Sociotechnical Obstacles to Archaeological Data Reuse

Adela Sobotkova

THE MARATHON OF DATA REUSE

As an early-career researcher working in landscape archaeology in southeast Europe I often reuse large regional datasets. Compiling such datasets is labor-intensive, requiring collaboration and teamwork. My colleagues and I search the internet eagerly in hopes of finding usable research data. Still, I have yet to find a colleague in my area who has encountered an archaeological dataset that is trivially reusable.

When I first read Jeremy Huggett's blog post "Digital Data Realities" (2016) in *Introspective Digital Archaeology* arguing that archaeologists do not reuse data enough, I thought: "He is exaggerating; of course I reuse old data all the time." But on further thought I see that this reuse is inconsistent and difficult to document through citations. In his example of Archaeology Data Service downloads, which number tens of thousands per year, Huggett argues that PDFs, not curated datasets, are at the center of archaeological interest. In a follow-up blog post, Huggett (2017) finds that citation measurement tools do not bring up many dataset DOIs, a problem for taxpayer-funded infrastructure curating data for future reuse. Whether the low dataset reuse is a function of poor measures or the difficulty of measurement or a reflection of actual practice is the question. If the measures are correct, why are digital datasets not seeing greater reuse, given that we now have massive archives for long-term digital data curation?

The short answer is that the digitization and preservation of data represent only the first hurdle to reuse. Additional obstacles include problematic data sharing and management practices and the great effort required to repurpose archaeological data. These problems are social as well as technical; they will only abate slowly through a combination of increased awareness of technical solutions and collaborations, better digital skills, and increased respect for data science in archaeology.

DATA SHARING: PRACTICES AND ATTITUDES

In order to characterize the rarity of data reuse, it is worthwhile to review current data sharing and reuse practices (Borgman 2012;

ABSTRACT

The ease of digital data capture and the proliferation of concepts such as the "data deluge" suggest that modern researchers are drowning in datasets. Yet citations of archaeological datasets are few and far between, pointing to low rates of data reuse. This article explores the difficulties that surround data reuse in large-scale regional research, including the cost and coordination necessary to extract useful data from digitized PDF reports. The amount of correction and enhancement matches the effort needed to undertake a small field survey project and can only be circumvented with a thoughtful application of computer-assisted text analysis. Missing data in excavation report PDFs are not only intractable but also insidious due to their concealed nature, leading to poor outcomes in terms of (re)use. Consequently, the degree of data reuse in archaeology has been overestimated.

La facilidad de captura de datos digitales y la proliferación de conceptos como el "diluvio de datos" sugieren que la investigación moderna se está ahogando en conjuntos de datos. Sin embargo, las citas de conjuntos de datos arqueológicos son escasas, lo que apunta a bajas tasas de reutilización de datos arqueológicos. Este artículo explora las dificultades relacionadas con la reutilización de datos en la investigación regional a gran escala, incluyendo el costo y la coordinación necesarios para extraer datos útiles de los informes digitalizados en PDF. La cantidad de corrección y mejora que se requiere iguala el esfuerzo necesario para llevar a cabo un pequeño proyecto de prospección de campo. Esto se puede evitar solo con una aplicación bien pensada del análisis de texto asistido por ordenador. Los datos faltantes en los reportes PDF de una excavación no sólo son intratables, sino también insidiosos debido a su naturaleza oculta, lo que lleva a malos resultados en términos de (re) uso. En consecuencia, se ha sobreestimado el grado de reutilización de datos en arqueología.

Advances in Archaeological Practice 6(2), 2018, pp. 117–124
Copyright 2018 © Society for American Archaeology
DOI:10.1017/aap.2017.37

https://doi.org/10.1017/aap.2017.37 Published online by Cambridge University Press

Faniel, Kriesberg, and Yakel 2012; Klump 2017; Shen 2016). A recent survey by Shen (2016) explores the perceived and actual barriers to data sharing and reuse and validates researchers' perceptions of data value and reusability with their practice of data sharing and reuse. Shen provides information on the practices of 423 staff members at Virginia Tech from eight colleges surveyed in November 2014. Some 57% of researchers reported the ownership of data with long-term value, and 44% believed that their data had reuse value (Shen 2016:161). The fraction of researchers actively engaging in data reuse, however, was only 6% across all disciplines. Another 55%–56% of the respondents had never or seldom reused existing data. The three top concerns about data reuse included difficulty finding or accessing reusable data, difficulty integrating data, and fear of misinterpreting data. When the findings were filtered by college, significant differences in how engaged researchers were with open-data communities were noted. Some 66% of respondents in the College of Liberal Arts and Human Sciences ranked their engagement in data sharing and access in the two lowest tiers ("nominal" and "pockets of activity") offered on the survey (Shen 2016:168). Virginia Tech has no Archaeology Department and thus provides only a broader picture of trends.

Focusing on archaeology, however, the picture does not improve much. McManamon et al. (2017:240) mention high numbers of page views and downloads in the Digital Archaeological Record but offer only citations of archaeofaunal data from Open Context as evidence of reuse. The 2011 Research Information Network (RIN) report on the Archaeology Data Service (ADS) lists a high rate of reuse and appreciation (RIN 2011:23–25). The ADS users, however, often download research data for their own use only when there is no need to share or cite it more widely (RIN 2011:29). Furthermore, the number of respondents is 83 (RIN 2011:21), pointing to a low level of engagement with the survey. If we take Aitchison and Edwards's (2008:12) estimate of 6,865 professional archaeologists working in the United Kingdom in 2007–2008, we have a 1.5% response rate. Even if these may be the most engaged innovators, the number is low compared with the perceived widespread nature of sharing and reuse.

Archaeologists were also specifically targeted in a survey of attitudes toward digital data sharing circulated by Frank Lynam in March 2015 as part of his doctoral research in the Department of Classics at Trinity College Dublin. In this self-selected sample of 246, the majority of respondents ascribed great value to data sharing and open access (http://linkedarc.net/surveys/arch-datasharing-results). A majority saw demand for greater data sharing but perceived that institutional regulations and a lack of expertise obstructed the sharing of digital data. Lynam's survey was designed to collect attitudes—perceptions and aspirations—rather than practices or patterns of behavior. The survey design was also less rigorous than Shen's. In Lynam's survey, 73.9% of respondents asserted to have "shared archaeological data that they have created." But the question does not let us evaluate the scope of this sharing activity. Was the sharing primarily among collaborators and personal contacts or with broader research communities? What was its frequency and primary incentive (personal request, institutional demand)?

Shen's careful interweaving questions provide ordinal measures for responses, allowing for a finer assessment of target behaviors. The answers to Shen's question of sharing reveal that "data are shared within limited scope or under limited conditions," such as among colleagues and personal contacts or "upon request" (2016:161). Shen's (2016:172) combination of questions interrogating engagement with data sharing and reuse allow researchers to reflect on these issues from both data producers' and data users' perspectives. Shen's survey design reveals significant discrepancies between the perceived value of reusable data and practices surrounding data sharing and reuse. The similarity between Shen's findings and results emerging from RIN and other sources indicate that archaeologists, despite their attitudes, share and reuse data within a limited scope. A number of factors contribute to this situation: from the low velocity of data acquisition (Borgman 2012) to the closed nature of research communities (Klump 2017:2–3). This article focuses on the costs and obstacles to archaeological data reuse.

What Are Archaeological Data?

Digital data and datasets in archaeology need to be defined or at least characterized. For data, I follow Uhlir and Cohen:

> The term "data" … is meant to be broadly inclusive. In addition to digital manifestations of literature (including text, sound, still images, moving images, models, games, or simulations), it refers as well to forms of data and databases that generally require the assistance of computational machinery and software in order to be useful, such as various types of laboratory data including spectrographic, genomic sequencing, and electron microscopy data; observational data, such as remote sensing, geospatial, and socioeconomic data; and other forms of data either generated or compiled, by humans or machines [cited in Borgman 2012:1061].

This broad definition includes PDFs and nontabular digital resources, which are my primary—and often only—source of information about the extinct landscapes that I study.

Borgman (2012:1061) also employs a broad definition of data in the social sciences, arts, and humanities. She recognizes that research data in "soft sciences" take many forms, are handled in many ways using many approaches, and are often difficult to interpret once removed from their initial context. This is true of archaeological datasets (Kansa and Bissell 2010; McNutt et al. 2016). Digital data in archaeology encompass a wide range of representations of objects unearthed in archaeological excavations, observations and descriptions relating to surface cultural heritage in its social and environmental settings, and many other representations and observations. These digital artifacts range from scanned texts and images, to digitally born tabular and geospatial data, to instrument data such as 3-D scans and remote sensing imagery.

While Huggett disqualifies PDFs from the status of data, I count them in, following Borgman (2012:1061). I use the term *dataset* when referring to structured, computer-readable information in tables or matrices. I extract such datasets from the scanned maps and PDFs of site gazetteers, by coding qualitative information into quantitative data and categorizing it according to content, relatedness, and purpose.

https://doi.org/10.1017/aap.2017.37 Published online by Cambridge University Press

Where Is the Deluge?

> The "dirty little secret" behind the promotion of data sharing is that not much sharing may be taking place [Borgman 2012:1059].

The absence of shared, relevant, and usable data is the primary impediment to data reuse. Huggett (2016) emphasizes the transparency and interoperability of digital systems when he asks what repositories can do to improve data discovery, reuse, and citation rates. Shen's (2016) survey captures nicely the bottlenecks in research data sharing and reuse among the owners and users of data. While researchers' choices regarding data curation (e.g., personal website, domain-specific repository, etc.) can impact upon data discoverability, it is the details of data sharing practices and the work that goes into digital data creation that most impact on reuse. Shen quantifies the actively sharing researchers at 6% (in other domains). If only a small fraction of researchers share datasets in an open and publicly accessible way, the likelihood of a relevant dataset being available is much reduced.

Why are data not shared more? Borgman lists a number of reasons:

> Researchers may lack the expertise, resources, or incentives to share their data. Data often do not exist in transferable forms. Some data are not sharable for ethical or epistemological reasons. In many cases, it is not clear what are "the data" associated with a research project [2012:1059].

Shen (2016) and Faniel, Kriesberg, and Yakel (2012) articulate additional limiting factors such as institutional policies and the lack of standards, time, and funding. Klump (2017) points to varied cultures of sharing among disciplines, which may cancel out even generational change. "Generation Y" doctoral students who behaved as "digital natives" in their private lives have reported following the behaviors of their role model supervisors in academic life (Education for Change 2012).

An important aspect that affects sharing is the amount of labor required for dataset collection (Borgman 2012:1066). Specifically, "the more handcrafted the data collection and the more labor-intensive the post processing for interpretation, the less likely that researchers will share their data" (Borgman 2012:1066). Low velocity is a defining feature of data from the arts, humanities, and social sciences as opposed to the hard sciences (Borgman 2015). "Low velocity" means that the data are slow to produce. Historians and archaeologists can devote months or years to extracting useful data from archives, trenches, artifacts, or landscapes. Interpretation requires experience, including deep knowledge of contexts, languages, and approaches, expertise requiring years of study. Labor-intensive approaches afford project directors flexibility and local control but generate datasets that are often not consistent in form or structure and thus hard to reuse by others. "Big Science" researchers using instruments (telescopes, automated sensors, etc.) may spend a lot of effort designing and developing their tools, but once deployed they can collect massive amounts of standardized data that can be used by many people (Borgman 2012:1065). Scientists produce data at a much higher velocity than archaeologists and develop data models and management plans in parallel with research design. The low velocity of archaeological data thus contributes to the poor availability of datasets and low rate of data sharing among researchers.

Open Data, Progressive Data, Reusable Data

Open digital data should make it possible to ask new questions. The respondents to Shen's survey, however, argued the opposite: "Data of others is rarely applicable to new problems" (2016:169). How are we to reconcile the open-data community's enthusiasm with this expression of frustration over data irrelevance?

Even though many studies underscore the tremendous value of open data and the fact that we are living in a data deluge (Australian National Data Service 2017), few truly open archaeological datasets exist that fulfill the FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson et al. 2016):

1. The work must be in the public domain or provided under an open license.
2. The work must be provided as a whole and at no more than a reasonable onetime reproduction cost and should be downloadable via the internet without charge.
3. The work must be provided in a form readily processed by a computer and where the individual elements of the work can be easily accessed and modified.
4. The work must be provided in an open (nonproprietary) format.

The FAIR principles facilitate data reuse by making data easier to obtain and cite and by helping users understand the original aims and purposes of data owners.

I have encountered a few datasets that fulfill these criteria, but most of them originate far from my area of study. Perhaps the best example of a well-documented survey dataset is the Intensive Survey Data from Antikythera, Greece, deposited in the ADS (Bevan and Conolly 2012). It is accompanied by a *Journal of Open Archaeological Data* essay that clarifies the context and methodology of the survey and points to dataset DOIs. The essay has been downloaded more than 250 times, but data download metrics are unavailable. I downloaded the dataset from the ADS and use it for teaching and reference. I also use datasets from other online sources such as the Comparative Archaeology Database from the Department of Anthropology at Pittsburgh (http://www.cadb.pitt.edu/) and the Digital Atlas of Roman and Medieval Civilization in the Harvard Dataverse (https://dataverse.harvard.edu/dataverse/darmc) for teaching, but none are suitable for my direct use.

The bulk of my data for Bulgaria are far from FAIR. My geospatial data have no license information, source, or author embedded within the files and come in a proprietary format (Esri geodatabases and shapefiles). Most column names are self-explanatory, but the year of creation, purpose, error, and resolution are unspecified, which concerns me when (e.g.) road vectors or contours of soil groups disagree with paper-based sources. I have recently discovered an online document clarifying the origin of some of the geodatabases—nine years after I first started using them. To this day I am not sure how to transform the mysterious 1970 projection (which uses a partially documented Krasovsky ellipsoid) common to old Bulgarian maps to a

https://doi.org/10.1017/aap.2017.37 Published online by Cambridge University Press

more modern one. I found a plausible transformation online, but my implementation of it in QGIS failed. I will be able to use the dataset once someone cracks this particular problem.

Among the scanned maps I use there is an offprint of a regional atlas that specifies no author, publisher, or year of publication. The PDFs of annual reports (*Arheologicheski Otkritiya i Razkopki*, or AOR) that I use are circulated among local practitioners but would be hard to find via a search by an outsider. Hard copies of the reports are accessible through Bulgarian libraries, but access to these libraries often requires language skills, letters of reference, or approval by a recognized archaeological authority. Bulgaria has a huge problem with looting of cultural heritage, and thus such limitations are understandable (Bailey 1998; Stoyanov and Lozanov 2008), but they severely limit the reuse of even hard copy information. Knowledge of these resources is implicit and passed from colleague to colleague. Data documentation is a massive problem. While a lot of sharing goes on, it is informal, ad hoc, and embedded in implicit knowledge rather than explicitly documented.

DATA REUSE: FORGING THE RIGHT DATASET

I study the past cultural evolution of communities in Bulgaria in their environmental context. I use a lot of legacy resources, especially scans of old maps, atlases, and site registers, since features I am interested in often no longer exist. I also use modern digitally born data, such as digital terrain models, satellite imagery, and other sensor-based resources. Bulgaria may seem an odd example given its former Eastern bloc country status, yet in data management it is ahead of its neighbors. Entwined with national identity, archaeological data have always been a priority (Bailey 1998). Since 1973, permit-holding archaeologists have had to publicly report on the previous year's campaigns in order to receive a permit for the next year. Annual meetings where these presentations happen provide the fastest way to learn who excavated what, where, and how. Details from these meetings are published in official annual reports (AOR). In addition to these reports, a comprehensive electronic register of sites was established in 1996 to provide a centralized and standardized resource for cultural heritage management (Nehrizov 2005). An ordinance of the Ministry for Culture that mandated the use of the National Register of Archaeological Sites (Arheologicheska Karta na Bulgaria, or AKB), a searchable online database, was issued amid general economic decline after the fall of the USSR. The AKB collects key data about each archaeological site in the country following national standards, similar to the OASIS records in the United Kingdom (Domaradzki et al. 1988; Richards 2017:228). As a result, Bulgarian archaeologists today submit site cards to AKB in addition to the annual oral and written archaeological reports. Access to AKB is limited to registered practitioners. While downloading data upon registration is possible, it is constrained by administrative region. I mostly find it tedious for large-scale projects (due to the absence of batch downloads). AOR and other regional reports (*Izvestia*) and stand-alone publications document fieldwork projects in more detail than site cards in the AKB. The AOR have been scanned in as PDFs and are easily accessible among insiders. Yet information from these reports needs to be extracted manually and interpreted by someone with good domain knowledge.

In my most recent project, I have been studying the spatial distribution of burial mounds on the Thracian Plain in southeast Bulgaria. I extracted thousands of burial mound locations from scans of old topographic maps, and a team of volunteers helped verify their existence and dimensions using Google Earth. I have complemented this information with mound dimensions, morphology, and chronology extracted from excavation reports and previous campaigns of ground verification. This project started a year ago, and I coordinated the teamwork remotely with the help of manuals, validation, and feedback loops. So far my team has verified the status and spatial information for 1,200 remotely sensed mounds and collected cultural information on about 900 excavated mounds.

The Cost of Assembling Bulgarian Burial Mounds

This work has consumed more than 1,100 hours of time and involved half a dozen student volunteers (Google Earth verification), me (geospatial data extraction, coordination, review), and three paid assistants (data extraction from reports). I had a budget of A$7,000 and spent most of it on data extraction from PDFs and geospatial tagging by Bulgarian-speaking and geographic information system–savvy assistants. The initial data extraction took circa 700 hours. The task of verifying and refining the x/y-coordinates in two columns of the final table alone took 125 hours by a highly skilled PhD student.

After a year of work my budget is gone, and I have my dataset. The dataset, however, is far from ready for use. The data extracted from PDF reports—my main cultural control sample—are neither clean nor flawless. Minor problems include the errors in transcription and formatting created by assistants (e.g., commas instead of periods for decimals, words in numeric fields, etc.), but these are automatically flagged, and I can fix them during data review. Missing or ambiguous mound locations (e.g., "2 km NW from the village") produced imprecise locations. Problems of subjective or fuzzy archaeological realities pushed to the surface during dataset creation. We struggled to distinguish the meaning of a "grave" from that of a "burial" in reused mounds with indistinct or disturbed grave boundaries. Likewise, intersecting clusters of ashy piles and skeletal remains prevented neat separation into burials and associated sacrifices. Given the often laconic nature of reports, we had to rely on the interpretation of the authors (if it was present). I suspect that much fuzziness had already been removed in the process of writing the reports and that problems were much more pervasive in reality.

Missing and incomplete information proved difficult to resolve, especially when it concerned basic parameters such as the dimensions of a mound. Core data omissions varied in scale with date and report type. Between 1980 and 1990, only 34% of AOR report mound dimensions. From 2000 on the dimensions are reported 80% of the time. When we look at the 1960–1975 *Izvestii* and other reports from the first half of the twentieth century, dimensions are mostly included (83% of reports have them). AOR, *Izvestii*, and stand-alone reports differ in reliability. Reports from the beginning of the twentieth century were high quality because they were published as stand-alone publications (Filov,

https://doi.org/10.1017/aap.2017.37 Published online by Cambridge University Press

TABLE 1. Extracted Mound Data.

		Have Diameter		Have Height		
Source	**Total Mounds**	***n***	**%**	***n***	**%**	**Hours on Task**
Izvestia and stand-alone reports	283	239	84.5	229	80.9	253
Arheologicheski Otkritiya i Razkopki: 2000–2016	372	299	80.4	294	79.0	273
Arheologicheski Otkritiya i Razkopki: 1980–1999	293	99	33.8	126	43.0	200
X/y-coordinates acquisition and assessment	948					125
Total	948	637	67.2	649	68.5	851

Velkov, and Mikov 1934; Škorpil 1925). The AOR did not have the status of final publications but served as synopses of past work for colleagues. By 1990 Bulgarian archaeologists recognized that final reports did not always happen, and so the standards for reporting in AOR increased. The standardization issue was also one of the catalysts for the creation of the AKB (Nehrizov 2005). In the end, only two-thirds of my mounds have the dimensions I need for my analysis (see Table 1).

Given my aim to explore the dependence of mound morphology on location, reuse, and chronology, missing dimensions and spatial definitions are criteria for data rejection. This means discarding well over 30% of my hard-acquired dataset. It seems wasteful to spend thousands of dollars processing resources that will have to be discarded in the end. While missing GPS coordinates can be estimated from a verbally reported location and supplied with an error radius, dimensions are irretrievable once a mound is destroyed. My assistants tried to estimate dimensions from available plans and figures, but this helped in fewer than a dozen cases. Chronology proved a minor problem as it was fastidiously reported. The precision and accuracy of the reported chronological definitions are sometimes openly problematized, but coarse chronology was sufficient for my purposes.

In the end, after I filter out the incomplete records (no dimensions), I can use 66% of the mounds collected by my assistants. The fact that circa 30% (250 hours; A$2,000) of student assistant time appears to have been wasted makes me wonder: Could I have done things differently?

Digital versus Digitized: Faster but Not Necessarily Better

Incidentally, while my team was cleaning data, I learned of a PhD student who was working on a dissertation on Thracian burial mounds in the United Kingdom. She shared an MS Access database with me that she had built for her PhD. It contained just over a hundred royal mounds, and I hurried to download it. I hoped that perhaps she had found a way to get over the limitations of PDFs.

The moment I opened the Access database, my expectations were crushed. Her entries were as incomplete as mine. Her conceptual categories mostly overlapped with mine, differing mainly in research emphasis. Her interest was in burial assemblages, while I focus on morphology and location. I could have used the information in her database, had I found it sooner and had I spent my time merging the relational database with my Google sheets and verifying the data. By this time, however, my assistants had covered the same ground. While some effort had been duplicated (50 to 100 hours possibly), my dataset has internal consistency, which is valuable.

The encounter provided two lessons. Encountering a database based on the same PDFs brought home the benefit of structured tabular data. In a few summaries I could spot the problems immediately, and I did in an afternoon. Even if the database had been an order of magnitude larger, the assessment would have taken a similar amount of time through the use of functions and sorting. A day spent on data review is a big difference from the 700 hours of work needed to tabulate decades of scanned PDF reports before I could arrive at a similar quantification of my data.

The experiment also confirmed that any derived dataset will inherit the limitations of its source regardless of its digital format. The core problem is the data missing in the source PDFs. While scanned PDF reports in my conceptual world occupy the position of data, they are very poor data. They hide problems and omissions. The deceptive nature of PDFs has a bizarre side effect. It increases PDF usage frequency. PDFs can conceal omissions as in my case, but there is also the chance that they conceal valuable information. If unavailable to semi-automated scrutiny or handled by unaware researchers like me, PDFs will be always find new readers hoping to discover new and interesting information.

Dealing with PDFs: Is There a Better Way?

Scholars such as Huggett do not consider PDFs of text as data unless they are enhanced through optical character recognition (OCR) and searchable, because there is no easy way to quickly see what a PDF contains and what it does not. PDFs of image resources, such as maps or plans, contain minimal text, a lot of symbols, legends, and metadata that allow the user to gauge their purpose and shortcomings. Printed text, scanned into digital format, conceals this information until you read it all.

My reports combined plain scans and OCRed PDFs. I knew that there were going to be problems with the reports. I was warned that the quality of information would decline as I moved back through older and older reports. I did not expect GPS coordinates before the 2000s, while I figured that "insignificant" finds such as secondary burials would be omitted. I was resigned to spending time to fix tractable errors, but I was surprised by the number of intractable ones—I was not expecting mound dimensions to be missing in over 60% of AOR reports from the 1980s. If I were dealing with decades of PDFs of excavation reports again,

https://doi.org/10.1017/aap.2017.37 Published online by Cambridge University Press

I might spend more time researching options for computer-assisted text analysis in order to guide my assistants to the fruitful resources that include the critical data.

Text is essentially data, and scholars in linguistics have tools to get it out, such as natural language processing, which obviates manual data extraction entirely (Tudhope et al. 2013). In my case, a less complex solution would have sufficed, as I wanted a human assistant to review and extract other attributes on top of the mound dimensions. Creating triaged lists of mound reports that contained key dimensions would have significantly reduced my assistants' load. The task could have been accomplished through the use of regular expressions on OCRed documents, using pattern recognition only.

While I can now conceptualize the technical solution, it did not even cross my mind before. OCRing of old documents is a fairly common practice, even though it always requires the attention of human reviewers. The next step would probably require someone writing regular expressions (the expert) and someone deciding on the keywords (me). What would the costs of such a tool and its fine-tuning be? To satisfy my questions I wrote up my needs, attached a sample PDF, and consulted a technical expert in residence in my department. I heard that my needs would require one day of setup once the requirements were nailed down. Manual work would still be necessary—finessing Cyrillic keywords, testing outputs, converting tables of content into usable author-title references with pages for the purpose of flagging productive reports. All of these would have increased my load. But a week of time might have been worth the 250 wasted hours of my assistants.

Running PDFs through OCR, fixing them, and using established tools of computational linguistics would have been labor-intensive. I am not a computational linguist, and so I would need to skill up or find a collaborator. Considering the ubiquity of PDFs, archaeologists may need to learn tools for text analysis; they are, after all, widely used in adjacent disciplines such as history and literary studies.

For one-off activities, collaboration is often a better option. However, collaboration would have placed additional demands of time on me upfront, in terms of developing an understanding of my own needs and of iterative and intensive testing. This work would be on top of later assistant supervision. The benefit would be that PDF triaging would have allowed me to save A$2,000 (minus the cost of technical help).

One reason I did not do this sooner is that while my assistants had 250 hours at their disposal, I did not have a week of time or mind space to run another subproject. I chose the traditional labor-intensive method because it was not my labor and because it was all I could coordinate while teaching and attending to my other responsibilities. I saw little risk in coordinating assistants, as the data extraction required relatively little expertise—a working knowledge of the Bulgarian language, the archaeological context, and my coding conventions—something I could explain to a Bulgarian student of archaeology in less than an hour. I would spend a couple hours a week reviewing the output and providing feedback to my assistants, but I could choose the time and was in control. Building a digital solution would have made me lose some of that control and would have imposed time, communication, and knowledge demands that were too much for me at the start of the project.

Mental Note to Future Self

No original data were created during my Bulgarian mound data mining stage. My team aggregated existing information in a format a computer could read and added missing information (geographic coordinates and status) so that I might proceed with my spatial regression and other analysis. The fact that all the resources were available in PDFs saved me from traveling to Bulgaria and made a regional assessment *possible*, but it has not made it effortless.

What would I do if I could do it again? I had the option to organize fieldwork with the budget but chose not to. Today, I would probably try going down the route of computer-assisted text analysis, as the experience would potentially provide a community-wide benefit for other scholars grappling with the same issue. It is easy to argue for a different route in hindsight. I did not feel the pain of reading useless PDFs. Deviating from a traditional method only makes sense to me in retrospect with the knowledge that 30% of the reports were not worth reading. If the percentage had ended up being only 10%, I might not have written this article.

CHOOSING BETWEEN DATA CREATION AND REUSE

The effort I spent on data reuse in this project could have easily been applied to a season of surface survey or legacy data verification. The hundreds of mounds excavated and recorded in my PDFs are gone, but there are tens of thousands of mounds still standing. While I would not get at the chronology of the mounds, I could, however, obtain a lot of precise information on the morphology and location of hundreds of mounds.

It is not only the cost that is similar between primary data collection and data reuse. The high amount of labor needed and the slow velocity of data production are also similar.

Both primary data collection and data reuse in my case require access (to data or the study area), research design, time to train and coordinate assistants, and time to liaise with local partners. While data collection includes project logistics, day-to-day management, and data curation, data reuse comprises data management, affordances for asynchronous multiuser collaboration, and avoidance of data loss. Both my fieldwork and desk work involve distributed teams of assistants with whom I communicate techniques for data collection and classifications and responsibilities and rights for reuse, analysis, and publication. Both projects take months to complete.

In terms of data processing, reusing the data of others is often more demanding than reusing one's own. The amount of labor increases with distance from the data source. A researcher who picks up a completely unknown digital dataset needs to invest considerable effort into testing and reconstructing the dataset's pedigree. Faniel, Kriesberg, and Yakel (2012) summarize the process of data reuse into three stages of understanding: (1) the conceptual model behind the dataset, (2) how qualitative data were

https://doi.org/10.1017/aap.2017.37 Published online by Cambridge University Press

transformed into quantitative data, and (3) how data might be matched and merged across multiple datasets. This process differs from primary data collection, where the researcher organizes observations of the world according to his or her own (implicit) data model.

When conceptualizing my project, I read dozens of AOR reports to assess how much they could be trusted and to what purpose they could be reused. I underestimated the decline in quality with time and failed to mitigate it. Natural language processing or other computer-assisted text analysis could have helped me by flagging quality issues as well as by doing the bulk of the preliminary assessment and leaving a smaller amount of valuable reports for laborious manual review. Learning to operate smoothly in a digital ecosystem (linguistic or other), however, poses a massive challenge to having the right digital skills. Marwick (2017:441) mentions three years of self-study for encoding a reproducible workflow for data analysis and publication in R. Few of us have months, and even fewer have years, and not all of us may want to go this far. Learning computer-assisted text analysis or natural language processing is, however, inevitable if we want to boost data reuse.

How are we to justify all this time spent on building expertise in data and technology when data reuse has none of the cachet and glamour of fieldwork? Without fieldwork, we all become data scientists. How sexy is that? Tongue-in-cheek comments aside, the archaeological community needs to work on the perceived sexiness of data reuse and value those who confront and cite the data of others. Perhaps funding bodies can assist us here by offering awards for projects without a significant primary data collection component or for following up on data publication for those projects that have received funding.

Digital data have the advantage of eliminating the tyranny of distance, and repositories and archives are to be much praised for this. Reusing digital data, especially if by data, we mean PDFs and other gray literature, however, consumes vast amounts of time and energy, making it into a self-contained archaeological project of its own. Reusing digital data is not automatically easier or faster than primary data collection.

CONCLUSION

Archaeological data are in the eye of the beholder. Given data's heterogeneity and the complexity of methodologies, approaches, and practices used for their acquisition, archaeological datasets pose difficulties for sharing and reuse. Well-documented and structured archaeological datasets are few and far between, because their production requires a lot of labor and because the rates of sharing labor-intensive data among researchers are low. Sensor-based data and unstructured data such as scanned PDFs are much more frequent than structured datasets but laborious to reuse due to the technical debt they entail. Manually processing large quantities of unenhanced PDFs for a landscape archaeology project has required an effort tantamount to a campaign of fieldwork. With 30% of reports missing basic, critical information, my PDFs underdelivered on their promise. The waste could have been avoided through the use of basic tools of linguistic computing, which would extract out only productive PDFs for manual review. As PDFs are often the only source of "real" data and their reuse in archaeology is labor-intensive, archaeologists need to learn computer-assisted text analysis if they want to increase the speed of data reuse and accelerate new knowledge production.

Domain-specific repositories curate thousands of PDFs, but many of these contain only poor information. Using the linguistic digital tool kit has the potential to make better use of these resources and inject some speed into the cycle of data creation. Rather than building new technical solutions, however, sociocultural change is needed. The archaeological community needs to commit to building digital literacy and rewarding data science and reuse. This change needs to come not only from funding agencies but from within the community itself (Klump 2017:5–6). Universities can help by offering training in tools that are emerging as the standard for reproducible research. Scholars who have experience with programming tools and who have ventured into the linguistic domain stand to bring about the future archaeological data deluge.

Acknowledgments

I thank Sarah and Eric Kansa for organizing the Society for American Archaeology forum "Digital Data Reuse" and Juan Pablo Maldonado Lopez for translating the abstract into Spanish. No permits were required for the research reported here.

Data Availability Statement

No original data are presented in this essay.

REFERENCES CITED

Aitchison, Kenneth, and Rachel Edwards
2008 *Archaeology Labour Market Intelligence: Profiling the Profession 2007/08*. Institute of Field Archaeologists, Reading, UK.

Australian National Data Service
2017 The Value of Research Data. ANDS: Working with Data. Electronic document, http://www.ands.org.au/working-with-data/articulating-the-value-of-open-data, accessed July 20, 2017.

Bailey, Douglas W.
1998 Bulgarian Archaeology: Ideology, Sociopolitics and the Exotic. In *Archaeology under Fire: Nationalism, Politics and Heritage in the Eastern Mediterranean and Middle East*, edited by L. Meskell, pp. 87–110. Routledge, London.

Bevan, Andrew, and James Conolly
2012 Intensive Survey Data from Antikythera, Greece. *Journal of Open Archaeology Data* 1(1): e3. DOI:http://doi.org/10.5334/4f3bcb3f7f21d

Borgman, Christine L.
2012 The Conundrum of Sharing Research Data. *Journal of the Association for Information Science and Technology* 63(6):1059–1078.
2015 *Big Data, Little Data, No Data*. MIT Press, Cambridge, Massachusetts.

Domaradzki, Miecsyslaw, C. Lisitsov, A. Kamenarov, and S. Goshev
1988 Arkheologicheska karta na Bulgaria. *Arheologicheski Otkritiya i Razkopki Sofia* 34:177–186.

Education for Change
2012 Researchers of Tomorrow – The Research Behaviour of Generation Y Doctoral Students. JISC, London, United Kingdom. URL: https://www.jisc.ac.uk/reports/researchers-of-tomorrow

https://doi.org/10.1017/aap.2017.37 Published online by Cambridge University Press

Faniel, Ixchel M., Adam Kriesberg, and Elizabeth Yakel
2012 Data Reuse and Sensemaking among Novice Social Scientists. Preprint. In *ASIS&T 2012 Annual Meeting Proceedings*, pp. 1–10. Baltimore, Maryland. Electronic document, http://www.oclc.org/content/dam/research/publications/library/2012/faniel-data-reuse-sensemaking.pdf, accessed July 20, 2017.

Filov, Bogdan, Ivan Velkov, and Vasil Mikov
1934 *Die Grabhügelnekropole bei Duvanlij in Südbulgarien*. Staatsdruckerei, Sofia, Bulgaria.

Huggett, Jeremy
2016 Digital Data Realities. *Introspective Digital Archaeology* (blog), June 29. Electronic document, https://introspectivedigitalarchaeology.wordpress.com/2016/06/29/digital-data-realities/, accessed July 26, 2017.
2017 Citing Data Reuse. *Introspective Digital Archaeology* (blog), May 23. Electronic document, https://introspectivedigitalarchaeology.wordpress.com/2017/05/23/citing-data-reuse/, accessed July 26, 2017.

Kansa, Eric C., and Ahrash Bissell
2010 Web Syndication Approaches for Sharing Primary Data in "small science" Domains. *Data Science Journal* 9:42–53.

Klump, Jens
2017 Data as Social Capital and the Gift Culture in Research. *Data Science Journal* 16(14):1–18.

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* 5:238–249.

McNutt, Marcia, Kerstin Lehnert, Brooks Hanson, Brian A. Nosek, Aaron M. Ellison, and John Leslie King
2016 Liberating Field Science Samples and Data. *Science* 351(6277):1024–1026.

Marwick, Ben
2017 Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory* 24(2):424–450.

Nehrizov, Georgi
2005 Carte archeologique de la Bulgarie (1994–2000). In *The Culture of Thracians and Their Neighbours: Proceedings of the International Symposium in Memory of Prof. Mieczyslaw Domaradzki, with a Round Table "Archaeological Map of Bulgaria,"* edited by J. Bouzek and L. Domaradska, pp. 267–268. BAR International Series. Archaeopress, Oxford.

Research Information Network
2011 *Data Centres: Their Use, Value and Impact*. Research Information Network, London. Electronic document, http://www.rin.ac.uk/system/files/attachments/Data_Centres_Report.pdf, accessed October 10, 2017.

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

Shen, Yi
2016 Research Data Sharing and Reuse Practices of Academic Faculty Researchers: A Study of the Virginia Tech Data Landscape. *International Journal of Digital Curation* 10(2):157–175.

Škorpil, Karel
1925 *Megalitni pametnitsi i mogilishta (starini v Chernomorskata oblast—chast 1).* Materiali za Arkheologicheska karta na Bulgaria (kniga 4). Drzhavna Pechatnitsa, Sofia, Bulgaria.

Stoyanov, Totko, and Ivaylo Lozanov
2008 Thracian and Classical Archaeology in Bulgaria in the Years of Transition (an Attempt for Synopsis). *Anamnesis* 1(1):1–13. Retrieved from http://www.anamnesis.info/broi1/TStoyanov_ILozanov_EN_no1.pdf.

Tudhope, Douglas, Keith May, Ceri Binding, and Andreas Vlachidis
2013 Connecting Archaeological Data and Grey Literature via Semantic Cross Search. *Internet Archaeology* 30. DOI:10.11141/ia.30.5, accessed October 15, 2017.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons
2016 The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3: 160018. DOI:10.1038/sdata.2016.18

AUTHOR INFORMATION

Adela Sobotkova ■ Department of Ancient History, Macquarie University, New South Wales 2109, Australia (adela.sobotkova@mq.edu.au)

https://doi.org/10.1017/aap.2017.37 Published online by Cambridge University Press