ETHNOBIOLOGY LETTERS

Perspective
Special Issue on Digital Zooarchaeology

Using Linked Open Data to Improve Data Reuse in Zooarchaeology

Sarah Whitcher Kansa

Author address: The Alexandria Archive Institute & Open Context, 125 El Verano Way, San Francisco, CA 94127, USA.
Email: skansa@alexandriaarchive.org

Received: August 20, 2015
Published: December 18, 2015

Volume: 6(2):224-231
© 2015 Society of Ethnobiology

Abstract: *The inability of journals and books to accommodate data and to make it reusable has led to the gradual loss of vast amounts of information. The practice of disseminating selected sub-sets of data (usually in summary tables) permits only very limited types of reuse, and thus hampers scholarship. In recent years, largely in response to increasing government and institutional requirements for full data access, the scholarly community is giving data more attention, and solutions for data management are emerging. However, seeing data management primarily as a matter of compliance means that the research community faces continued data loss, as many datasets enter repositories without adequate description to enable their reuse. Furthermore, because many archaeologists do not yet have experience in data reuse, they lack understanding of what "good" data management means in terms of their own research practices. This paper discusses Linked Open Data (LOD) as an approach to improving data description, intelligibility and discoverability to facilitate reuse. I present examples of how annotating zooarchaeology datasets with LOD can facilitate data integration without forcing standardization. I conclude by recognizing that data sharing is not without its challenges. However, the research community's careful attention and recognition of datasets as valuable scholarly outputs will go a long way toward ensuring that the products of our work are more widely useful.*

Keywords: Annotation, Data publishing, Integration, Data modeling, Zooarchaeology

Introduction

Access to rich, well-described datasets can enable large-scale analysis, drawing on multiple data sources to address "big picture" research questions. Recognizing the research potential of multiple datasets, many public and private funders of archaeology now mandate data management plans as part of the research they fund.[1] As digital data increasingly play a key role in all forms of archaeological observation and recording, professional practice must emphasize rigorous and effective data management. Unfortunately, without examples of how standards, metadata, and data quality impact research outcomes, field archaeologists will have little motivation to improve their data creation and management practices. Furthermore, if scholars only see data sharing as a matter of bureaucratic compliance, there is the risk of filling data repositories with poorly documented, poor quality, and nearly useless data. To avoid this, researchers need clear examples of how to align data creation and management with reuse and understanding. This paper discusses one such approach by describing how zooarchaeology can benefit from linking faunal data with data curated by a much wider research community using Linked Open Data (LOD) methods.

The Perils of Current Data Sharing Practices

The zooarchaeological community has long recognized that data sharing is a critical part of communicating research outcomes (see Clason 1972; Driver 1992; Grigson 1978). However, while comprehensive datasets used to commonly accompany monographs, in particular, print has become an increasingly difficult format for accommodating the complexity and size of today's datasets (Marwick 2015). The fact that most journals and books are unable to accommodate datasets in full has led to the gradual loss of vast amounts of information, and has prevented "computational reproducibility that might lead the work to have greater impact and reuse" (Marwick 2015). Researchers select (whether by necessity or by choice) what they see as the most important data to disseminate or provide summarized data tables to support arguments. This has done a disservice to archaeology (and anthropology and ethnobiology, more generally) by permitting only certain kinds of reuse, bounded by the reporting format chosen by the

original author. Data tables in print format require manual transcription for reuse, leading to a high rate of error (Dibble, this issue). Furthermore, printed data tables, even if complete, cannot be searched or sorted, and thus lead to a painfully slow process of transcription by the person seeking to use them. Finally, this work is done by one person, and all other researchers seeking to use the data will have to transcribe it themselves, again potentially making errors, and again spending hours of precious research time on a tedious task. In some cases, full data tables are provided on DVDs that accompany a print publication. While easier to manipulate, these datasets are still problematic because DVDs degrade over time and are often broken, scratched, or lost.

Another persistent data sharing practice in zooarchaeology involves the one-to-one exchange of information, usually over email (Faniel et al. 2013). For example, a colleague contacts me indicating interest in a subset of data, and I share that data with him, often over several emails explaining the nature of the dataset, the methods, errors, etc. This type of one-to-one transaction leads to information loss because data description and clean up is not formally documented (if it occurs at all). The dataset is shared with one person, and any future sharing requires the same process. Handing out batches of data piecemeal in such a manner does not lead to full data preservation and does not enable reuse. These practices also promote "choosing favorites" by allowing sharing with only certain people, and lead to fears of "scooping" because of the informal nature of the communication.

Another entrenched data sharing practice is through summary tables in the published literature. While summary tables are an acceptable and effective approach to support the interpretive perspective being advanced in the paper in which they appear, they are of limited analytical use to those who want to reuse these data. Table 1 is an example of data presentation that may sufficiently support an author's argument, but leaves the reader with no means to leverage that data in future research. For instance, a researcher may like to know which specific skeletal elements were burned, or which were fused. What is the basis for calling a skeletal fragment "juvenile"? What was the nature and location of the cut marks on the bone surface? There are infinite future research questions that this dataset could inform, but the data are not shown when in summary form. Unless the full dataset is available elsewhere (ideally, in an institutional archive), such data presentation is not sufficient stewardship because reuse of the data is extremely limited. It is important to note that Table 1 is not necessarily a poorly constructed table, it may suit an argument built in the paper; this example simply illustrates the limitations not reporting datasets in ways that make them available for future use.

Table 1. A hypothetical summary data table.

Sheep (*Ovis aries*)	Juvenile	Adult
NISP	238	459
MNI	6	11
Burned	3%	5%
Cut	12%	22%
Gnawing (rodent)	1%	1%
Gnawing (carnivore)	2%	4%

Several recent studies have used summary tables in the published (and gray) literature to explore new research questions that may be better addressed with access to large corpora from multiple sites (among others, see Conolly et al. 2011; MacKinnon 2004, McKechnie et al. 2014, Sasson 2010, Thomas et al. 2013). Though useful for addressing certain questions, these meta-analyses run the risk of leading to misinterpretations simply because comparing summary data across a large number of sites requires finding such a broad basis for comparison ("present / absent", "many / few" or, as above, "juvenile / adult") that researchers are unable to see or incorporate any higher-resolution observations that may be very important to the broad interpretations. In short, when primary data and detailed documentation about how the data were collected and analyzed are not available, considerable caution must be taken in aggregating data from multiple studies (Jones and Gabe 2015).

Part of the greater public policy interest in research data management comes from recognition that researchers do a poor job as stewards of their own data, where many datasets maintained by individual researchers are lost entirely after only a few years, while others are useless because of a lack of detailed data description (Vines et al. 2014). Clearly, we need to identify better data management practices and find incentives to encourage zooarchaeologists to adopt better practices. The sections below discuss how current technologies and emerging data sharing practices promise to change common out-dated practices to make data more useful to others.

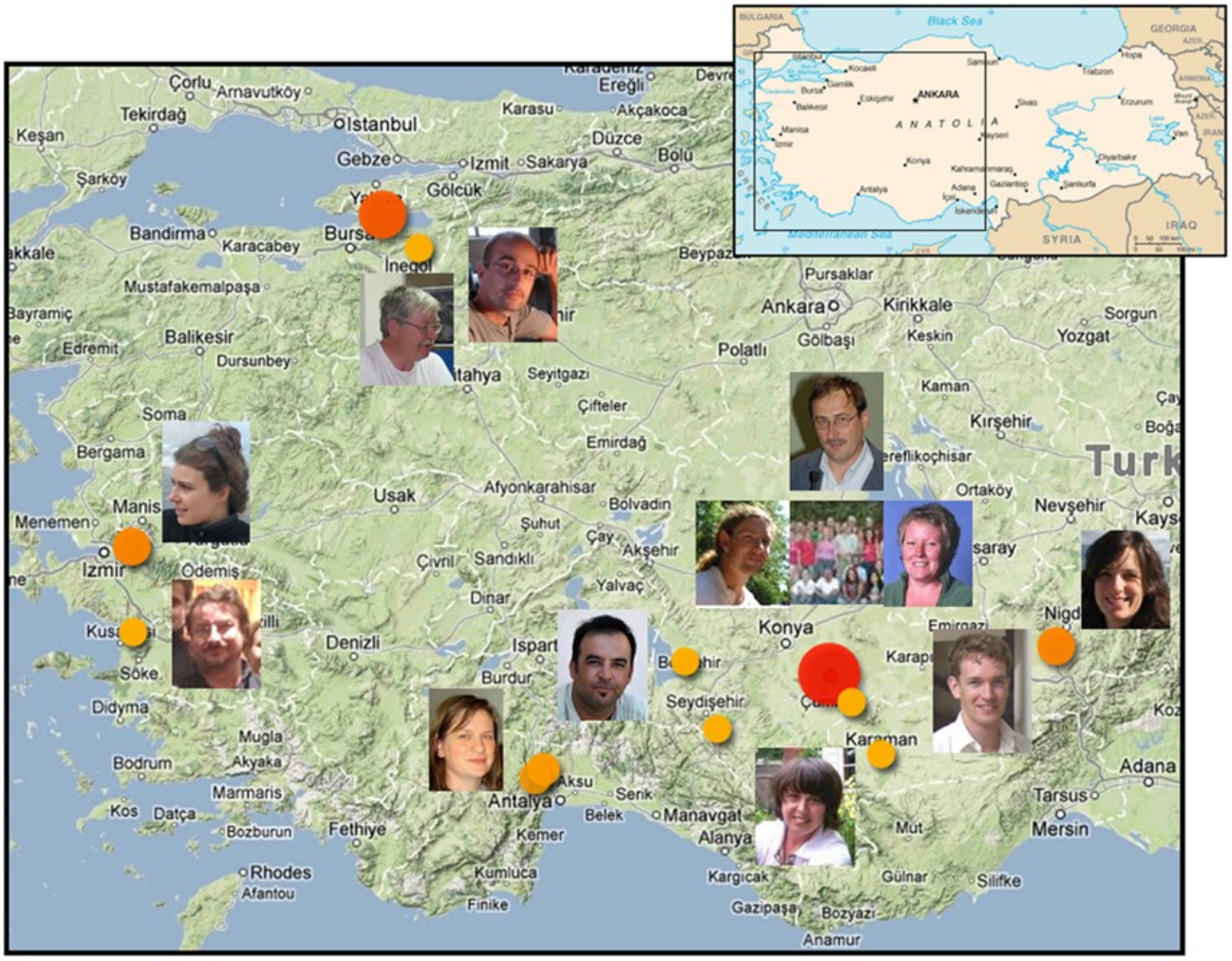

Figure 1. Participants in the Central and Western Anatolia Neolithic Working Group, who collaborated to integrate and analyze multiple faunal datasets from archaeological sites in Turkey (see Arbuckle et al. 2014).

A Linked Open Data Approach to Zooarchaeological Data Sharing and Integration

Advances in technology now offer opportunities to share and document data in full. However, "sharing data" is not simply a matter of dropping a spreadsheet onto a website or into an archive. I participated in a recent study that explored issues in data access and reuse by working directly with researchers to gain first-hand experience of the challenges data reuse presents (Arbuckle et al. 2014). This study, funded by the Encyclopedia of Life and the National Endowment for the Humanities, brought a group of scholars together to integrate data from one dozen archaeological sites (Figure 1) and to collaborate on a research topic using those data. This group, led by Benjamin Arbuckle (UNC Chapel Hill) represents a rare collaborative effort to publish and integrate open data in archaeology. Project participants shared faunal datasets in the open access data publishing platform, Open Context. These datasets, from archaeological sites in Turkey that span the Epipaleolithic through the Chalcolithic, were used to explore how integrated datasets can inform archaeologists about the spread of early domestic animals westward across Turkey. The project highlighted a complex regional picture in the spread of agriculture, with particularly notable differences between different coastal and inland regions (Arbuckle et al. 2014). It also highlighted critical differences in the way different zooarchaeologists describe data (Kansa et al. 2014).

Part of my role in this project, as editor for Open Context, was to work with data authors to clean up and document their datasets in preparation for publication and for integration for analysis by the group. Data clean up proved to be challenging, but it was not without its rewards. Some datasets documented over 100,000 specimens, sometimes using more than 100 fields. Each dataset used a largely idiosyncratic system of organization and terminology. Integrating these diverse datasets entailed making sure terms were consistent and all fields and terms were clearly described. To complicate matters, many of the datasets were either fully or partially coded, requiring specialist knowledge and ten times the effort to clean up and decode than other datasets. In one case, the project codebook was a 90-page PDF; in another, codes had been added later and not included in the codebook, making contact with the data author critical to making the dataset intelligible. This exercise convinced me that clean-up and additional documentation through a formal editing process creates datasets that are of far greater quality and that have vastly increased potential for reuse than simply uploading a spreadsheet to an archive. Though this documentation requires substantial time and effort, it is a one-time job that benefits from direct interaction with other analysts to create a more robust and appropriately described dataset.

Once the datasets were cleaned and richly documented, the next editorial step was to prepare them for integration by annotating them with Linked Open Data (LOD). Essentially, LOD boils down to using stable Web identifiers or "URIs" (Uniform Resource Identifiers) to reference shared concepts and other information resources.[2]

LOD can help zooarchaeologists aggregate data at larger scales without necessarily forcing everyone to adopt the same predetermined recording standards. In the Anatolian example above, we used LOD to annotate data to relate different idiosyncratic terminologies to common controlled vocabularies. The example in Figure 2(a-c) shows the different ways various analysts might describe *Ovis orientalis* Linnaeus Bovidae in their database. Rather than require analysts to change the way they document their data, the data publication process can add links to shared concepts to help describe data and relate different terminologies across datasets. This example shows how the Encyclopedia of Life (EOL) can be used to integrate taxonomic descriptions across datasets. EOL publishes a webpage, with a stable identifier and address, for every taxonomic group defined by the life sciences. A researcher can go to their page that describes the concept of "wild sheep," grab that address and paste it into a spreadsheet. This tells everyone "this is the animal my term is describing." By linking all the different ways the analysts describe *O. orientalis* to the authoritative concept, different datasets are integrated around taxonomic concepts. Furthermore, this provides a common point of reference for all other data on the Web that reference this concept, allowing for discovery and large scale integration.

Another example of the benefits of a Linked Open Data approach to zooarchaeological data management can be structured around the "adult / juvenile" problem mentioned in the previous section. Though "adult" or "juvenile" may be the only terms that serve as the "least common denominators" across multiple projects, annotating the epiphyseal fusion data with these terms enables linkages across datasets while still maintaining the original researcher's descriptions. This allows for a much more transparent research process, where the annotations allow for integration across multiple datasets, but the original data can still be seen. That is, more refined categories (e.g., "newborn", "old adult", "fusing", etc.) present in certain datasets will remain visible, allowing for more nuanced interpretations.

Given the exponential growth of LOD on the Web (see Figure 3), the potential of LOD in facilitating the discovery and use of relevant, quality research data is vast. The pioneering efforts of the Perseus Digital Library[3], Pleiades[4], Pelagios[5], Arachne[6], DINAA[7], FASTI Online[8], and the Portable Antiquities Scheme[9], to name a few, as well as increasing openness of museums in sharing collections data and metadata (especially the British Museum) creates many research opportunities for digitally enabled scholarship. Zooarchaeology is "low-hanging fruit" in the world of LOD and data integration. Several authoritative sources of LOD already exist, including the EOL described above and UBERON, an anatomy ontology that can be used to describe skeletal elements. LOD is extremely easy for zooarchaeologists to build into their data collection protocols: in many cases, one can simply add a field to a databases or spreadsheets where they insert a link (URI) to the taxon or skeletal element referenced. This immediate disambiguation of terms will begin to consolidate more intelligible and reusable data that will have wide benefits to zooarchaeologists. However, in order for this to work, the community as a whole must change expectations

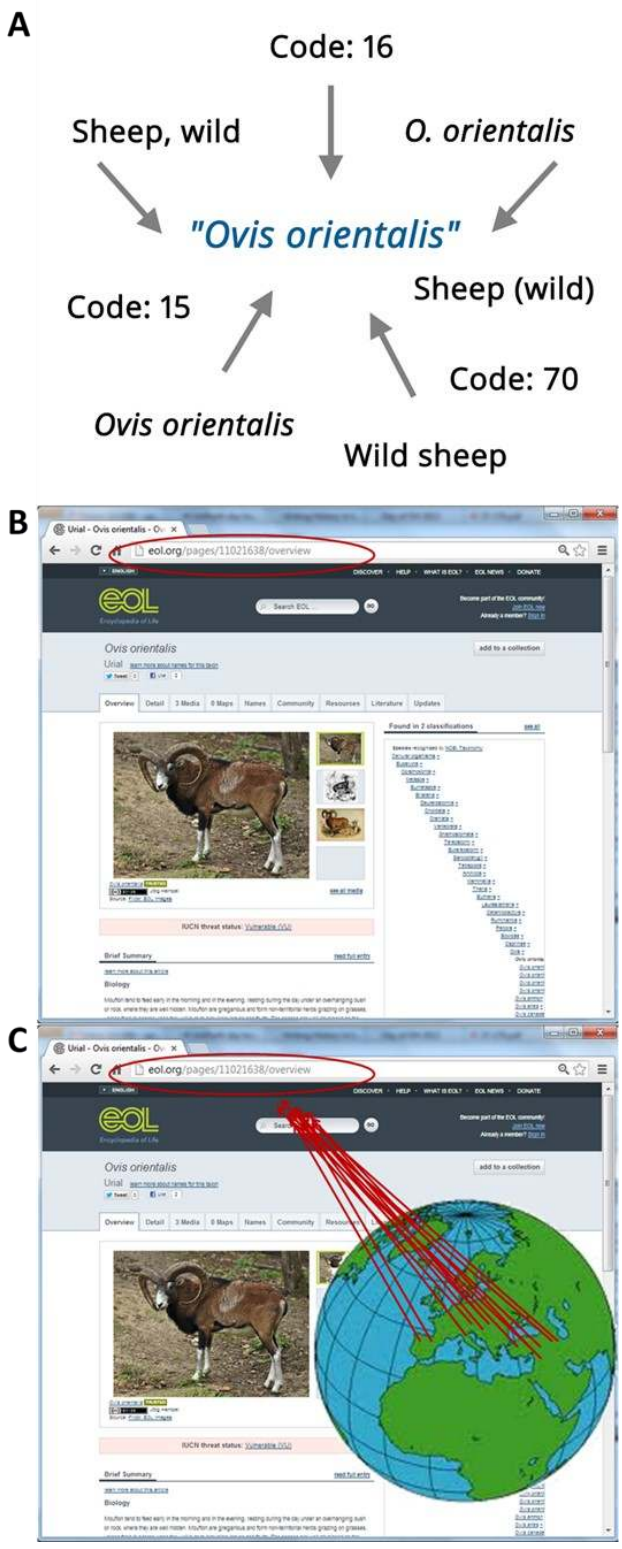

Figure 2. A) A small sample of the many terms analysts may use to describe a specimen from *Ovis orientalis* in their databases and spreadsheets. Entrenched data collection practices, different data description conventions, and a reluctance to adopt standards, mean that researchers will continue to collect disparate data. Linked Open Data approaches allow us to embrace the diversity of our data collection practices by offering an external source of data integration. B) The EOL URI for the term "*Ovis orientalis*" provides an authoritative and unambiguous description of this species, as well as additional descriptive content from authoritative resources across the Web. Linking terms that mean "*Ovis orientalis*" to this URI provides a common language to integrate many data sets without forcing analysts to adopt standard terminology. C) Linking data in this way is an essential step to enable future research that draws on multiple data sets.

around data. While data archiving is needed, we should also encourage additional steps toward contextualizing our data, such as the LOD approaches introduced above. Linking data is about networking data across datasets, across systems, and across communities. As the network grows and diversifies, it offers more opportunities not just for larger scale forms of analysis, but also for new collaborations that may result from linking our data to the data curated by other expert communities. While LOD offers many exciting and open-ended possibilities, as discussed in the next section, realizing these opportunities requires that we make important changes in our research practices even before we begin data collection.

Data Sharing Is Not without Its Challenges

While it is clear that linked data annotation was invaluable in the data sharing project described above, the participants were surprised to note how certain limitations in source datasets themselves impeded annotation and thus limited comparative analysis. Zooarchaeological taxa and skeletal elements were easy to align because most people find these characteristics easy to model and represent in a spreadsheet, usually with fields for "taxon" and "element." Certain other characteristics proved more difficult to align. For example, all participants took measurements according to guidelines provided by von den Driesch (1976). However, since von den Driesch gives many different measurements for different elements, these are difficult to represent in a single-table spreadsheet which many zooarchaeologists use. As a result, Open Context's editors needed to expend significant editorial effort to align bone measurements to a common measurement ontology so that they could be compared.

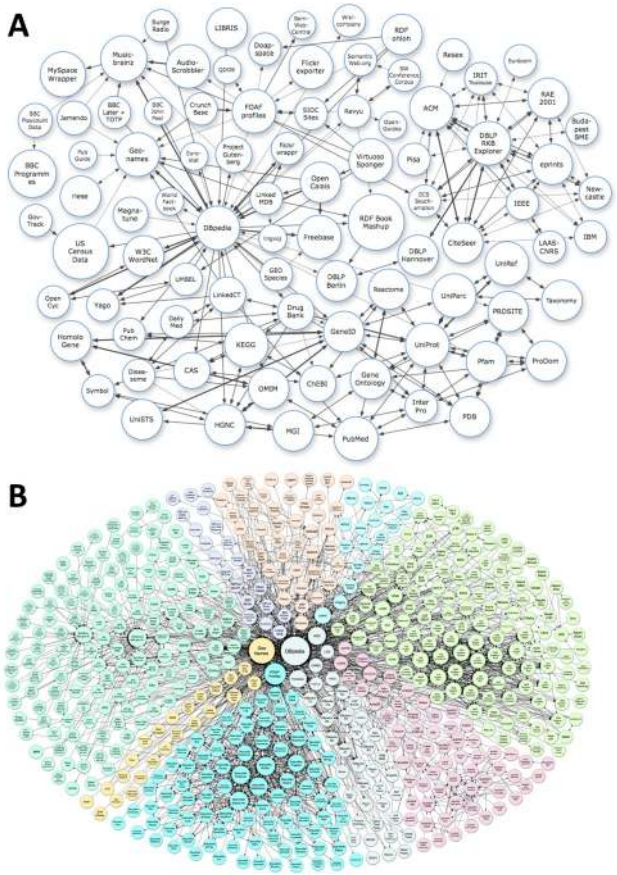

Figure 3. The growth of Linked Data on the Web, from 2009 (A; 89 data sets) to 2014 (B; 570 data sets). These images show datasets from all domains (i.e. not just archaeology) that have been published in Linked Data format by contributors to the Linking Open Data community project and other individuals and organizations. [Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/] For information about the colors and text, see the Web versions of the diagrams: 2009 here: http://lod-cloud.net/versions/2009-03-05/lod-cloud.png ; 2014 here: http://lod-cloud.net/versions/2014-08-30/lod-cloud_colored.svg.

Similarly, tooth eruption and wear data recorded by project participants proved very difficult to integrate and compare. Though all participants used the system for recording tooth eruption and wear developed by Payne (1973), the manner by which they recorded observations varied greatly. Again, the limitations of spreadsheets to organize and model complex data played an important role in impeding data reuse. For example, one analyst noted the tooth number in the column heading ("Molar 1") and listed the tooth wear stage in the cell. Another analyst noted the tooth number in a "Tooth Number" column and the wear stage in a "Wear Stage" column. Though both approaches record information according to Payne's system, the splitting of data across different fields, including free-text comments fields, makes integration very labor-intensive.

Although these two examples are Near East-specific (where researchers tend to use the two recording systems discussed), they illustrate more generally how data modeling (data organization), plays an important and largely ignored role in interpretation. Even small differences in recording, or in the structure of databases and spreadsheets, can have significant impacts on interpretation. These recording and modeling discrepancies become apparent when data authors begin looking "under the hood" at each other's datasets. Data sharing is important, then, not only in terms of getting access to data, but also in terms of getting access to each other's data models and systems of organization. Data modeling issues play a huge role in how data can be interpreted, especially in integrative studies, and this issue needs more attention.

Thus this discussion of data modeling illustrates how zooarchaeologists need to invest more thought and effort in describing and modeling their data, well before data collection, if they are to create data of lasting value to a wider community. While LOD offers powerful methods, LOD needs to be coupled with improved data modeling practices.

Conclusions

Data management in the 21st century is still a new frontier, and considerable research and perspectives are needed on how to integrate data dissemination and preservation meaningfully into the research process. A good starting point is to avoid perceiving data management as only as a byproduct or a residue of research, to be quickly filed away in an archive to comply with a grant requirement. If researchers want to unlock new opportunities with data, data need to be treated as "first class citizens" in scholarly communication. Achieving this requires several things. One is a shift in perspective on archiving practices. Most digital repositories focus on the quality of the metadata, with an end goal being archiving. However, from working first-hand with data reuse and integration, we have learned that investing more effort into the

individual data themselves is essential to understanding and reuse. Meaningful data preservation also means providing access to full datasets, not just summary tables. Though summary tables are useful to support the theoretical perspective being advanced in a given paper, their summarized format precludes many uses that address a number of vital research questions. Sharing summary tables without sharing the original, ungrouped data, often means immediate loss of information. Finally, researchers must commit a level of intellectual effort to data. Such a level of effort entails professionalism and dedicated expertise on par with current print publication practices. Unless data dissemination sees similar rewards, with regard to professional recognition and advancement, as conventional publishing, scholars will not find the time or motivation to share their data, and datasets amounting to years of work and (often public) funding continuing to languish on hard drives and in file cabinets around the world.

Acknowledgments
The research reported would not have been possible without funding from the Encyclopedia of Life and the National Endowment for the Humanities, as well as hard work on the part of the project participants, Ben Arbuckle, and Eric Kansa. I would like to thank Iain McKechnie for working with me so enthusiastically in organizing another ICAZ session on the theme of digital data in zooarchaeology. Heartfelt thanks also to the conference organizers and sponsors for a very successful ICAZ 2014 in San Rafael, Argentina.

Declarations
Permissions: None declared.

Sources of Funding: The Encyclopedia of Life and The National Endowment for the Humanities.

Conflicts of Interest: None declared.

References Cited

Arbuckle, B. S., S. W. Kansa, E. Kansa, D. Orton, C. Çakırlar, L. Gourichon, L. Atici, A. Galik, A. Marciniak, J. Mulville, H. Buitenhuis, D. Carruthers, B. De Cupere, A. Demirergi, S. Frame, D. Helmer, L. Martin, J. Peters, N. Pöllath, K. Pawłowska, N. Russell, K. Twiss, and D. Würtenberger. 2014. Data Sharing Reveals Complexity in the Westward Spread of Domestic Animals across Neolithic Turkey. *PLoS ONE* 9:e99845. Doi: http://doi.org/10.1371/journal.pone.0099845.

Clason, A. T. 1972. Some Remarks on the Use and Presentation of Archaeological Data. *Helinium* 12:139-53.

Conolly, J., S. Colledge, K. Dobney, J. -D. Vigne, J. Peters, B. Stopp, K. Manning, and S. Shennan. 2011. Meta-Analysis of Zooarchaeological Data from SW Asia and SE Europe Provides Insight into the Origins and Spread of Animal Husbandry. *Journal of Archaeological Science* 38:538-545. Doi: http://dx.doi.org/10.1016/j.jas.2010.10.008

Driver, J. C. 1992. Identification, Classification and Zooarchaeology. *Circaea* 9:35-47.

von den Driesch, A. 1976. A Guide to the Measurement of Animal Bones from Archaeological Sites, Peabody Museum Bulletin 1, Cambridge, MA.

Faniel, I., E. Kansa, S. W. Kansa, J. Barrera-Gomez, and E. Yakel. 2013. The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse. JCDL 2013 Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries: 295-304. New York, NY: ACM. Doi: http://doi.org/10.1145/2467696.2467712 . [Preprint available online at http://www.oclc.org/content/dam/research/publications/library/2013/faniel-archae-data.pdf].

Grigson, C. 1978. Towards a Blueprint for Animal Bone Reports in Archaeology. In *Research Problems in Zooarchaeology*, edited by D. R. Brothwell, K. D. Thomas, and J. Clutton-Brock, pp. 121-128. Institute of Archaeology Occasional Papers 3, London.

Jones, E. L. and C. Gabe. 2015. The Promise and Peril of Older Collections: Meta-Analyses and the Zooarchaeology of Late Prehistoric/Early Historic New Mexico. *Open Quaternary* 1:Art. 6. Doi: http://doi.org/10.5334/oq.ag.

Kansa, E., S. W. Kansa, and B. Arbuckle. 2014. Publishing and Pushing: Mixing Models for Communicating Research Data in Archaeology. *International Journal of Digital Curation* 9:57-70. Doi: http://doi.org/10.2218/ijdc.v9i1.301.

Marwick, B. 2015. Geoarchaeology of Aboriginal Landscapes in Semi-Arid Australia. S. J. Holdaway and P. C. Fanning. 2014. *Geoarchaeology* 30:459-461. Doi:10.1002/gea.21522.

MacKinnon, M. 2004. Production and Consumption of Animals in Roman Italy: Integrating the Zooarchaeological and Textual Evidence. *Journal of Roman Archaeology*, Supplement 54.

McKechnie, I., D. Lepofsky, M. L. Moss, V. L. Butler, T. J. Orchard, G. Coupland, F. Foster, M. Caldwell, and K. 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: http://doi.org/10.1073/pnas.1316072111.

Payne, S. 1973. Kill-Off Patterns in Sheep and Goats: The Mandibles from Aşvan Kale. *Anatolian Studies* 23:281-303.

Sasson, A. 2010. Animal Husbandry in Ancient Israel: A Zooarchaeological Perspective on Livestock Exploitation, Herd Management and Economic Strategies. Equinox, London.

Thomas, R., M. Holmes, and J. Morris. 2013. "So Bigge as Bigge May Be": Tracking Size and Shape Change in Domestic Livestock in London (AD 1220–1900). *Journal of Archaeological Science* 40:3309-3325.

Vines, T. H., A. Y. K. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J. -S. Moore, S. Renaut, and D. J. Rennison. 2014. The Availability of Research Data Declines Rapidly with Article Age. *Current Biology* 24:94-97. Doi: http://doi.org/10.1016/j.cub.2013.11.014.

Notes

[1]See recent policies by US National Science Foundation (http://www.nsf.gov/sbe/bcs/arch/archaeom.jsp) and US National Endowment of the Humanities (http://www.neh.gov/files/grants/data_management_plans_2015.pdf)

[2]Unlike most URLs, Web URIs not only serve as addresses to retrieve content, but URIs also as globally unique and unambiguous identifiers, backed by an institutional commitment for long-term curation. While URLs are simply addresses that can point to changing content (and those addresses themselves can come and go), using well curated and institutionally backed Web URIs provides much greater stability and clarity in identifying (and usually accessing) data across the Web. This makes it possible to network together widely distributed data, curated in different systems by different professional communities and different disciplines.

[3]http://www.perseus.tufts.edu/hopper/

[4]http://www.pleiades.stoa.org/

[5]http://www.pelagios-project.blogspot.com/

[6]http://www.arachne.uni-koeln.de/drupal/

[7]http://ux.opencontext.org/archaeology-site-data/

[8]http://www.fastionline.org

[9]https://www.finds.org.uk/

Biosketch

Sarah Whitcher Kansa directs the non-profit Alexandria Archive Institute, working with researchers to publish open access data with Open Context.