BEYOND THE ARCHIVE: BRIDGING DATA CREATION AND REUSE IN ARCHAEOLOGY

Ixchel M. Faniel OCLC Research, 6565 Kilgour Place, Dublin, OH 43017, USA (fanieli@oclc.org, corresponding author)
Anne Austin Department of History, Stanford University, Lane History Corner, 450 Serra Mall, Stanford, CA 94305, USA
Eric Kansa and **Sarah Whitcher Kansa** The Alexandria Archive Institute, Open Context, 125 El Verano Way, San Francisco, CA 94127, USA
Phoebe France Department of Anthropology, University of Hawaii, College of Social Sciences, 2424 Maile Way, Honolulu, Hawai'i 96822, USA
Jennifer Jacobs The Alexandria Archive Institute, Open Context, 125 El Verano Way, San Francisco, CA 94127, USA
Ran Boytner Institute for Field Research, 2999 Overland Ave. #103, Los Angeles, CA 90064, USA
Elizabeth Yakel School of Information, University of Michigan, 105 S. State St., Ann Arbor, MI 48109, USA

This version of the article was accepted for publication and appears in a revised form, after peer review and/or editorial input by Cambridge University Press, in *Advances in Archaeological Practice* published by Cambridge University Press.

Suggested citation: Faniel, Ixchel M., Anne Austin, Eric Kansa, Sarah Whitcher Kansa, Phoebe France, Jennifer Jacobs, Ran Boytner, and Elizabeth Yakel (2018). Beyond the Archive: Bridging Data Creation and Reuse in Archaeology. *Advances in Archaeological Practice*, 6(02), 105-116. https://doi.org/10.1017/aap.2018.2.

Abstract

This paper presents research on archaeological data creation and management practices at two excavations in Europe in order to gain a better understanding of how to align these practices with the data reuse needs of a broader research community. The Secret Life of Data (SLO-data) project follows the lifecycle of data from the field to the digital repository to better understand opportunities and challenges in data interpretation, publication and preservation. Our "slow data" approach focuses not on maximizing the speed and quantity of data, but rather on emphasizing curation, contextualization, communication, and broader understanding. Through a mixed-methods approach of interviews, field observations, and excavation data assessments, we recommended changes (both technical and organizational) to improve data creation and management practices. We report our findings and offer readers guidance on streamlining data collection for reuse during their excavation projects.

Resumen

Este artículo recoge la investigación realizada sobre la creación de datos arqueológicos y su gestión de aplicada en dos excavaciones en Europa con el objetivo fundamental de lograr un mejor conocimiento sobre cómo alinear dicha gestión con las necesidades de re-utilización de información en comunidades más amplias. El proyecto "La vida secreta de los datos" (SLO-data su acrónimo en inglés), hace un seguimiento del ciclo de vida de éstos desde el ámbito del Repositorio Digital para comprender en profundidad las oportunidades y los retos de su interpretación, publicación y preservación. Nuestro enfoque "Slow data" no busca maximizar la velocidad y cantidad de datos, si no que pone el énfasis en la curación, contextualización, y mejora del conocimiento. Gracias a un planteamiento metodológico combinado de entrevistas, observación de campo y evaluación de datos de excavación, hemos elaborado una serie de recomendaciones sobre cambios (técnicos y organizativos) que optimizan la creación y gestión practica de datos. Presentamos a continuación un informe de los hallazgos encontrados y ofrecemos al lector una guía con la que reconsiderar las líneas de trabajo a seguir en la recogida y re-utilización de datos en sus propios proyectos.

Digital recording of material culture has played an increasingly important role in archaeology (see Gordon et al. 2016). From spreadsheets to digital photography; GIS to portable XRF readings, the shift to digital is accelerating and the number of digital formats expanding. The transition to digital brings both benefits and disadvantages to archaeology. Digital recording reduces human error, dramatically increases the amount and granularity of data collected and analyzed, and allows for a significant increase in statistical and analytical capabilities (Austin 2014; Roosevelt et al. 2015). It also allows scholars to answer different types of research questions and potentially the more rapid publication of results with accompanying primary data, thus improving peer review and facilitating replicability of results.

Yet a number of issues still remain unresolved, which jeopardizes the potential of digital. First, longevity of digital data is extremely short – usually in years and rarely in decades. Books and other paper products can be in circulation for centuries. Second, the promise of comparative digital data – the ability to create synthetic analysis of datasets from different sites excavated by different scholars – akin to "big data" analysis (1), has not yet materialized. A lack of shared vocabularies and recording protocols limits comparison between projects (but see Arbuckle et al. 2014; Kansa et al. 2014). Third, data literacy among many practicing archaeologists – and especially among academic archaeologists – is significantly lacking. Fourth, it is unclear who will maintain and migrate digital data over time as libraries – the traditional repository of academic knowledge – are themselves grappling with fiscal austerity and escalating costs. Although digital data archiving facilities exist (such as tDAR and Open Context, in collaboration with the California Digital Library) and some US federal granting agencies require data management plans, digital repositories still generally lack secure forms of institutional financial support. Fee-for-service and short-term grant financing still underwrite most dedicated programs for digital data dissemination and archiving in the US (Kansa 2016:447-453; Kintigh and Altschul 2010).

Finally, as we explore below, there is a tension between what data creators do and what data reusers

need (Faniel and Yakel 2017). As archaeology can be a destructive practice, and given the ever-growing threats of development, urbanization, war, and the antiquities trade that erase so much of the record of the past, the data that archaeologists collect may be the only evidence that remains. In order to be good stewards of the past and make it available and useful to others, we need ways to better align data creation and reuse.

Background: The SLO-data Project

The Secret Life of Data (SLO-data) project (https://alexandriaarchive.org/secret-life-of-data/) explores how data creation practices shape data reuse. Data management has emerged as a key concern in archaeological practice, both in academic and heritage management settings. Many granting programs, including the National Science Foundation (NSF) and the National Endowment for the Humanities (NEH), now require data management plans as part of applications. These policy changes reflect technological advances in data capture and storage, as well as increasing recognition of the strategic need for cost-effective means to share, preserve and reuse research data.

Regardless of the research question or theoretical approach, methodological rigor must underpin all archaeological endeavors. Unfortunately, many archaeologists lack awareness of the downstream research uses of digital data. Thus, they lack understanding of what good data management means in terms of their own research practices. A host of complex issues, including costs, technological capabilities, data documentation challenges, professional incentives, and legal considerations hinder the ability of archaeologists to better manage, use, and share their data (Averett et al. 2016; Kansa et al. 2011). Without examples of how standards, metadata, and data quality impact research outcomes from sharing data, field archaeologists will have little motivation to improve their data creation and management practices. Without clear professional incentives and rewards, field archaeologists will continue to regard data

management as a compliance issue, of only secondary or tertiary importance to core research goals.

Prior research has examined data and knowledge production practices during archaeological excavations (Edgeworth 2006; Khazraee and Gasson 2015; Mickel 2015) as well as the reuse of archaeological data and documents (Faniel et al. 2013; Huvila 2011). The SLO-data project takes a holistic approach to data, considering every aspect of data's lifecycle, including research design and planning, data creation or capture, use, dissemination, and preservation. The project aims to improve the practice of archaeological data management by developing cost-effective strategies to align data creation with reuse. It leverages existing best practice guidance by considering how data creation activities impact downstream reuse of data. Our study of both data creators and data reusers informs an understanding of these groups different needs across the entire data lifecycle (Figures 1 and 2). Only by considering how data flows in a research information ecosystem, before the tip of the trowel even touches the ground or a survey begins (Austin 2014), can we better meet the demands of data-intensive, twenty-first century research programs.

The emerging discipline of "digital data curation" has a growing body of literature documenting disciplinary data practices (among others, see Buchanan 2016; Faniel and Yakel 2017; Yakel et al. 2013). Until recently, most of the data curation literature has focused on digital research data archiving needs and practices (general overview: Borgman 2007; archaeology: McManamon and Kintigh 2010; Richards 1997). Researchers assume that repositories will do the heavy lifting preparing data for reuse. Unfortunately, reusable data also depends on the data producers (researchers) own practices. Data producers cannot simply handoff data to repositories to disseminate and preserve.

Although the NSF, NEH and others currently mandate the management of digital data, they have no specific requirements; leaving data management review criteria up to the discretion of review panels and

disciplinary practices up to specific fields. As review panels can be multidisciplinary, reviewers may lack knowledge of what constitutes a good data management plan in a discipline. To help fill this void, several university libraries and disciplinary repositories have come together to give the research community better guidance in grant-mandated data management. For instance, the DMPTool (https://dmptool.org) is an online system to aid the creation of project-specific data management plans. While useful in general terms, the DMPTool does not offer discipline-specific standards or guidance on good database design and organization (i.e., data modeling).

While data archiving has attracted funding and assumed greater policy importance, calls to archive data may not sufficiently motivate the changes in professional practice. In general archaeologists are not professionally rewarded for data sharing, grants usually do not fund data preparation to make shared data widely useful, and data reuse is not universally valued as a research approach. In terms of data preparation for reuse, archaeologists still invest little professional discussion or scrutiny in database design and documentation. Surveys and interviews conducted by the DIPIR project (http://www.oclc.org/research/themes/user-studies/dipir.html) reveal that archaeologists often ignore or see extant guidelines (such as the valuable guide co-published by the Archaeology Data Service (ADS) and Digital Antiquity (http://guides.archaeologydataservice.ac.uk/) as unhelpful (Faniel et al. 2013). Although they believe preserving archaeological data is important, they question whether it is viable, given such things as the skills, time, and money required (Frank et al. 2015).

The failure to align data management with research needs and outcomes undermines the whole point of data preservation (Huggett 2015). Data require substantive intellectual investment in design and validation to be usable in the wider community (Kansa 2015; see also Kratz & Strasser 2014). The data curation literature notes that the actual reuse of data remains rare in many fields (Peer et al. 2014; Wallis 2014; Wallis et al. 2013). Addressing the issue of data reuse has assumed greater urgency, given the

substantial investments flowing into disciplinary repositories (Faniel and Jacobsen 2010; Faniel et al. 2013). Case studies in archaeological data reuse are still rare, but as Kansa and colleagues (2014) illustrated, data reuse can require significant investments in labor in order to adequately prepare data for comparative study.

SLO-data Methods

Best practice guidance without reference to concrete and specific examples of research applications and outcomes may seem too abstract and irrelevant to the priorities of many practicing field researchers. The SLO-data project aims to complement existing best practice guidance by, for the first time in archaeology, formally documenting how data management impacts research practice for both the primary users and secondary data reusers. To do so, the SLO-data project uses a variety of qualitative research methods, including interviews and direct observations of archaeologists working in the field and in their labs (2). This paper focuses on data collected from two excavation sites in summer 2016 – Europe Project 1 and Europe Project 2 (3). Site descriptions for each including a description and status of the excavation and composition of the team follows.

Europe Project 1 Background

Europe Project 1 has been excavated before, but the project directors have only been involved since 2013. They took small groups of students to the site in 2013 and 2014 to excavate test pits, do geophysics, understand how the soil worked and what they might recover, and build relationships with the local residents. The first main trench was not opened until 2015, which the project directors considered the first season. They took it slowly trying to make sense of the site, and determining if it had been excavated previously. They undertook limited excavations, invited several specialists to visit, and started to focus their research questions.

In 2016, the project directors began excavations in earnest, spending one month at the excavation site with a core team of 45 people, primarily comprised of students from Europe, Australia, and North America. Each student group had particular objectives given their experience. Undergraduates who finished their first year were to experience excavating from beginning to end, often for the first time. Those who finished the second year were training to be supervisors next year, while third years were supervising and assisting students and managing data and masters' graduates were involved in decision making. Other experienced, returning students were assigned particular roles given their interests; a PhD student and two master's students were assigned to be the finds manager, data manager, and volunteer coordinator respectively. High school students and volunteers also participated in the project, rotating through in teams of four/week and four/day. The geophysical surveyor and director of the local museum also comprised the core team. In addition, different specialists were present year to year depending on what was found. In 2016, an archaeobotanist was on site.

Europe Project 2 Background

Europe Project 2 has been an active research and excavation site for over 50 years, under the supervision of three successive project directors. The research questions and approaches have varied over time as key personnel changed. By collecting general data associated with daily life in a community that spanned two centuries, the current project director discussed exploring issues surrounding social structure and "the economics of political identity". The 2016 season focused on refining understanding related to the chronology of the site, including the emergence of the elite and the circumstances of the site's origin and destruction.

The project director has worked on Europe Project 2 for 30 years. In 2016, he and his team of 52

people spent five weeks in the field. The majority of the students on the team were from North America, and these ranged from new students on their first excavation to returning students, cultivated over time to supervise, and create, manage and document data. In addition, there were fine arts and design students brought on the team to create plans and illustrations of artifacts architecture. Like the project director, there were team members with long tenures. Some students stayed on to work on their masters' theses, PhDs, and beyond, and some senior leaders, such as the field director and the epigrapher, have worked on the project almost 15 years. This project had a wider array of core staff than Europe Project 1. In addition to the database manager there was a lead conservator, an assistant conservator, and a cataloger. Like Europe Project 1, this project had specialists visit each year to study or advise on a particular aspect of the project (weaving tools, coins, ivory objects, human bones). In 2016 there was a zooarchaeologist studying faunal remains and an environmental team conducting flotation and wet-sieving.

Data Collection

Before arriving at each excavation site to interview and observe project members, the SLO-data team conducted semi-structured interviews with project directors to learn more about the site. Interviews included questions about the director's time and responsibilities on the project, the team, past excavation activities and plans for the next three years, research questions, data being created, data standards and procedures in use, and tools and software being used to create, record, and manage data in the field. In some cases, key team members also were interviewed in advance to learn more about their roles on the project. Although there was some variation in the questions asked given roles, the topics of inquiry were similar to those of the project directors.

During summer of 2016, one SLO-data observer was sent to each site. Upon arrival, the project director(s) introduced the SLO-data observer to the team and gave them an opportunity to describe the

study, including team members' rights as study participants, and then to request their participation. One student at Europe Project 1 declined to participate, citing strict rules around participation in such studies at their university. At the excavation sites for two weeks, each SLO-data observer interviewed and observed project members engaged in data events, as they worked both individually and in teams. The interviews and observations focused on team members who collected, documented, managed, and analyzed data at the excavation site. Interviews were audio recorded and later transcribed by an outside transcription service. The SLO-data observers wrote their observations in notebooks and later typed them up along with follow-up questions and general thoughts and reflections. They also took photos of the site, including data practices and materials used to support them, such as guidelines, notebooks, forms, tools, technology, software, etc. (Figure 3). In addition, project directors gave the SLO-data team access to the databases they used to store, manage, and discover excavation data.

At the end of two weeks, the SLO-data observer for Europe Project 1, had conducted 8 interviews, lasting approximately 10-90 minutes and 29 observations lasting approximately 2-30 minutes. The interviews and observations included key personnel such as the project director, finds manager, data manager, plans manager, environmental archaeologist, specialists, and students. For Europe Project 2, the SLO-data observer conducted 11 interviews, approximately 30-60 minutes and 21 observations lasting approximately 2-220 minutes. Key personnel included the project director, current and former catalogers, catalog supervisor, data manager, conservator, operations director, field director, trench supervisors, specialists, and students.

Data Analysis

Our team analyzed interviews and observations using NVivo, a qualitative data analysis software. The SLO-data team worked together to create the initial codeset based on interview and observation

protocols and samples of text from the interviews and observations. For instance, interviews and observations were coded for mentions of team members' responsibilities, facilities and infrastructure, tools used during an excavation, data descriptions and standards, training, identifiers, workflows, data updates, links, validations, etc. Next, two members of the team each coded the same interview, calculated their agreement, discussed and resolved their discrepancies, and refined existing codes or added new ones that arose from ongoing analysis. After several rounds, the coders reached a reliability of .81 using Scott's Pi, a statistic showing high inter-rater reliability among the two coders.

In addition to interviews, SLO-data team members with database expertise examined the databases created by the different excavations. Among other issues, they documented schemas (the organizational structures of databases), consistency and validation measures, and use of identifiers. Both Europe Projects 1 and 2 used fairly complex (with more than 12 data tables each) relational databases to document excavation contexts and catalog objects. Neither project made much use of controlled vocabularies and both emphasized unconstrained free-text description. In both Europe Project 1 and 2, specialists working at these projects created and used their own databases (or spreadsheets), independent of the primary excavation databases. As discussed below, this issue highlights some of the challenges in integrating specialist data contributions into the larger record of an excavation.

In preparation for year 2 data collection, the SLO-data team selected a subset of codes to examine for themes and patterns. The objective was to develop a set of recommendations for the project directors that might improve data documentation and management within the team. The codes selected were related to key data practices and flows at the excavation sites and included the following: codes highlighting mentions of data linking, transferring, updating, and validating; workflow codes that highlighted mentions of satisfaction versus problems with processes; workarounds vs. formal changes in processes and timing or sequencing of processes; and codes reflecting mentions of the use of local vs. global data standards and

names (i.e. identifiers) to reference items for the excavation. The team analyzed the data individually and together to identify key themes that were common across the projects. Next, the team met with project directors to discuss the findings and recommendations. Project directors from both projects agreed with the findings and accepted most of the recommendations. In the sections that follow, we discuss both.

Year 1 Findings

Analysis of observations and interviews from 2016 yielded numerous topics across both sites surrounding work practices that impacted data management. This paper focuses on three key themes – data management training, managing identifiers, and communicating with specialists about data expectations.

Data Management Training

In both Europe Projects 1 and 2, students were expected to create and manage data during excavation by recording observations in the field via hand written documentation, photographs, and database entries. Both projects approached data management training using a learn-by-doing approach and experts trained students one-on-one. While this is expected in a field school setting, findings showed there were some inefficiencies, which put a strain on experts and frustrated students. In both cases, experts provided one-on-one training, where group training would have been more efficient. Moreover, findings indicated written guidelines existed to supplement training, but in several cases the guidelines were insufficient or inconsistently used. This meant experts expended additional time and effort re-training students and correcting errors, which impacted the time they spent on other project responsibilities.

The database manager for Europe Project 1 sat down with each student doing database entries for the first time to discuss what to enter in each field. Although there was a cheat sheet to guide students through entering data from their context sheet to the database, it was not sufficient. Observations showed that there was a translation process that had to take place from hand written field notations to database entries that students did not always pick up and remember after one training session. As a result they encountered errors when inputting data incorrectly or created database inconsistencies by entering data in multiple formats for the same field. For instance, multiple photograph identifiers that were hand written as a sequence on the context sheet ("N189-N194") had to be entered into the database in separate fields (Figure 4). Database entry errors led to frustrations for some students, because there were no adequate error messages or written guidelines about how to resolve them. After explaining this to one of the project directors, he said, "What students don't know is that you can't search for it then, because actually…" and the interviewer responded "well, people don't think like a database" and the interviewer agreed. Yet, the students were not trained to think like a database and consider how changing the order, format, or spelling of a field impacted later use of the database.

In the absence of written guidelines, the project director for Europe Project 2 demonstrated the photograph editing process to each new student. Using guidelines in his head, the project director told the student how to perform image cleanup and documentation as the student wrote it all down. Some students expressed the desire for clearer guidelines, beyond informal verbal training. For example, one of students wanted more instructions so that the drawings produced followed standard conventions. Other students developed their own guidelines based on what they were taught. A previous cataloger described why she wrote the first cataloging manual in 2004/2005, after the project director taught her how to do it. "And there really wasn't anything written down prior to that. It was all just kind of well, [the project director] teaches you, or somebody else teaches you what to do". Interestingly, when the current cataloger joined the project in 2014, she also learned the job via one-on-one interactions, even though a cataloging manual

existed. She explained how the database manager and project director quickly explained how to enter data and the important things not to do. She also sat with the previous cataloger to have what she described as "a description and condition talk, less like data management and more cataloging conversation." There were additional things the current cataloger needed from the former cataloger that were profession specific and neither the project director nor the database manager could provide the information and it was not written down in the cataloging manual.

In all the examples discussed across the two projects, findings showed that the project team benefited from having data management training provided via direct interactions but also needed specific documentation on data management to supplement the interactions. Given that the experts providing the training were senior project members with responsibilities in addition to training, written documentation became all the more important, because it would serve to not only reinforce training and create consistency in student performance, but also provide data management memory allowing experts more time on other tasks and project directors less memory loss with staff turnover.

Managing Identifiers

The SLO-data team found that identifier management was a challenge across both projects. Identifiers are unique names given to artifacts, contexts, or other archaeological entities. They can be created in a variety of ways, from simple numeric sequences (e.g., artifact 001, 002, 003) to more complex site-specific protocols. The use of identifiers in archaeological data management is common, but standards for the implementation of identifiers vary greatly. Generally, for identifiers to be effective, there needs to be some control over how they are created in order to maintain consistency in naming and to prevent "collisions" when non-unique identifiers create ambiguity in naming.

One approach to this is to empower one person with the authority to mint all new identifiers, therefore ensuring data integrity. However, this caused bottlenecks during excavation (Figure 5), as seen in interviews and observations from Europe Project 1, where unique identifiers for excavation contexts, special finds, and elevations were all controlled by the database manager. He explained that wait times for identifiers were particularly problematic when he was busy doing something else or there was a queue. In the former case, team members short on patience took sheets from the registry into the field which caused rework after he found out he and another team member were assigning the same sequence of identifiers to different things. In the latter case, he explained that queues typically formed, “because people make two or three trips. So they'll come for the small find number, which is fair enough because that's what they need, and then they'll go back, and then they'll come back with the context number if they forgot it, and then that'll take them, I don't know, 10 minutes, because there's always a queue for the levels [identifier numbers].”

Europe Project 2 took the opposite approach to identifiers by allowing different team members to create them. Consequently, an artifact could be given several different identifiers depending on how it was treated upon excavation. Different identifiers were assigned by the original excavator, the conservator, the cataloger, and different specialists, because they all kept separate databases or spreadsheets to track their work. Many of these identifiers involved different rules for creation. For example, excavation supervisors assigned integer values, unique only within a specific excavation unit and date of excavation. These “small finds” identifiers were only unique if associated with their excavation date, unit information, and the page of the trench book where they were described. Without these three pieces of information, these “small finds” identifiers would be ambiguous, in part because the assigned integer value in every excavation area started at “1” each day. In addition, assigning many different identifiers to an object made it difficult to cross-reference all the information recorded about it for common retrieval. This was especially true for the small finds identifier that the original excavators

assigned in the field. Field-assigned small finds identifiers required date information to resolve and the date information was expressed, sometimes incompletely, on hand-written tags in a variety of styles and formats (Figure 6). A specialist expressed concern about the small finds identifiers for objects not being unique, noting that any expert observations she made in the lab would not necessarily make it back to the original excavators. In other words, the excavators would not be able to take her analysis of an artifact and trace it back to their original recording of it to correct mistakes in identification in the field trench books.

Specialist Communications

In both projects, observations and interviews indicated that there were no clear guidelines for how specialist should integrate their datasets with the project database. In Europe Project 1, this was due in part to project directors not wanting to dictate how specialists should record their data. One of the project directors from Europe Project 1 did not think it was a good way to collaborate, even if that meant specialists data didn't make it into the excavation database. "I'd rather people just record what they... The ways they have to do it, and then we just have to deal with that and put it in somehow. It might be that some data doesn't get put in. So, it might just be that some of the specialist information doesn't get into the database, it just remains as a report."

Interestingly, a specialist from Europe Project 2 wanted clearer communications about how her data should be integrated into the excavation data workflows and records. "They seem to have a workflow worked out where, you know, they clean the thing, it goes and gets cataloged, the catalog number comes back to conservation so they know what the catalog number was, so they've closed that loop. But the identifications of the animal bones don't seem to be part of that, that process yet...So far my, the data that I produce, I produce in an Excel spreadsheet, that's it, every summer, and I leave like a copy of it with

[the project director]. And so far, the database doesn't accommodate that... I'm not sure where it ends up."

The project director for Europe Project 2 expressed similar sentiments, when asked about post-excavation data processing. He described his experience working with a GIS specialist who took all the geospatial information to create a very useful mapping tool, but it had not been integrated back in with the project's data system. He mentioned experiencing not only issues enticing others to help him, but also difficulties related to integrating their work into the broader project. He explained that oftentimes what he ended up with were "cul du sacs of experience" or "simply a one-off project" developed for a specialist's own purpose. In these cases, clearer communications between specialists and project directors prior to fieldwork could aid in integrating specialist datasets into the main project database.

Recommendations

The SLO-data project documented how the digital databases in both Europe Project 1 and Europe Project 2 play roles in the description and recording of excavations, contexts, and material culture. In both cases, the SLO-data project documented issues with data consistency, identifier management, and integrating specialist data. Based on our understanding of these challenges, the SLO-data team formulated the following recommendations, which are broadly applicable to archaeological excavations, particularly those with field school components.

How to "Think Like a Database"

In both Europe Project 1 and 2, excavation staff and field-school participants often faced knowledge gaps in creating new database records. Participants of both projects lacked clarity in what constituted data quality. High staff turnover rates on excavations also contributed to this problem. To remedy this issue, field schools should allocate time and resources in advance of field work to the following tasks:

- Formalize data management training through written training manuals in conjunction with training sessions for groups of students, as well as individual staff.
- Consider organizing group training sessions to cover the fundamentals and only following up with individual training sessions as needed to develop deeper, more nuanced expertise or provide additional feedback.
- Provide enough site-specific details in written guidelines so they can be referenced after training sessions and reduce the common questions, errors, and rework that can overburden supervisors.
- Document the data management responsibilities and procedures during the field season and update them annually to maintain data management memory and minimize the impact of staff turnover in key data positions (e.g., catalogers, photographers, illustrators, specialists, and any others needing to manage data identifiers).

While documentation and additional training can help clarify some aspects of data creation, project participants also need a more holistic perspective on the role of databases in archaeological interpretation. For the most part, the people creating new records of data make only limited use of databases for analysis or interpretation. For this reason, most project participants have little direct experience with how data creation processes impacts later data retrieval and analysis. Thus, rather than merely training archaeology students to "think like a database", field schools need to provide more opportunities for students to actually make use of project database records. As students gain practice in working with data, they will gain first-hand experience about the vital importance of issues like identifier management and data consistency.

Procedural manuals also provide important context for data reusers into the research design. How data were entered and perhaps more importantly what information about sites, finds, or loci was not deemed

important to document are critical for data reuse. Likewise, thinking like a database is important for data reusers. To get the most out of a data, reusers need to understand the construction behind any data management system to more effectively retrieve the data. This data literacy goes beyond just being able to use a specific system to also being able to query underlying databases to fully mine data.

Context and Better Management of Identifiers

Identifiers have fundamental importance in data management to unambiguously name the main subjects of archaeological inquiry (artifacts, ecofacts, contexts, etc.). Furthermore, identifiers provide the basis for linking together disparate records of information. For example, one needs good identifier management to associate a database record of an object with a database record of its archaeological context. Indeed, researchers use database identifiers to model and record archaeological context. In other words, good documentation of archaeological context requires good identifier management practices. However, in recommending the following we recognize that this is a difficult task, especially for ongoing excavations. We found changing identifier management practices after a few years let alone 50 years had implications for compatibility with existing data in the database as well as the paper archive. In addition, there were other issues that were out of the control of the project directors, such as the lack of wifi connectivity and institutional regulations. This is why we recommend the following:

- Delegate a "namespace" for excavators and other project staff members who need to assign identifiers. A namespace is a symbolic container for identifiers that enables multiple authorities to assign identifiers independently without risk that these authorities will assign duplicates in error. For example, an area "A" excavation supervisor may assign identifiers to deposits like "Unit A-1" and "Unit A-2" without risk of clashes with an area "B" supervisor assigning identifiers like "Unit B-1" and "Unit B-2". This approach can reduce the friction of having one central authority to assign all identifiers.
- Archaeologists often want to make identifier assignment sequential, because it helps track the

order of discovery and identification of contexts, finds, etc. Integer identifiers automatically assigned by many relational database applications makes this easy. However, if pairing automatically generated identifiers with the use of namespaces to accommodate multiple team members, identifiers will need to include alphanumeric characters for the namespace. For example, one of the identifier systems deployed in Europe Project 2 uses a namespace system along with leading zeros to create sortable sequential identifiers that read like "Prefix-20150020" (meaning, the 20th cataloged find of 2015).

- Automatically generate and manage identifiers via the database, if possible. Interestingly neither Europe Project 1 or 2 could implement this recommendation, but for different reasons. Europe Project 1 was required by local authorities to have a paper archive, and an entirely digital solution would violate that requirement. Europe Project 2 had teams working in multiple locations without network connectivity, so software-created identifiers could not propagate to all members of the team. Fortunately, certain new archaeological field data collections tools like FAIMS (https://www.fedarch.org/) can automate identifier creation with special algorithms that guarantee uniqueness, even without network connectivity.

As discussed above, workflow demands, network connectivity, and the desire to ensure sequential ordering of identifiers can require different strategies for identifier management. The strategies and rules behind identifier management also need to be communicated with specialists to enable future integration of specialist databases with other project databases. Without such associations, the data managed by specialists misses key elements of archaeological context.

Overcoming Specialist "Silos"

As described above, specialists data were not integrated with excavation data captured in the project

databases. Rather, specialist studies and datasets were siloed bodies of inconsistently managed data and documentation. While the silos may be due in part to specialists' pursuit of particular research interests and development of local data standards and best practices, we believe the data flows between specialists and the rest of the excavation team can be improved and recommend the following:

- Develop a written data policy for data analysis and have specialists review and agree to it as part of their participation on the project. In the policy consider including the project's expectations for sharing data within and outside of the project, a timeline for specialists to complete their analysis and plans for any data or conventional publications based on the data from the project.
- Discuss how to integrate specialists' data with the primary excavation data set, including conversations about identifier management, file formats, and documentation and metadata.
- Request that specialists provide a summary of the work they did before they leave the site each year and submit a draft of their data to the project director.

As is the case with student training, the best practices developed for specialist data will be more meaningful if specialists draw on other data from the excavation in their own studies. Project directors and specialists should therefore consider more integrative research questions, that make more holistic use of project data. Doing so will make specialist studies more contextually informed and integral to the overall project (see, for example, chapters addressing integration of zooarchaeogical and paleoethnobotanical data in Maltby 2006 and in VanDerwarker and Peres 2010). For data reusers, combining data in one system will enable better evaluation of data, such as data relevancy and completeness, among other factors in reuse.

Conclusions and Next Steps

To begin to satisfy the ethical and legal responsibilities of digital data preservation, Clarke (2015) contends that archaeologists must be better stewards of their data at the point of data's creation and suggests archaeologists improve their project workflows such that metadata are integrated into their private archives and storage systems. Although metadata creation is important, findings from this study showed good data management demands much more, such as workflows for creating and propagating identifiers and measures to ensure consistency in recording. In short, data management extends to all aspects of archaeological practice, including education and training, human resource issues, and coordination between specialists. This broad set of workflow issues needs more attention and peer-review consideration in grant data management plans.

The SLO-data project is documenting how archaeological databases serve different purposes and functions. For both the Europe 1 and 2 excavations, findings suggests that the primary function of their databases is to document and archive field observations and material culture. This supports prior research, which found archaeologists' answer to the destruction that occurs during an excavation is to document data to contextualize the finds so that data can be reconstructed for later analysis (Frank et al. 2015). However, on the projects discussed in this paper, findings also indicated less emphasis was placed on building databases that enabled the search, discovery, and integration of data for quantitative analysis and interpretation within their own project team, let alone across project teams. This other role of the database can only be accomplished with more formalized and explicit data modeling that uses controlled vocabularies and more consistent recording, as opposed to the free text (unstructured data) we observed commonly in use. A more formal approach to database design and use would unlock uses of a database beyond the retrieval of individual records (to read), and would allow archaeologists to more fully realize the analytic affordances that digital data can bring.

In general, data management practices will likely improve if data creators start making more data analysis demands on their own databases. Instructional uses of project databases and more integrative specialist studies will promote changes that improve data practices. Furthermore, if archaeological reporting and publishing demands more "reproducibility" by requiring disclosure of data behind interpretative claims (see Marwick 2016), then researchers will face more incentives to create more analytically capable databases.

The changing landscape of how researchers want to use and reuse data also will promote practices that support wider interoperability not just within, but also across, projects. As discussed above, identifiers play a central role in overall data management. Identifiers enable linking and relating different records in and between databases. Identifiers can also reduce ambiguity in describing archaeological observations. For example, the string of letters "bronze" can mean a color or a type of material. But an identifier can be used to reference a controlled vocabulary or ontology that more precisely specifies the meaning of "bronze". Using identifiers to reference controlled vocabularies and ontologies, especially "standards" curated by various expert communities, represents a key aspect of larger-scale data interoperability and data integration (see example in Binding and Tudhope 2016). Indeed, "Linked Open Data" practices used to integrate data for meta-analyses (see Kansa et al. 2014) emphasize the use of Web identifiers to globally reference concepts in controlled vocabularies and ontologies.

As the SLO-data project continues, we will further explore project use of controlled vocabularies with an aim toward identifying practices that promote wider interoperability. However, we should expect that not all data will see the same levels of reuse, nor will all data see similar demands for interoperability. In this regard, an important goal for the SLO-data project will be to better understand how the management and reuse needs of specialist data may diverge from excavation databases like those managed in Europe Project 1 and 2.

Acknowledgments. The SLO-data project has been made possible by a major grant from the National Endowment for the Humanities (grant # PR-234235-16). Any views, findings, conclusions, or recommendations expressed in this article do not necessarily represent those of the National Endowment for the Humanities.

Notes

1. While “Big Data” gets most attention, data can have value in ways other than scale, as explored with ideas of “Slow Data” (Kansa 2016; see also Caraher 2016) and “Thick Data” (Shawn Graham cited Tricia Wang 2013, to introduce this concept to archaeology).

2. On April 11, 2016, the Stanford University Institutional Review Board (IRB) approved the SLO-data methods and practices with a “Notice of Exemption” of further review regulations.

3. The SLO-data project also has additional comparative field sites in South America and East Africa now undergoing data collection and analysis to be presented in future publications.

Figures

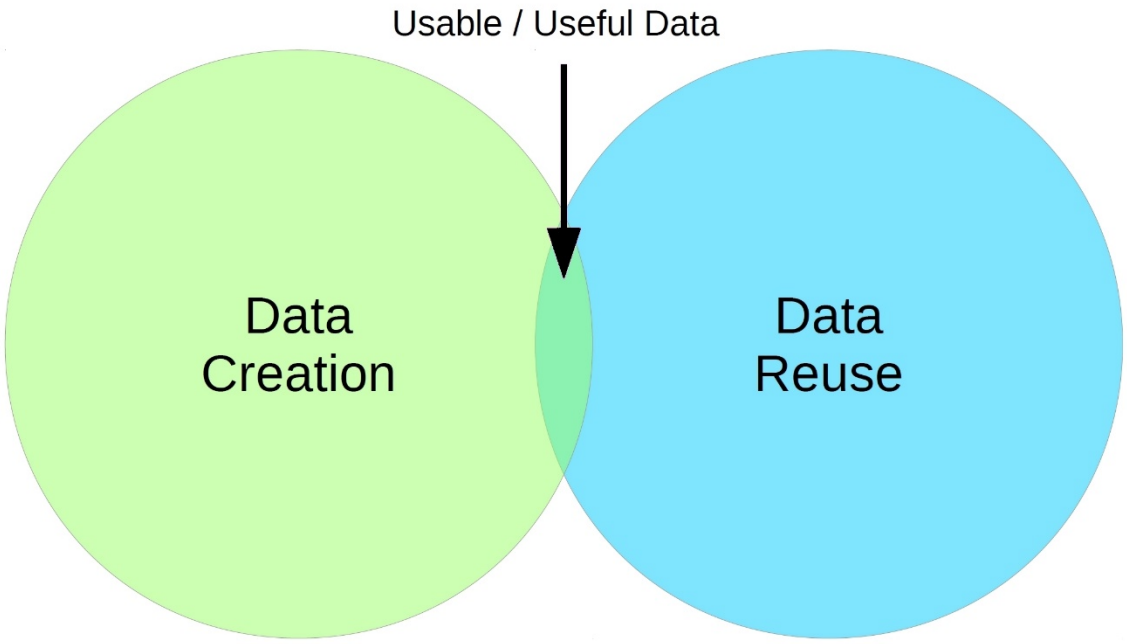

Figure 1: Venn diagram illustrating the tension between what data creators provide and what data reusers need. The SLO-data project aims to significantly increase the overlap.

Figure 2: The SLO-data lifecycle

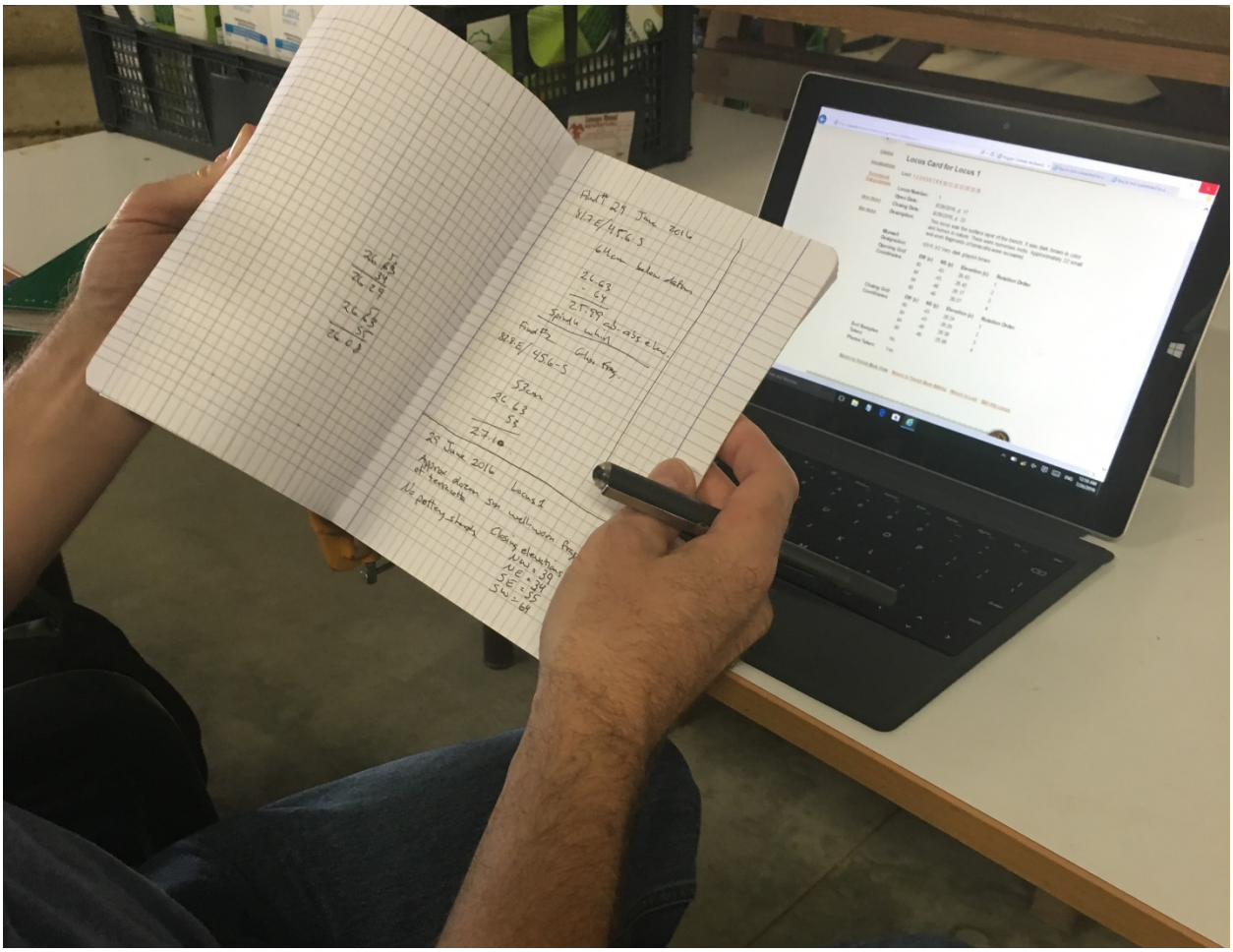

Figure 3: A team member transcribing paper trench notes into the project database

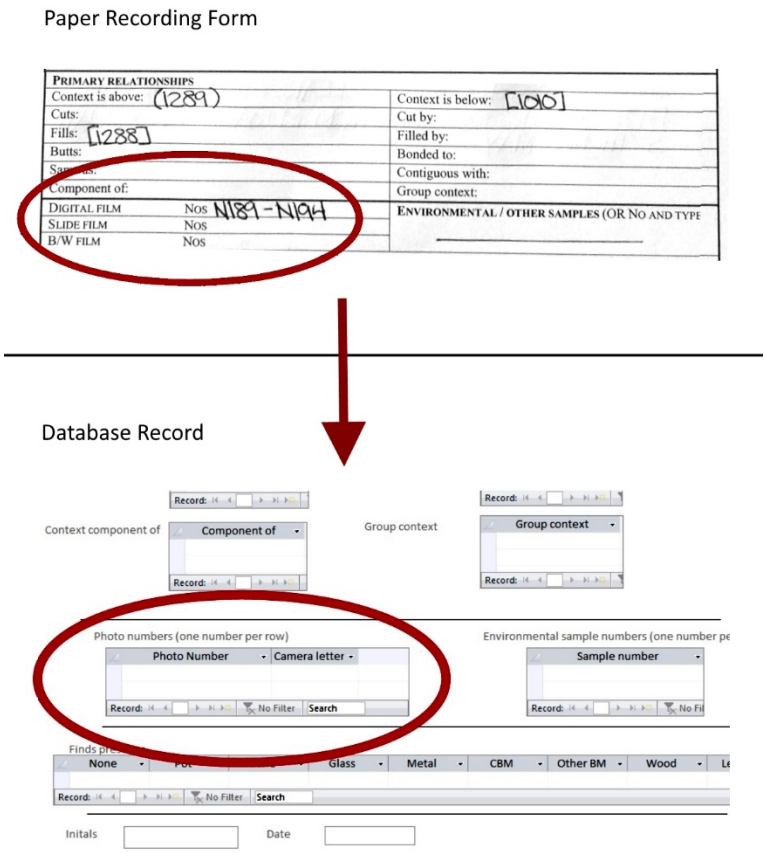

Figure 4: A portion of a context sheet (at top), showing the sequence of photos related to that context. The sequence is listed as “N189-N194” on the context sheet. Based on our findings and recommendations to the project directors, the database (at bottom) was updated to clarify how to enter photo names, given it differed from how photo names were written on the context sheet.

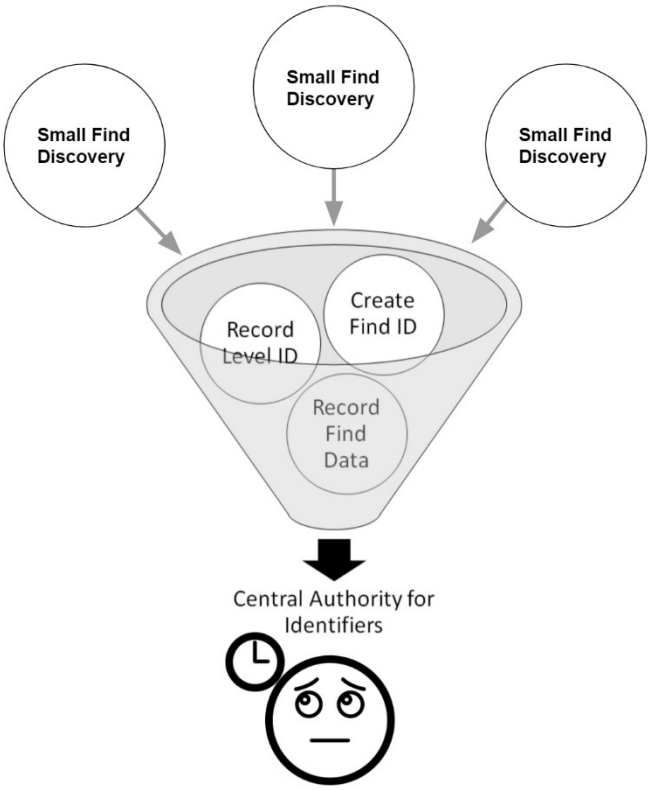

Figure 5: In Europe Project 1, the use of a registrar as the central authority to receive, respond, and record identifiers led to a bottleneck at the site. This placed time and stress on the registrar and led to longer delays between the discovery of a special find and its documentation.

Figure 6: Hand-written “small finds” tags, showing key information such as date and find number recorded in a variety of styles and formats across different units. This variability makes linking non-unique identification information difficult.

References Cited

Arbuckle, Benjamin S., Sarah Whitcher Kansa, Eric Kansa, David Orton, Canan Çakırlar, Lionel Gourichon, Levent Atici, Alfred Galik, Arkadiusz Marciniak, Jacqui Mulville, Hijlke Buitenhuis, Denise Carruthers, Bea De Cupere, Arzu Demirergi, Sheelagh Frame, Daniel Helmer, Louise Martin, Joris Peters, Nadja Pöllath, Kamilla Pawłowska, Nerissa Russell, Katheryn Twiss, and Doris Würtenberger

2014 Data Sharing Reveals Complexity in the Westward Spread of Domestic Animals across Neolithic Turkey. *PLOS ONE* 9(6):e99845. https://doi.org/10.1371/journal.pone.0099845.

Averett, Erin Walcek, Jody Michael Gordon, and Derek B. Counts

2016 *Mobilizing the Past for a Digital Future: The Potential of Digital Archaeology*. The Digital Press at the University of North Dakota, Grand Forks, North Dakota. http://dc.uwm.edu/arthist_mobilizingthepast/1.

Austin, Anne

2014 Mobilizing Archaeologists: Increasing the Quantity and Quality of Data Collected in the Field with Mobile Technology. *Advances in Archaeological Practice* 2(1):13–23.

Binding, Ceri, and Douglas Tudhope

2016 Improving Interoperability Using Vocabulary Linked Data. *International Journal on Digital Libraries* 17(1):5-21. https://doi.org/10.1007/s00799-015-0166-y.

Borgman, Christine L.

2007 *Scholarship in the Digital Age*. MIT Press, Cambridge, Massachusetts.

Buchanan, Sarah A.

2016 A Provenance Research Study of Archaeological Curation. PhD dissertation, Department of Information, The University of Texas at Austin, Austin, Texas.

Caraher, William

2016 Slow Archaeology: Technology, Efficiency, and Archaeological Work. In *Mobilizing the Past for a Digital Future: The Potential of Digital Archaeology*, edited by Erin Walcek Averett, Jody Michael Gordon, and Derek B. Counts, pp. 421–441. The Digital Press at the University of North Dakota, Grand Forks, North Dakota.

Clarke, Mary

2015 The Digital Dilemma: Preservation and the Digital Archaeological Record. *Advances in Archaeological Practice* 3(4):313-330. https://doi.org/10.7183/2326-3768.3.4.313.

Edgeworth, Matt

2006 Acts of Discovery: An Ethnography of Archaeological Practice. PhD dissertation, Department of Archaeology and Anthropology, University of Durham, Durham, United Kingdom.

Faniel, Ixchel, Eric Kansa, Sarah W. Kansa, Julianna Barrera-Gomez, and Elizabeth Yakel

2013 The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 295–304. ACM, New York. http://dx.doi.org/10.1145/2467696.2467712.

Faniel, Ixchel M., and Trond E. Jacobsen

2010 Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Computer Supported Cooperative Work* 19(3-4):355–375. http://dx.doi.org/10.1007/s10606-010-9117-8.

Faniel, Ixchel M., and Elizabeth Yakel

2017 Practices Do Not Make Perfect: Disciplinary Data Sharing and Reuse Practices and Their Implications for Repository Data Curation. In *Curating Research Data, Volume One: Practical Strategies for Your Digital Repository*, pp. 103–126. Association of College and Research Libraries, Chicago.

Frank, Rebecca D., Elizabeth Yakel, and Ixchel M. Faniel

2015 Destruction/Reconstruction: Preservation of Archaeological and Zoological Research Data. *Archival Science* 15(2):141-167. http://dx.doi.org/10.1007/s10502-014-9238-9.

Gordon, Jody Michael, Erin Walcek Averett, and Derek B. Counts.

2016 Mobile Computing in Archaeology: Exploring and Interpreting Current Practices. In *Mobilizing the Past for a Digital Future: The Potential of Digital Archaeology*, edited by Erin Walcek Averett, Jody Michael Gordon, and Derek B. Counts, pp. 1-30. The Digital Press at the University of North Dakota, Grand Forks, North Dakota.

Huggett, Jeremy

2015 Challenging Digital Archaeology. *Open Archaeology* 1(1):79-85. http://dx.doi.org/10.1515/opar-2015-0003, accessed November 15, 2017.

Huvila, Isto

2011 The Politics of Boundary Objects: Hegemonic Interventions and the Making of a Document. *Journal of the Association for Information Science and Technology* 62(12):2528-2539. 10.1002/asi.21639.

Kansa, Eric C.

2016 Click Here to Save the Past. In *Mobilizing the Past for a Digital Future,* edited by Erin Walcek Averett, Jody Michael Gordon, and Derek B. Counts, pp. 443–474. The Digital Press at the University of North Dakota, Grand Forks, North Dakota.

Kansa, Eric C.

2015 Reimagining Archaeological Publication for the 21st Century. In *Across Space and Time: Papers from the 41st Conference on Computer Applications and Quantitative Methods in Archaeology, Perth, 25-28 March 2013*, edited by Arianna Traviglia, pp. 367–378. Amsterdam University Press, Amsterdam.

Kansa Eric C., Sarah Whitcher Kansa, and Benjamin Arbuckle

2014 Publishing and Pushing: Mixing Models for Communicating Research Data in Archaeology. *International Journal of Digital Curation* 9(1):57–70.

Kansa, Eric C., Sarah Whitcher Kansa, and Ethan Watrall (editors)

2011 *Archaeology 2.0: New Tools for Communication and Collaboration*. Cotsen Institute of Archaeology Press, Los Angeles.

Khazraee, Emad, and Susan Gasson

2015 Epistemic Objects and Embeddedness: Knowledge Construction and Narratives in Research Networks of Practice. The Information Society 31(2):139-159. http://dx.doi.org/10.1080/01972243.2015.998104.

Kintigh, Keith, and Jeff Altschul

2010 Sustaining the Digital Archaeological Record. *Heritage Management* 3:264–274.

Kratz, John, and Carly Strasser

2014 Data Publication Consensus and Controversies. *F1000Research* 3. http://f1000research.com/articles/3-94/v1.

McManamon, Francis P., and Keith Kintigh

2010 Digital Antiquity: Transforming Archaeological Data into Knowledge. *The SAA Archaeological Record* 10(2):37-40.

Maltby, Mark (editor)

2006 *Integrating Zooarchaeology. Proceedings of the 9th ICAZ Conference*. Oxbow Books, Oxford.

Marwick, Ben

2016 Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory* 24(2):424–450.

Mickel, Allison

2015 Reasons for Redundancy in Reflexivity: The Role of Diaries in Archaeological Epistemology. *Journal of Field Archaeology* 40(3):300-309. http://dx.doi.org/10.1179/2042458214Y.0000000002.

Peer, Limor, Ann Green, and Elizabeth Stephenson

2014 Committing to Data Quality Review. *International Journal of Digital Curation* 9(1):263–291. http://dx.doi.org/10.2218/ijdc.v9i1.317.

Richards, Julian

1997 Preservation and Re-Use of Digital Data: The Role of the Archaeology Data Service. *Antiquity* 71(274):1057–1059.

Roosevelt, Christopher H., Peter Cobb, Emanuel Moss, Brandon R. Olson, and Sinan Ünlüsoy

2015 Excavation is Destruction Digitization: Advances in Archaeological Practice. *Journal of Field Archaeology* 40(3):325-346. http://dx.doi.org/10.1179/2042458215Y.0000000004.

VanDerwarker, Amber M., and Tanya M. Peres (editors)

2010 *Integrating Zooarchaeology and Paleoethnobotany: A Consideration of Issues, Methods, and Cases*. Springer, New York.

Wallis, Jillian, Elizabeth Rolando, and Christine Borgman

2013 If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLOS ONE* 8(7):e67332. http://dx.doi.org/10.1371/journal.pone.0067332.

Wallis, Jillian

2014 Data Producers Courting Data Reusers: Two Cases from Modeling Communities. *International Journal of Digital Curation* 9(1):98–109. http://dx.doi.org/10.2218/ijdc.v9i1.304.

Wang, Tricia

2013 Big Data Needs Thick Data. *Ethnography Matters* (blog), May 13, 2013. http://ethnographymatters.net/blog/2013/05/13/big-data-needs-thick-data/.

Yakel, Elizabeth, Ixchel Faniel, Adam Kriesberg, and Ayoung Yoon

2013 Trust in Digital Repositories. *International Journal of Digital Curation* 8(1):143–156. http://dx.doi.org/10.2218/ijdc.v8i1.251.

Data Availability Statement

The findings we report here result from analysis of qualitative data our team collected from interviews and observations with archaeologists in 2016. We have used strict naming conventions to manage the interview and observation data, according to the project IRB. To the extent possible, in consultation with the IRB, we will make the available on the web after the conclusion of the grant and after we have concluded analysis and prepared the data for sharing.