Data Beyond the Archive in Digital Archaeology

An Introduction to the Special Section

Sarah Whitcher Kansa and Eric C. Kansa

INTRODUCTION AND BACKGROUND

In May 2017, the editors of *The Economist* declared, "The world's most valuable resource is no longer oil, but data" (2017). This encapsulates a widely held belief among policy makers in government, the commercial sector, and universities that mastery over data represents a key strategic need in the twenty-first century. We increasingly see political contests over data, including scientific research data. The political struggles of 2017 included "data rescue" efforts that resisted Trump administration efforts to delete climate science information from federal information systems (Molteni 2017). This highlights how "research data management" has significance well beyond compliance with National Science Foundation granting policies and how scientific data can play an integral role in twenty-first-century power politics.

Given that data seem so self-evidently important in so many institutional settings, we may be tempted to assume that data would be a central issue of concern in a discipline such as archaeology. Indeed, the past several years have seen high-profile investments in programs to manage digital data in archaeology, as pioneered by the Archaeology Data Service, followed by tDAR: The Digital Archaeological Record, Open Context, and others. How much impact have these programs had in shaping archaeological

ABSTRACT

This special section stems from discussions that took place in a forum at the Society for American Archaeology's annual conference in 2017. The forum, Beyond Data Management: A Conversation about "Digital Data Realities", addressed challenges in fostering greater reuse of the digital archaeological data now curated in repositories. Forum discussants considered digital archaeology beyond the status quo of "data management" to better situate the sharing and reuse of data in archaeological practice. The five papers for this special section address key themes that emerged from these discussions, including: challenges in broadening data literacy by making instructional uses of data; strategies to make data more visible, better cited, and more integral to peer-review processes; and pathways to create higher-quality data better suited for reuse. These papers highlight how research data management needs to move beyond mere "check-box" compliance for granting requirements. The problems and proposed solutions articulated by these papers help communicate good practices that can jumpstart a virtuous cycle of better data creation leading to higher impact reuses of data.

Esta sección especial nace de las discusiones que tuvieron lugar en uno de los foros del Congreso Anual de la Society for American Archaeology en 2017. El foro, *Beyond Data Management: A Conversation about Digital Data Realities* ("Más allá de la gestión de datos: Conversaciones sobre las realidades de los datos digitales"), abordó los retos que se plantean al fomentar una mayor reutilización de los datos arqueológicos digitales actualmente conservados en repositorios. Los participantes del foro sostuvieron que la arqueología digital va más allá de su interpretación tradicional como mera gestión de datos, argumentando que es necesario situar de manera mejor el intercambio y la reutilización de datos en la práctica arqueológica. Los cinco textos que conforman esta sección especial abordan temas clave que surgieron de estas discusiones: el desafío de ampliar el alfabetismo de datos mediante el uso de los mismos como herramientas de instrucción; estrategias para lograr que los datos sean más visibles, mejor citados y más integrados en el proceso de revisión por pares; y formas de crear datos de mayor calidad que se presten mejor a la reutilización. En estos trabajos se destaca además cómo la gestión de datos de investigación debe ir más allá del simple cumplimiento del requisito de "rellenar casillas" para su verificación. Los problemas y las propuestas articulados en estos comunicaciones pueden ayudar a implementar mejores prácticas de creación de datos, que a su vez resultarán en un mayor impacto en la reutilización de los mismos.

Advances in Archaeological Practice 6(2), 2018, pp. 89–92
Copyright 2018 © Society for American Archaeology
DOI:10.1017/aap.2018.7

practice? Are digital data really the new "oil" in archaeology? And if so, how are we managing our supply?

In a recent blog post, Jeremy Huggett (2016) asked frank questions about the feasibility of reusing data that archaeologists archive in digital repositories. Huggett asked whether data are still too siloed, with too little linking for effective discovery and reuse. If so, what measures can we take to better capitalize on research data management so that data reuse becomes more commonplace? Huggett's blog post served as the inspiration for a forum that we organized at the 2017 Society for American Archaeology conference in Vancouver, British Columbia. Forum participants Anne Austin, Adam Brin, Ixchel Faniel, Timothy Goddard, Jeremy Huggett, Eric Kansa, W. Fredrick Limp, Ben Marwick, Julian Richards, and Adela Sobotkova discussed how to move "beyond the archive" to consider how archived and accessible data can be used and reused in archaeology. We were fortunate to have several audience members and participants tweet during the forum, and we collected the stream of tweets on Storify (Kansa 2017).

Themes that emerged during the forum discussions include the following, which we present as considerations for readers to keep in mind when reading the articles in this special section and when working with digital data:

- What is the value of data reuse? Why should data reuse be encouraged?
- How do we measure and track data reuse?
- What aspects of professional culture and incentives need to change to motivate more public data practices? How do we promote greater participation in "reproducible research" workflows?
- What kind of data should we reuse? How should these data be structured to encourage greater reuse?
- What kinds of skills do we need to promote to prepare archaeologists to better engage with research data?

OVERVIEW OF THE CONTRIBUTIONS

The five essays in this special section result from the Society for American Archaeology forum discussions. Four are authored by forum discussants, and one (Cook et al.) is an article we solicited in order to include insights from colleagues who teach with digital archaeological datasets (including, as a coauthor, Timothy Goddard, who was a named panelist but was unable to attend the forum). We would also like to highlight two additional publications, in press at the time of the forum and authored by forum participants (Richards and Brin), which, in discussing the Archaeology Data Service and tDAR, touch on some of the themes discussed in this special section (McManamon et al. 2017; Richards 2017).

This special section considers digital archaeology beyond the current status quo of "data management" to better situate the sharing and reuse of data in archaeological practice. Within this theme, articles discuss data stewardship and preservation, new pathways for interpretation and science, the place of "Big Data" in archaeology, public engagement, transparency, public policy, compliance, and improving digital literacy. Contributions grapple with the realities of working with digital data, ensuring access to archaeological data in the future, and considering the ethical implications of data access.

Since Huggett's blog post helped inspire the Society for American Archaeology forum, we begin this special section with his contribution. Huggett, Marwick and Pilaar Birch, and Sobotkova all provide invaluable discussions about what we mean by data, use, and reuse. Different definitions of data help shape our understanding about reuse practices. Huggett highlights the fact that reports (presumably in PDF format), and not data, dominate content and use in many digital repositories. He sees far greater use of these reports than of files containing structured data (databases, spreadsheets, tables, and the like). To Huggett, the greater use of reports over structured datasets illustrates how data reuse remains limited in archaeology. In contrast, reports (as PDF documents) represent a key source of "data" to Sobotkova, even though this format requires a great deal of largely manual effort to process in order to extract analytically usable information.

The (stubborn) centrality of the article or report seems evident in citation practices. As noted by Huggett and by Marwick and Pilaar Birch, archaeologists are habituated to citing literature, not structured, readily computable data. If, in citing data, researchers fail to use digital object identifiers (DOIs), a persistent identifier tracked in library and publishing systems, then reference to datasets' citations will hide in obscurity. Even altmetrics, which typically measure references to DOIs, may miss data reuse. Interestingly, an important reuse of Open Context (the system we manage) published data centers on aggregation in another scholarly information system, Pelagios, as "Linked Data." The Linked Data paradigm emphasizes Web identifiers (Uniform Resource Identifiers) more than DOIs. Similarly, in order to contextualize data with rich metadata, Open Context regularly references Uniform Resource Identifiers for concepts in data stores such as PeriodO (for chronology), the Getty Art and Architecture Thesaurus (for general typology), Pleiades (for ancient Mediterranean places), and more. Yet these linked data methods of reuse would not register in DOI-centered data citation metrics. Regardless, the lack of widely applied data citation practices reflects the relative novelty and rarity of citing data.

At least in the context of the United Kingdom, Huggett notes that archaeologists are much more likely to deposit digital data in a repository than they are to make use of digital data in a repository. Marwick and Pilaar Birch report that fewer than 1% of archaeological datasets have been cited. Digital data management has come into archaeology as an "archival" activity. Data repositories archive and preserve data. It may be that many archaeologists see their main obligation with respect to data to center on preservation. A perspective of "data are for preservation" may impede reuse. As Huggett notes, if conventional physical repositories see little use, then why should we expect digital repositories to see much use?

Faniel and colleagues also touch on this issue of archival vs. analytic purposes of data. They report findings from two excavation sites where excavation teams created data to record field observations without necessarily optimizing their data creation practices to support analysis. They document problems managing identifiers, ensuring consistency in recording (including heavy use of "free-text" fields), and difficulties in integrating specialist data.

To Faniel and colleagues, data quality and modeling problems can be better identified and addressed if archaeologists prioritize making more analytic use of their own databases.

Although both excavation sites provided field school training, Faniel and colleagues find that training in data management needs to be more systematic and formalized. They also note how field schools should encourage student use of excavation databases to better teach how data creation practices impact later reuse. Cook and colleagues expand on the instructional and training challenges of archaeological data management. They highlight that although students are habituated to digital social media, undergraduates need to develop basic computational skills before they can work with even simple archaeological datasets, much less sophisticated linked open data resources and Web application program interfaces (enabling software interactions with an information system). In their experiences teaching with Open Context, the Archaeology Data Service, and Data Archiving and Networked Services (the Dutch national digital research archive), the steep learning curves and complexity of the data and information systems made student engagement with archaeological data very difficult.

The points Cook and colleagues make about complexity barriers to reuse help us better understand some instances of data reuse that have occurred. For example, the past few years have seen the tremendous success of the Portable Antiquities Scheme in catalyzing many research programs, including more than 100 doctoral dissertations and 20 major peer-reviewed publications (see https://finds.org.uk/research). tDAR has powerful data integration tools (McManamon et al. 2017), as demonstrated in a recent publication discussing research outcomes of the analysis of aggregated zooarchaeological data (Kintigh et al. 2017). Open Context has used "Linked Open Data" approaches to integrate zooarchaeological datasets, as well as government-created archaeological site file data from the Digital Index of North American Archaeology (Anderson et al. 2017). Some of these data have already supported multiple publications by different teams of researchers (Arbuckle et al. 2014; Atici et al. 2017) as well as instructional use described by Cook and her coauthors. These examples of reuse seem to center on data that have sufficient scale and consistency to attract attention and to motivate researchers to invest time and effort in understanding and analysis. Researchers will probably be more willing to invest time in navigating a complex information resource if it has sufficient relevance and significance.

Engaging with complex data requires time, appropriate skills (especially for technically challenging resources), and motivation. Sobotkova had the motivation to reuse "data" from PDFs, and she had the background in Bulgarian archaeology required to understand these reports. Sobotkova nonetheless regrets that archaeological training did not provide her with the skills to extract information out of large numbers of PDFs using computer-assisted text analysis, a lost opportunity for archaeology since so much information and infrastructure is built around long-form PDF reports. Unfortunately in the case of archaeology, the computer skills required to work with structured data, or text as "data," seem more rarified. Conventional literacy, the kind of literacy required to read PDFs, is widespread, while the data literacy required to unlock value from structured data that would allow us to treat large amounts of PDFs as corpora for analysis is not. We cannot just assume that this skill gap will be solved as "digitally native" students advance in the profession.

On this last point, Cook and colleagues also note how students quickly adopt faculty's tacit opinions about the relative value and prestige of reusing other people's data. With so few examples of data reuse to emulate, students may doubt that they will gain much by investing the time and effort to develop the skills and background needed to work in data reuse. Indeed, every contributor notes that data still fit awkwardly into archaeology's system of professional rewards and evaluation. Marwick and Pilaar Birch note poor enforcement of the data availability statements required in one of archaeology's most prestigious journals. They also note a widespread reluctance to share data, even data of low risk and low sensitivity to stakeholders (including indigenous communities).

Marwick and Pilaar Birch note that an impediment to the uptake "of data sharing in archaeology and other 'small' sciences is that [it] is unfunded, unrewarded, and only rarely required." While this is not a new observation, they do suggest several new strategies to address the problem. They suggest a formal citation structure and reforms to editorial policies, as well as stricter enforcement of such policies, to make data disclosure a routine requirement for peer-review publication. A focus on enforcing data availability and consistent data citation practices in conventional journal publication will indeed change the incentive landscape and help drive expectations and cultural changes around the role of data in scholarly communications.

CONCLUSIONS

As the managers of Open Context, we have a strong professional interest in understanding and encouraging data reuse. It may be, as Huggett discusses, that many of the obstacles to reuse will resolve as repositories amass more data. After all, an individual dataset may be of relevance to only a small number of niche research questions. It will take time to archive enough highly specialized datasets to sustain wide community interest. Even so, Faniel and colleagues show that data creation practices may not encourage reuse, because data creation may emphasize archival rather than analytic goals. Thus, even if our repositories had enough data to meet researchers' specialized interests, these data might be too cumbersome to use without the level of labor investment Sobotkova describes in her use of PDF reports.

These contributions highlight some common themes about incentives, complexity and skills barriers, and data creation practices that poorly align with reuse. The steep learning curves associated with software and data, especially those required to fully engage in reproducible research workflows, represent real obstacles to many archaeologists at all career stages. The complexity barriers to reuse may highlight tensions between the interests of data creators and data reusers. Data reusers will probably want relatively simple, consistent data at a scale and significance that make investing time in understanding and analysis worthwhile. Scale and consistency can best come when data conform to common standards. On the other hand, data creators typically want freedom to adapt and shape data recording practices to meet their immediate needs, ingrained habits, and particular research agendas. This can result in a great deal of variety and inconsistency in how data are created. Currently, digital repositories try to reduce barriers for data creators, by making deposit as easy as possible. However, the lack of consistency between datasets makes it harder for repositories to meet the needs of data reusers. Achieving greater interoperability (consistency) across diverse datasets comes at the cost of greater complexity. In the case of tDAR, that interoperability requires ontology mapping together with sophisticated software. In Open Context's case, interoperability comes about through schema mappings and annotations to common controlled vocabularies and ontologies via linked open data. Neither approach is simple or broadly understood by the larger archaeological research community.

Moving forward, how should our discipline reconcile the divergent interests of data creators and data users? What are the costs and benefits of creating data conforming to some sort of standard for straightforward interoperability? The forum discussion and these articles highlight such tensions and a multifaceted "chicken and egg" challenge in bootstrapping a virtuous cycle of data creation and reuse. In order to encourage more data reuse, we need examples of high-impact data reuse to inspire and guide students (see Cook et al.) and other researchers. We hope that these essays provoke more and wider engagement in digital data. A broader and more diverse community engaged with these issues will help bring fresh perspectives, new experiments, and more insight into how we can best make use of the data contributions created by our colleagues.

Acknowledgments

We would like to thank the Society for American Archaeology forum panelists, whose insightful comments inspired this set of essays, as well as all the authors who contributed to this special section. The Society for American Archaeology forum was made possible by a grant from the National Endowment for the Humanities to the Alexandria Archive Institute (PR-234235-16). Any views, findings, conclusions, or recommendations expressed in the forum or in the resulting publications presented here do not necessarily represent those of the National Endowment for the Humanities.

REFERENCES CITED

Anderson, David G., Thaddeus G. Bissett, Stephen J. Yerka, Joshua J. Wells, Eric C. Kansa, Sarah W. Kansa, Kelsey Noack Myers, R. Carl DeMuth, and Devin A. White
2017 Sea-Level Rise and Archaeological Site Destruction: An Example from the Southeastern United States Using DINAA (Digital Index of North American Archaeology). *PLOS ONE* 12(11): e0188142. DOI:10.1371/journal.pone.0188142.

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. DOI:10.1371/journal.pone.0099845.

Atici, Levent, Sarah Whitcher Kansa, Justin Lev-Tov, and Eric C. Kansa
2012 Other People's Data: A Demonstration of the Imperative of Publishing Primary Data. *Journal of Archaeological Method and Theory* 1(3):1–19.

Atici, Levent, Suzanne E. Pilaar Birch, and Burçin Erdoğu
2017 Spread of Domestic Animals across Neolithic Western Anatolia: New Zooarchaeological Evidence from Uğurlu Höyük, the Island of Gökçeada, Turkey. *PLOS ONE* 12(10): e0186519. DOI:10.1371/journal.pone.0186519.

The Economist
2017 The World's Most Valuable Resource Is No Longer Oil, but Data. May 6. Electronic document, https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource, accessed January 21, 2018.

Huggett, Jeremy
2016 Digital Data Realities. *Introspective Digital Archaeology* (blog), June 29. Electronic document, https://web.archive.org/web/20180128212126; https://introspectivedigitalarchaeology.wordpress.com/2016/06/29/digital-data-realities/, accessed February 2, 2018.

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, Sarah W. (@skansa)
2017 Beyond Data Management: A Conversation about "Digital Data Realities": Tweets Covering a Forum at the 2017 Society for American Archaeology Conference (March 31, 2017, Vancouver, BC). Storify.com, April 6. Electronic document, https://web.archive.org/web/20180121191614; https://storify.com/skansa/digital-realities, accessed February 2, 2018.

Kintigh, Keith, Katherine A. Spielmann, Adam Brin, K. Selçuk Candan, Tiffany C. Clark, and Matthew Peeples
2017 Data Integration in the Service of Synthetic Research. *Advances in Archaeological Practice*. DOI:10.1017/aap.2017.33.

McManamon, Francis P., Keith W. Kintigh, Leigh-Anne Ellison, and Adam Brin
2017 tDAR: A Cultural Heritage Archive for Twenty-First-Century Public Outreach, Research, and Resource Management. *Advances in Archaeological Practice* 5(3):238–249. DOI:10.1017/aap.2017.18.

Molteni, Megan
2017 Diehard Coders Just Rescued NASA's Earth Science Data. *Wired*, February 13. Electronic document, https://www.wired.com/2017/02/diehard-coders-just-saved-nasas-earth-science-data/, accessed February 2, 2018.

Richards, Julian D.
2017 Twenty Years Preserving Data: A View from the United Kingdom. *Advances in Archaeological Practice* 5(3):227–237. DOI:10.1017/aap.2017.11.

AUTHORS INFORMATION

Sarah Whitcher Kansa ■ Alexandria Archive Institute, Open Context, 125 El Verano Way, San Francisco, CA 94127, USA (sarahkansa@gmail.com, corresponding author)

Eric C. Kansa ■ Alexandria Archive Institute, Open Context, 125 El Verano Way, San Francisco, CA 94127, USA