wapnish, P. (1985). Animal Bone Archaeology: From Objectives to Analysis. Manuals on Archaeology 5

Washington: Taraxacum.

Kansa, E.C. (2010). Open Context in Context: Cyberinfrastructure and Distributed Approaches to Publish and Preserve Archaeological Data. The SAA Archaeological Record 10, 12-16.

Kansa, E.C., Schultz, J., Bissell, A.N. (2005). Protecting Traditional Knowledge and Expanding Access to Scientific Data. International Journal of Cultural Property 12, 285-314.

Kansa, S.W., Kansa, E.C. (2011). Beyond BoneCommons: Recent Developments in Zooarchaeological Data Sharing. The SAA Archaeological Record 11, 26-29.

Kansa, S.W., Lev-Tov, J., Atici, L., Kansa, E.C. (In Preparation). Guidelines for Collecting and Disseminating Zooarchaeological Data in the Digital Age.

Kintigh, K.W. (2006). The Promise and Challenge of Archaeological Data Integration. American Antiquity 71, 567-578.

Klein, R.G. (1989). Why does skeletal part representation differ between smaller and larger bovids at Klasies River Mouth and other archaeological sites? Journal of Archaeological Science 16, 363-381.

Klein, R.G., Cruz-Uribe, K. (1984). The Analysis of Animal Bones from Archaeological Sites. Chicago: University of Chicago Press.

Lev-Tov, J., Atici, L., Kansa, S.W. (In Preparation). A Cooperative Study of Faunal Remains from Chogha Mish, Iran after 40 Years of Data in the Wilderness.

Lyman, R.L. (1994a). Quantitative units and terminology in zooarchaeology. American Antiquity 59, 36-71.

Lyman, R.L. (1994b). Vertebrate Taphonomy. Cambridge: Cambridge University Press.

Lyman, R.L. (2008). Quantitaive Paleozoology. Cambridge: Cambridge University Press.

Meadow, R.H. (1978). "Bonecode" A system of numerical coding for faunal data from Middle Eastern sites. In: Meadow, R.H., Zeder, M.A. (Eds.), Approaches to Faunal Analysis in the Middle East (pp. 169-186). Cambridge: Peabody Museum, Harvard University.

Meadow, R.H. (1980). Animal bones; problems for the archaeologist together with some possible solutions. Paleorient 6, 65-77.

O'Connor, T.P. (2003). The Analysis of Urban Animal Bone Assemblages: A Handbook for Archaeologists. York: Council for British Archaeology.

Onsrud, H., Campbell, J. (2007). Big Opportunities in Access to "Small Science" Data. Data Science Journal 6, 58-66.

Reitz, E.J., Wing, E.S. (2008). Zooarchaeology, Second ed. Cambridge: Cambridge University Press.

Richards, J. (2004). Online Archives. Internet Archaeology http://intarch.ac.uk/journal/issue15/richards_index.html (Accessed March 18, 2008).

Ringrose, T.J. (1993). Bone Counts and Statistics: A Critique. Journal of Archaeological Science 20, 121-157.

Silver, I.A. (1969). The ageing of domestic animals. In: Brothwell, D., Higgs, E. (Eds.), Science and Archaeology (pp. 283-302). London: Thames & Hudson.

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, 958-959.

Speth, J.D. (1983). Bison Kills and Bone Counts. Decision Making by Ancient Hunters. Chicago and London: The University of Chicago Press.

Thomas, K.D. (1996). Zooarchaeology: Past, Present, and Future. World Archaeology 28, 1-4.

Turner, A. (1989). Sample Selection, Schlepp Effects and Scavenging: The Implactions of Partial Recovery for Interpretations of the Terrestrial Mammal Assemblage from Klasies River Mouth. Journal of Archaeological Science 16, 1-12.

Uerpmann, H.-P. (1973). Animal bone finds and economic archaeology: a critical study of 'osteo-archaeological' method. World Archaeology 4 (3), 307-322.

Uerpmann, H.-P. (1978). The KNOCOD System for Processing Data on Animal Bones from Archaeological Sites. In: Meadow, R.H., Zeder, M.A. (Eds.), Approaches to Faunal Analysis in the Middle East (pp. 149-167). Cambridge: Peabody Museum, Harvard University.

Wheeler Pires Ferreira, J., Atici, L., Lev-Tov, J., Kansa, S.W. (2010). Chogha Mish Fauna (Released 2010-08-24). In: Atici, L., Lev-Tov, J., Kansa, S.W. (Eds.), Table generated by: Open Context Editors. Open Context. <http://opencontext.org/tables/39fd14fe7196aea0821ce8c7e08251f8> California Digital Library Archival Identifier <ark:/28722/k2c824d31>

Table and Figure Captions

Table I: Criteria for evaluating and improving data integrity

Table II: Archaeological data sharing criteria ("essential" criteria are in bold)

Table III: Common variables for zooarchaeology-specific data sharing ("strongly recommended" criteria are in bold)

Figure 1 A-C: Relative proportion of animals in the Chogha Mish assemblage, showing different researcher choices in aggregation of cultural periods and taxa.

Figure 2 A-C: Sheep and goat kill-off patterns in the Chogha Mish assemblage, showing different researcher choices in aggregation of cultural periods and age at death estimates.