

Other People's Data:

Blind Analysis and Report Writing as a Demonstration of the Imperative of Data Publication

Levent Atici

University of Nevada, Las Vegas
Levent.Atici@unlv.edu

Justin S.E. Lev-Tov

Statistical Research, Inc.
jlev-tov@srircrm.com

Sarah Witcher Kansa

Alexandria Archive Institute
skansa@alexandriaarchive.org

Introduction

Scholars increasingly rely on digital data for their research, from recording to data analysis and presentation, and routinely publish in journals that can be accessed both online and offline. While the scholarly community has embraced the Web for accessing published papers, researchers still tend not to share or publish comprehensive "raw" data. However, recent policy changes, such as the National Science Foundation requiring "data access plans" of all grant-seekers, promise to raise the professional stakes in data sharing.

Despite wide acknowledgement that approaches to data collection, recording, analysis, presentation, and interpretation vary among zooarchaeologists, few studies have documented the challenges faced when multiple zooarchaeologists work on the same dataset. Moreover, if the analysts involved have not generated the dataset in hand, the problem becomes multifaceted.

In light of the increasing significance of data sharing, this poster explores the fundamental challenges behind using data generated by others. We focus on two key questions:

What is the research benefit to sharing raw data?

How do we confront the challenges in using other people's data?

Objectives

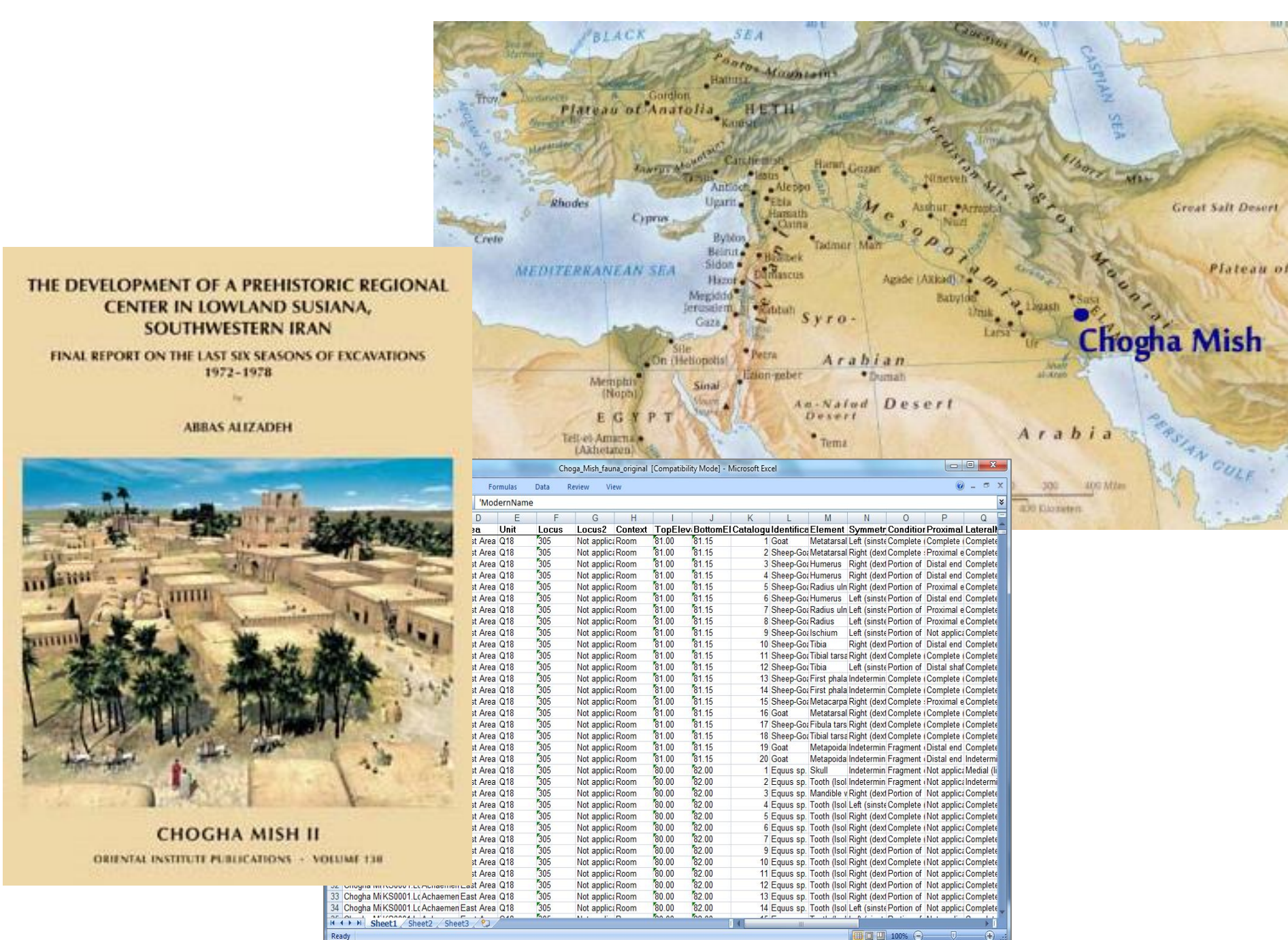
This study aims to highlight the importance of publishing original datasets, ideally alongside the published syntheses, but also in cases where no analysis was undertaken. We do this by demonstrating the diversity of interpretive outcomes when different researchers with similar expertise analyze the same dataset independently of each other. Our study also demonstrates the need for datasets to be adequately documented in certain key areas to enable reuse.

Methods & Materials

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 though she never analyzed the data or produced a report, her identifications were saved and later transferred to punch cards and then to Excel. This "orphan" 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 II*.



Three researchers, each with 15+ 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 possible with the given dataset. They documented the full process, from data cleaning to interpretation. The analysts had no contact with each other or discussion of the project until they concluded their independent analyses. They then met in person to compare their methodological approaches, discuss their findings, and develop a collaborative analysis plan.



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 metrical data and methodological background.

The faunal analysts took widely different approaches to analyzing the same dataset, which led to different interpretive results. Inter-analyst variation included: decisions about aggregation of phases and taxa; judgments about data reliability, consistency and comparability of the data; and the "threshold" at which the researcher decided they had made too many assumptions and could not conduct further analyses.

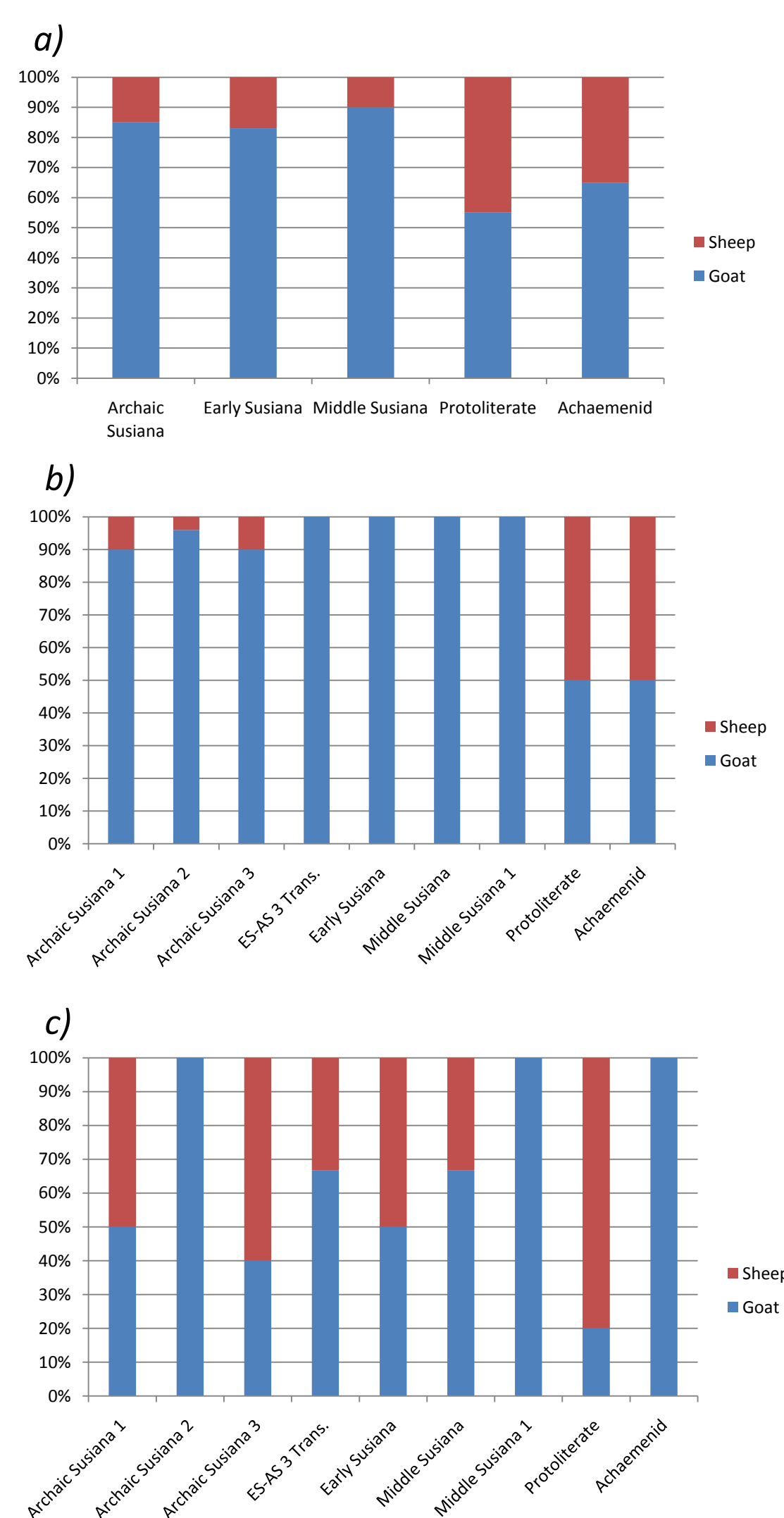
Quantification and Periodization

These choices about data aggregation led to discrepancies even at the most fundamental level. 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.

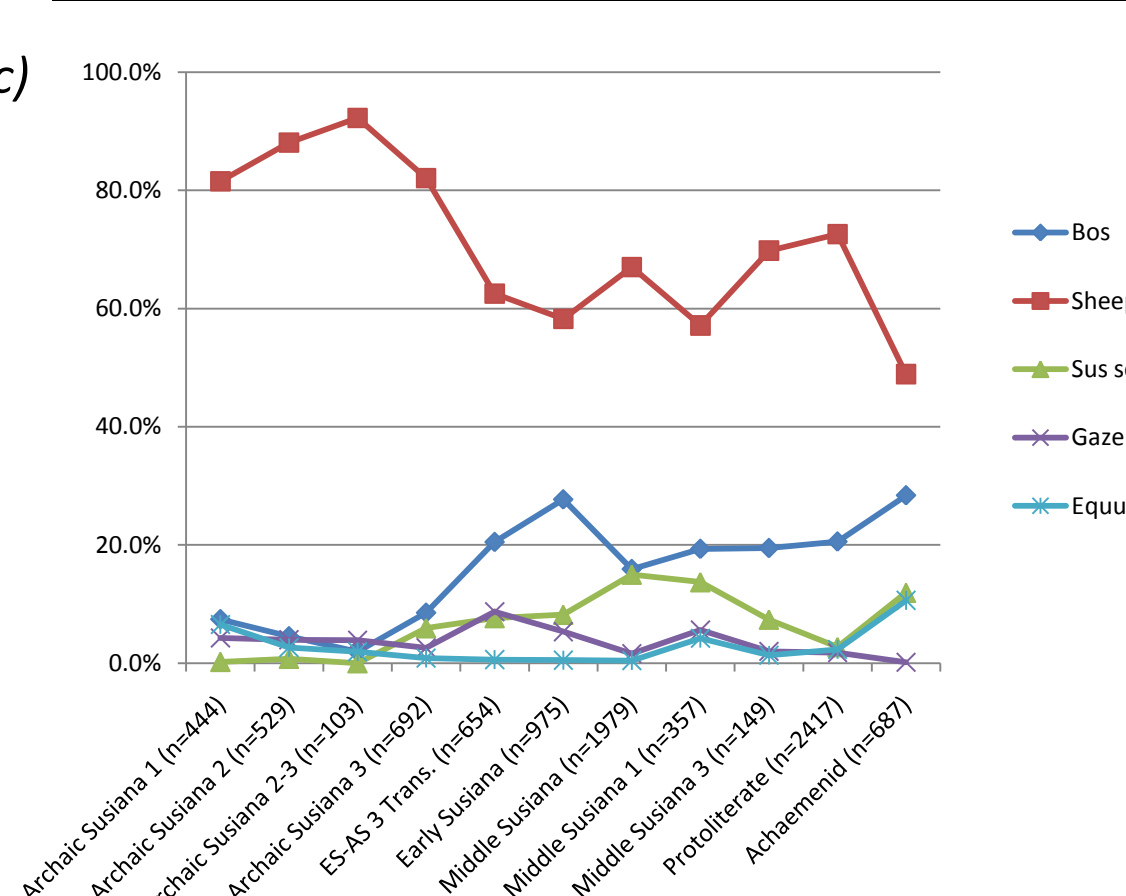
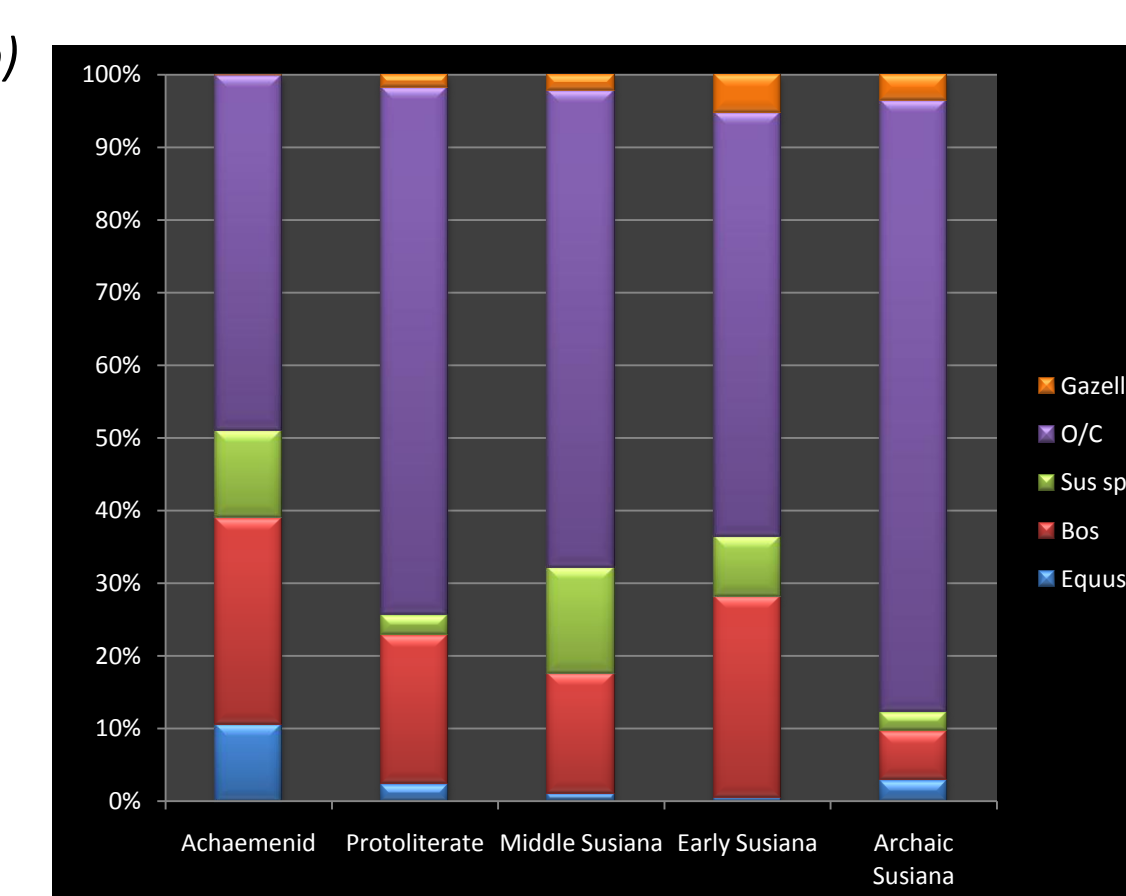
As shown in Figure 1 (at right), each analyst made different decisions about how to aggregate taxa and periods before beginning with basic analytical tasks. One analyst included a broader range of taxa, but consolidated the periods. Two analysts chose to focus on the predominant taxa, but combined cultural periods differently, which led to different results. For example, Figures 1a and 1b both show an increase in cattle in the Early Susiana. However, both omitted the "Archaic Susiana - Early Susiana Transition," the period in which the increase in cattle is documented by the analyst of Figure 1c. These seemingly small choices can lead to vastly different interpretive results.

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

Finer-Grained Analyses:



Taxa (relative % by period)	Achaemenid	Protoliterate	Middle Susiana	Early Susiana	Archaic Susiana
cattle	17	8	9	8	3
sheep	1	1	0	0	0
goat	1	1	1	1	2
sheep/goat	27	26	32	16	30
sheep/goat/gazelle	1	2	6	4	7
PG	7	1	7	2	1
deer	0	0	0	0	0
gazelle	0	1	1	2	1
artiodactyl	0	0	0	0	0
camel	0	1	1	0	1
carnivores	3	1	2	2	1
construal	0	0	2	1	0
large mammal	8	13	19	13	5
medium mammal	28	42	18	50	45
small mammal	0	0	1	0	1
indeterminate	1	3	2	1	2
Totals (%)	100	100	100	100	100
Totals (N)	1170	6299	4898	3417	4747



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.

As a simple example of this, Figure 2 (at left) shows the widely different results for the changing sheep / goat ratio through time at Chogha Mish. In Figure 2a, the analyst included all of the original identifications, regardless of element, and aggregated the data into major cultural periods. The results are vastly different than Figure 2b, where the analyst used only sheep/goat distinctions made on metapodials and displayed the data for all sub-periods. These results are different, again, from Figure 2c, which shows distinctions made on horn cores alone (which favor goats, perhaps due to hornless female sheep in some periods).

Figure 2: Ratio of sheep to goats by period, where taxonomic distinction was made based on (a) horn cores and (b) metapodials

Conclusions

Any data set will see multiple interpretations depending on analyst perspective. It is imperative to publish original datasets along with syntheses to allow for multiple interpretations and replication of results.

Choices about data aggregation and splitting will depend on the research question being asked. For example, regional syntheses differ from comparisons of faunal data with material culture at one site. Access to raw data is needed to make aggregation options available.

Contextual information (time and place) must be provided with raw datasets in order to make them useful. Other critical information includes: the name and background of the original analyst, decoded data (or, at a minimum, use of a published code), and identification basics (taxon, element, portion, side, fusion, sex).

Even in cases where no analysis was undertaken, it is essential to share raw datasets so that future generations of scholars can benefit from the work.

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.

Recommendations

- 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. Comprehensive data sharing should be adopted as an integral aspect of zooarchaeological scholarship.
- Need Adequate Documentation:** Datasets should be published, fully described, in their entirety. This involves the task of determining the "fundamental" information needed for data publication (i.e. the information that every faunal analyst must collect, record, analyze, interpret, and present). For example, published data should include enough discussion of methods, research aims, and data collection practices as required to facilitate informed reuse.
- Data sharing as Publication:** Finally, adequately describing a dataset requires thought and effort and possibly even editorial guidance. Therefore, we recommend that data dissemination should take on more of the formal (and, hopefully, rewarded) trappings of "publication," rather than informal "sharing." Thus, instead of advocating for the dissemination of "raw data" we should advocate for comprehensive publication of cleaned up, properly documented, and usable data.

Next Steps

We will publish our collaborative Chogha Mish faunal analysis and interpretations in print and online, linked to both the original and the revised datasets (Lev-Tov, lead author). We will also publish the results of this blind analysis exercise and address some of the rewards and challenges associated with using other peoples' data (Atici, lead author). Finally, we will produce a set of detailed guidelines for documenting and maximizing the usefulness of published zooarchaeological datasets (Kansa, lead author).

Acknowledgements

This study is part of a broader endeavor exploring user experience 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.

