Post-print of Gupta, Neha. 2020. Preparing Archaeological Data for Spatial Analysis. In *Archaeological Spatial Analysis: A Methodological Guide*, M. Gillings, P. Haciguzeller and G. Lock (eds.), 17-40, Routledge. https://doi.org/10.4324/9781351243858-2

Preparing archaeological data for spatial analysis

Neha Gupta (neha.gupta@ubc.ca)

Preparation of archaeological data for spatial analysis and the documentation of these procedures is now seen as key for effective management, analysis, interpretation and potential re-use of digital archaeological data. In an era of more user-friendly cyber-infrastructures in the field of archaeology, and with growing interest in the integration of information from diverse sources, greater efforts are now made to consider the quality of archaeological data, how best to document workflows and versions, as well as facilitate collaborative research. This chapter presents an overview of these issues, and offers guidelines and best practices in terms of scripted workflows, version control for managing data and open and collaborative research in archaeological spatial analysis.

Introduction

In the final pages of *Spatial Analysis in Archaeology*, Ian Hodder and Clive Orton (1976, p. 245) remarked that the 'slow collection of large bodies of reliable data, [. . .] will allow spatial processes to be better understood'. This scholarly work made explicit spatial concepts in the field of archaeology, and drew attention to cross-disciplinary conversations that archaeologists can have with geographers and social scientists. Published in 1976, Hodder and Orton's remarks might seem simple and banal, yet they underscore two key facets in archaeology that hold true in today's digital data-rich environment; first, that archaeologists will re-use archaeological data that were collected by other scholars at different times, who employed different methods, tools and technologies; and second, that for archaeology to engage in meaningful conversations on complex phenomena, we must have 'large bodies of *reliable* data' (emphasis mine) which can be understood as a reliable archaeological database (Gupta & Devillers, 2017, p. 857). Hodder and Orton (1976, p. 244) further note that spatial analytic techniques go 'hand-in-hand' with the 'collection of better data' and remark on the value of 'very detailed information' but they do not explicitly describe what reliable means, and how research design and the goals of a particular

project are linked to data quality and what role data quality might play a role in the analysis, interpretation and re-use of archaeological data.

Archaeologists increasingly face a scenario where the re-use of archaeological data, particularly the processing and analysis of digital archaeological data, is posing pointed challenges to the practice of archaeology (Kansa, Kansa, & Arbuckle, 2014; Huggett, 2015). The social life of archaeological data typically extends beyond specialists and is best understood in relation to local communities and society as a whole. The changing relationship between archaeology and society is reflected in scholarship on the abuse and misuse of archaeology (Silberman, 1989; Kohl & Fawcett, 1995; Meskell, 2005; Kohl, Kozelsky, & Ben-Yehuda, 2007), as well as on challenging inaccurate views of the human past (Wylie, 2002; Trigger, 2006). Archaeology (and archaeological data), therefore, can serve both public and scholarly goals.

Recent interests in the preparation of archaeological data for further use, and the quality of archaeological data are influenced by two broader developments, namely; the growing use of digital and geospatial tools and technologies in data acquisition (Dibble & McPherron, 1988; Levy & Smith, 2007); and second, the exponential growth of communication tools, particularly Web 2.0 technologies that facilitate collaboration and can encourage exchange and sharing of data between scholars, institutions and non-specialists (Kansa, 2011). The apparent democratization of archaeological data has renewed concerns over the privacy of archaeological sites and the sharing of sensitive locational information (Bampton & Mosher, 2001; Sitara & Vouligea, 2014). Growing awareness of a digital data-rich environment (Bevan & Lake, 2013) has spurred calls for Open Science in archaeology (Marwick, 2017), practices that aim to enable 'reproducibility', generate 'scripted workflows', 'version control', 'collaborative analysis', and encourage public availability of preprints and data (Marwick et al., 2017, p. 8). Efforts in open archaeology are premised on the belief that the 'analytical pipeline' in archaeological research has not been available for scholars to examine, critique and re-use, a situation that impacts the range and scope of archaeology (Marwick, 2017, p. 424).This situation is reflected in the prevailing use of 'point-and-click' commercial software that obscures underlying algorithms and assumptions. Moreover, archaeologists typically do not document (or do not report) sufficient information on how and why particular decisions were made during cleaning and analysis, a

situation that presents significant challenges in replicating analytical methods and results, even when data are available for re-use.

Understanding the role of quality information in archaeology is pressing as digital geospatial data acquired in the field are increasingly compiled with existing digitized information and together they are combined into computational pipelines (Snow et al., 2006; Kintigh, 2006). Archaeologists are accumulating large amounts of data through 'real-time' digital documentation in the field (Vincent, Kuester, & Levy, 2014), 'mobilizing the past' (Averett, Counts, & Gordon, 2016) and promoting 'transparency' in field collection (Strupler & Wilkinson, 2017). Typically paperless, these efforts are thought to minimize redundancy and human-introduced errors in the recording of archaeological sites and archaeological data (Austin, 2014) and potentially shorten the time interval between stages in the archaeological workflow (Roosevelt, Cobb, Moss, Olson, & Ünlüsoy, 2015).To manage, store and analyse these large amounts of digital archaeological data, archaeologists typically harness geospatial technologies such as commercial Geographic Information Systems (GIS). However, these spatial databases are known to have poor error management, a situation that can result in error propagation that impacts subsequent analysis and the final result (Figure 2.1) (Hunter & Beard, 1992). The widespread use of GIS in archaeology therefore can constrain broader assessments of archaeological methods and the quality of data in terms of interpretation and re-use. Moreover, processing of data and analysis within computational pipelines is rarely documented and shared (Costa, Beck, Bevan, & Ogden, 2013, p. 450), limiting what is known on procedures and transformations and the overall quality of sources (Evans, 2013).

Figure 2.1 Sources and types of errors in data collection and compilation, data processing and data usage that result in final global error, adapted from Hunter and Beard (1992).

In this chapter, I discuss the preparation of archaeological data for spatial analysis and re-use from the perspective of spatial data quality in the archaeological workflow. I draw from scholarship on geospatial data quality to shed light on data-centric issues in archaeology. I show that archaeological spatial analysis can be improved through better documentation procedures and transformations of archaeological data and discuss how these practices can facilitate deeper understanding of archaeological methods and practice and open new forms of research in

archaeology. I argue that documentation on data cleaning and tidying procedures, and version control can enable more rigorous research practice, and attune archaeologists to data-centric imperfections in archaeological data.

Uncertainty in geographical data is the recognition that there exists a difference between a complex reality and our conceptualization and measurement of that reality (Plewe, 2002). Our conceptualization of reality is necessarily a generalization and abstraction, and thus an imperfect model. Uncertainty in this model can be described as having three dimensions; space, time and theme. A map, for example, is an imperfect model of a complex reality; a full scale (1:1) map of a town could never be rolled out, nor would we make good use of such a document. Within this framework, we can better understand for each of the three dimensions, elements of quality, including error, accuracy, precision, consistency and completeness. Greater awareness of sources and causes of error and uncertainty can enable us to represent and manage imperfections within spatial databases, which in turn can facilitate greater confidence in the interpretation of digital archaeological data.

Quality issues are present in all data and throughout the research process, a situation that impacts the interpretation of archaeological data and decision-making (Figure 2.2). The preparation of archaeological data for analysis is tied to data quality which, in turn, is related to research design and the archaeologist's intended purpose for those data. Data quality can be understood in terms of internal and external. Internal quality is the 'level of similarity between data produced' and the 'ideal data' (data without error) or 'control data'. The ideal data are based on a set of specifications or rules and requirements that define how objects will be represented, which geometries will represent each type of object, the attributes that describe them and possible values for these attributes (Devillers & Jeansoulin, 2006, p. 38). Therefore when a result differs to some degree from what a theory was expecting, we have error, imprecision and incompleteness. External quality relates to data that were produced and how they meet the needs of a particular user. Whereas data quality elements can be measured separately for space, time and theme, an assessment of overall data quality requires careful consideration on all three dimensions because they are interdependent. Greater emphasis is now placed on 'fitness for use', which shifts focus to the needs of particular users and their intended use of data (Chrisman, 2006), which I will discuss in detail in a later section.

Figure 2.2 The archaeological workflow in terms of a computational pipeline from data acquisition to unpublished data, and re-use. Problems of quality impact data in each stage. Black boxes in this workflow occur wherever archaeologists employ software and tools whose code are unavailable to review and modify and that do not enable documentation of transformations.

Scholarship on quality issues in archaeology ranges from 'quality assurance' (Banning, Hawkins, Stewart, Hitchings, & Edwards, 2017), and 'quality standard' (Willems & Brandt, 2004) in field surveys to verifying and validating the quality of computational models (Burg, Peeters, & Lovis, 2016), the use of statistical techniques to address spatio-temporal uncertainty (Zoghlami, de Runz, Akdag, & Pargny, 2012; Kolar, Macek, Tkáč, & Szabó, 2015; cf. Fusco & de Runz this volume) and temporal uncertainty (Green, 2011; Bevan, Crema, Li, & Palmisano, 2013; Crema, 2012) to the quality of 3-dimensional photogrammetric models (Porter, Roussel, & Soressi, 2016). While fruitful, these efforts typically offer immediate solutions for project-specific problems and emphasize quantification techniques themselves, overlooking data quality issues in processing digital archaeological data and their future usage.

Recent interest in the quality of digital archaeological data shifts the focus away from positional accuracy as the primary concern in archaeology, as is reflected in works such as Dunnell, Teltser, and Vercruysse (1986), Dibble and McPherron (1988), Wheatley and Gillings (2002), Heilen, Nagle, and Altschul (2008), Atici, Kansa, Lev-Tov, and Kansa (2013), Evans (2013), Kansa et al. (2014), Wilshusen, Heilen, Catts, de Dufour, and Jones (2016), Cooper and Green (2016) and McCoy (2017). Managing and sharing digital geospatial data are encouraging archaeologists to think in terms of data-intensive methods and 'big data' i.e. data that are characterized by volume, velocity, variety, veracity, visualization and visibility (McCoy, 2017, p. 76; Green, this volume). Large sets of data, such as those that are generated through multiple field seasons, are greatly impacted by error (Dunnell et al., 1986). A similar scenario might be described for data collected through highly complex projects that involve the integration of multiple sources of information that are heterogeneous spatially, temporally and thematically.

To improve data documentation and quality, Kansa et al. (2014) place emphasis on editorial and collaborative review of archaeological collections. They suggest that data cleaning early in the archaeological workflow can avoid costly investments in terms of person hours, and publication delays later in the process, particularly when the archaeologist who collected the data

is not available to encode and link individual field documents. Re-use and comparison of existing digital archaeological collections is highlighting data 'accuracy, reliability and completeness' (Evans, 2013, p. 20), and the challenges in integrating diverse data that do not have clear documentation (Cooper & Green, 2016). Most importantly, recent scholarship greatly expands what archaeologists consider pertinent to the quality of data, and draw attention to elements such as error, accuracy, precision, consistency and completeness in digital archaeological collections. These efforts reflect growing intellectual interest in re-use of research data.

Research data management initiatives are now supported by governments. In the United States and Canada as well as other countries, publically funded research projects are required to lay out data management plans that ensure academic outputs are prepared for preservation and re-use (National Science Foundations, 2017; Tri-Agency, 2016). The Canadian Tri-Agency Statement for Principles on Digital Data Management, for example, includes an overview of the responsibilities of researchers, research communities, research institutions and research funders, as well as best practices in data management planning throughout the research project lifecycle. These broader developments are encouraging archaeologists in enabling sharing of digital archaeological data over Web-based platforms. Data publishers such as Open Context, and digital repositories such as US-based, the Digital Archaeological Record (tDAR), UK-based Archaeology Data Service (ADS), and the Advanced Research Infrastructure for Archaeological Dataset Networking in Europe (ARIADNE) offer new opportunities for data re-use. These efforts reflect greater control over metadata in archaeology and the potential for new forms of collaborative research.

Method

Consideration of data quality is invariably linked to research design and the goals of a particular project. Until recently, spatial data quality in archaeology seemed to refer to positional accuracy which is commonly associated with tools and technologies such as Global Positioning Systems (GPS) and remotely sensed imagery (Wheatley & Gillings, 2002). Emphasis on locational information comes as no surprise, given archaeological interests in field-based research and because of requirements in Cultural Resource Management (CRM) and planning to inventorise archaeological sites (Heilen et al., 2008). Archaeologists made great efforts to better model the

spatial dimension in digital archaeological data, overlooking the temporal dimension or chronology (Llobera, 2007; Rabinowitz, 2014). Yet archaeological data have spatial, temporal and thematic dimensions, all of which must be considered in any evaluation of data quality, and especially when data-intensive methods are employed.

Archaeological field data are typically complemented with terrestrial imagery (e.g. ground penetrating radar, aerial, and satellite imagery) and the recovery of portable artefacts such as potsherds, tools and skeletal material. Archaeological documentation of surface features such as earthworks, field walls, monuments, pathways, rock images, and subsurface ones such as hearths, camps, and dwellings and their spatial relationships with other recovered material culture can be thought of as a collection. In this conceptualization, the *archaeological database* is differentiated from the *archaeological record*. The latter refers to material culture that exists, whether it has been recovered or awaits investigation (Gupta & Devillers, 2017). The archaeological database consists of collections that archaeologists have successfully recovered at different times and places, and can be thought of as an imperfect model of a complex reality. A growing, more reliable archaeological database can facilitate insights into human history.

In practice, archaeologists are increasingly digitizing and integrating new archaeological data with archaeological collections stored in local and national repositories for combined analysis (Kintigh et al., 2015). Yet repositories are themselves a product of the society in which they were created, and thus, social, political, cultural and historical circumstances influence them. One might consider, for example, why a particular collection is chosen for digitization, and how and why specific classes of data within that collection are preserved and curated. These decisions can impact subsequent study of these research data. Data-intensive methods that integrate different collections take on their assumptions and limitations (Atici et al., 2013), in addition to uncertainties in any new data (Allison, 2008).

Moreover, at some point in the life of archaeological data, regardless of their acquisition through research or regulatory projects, they will be in the hands of experts who do not have direct access to the original data collectors, their ‘contextual knowledge’ and field journals. In practice, the person who acquired data on-site during archaeological fieldwork typically also encodes them for further use. Therefore, the encoder had pre-existing knowledge of the spatial relationships in the data that enable linkages between individual documents. Ideally, the same archaeologist analyses, interprets and presents results, a situation that is typical in small

academic projects. In the case of regulatory or CRM archaeology, digital archaeological data, once acquired, might be transferred to data analysts and data managers.

With site information, aerial photographs, geophysical readings, and topographic surveys, an archaeologist might prepare derived data products such as digital elevation models and files that store the location, shape and attributes of archaeological features (points, lines, polygons). These data are processed and analysed within a computational pipeline, the results of which are used to produce a synthetic document that receives some form of peer review either as a technical report, or a scholarly publication (Van der Linden & Webley, 2012). Such documents, particularly those produced under regulatory frameworks, in turn can be the basis upon which scholars and policy makers make decisions that impact local communities and society as a whole. Yet, in many cases, although not all, the data themselves are not subject to review (Gobalet, 2001; Roebroeks, Gaudzinski-Windheuser, Baales, & Kahlke, 2017), nor is quality information on research data necessarily made explicit (McCoy, 2017, pp. 4–5). This oversight can 'hide serious logical and empirical faults in the underlying assumptions' in archaeological practice (e.g. in CRM archaeology, the failure to detect archaeological sites despite 100% or full-coverage survey) (Heilen et al., 2008, p. 1.1). This situation, however, does not mean that the quality of data did not matter or that these data cannot be repurposed. Rather, recent scholarship suggests that archaeologists are concerned about the quality of data, and have criteria upon which they base their level of confidence. It should come as no surprise that insights into data management and field methods are often gained through repurposing of existing data.

For example, Wells (2011) presents the integration of archaeological information from four state historic preservation offices (SHPOs) in the United States. The SHPOs included in the study were Kentucky, Illinois, Indiana and Missouri and each office stored and maintained archaeological data in a GIS. The author notes that archaeological site records in each spatial database included similar basic information. Wells examines the format, projection and coordinate system of location information (e.g. polygon shapefile in Lambert conformal conic, measurements in feet) to assess interoperability across the four sources. To bridge the four sources, the author devised six categories of attributes, including location information, site identification, site type, definitions of one specific cultural affiliation, the quality of previous investigations and an assessment of site informational quality (2011). Although Wells does not explicitly define 'quality', he had clear criteria upon which to evaluate archaeological

information such as cultural affiliations and Mississippian culture change. Specifically, he bases the strength of these 'ontological definitions' on two factors, namely; the level of investigation at a site, as it offers consideration on how far an ontological characterization can be extended (maximum intensity of previous investigations), and second, the diversity of data structures used to represent the diversity of investigative approaches. Most fundamentally, this approach highlights thematic information in assessing overall data quality. Wells shows that careful evaluation of spatial and thematic accuracy enables thoughtful integration and meaningful repurposing of archaeological site records.

In an era of cyber-infrastructures, scholars are increasingly interested in 'grey literature', unpublished reports prepared by professional archaeologists under regulatory frameworks as a source of archaeological information. Some scholars have shed light on the 'accuracy, reliability and completeness' (Evans, 2013, p. 20) of these unpublished documents. In his examination of three sources on archaeological field investigations in England ―the National Monuments Record Excavation Index, the Archaeological Investigations Project, and Online Access to the Index of Archaeological Investigations―Evans (2013) suggests that greater efforts are necessary to understand the limitations of unpublished reports. The author examines and compares the three national spatial databases, and like Wells, considers archaeological site records or 'events' within them. In his study, Evans (2013, p. 26) devised overarching nomenclature to incorporate the range of terminology that describes on-site investigations such as 'post-determination/research', 'evaluation' and 'excavation'. The author then analysed the frequency of reporting across these investigative approaches between 1990 and 2007. Evans also examined each source for records on one county (Staffordshire) to ascertain gaps in coverage between them, challenging perceptions that national databases are complete and 'authoritative' (2013, p. 32). He concluded that meta-analysis highlight the uneven distribution of archaeological investigations, identifying regions where investigations have been overlooked and where accepted data standards have not been implemented (2013, pp. 21–22).

Similarly, in his examination of 'integrative databases', McCoy (2017, p. 77) remarks that 'clear biases' are evident when distribution of site records is presented on a map. The author defines integrative databases as those that 'continuously take in new information', usually from a variety of sources, distinguishing them from 'archival databases'. Archival databases are those that 'grow by accretion of distinct datasets' (e.g. tDAR, ADS) (McCoy, 2017, pp. 75–76). The

author suggests that the quality of geospatial data can be thought of as 'how well the dataset conforms to established best practices' (2017, p. 78). To this end, McCoy (2017, pp. 91–92) has proposed a 'standalone quality report' that describes the archaeological geospatial data and how they were derived for tasks such as research, assessment and documentation. This quality report would be a supplement to technical information in metadata. While potentially fruitful, we currently do not know how effective metadata and quality reports are in archaeology, to what degree quality information minimizes misuse of digital archaeological data and/or potentially enables their re-use.

In the English Landscapes and Identities project, Cooper and Green (2016, p. 289) seek to integrate diverse 'secondary digital datasets' from four national archaeological repositories (Green, this volume). These data include GIS-based vector files, associated documents in portable document format and spreadsheets, and in one case, records that were downloaded from a website (Cooper & Green, 2016, p. 300). The authors treated 'multiple and varied representations' of an archaeological entity within different sources 'as if it is accurate' (Cooper & Green, 2016, p. 292). The authors note that the source data have 'diverse histories, contents and structures' and are 'riddled with gaps, inconsistencies and uncertainties' (Cooper & Green, 2016, p. 294). The authors do not offer insight into what they consider to be 'inconsistencies and uncertainties' or what impact error and imperfections have on potential re-use of these data. They do, however, suggest that thematic information in site records can shed light on spatial relationships between archaeological sites, particularly when analysed at the 'national level'. The authors remark that at this scale of analysis, spatial precision becomes less important and emphasis shifts to the 'spatial character' of structures such as field systems including their length and orientation.

Heilen et al. (2008) examine data quality in American archaeology from the perspective of CRM projects. In their study of military installations for the Department of Defense, they focused on survey reliability, site location recording and site boundaries. They note that whilst overall accuracy of site location recording improved with the use of GPS, this brought other concerns to the fore. They suggest that with definitions and standards 'came the expectation that data collected at different times, by different contractors, would be equivalent in quality' (p. 5.1). The authors remark that this assumption has resulted in sites being 'mischaracterized'. For example, small artifact scatters that were recorded at a location were later identified as large

village sites, and some sites were missed entirely. Furthermore, they highlight key issues in the management of inventory data; specifically, that location information on archaeological sites can be accurate, yet, the 'size, shape, depths and importance' changes with 'environmental conditions' and 'academic debate', suggesting the complexity of delineating these attribute values (2008, pp. 5.6–5.7). They observe the problematic practice of deletion of 'repeated' site records in favour of the most recent site inventory record. In this context, the authors recommend detailed records on the history of site discovery and recording that include the equipment used, as well as details on field methods such as survey intervals, transect size and shapes, shovel-test design, and observations on erosion and visibility at the time of field documentation. While the authors do not discuss how data managers would interpret this information or how such an evaluation would impact the quality of inventory data, they do draw attention to thematic and temporal accuracy in digital archaeological data. These efforts underscore institutional practices as a factor in data quality in archaeology.

Data preparation is based on data quality, which in turn is fundamentally tied to research design and a user's intended purpose, as I have described above. In order for archaeologists to share the preparation of high quality data, methods and analytic techniques, we must document how our data transformed from one state to another. As noted, in digital environments, the use of commercial software within the archaeological workflow often means that we impose 'black boxes' that prevent us from examining and modifying underlying algorithms and code. In this context, a black box can refer to an instrument, device or software that receives an input and delivers an output, yet its internal workings are unknown or poorly understood by a scholar. This situation can cast doubt on the received output. When archaeologists give up the opportunity to critically evaluate and improve existing tools and technologies, we, in effect, limit the aims and scope of archaeology (Marwick, 2018). Much effort is put toward cleaning and tidying data to make them machine understandable and re-usable, yet these investments are lost in a computational pipeline that is closed to scholarly review, modification and development.

In data-intensive archaeology, three concepts are of prime importance; namely, scripted workflows, versioning, and open and collaborative research processes. These concepts are central in preparing archaeological data for data analysis and can be facilitated by data cleaning tools and techniques that offer 'recipes' or replicable steps. Some of techniques are applicable specifically to geospatial data while others have a broader scope. These tools and techniques in

turn, can create opportunities to disassemble black boxes in archaeology's computational pipeline. Documentation of data transformation and code sharing can enable more rigorous archaeological research, while also opening intellectual space for collaboration across disciplinary boundaries. I show how archaeologists might employ tools such as OpenRefine, languages such as Python and versioning systems such as Github to document, manage and share digital archaeological data and code. I emphasize that whilst these technologies might change, the goal remains the same: dismantling black boxes in archaeology.

Scripted Workflows

Scripted workflows are a way to document the research process, which in turn, can disable black boxes in archaeology. A script is typically a simple text file that consists of instructions to initiate and complete tasks in a computational environment. These instructions can be combined with other instructions to complete different tasks with the archaeological research process, or workflow. For example, an archaeologist might write a script to transform geographic coordinates (latitude, longitude) to Universal Transverse Mercator (northing, easting) coordinates based on some specifications, save these transformed data to a new file and display them on a map. In this case, the script serves not only as instructions for computational tasks, but also as a 'very high-resolution record' of the research process that can be shared, examined, modified and re-used multiple times and by different scholars (Marwick, 2017, p. 432). This is crucial as most commercial software do not enable documentation of the research process.

Scripted workflows have been utilized in different fields, including processing of geospatial information from different sources for land classification (Leroux Lemonsu, Bélair, & Mailhot, 2009). In a scripted workflow, the 'process becomes public, transparent and reproducible' (Thompson, Matloff, Fu, & Shin, 2017). A scripted workflow can contain instructions for several, often sequential tasks within, and throughout data processing, visualization and analysis and presentation. The key facet in a scripted workflow is its explicit description of process and code to enable transformation of data, a situation that can facilitate insights into decisions that were made during processing, their potential impact on results, and how best the data and code might be re-used.

Versioning as data management

Version control can offer a way to manage digital archaeological data. Versioning presents a history of digital objects and documentation on the creation of, and subsequent changes made on that digital object (Leeper, 2015). The concept draws heavily from software development practices that enable large teams of programmers in collaborating on code writing and keeping track of who has made which changes, as well as when and why these changes were made.

In research contexts, scholars increasingly employ data versioning, whereby a new version of a dataset is created when the structure, content or condition of an existing dataset is changed (ANDS, 2018). These changes include reprocessing, corrections and when additional data are appended to existing ones. A unique digital object identifier is assigned to each version of a dataset, enabling scholars to process and cite specific versions and compare across different versions if necessary. While effective in documenting changes in a dataset (e.g. lines in a text file), version control typically does not handle modifications in metadata, that is, the explanation of why a particular data value was modified.

Archaeologists generally agree that data have 'versions', and that these collections should be managed for future re-use. Yet at present, there is little agreement on when a collection has changed, warranting a new version, how best to document these changes and which versions to make publically available. These are pressing issues when it comes to government-sponsored data providers. Nonetheless, archaeologists are encouraging the use of version control throughout the phases of the archaeological workflow, as I will discuss in greater detail in a later section.

Open and collaborative research

Open science is a social development that is impacting the way scholars and scientists carry out research and communicate their findings in the 21st century. Facilitated in part by Web 2.0 technologies, Open Science aims to promote transparency, openness and reproducibility across scientific disciplines and change the culture of research publication (Nosek et al., 2015). To that end, Nosek et al. (2015, p. 1424) envision more rigorous publication policies, and they propose eight standards that open research communication aspires to, including citation standards, data transparency (sharing), analytic methods (code) transparency, research materials transparency, design and analysis transparency, preregistration of studies, preregistration of analysis plans and replication. Recognizing that journals vary across disciplines and that there are barriers to

adopting the standards, each standard is measured on three different levels (Nosek et al., 2015, p. 1425). The levels, increasing in stringency, are meant to facilitate gradual adoption of the eight standards. Implementation is recognized by 'badges'.

A growing number of scholars are interested in open research data as a way to practice 'better science' (Molloy, 2011; Foster & Deardorff, 2017). They draw attention to barriers to 'maximum dissemination of scientific data' such as inability to access data, restrictions by publishers on data usage, and difficulties in re-use due to poor annotation, as well as cultural concerns over losing control over data and the lack of incentives to make data re-useable. While informative, these efforts tend to address communication issues in research, overlooking deeper structural inequalities in academia and in society.

For example, in his call to humanize open science, Eric Kansa (2014, p. 32) draws attention to 'underlying causes' of dysfunction in research, beyond technical and licensing issues. He argues that broadening the boundaries of open science to encompass 'systematic study' creates intellectual space for social science and humanities scholars, enabling them to meaningfully engage with efforts in reforming research. Kansa (2014, p. 36) rightly observes that archaeology relies on primary research data, and that recovered material culture is not replaceable or renewable, yet archaeologists are often reluctant to disseminate and archive research data. He suggests that these challenges reflect neoliberal values and problematic institutional practices (Kansa, 2014, pp. 50-51). Most crucially, Kansa (2014, p. 52) remarks that a 'high level of collegiality and trust' are necessary for truly opening the research process to a wider community, a situation to which archaeologists can certainly relate. He suggests that open science can succeed when real efforts are made to 'dismantle a powerful and entrenched set of neoliberal ideologies and policies' (Kansa, 2014, p. 54).

In this context, collaborative research, particularly with Indigenous and descendent communities is an overarching theme in archaeology of the 21st century. Ownership of the past, including digital archaeological data is emerging as a key concern amongst equity-seeking groups in the United States, Canada, Australia and New Zealand. In this context, Indigenous peoples want to generate knowledge about their ancestors, and they are increasingly engaging with digital tools and technologies to challenge colonial practices that prevented them from access to, and control of archaeology. This is particularly pressing in scenarios where archaeology is practiced within a regulatory framework that privileges government-and/or CRM-

led field collection. Barriers to accessing primary research data persist for many Indigenous peoples and archaeologists, and these social issues are impacting how archaeological research is carried out (Gupta, Nicholas & Blair, n.d). These tensions will continue to influence the way 'openness' is practiced in archaeology.

Case studies

Iterative data cleaning

Data cleaning has gained currency in recent years with the growth of data science and studies (Rahm & Do, 2000; Van den Broeck, Cunningham, Feckels, & Herbst, 2005; Osborne, 2013). While archaeologists are familiar with performing checks on digital data, there are, at present very few journal articles that describe these procedures, suggesting that archaeologists generally overlook reporting having screened for extreme values, duplicate records, misspellings, missing values and other input errors. It should come as no surprise then that archaeologists generally do not document the transformation of data in the computational pipeline, although this situation is changing (Kansa et al., 2014; Marwick et al. 2017; Stupler & Wilkinson, 2017; Marwick 2018; more broadly, see Shawn Graham's open lab notebook). The common thread in each of these works is the aim to lay bare 'point-and-click' procedures, while documenting what worked and what did not.

A key aspect in data cleaning is its iterative nature, i.e. that the analyst must go through a number of transformations and cleaning routines that are often non-linear, and tailored to specific analytic goals and quality specifications. Interactivity and visualization are important as an analyst works through cleaning routines, and data cleaning systems typically offer user interfaces that enable an analyst to write cleaning sequences, preview them on a portion of the data, and then apply these instructions to whole sets of data. The instructions are saved and can be un-done or extracted at any step. The cleaning sequences can also be applied to other data and offer a real-time history of transformations. This kind of documentation can be readily reviewed, shared, modified and repurposed. Below, I offer an example through OpenRefine, an open source, standalone, desktop application that supports iteration with a spreadsheet style interface.

In his study of data quality, privacy and 'geospatial big data', McCoy (2017, p. 79) examines the case of publically available and 'professional' or privately maintained and restricted archaeological site records. Specifically, he evaluates the frequency and density of reported

fortifications across New Zealand in three sources, and complements them with LiDAR images for one particular fortification called Puketona Pa. The author employed spreadsheet software and ArcGIS for his analysis. The professional database of archaeological site records, developed and maintained by the New Zealand Archaeological Association, is available only through prior authorization and is not considered here. The two publically available sources are a radiocarbon database maintained by the Waikato Radiocarbon Lab with 1671 records, and location information on fortifications maintained by Land Information New Zealand. In a GIS, these data are represented as a point, a single location defined by a set of geographic coordinates. Thematic information such as the name of an archaeological site, the site identification number, site type, the radiocarbon date, the material that was sampled, and the source of the sample are added as 'attributes' to the point.

For each source, McCoy describes the methods and analytic techniques he employed, yet there is limited documentation on his data cleaning and processing. This is somewhat surprising given his remarks that 'filtering, classifying and coding temporal information' was the most time-intensive part of the analysis (McCoy, 2017, p. 84). Elsewhere, the author notes that the radiocarbon data were downloaded as a Google keyhole markup zip (kmz) and then 'transferred' to the commercial GIS software, ArcMap 10.3. However, McCoy (2017, p. 83) notes that information on 'site type, and material dated did not migrate smoothly', a problematic situation because these two fields were central in processing temporal values. To correct this, the author manually searched the online database for lab identification numbers and 're-attach[ed]' the missing information to all 1,671 site records.

I offer an alternative processing sequence on OpenRefine for McCoy's radiocarbon data that transforms the information for use in a GIS without manual searching and re-attachment of missing data fields. I note that ArcMap and other GIS software, such as QGIS have available tools that convert between Google's keyhole markup language (kml) and shapefiles (shp). These procedures are 'point-and-click' within GIS software, and by default, the software does not retain a history of transformations in a project. In OpenRefine, the cleaning sequence created can be applied to other data in need of similar processing and most importantly, for the purpose of this study, it serves as documentation of data transformations that are typical in data-intensive methods. For example, a recipe can be created for identifying missing values, extreme or anomalous values, and resolving them. More complex tasks such as linking codes or shorthand

(e.g. LBK for Linearbandkeramik and ‘grv’ for grave) from field journals and code books can be facilitated through a cleaning sequence. The cleaned data can be exported in Comma Separated Value (csv) format which are easily read in GIS software.

The radiocarbon data were directly accessed through the Web link supplied in McCoy (2017, p. 82), (www.waikato.ac.nz/nzcd/C14kml.kmz). OpenRefine enables parsing of data in extensible markup language (xml), which is a standard used in Google’s keyhole markup language (kml) (Google Developers, 2018) and is therefore interoperable (Figure 2.3). The placemark is an object that contains three elements; namely a name, a description and a point that specifies the position of the placemark on the Earth’s surface using a pair of coordinates (longitude and latitude). Additional thematic information is added to the placemark as ‘description’, as well as styling for the icon and text. Thus, each placemark object contains information that is of greatest interest.

Figure 2.3 Parsing options in OpenRefine for a file in keyhole markup language.

Source: Note that the information of interest is within the placemark tags.

Figure 2.4 A spreadsheet style interface on OpenRefine that shows information in columns.

Source: Note that radiocarbon and site information is within tags and will require cleaning.

Once parsed, the data are displayed within a spreadsheet-style interface with rows and columns where they and the values within them can be cleaned (Figure 2.4). Column names are based on tags <tag> </tag> within the placemark object, and include information that is not relevant for further analysis. Visual inspection of the data show empty rows and unnecessary columns that can be removed. More importantly, thematic information (e.g. site name, site type, etc.) are all parsed into one column (Placemark – description), and they have tags that will cause problems in further analysis. But information within the tags is needed and must be separated into individual columns for use in a GIS. When performed manually on over 1600 records, such an undertaking can easily result in input error and unintended modifications and deletions. With OpenRefine, it is possible

to write a cleaning sequence that can be previewed on part of the data, and then applied to all records. In this case, the cleaning sequence is shown in Box 2.1:

1) Remove empty rows

2) Remove empty columns

3) Remove tags and replace with a ';' #this replaces closing tags with a semicolon

4) Replace
, with a space

5) Split columns based on separator ';' (semicolon) #this creates individual columns from description column

6) Text transform based on split at ':' (colon) #this retains values after the colon, applied to all columns

7) Column addition based on text length #this creates a new column called Easting based on derived values from coordinates column, repeated for Northing

Once the sequence is implemented, the data are manipulated accordingly. The figure below shows the resulting data (Figure 2.5), along with the history of the cleaning sequence (Figure 2.6). Note that the cleaning sequence is available as description and as code. The code can be extracted and applied to other data that need similar processing. The 'clean' data can be exported as a cross-platform spreadsheet format (csv) that can be read routinely by GIS software.

Figure 2.5 The cleaned version of the file ready with coordinates for mapping.

Figure 2.6 The cleaning sequence or 'recipe' for converting kml into comma separated value (csv) format. The code can be exported, modified and re-used

This brief case study did not replicate McCoy's manual processing and re-attachment of values and thus, it cannot offer any specific measure of duration of that task nor compare it with processing time in OpenRefine. Automating transformations can reduce the potential for mistaken entry or deletion within a spreadsheet. Moreover, in using a platform that enables

writing of a cleaning sequence, we gain clear documentation of data transformations and facilitate potential re-use of these procedures. This resulting routine can be reviewed, modified and repurposed for processing other geospatial data.

Processing field data

As suggested above, archaeologists routinely use a vast range of survey tools to produce local maps and to document the spatial and stratigraphic relationships between archaeological features and artefacts. A total station or electronic theodolite takes high-precision distance and angle measurements that enable 3-dimensional recording. The measurements are based on a 'local' or arbitrary coordinate system where the origin (0,0) is the point location of the total station. Once set up in the field, archaeologists can collect information relatively quickly and can encode each point with a description (thematic information). Accuracy of measurements is highly dependent on levelling the instrument and recent models contain additional software to perform levelling and adjustment calculations. The instrument typically comes with propriety software that enables recording, management and calculation of distance and angle measurements. Until recently, measurements recorded on a total station could not be automatically tied to geographic space, i.e. to a location on the Earth's surface and as a result, control points were required to enable additional processing that would tie measurements to real geographic space. Real-Time Kinematic systems now bring Global Navigation Satellite System positioning with total station surveying and a suite of sophisticated point-and-click software to calculate position data derived from satellites. These processing software are typically available only for licensed users and are not available for review.

I present the case of a survey on the island of Saint-Pierre, France, where initial archaeological fieldwork was carried out using a total station and a handheld Global Positioning Systems (GPS) unit. I document efforts to transform data collected in a local coordinate system to a global system in a situation where only two control points are available. Transformation of locational information into a global coordinate system can facilitate the integration of field data with other sources, and can enable spatial analysis of archaeological data.

The study area is located on the eastern coast of the island of Saint Pierre, France (Figure 2.7). A field survey with a total station and handheld GPS unit was carried out as part of an

archaeological project at Memorial University of Newfoundland, Canada, to identify historical (18th century – onwards) settlement on this part of the island. The initial survey team consisted of three archaeologists. Field collection in the study area (measuring approximately 100 x 200 m), was organized into two surveys; one focused on archaeological features visible on the surface, and the second focused on recording topography at regular intervals, with the intention to bring these data into a GIS and examine them with historical maps and other documents. For example, the archaeological features can be made into polygons (where appropriate) with thematic information, enabling measurement of size and shape of surface features. This information can be used to assess survey strategy, and offers a historical document prior to site excavation. Therefore, a geographically referenced model of the topography and archaeological features was highly desirable.

Figure 2.7 An overview of the location and estimated size of the study area in Saint-Pierre, France.

The first survey on archaeological features (features) resulted in 178 points, and the second survey on topography (landscape) consisted of 343 points (Figure 2.8). All measurements were made in metres. Both surveys had the same origin and back sight for registration. Coordinates for the origin and back sight were recorded on the handheld GPS unit with an error of +/-5 metres. During initial processing of the data in a local coordinate system, we immediately identified a significant problem with the first survey. The survey points were rotated to some degree and had to be corrected prior to being transformed to a global coordinate system. The team recognized that a mistake in registering the total station set-up was likely the source of this error, yet the second survey did not share this dislocation. Because the team had used an old model total station that came with limited support for processing survey measurements, the project directors decided it was necessary to develop an equation to adjust the archaeological features based on known locations in geographic space i.e. the origin and back sight.

Figure 2.8 A map showing points from two surveys that were collected on a total station. Location of the total station or origin is represented as a star, survey on archaeological features on surface is marked in green, and the survey of topography is in brown. (Also Colour Plate 2.8) Source: *Note that the feature survey data is rotated*.

The situation, however, was not ideal as there were only two control points (origin and back sight) available and most transformations require a minimum of three control points. For example, the CHaMP Topo Processing tool developed by Wheaton, Garrard, Whitehead, and Volk (2012) at Utah State University for use in ArcMap is available for local to global coordinate transformations. However, this tool was not used because of its three control point prerequisite. To transform the survey data, Maria Yulmetova (2018), a student at Memorial University of Newfoundland developed a Python script that calculated the rotation factor to transform local coordinates into Universal Transverse Mercator (northing and easting) using two control points. In practice, the script generated modified UTM coordinates that could be aligned with two different sources: a scanned topographic map from the National Institute of Geographical and Forestry Information (IGN-F), France, (Figure 2.9) and on imagery from Google Earth. The validation was based on visual inspection of the overlap between survey measurements and features visible on the topographic map. With survey data corrected, it was possible to more closely examine archaeological features, and their estimated size and shapes alongside historical documents.

Figure 2.9 The survey points overlaid on a scanned map that is geo-rectified to WGS-UTM 21. A Python script was developed to enable rotation and transformation of points in a local coordinate system to a global coordinate system (UTM) using two known coordinate pairs. (Also Colour Plate 2.9)

The script reflects a step towards creating a tool that archaeologists can use for transformations of survey data that have a limited number of control points. Because most tools for processing survey data are proprietary, they are not available for scholarly review or modification, as was needed in this scenario. When the underlying code is available, it is possible to alter and customize default criteria and this in turn, can enable more appropriate decision-making and more rigorous research practice in archaeology. More fundamentally, scripted workflows offer archaeologists the chance to engage more deeply with the range of tools and technologies they employ throughout the archaeological workflow, and open cross-disciplinary collaborations with geographers, cartographers and computer scientists in disabling black boxes. By sharing their scripted workflows, archaeologists can encourage re-use, modification and refinement of 'recipes' and data processing tools and technologies. Furthermore, greater attention to examining and modifying code can create intellectual space for training and empowering undergraduate and graduate students for archaeology of the 21st century (Marwick 2017).

Data management with version control

In this final section, I discuss data management and version control through platforms such as GitHub. Strupler & Wilkinson (2017) offer an implementation of a 'distributed version-control data management platform' for field survey using Git. Git is a dedicated version control system for tracking changes in digital files, and facilitates coordination of tasks amongst multiple collaborators. Version control systems are common in programming contexts where they are used to trace errors and track any updates in code. In a version control system, the repository is where files and their histories live. The person who creates the repository is its owner and is responsible for integrating merges and resolving any conflicts into that repository. Each file in a repository is authored, and because each change (however minute) is logged, it becomes the responsibility of a repository owner to provide a description of that change. For example, after correcting a series of misspellings in a spreadsheet called 'artifacts', the repository owner is ready to close or 'commit' those changes. The owner will ideally describe these changes

‘corrected misspellings’ as the record of this particular commit. When the history of ‘artifacts.csv’ is examined, the owner can review each commit through time, and select it by name and ‘revert’ to a previous version of ‘artifacts’.

Versioning functionality proves highly effective in projects that have multiple authors, each of whom submits changes to a single file. In that case, authors will clone or ‘fork’ the original repository, where they make changes on their own copy of ‘artifacts.csv’ and then request to merge their versions into the original repository. Onus falls on the owner of the original repository to review the changes before accepting (or rejecting) a merge request. For instance, this form of data management can be most helpful for archaeologists who manage site inventory records that are updated from time to time, but require a history of site discovery (Heilen et al., 2008).

In Strupler and Wilkinson’s survey, the data were ‘born-digital’ and the authors employed the versioning system, Git, to manage their field collection records. They remark that these data ‘must have history’ (2017, p. 283), as it is a necessary part of ‘improving the quality’ of the study and overall results. This is key as the authors argue that error correction and other modifications are better tracked within processing, and can be linked to individual contributors. The functionality further enables comparison between any two or three file versions which can highlight the source of duplicate entries, for example, or conflicts in particular values. These issues must be reconciled at source and therefore potentially minimize the likelihood of undetected errors propagating through a computational pipeline.

The authors organized their field collection into sub-projects in a main repository, such as ‘survey’, ‘survey-design’, ‘survey-data’, ‘gis-static’, ‘gis-tools’, ‘photo-archive’ and ‘team’ (2017, p. 290). The sub project ‘survey-data’, for example, consisted of two sources of data, a direct input from a field walker’s GPS unit, and digital forms in which a field walker reported points of interest observed during a walk. The advantage of this data management is the relative ease with which digital data can be processed off-site, while minimizing accidental loss, corruption or deletion of those data. Harnessing the history of changes in any file or sub-project

with a repository can enable authors to more easily detect errors in data, correct them and track subsequent work. That said, most versioning systems (e.g. Git, GitHub) have a learning curve and would require each team member to have some familiarity with the platform and management practices. More fundamentally, for use in the field and off-site, a server and institutional support (i.e. financial support and expertise) are often prerequisites (Strupler & Wilkinson, 2017, p. 301). These are especially necessary to enable long-term use of collected data, and their re-use. Strupler & Wilkinson (2017) do not discuss the long-term use of their survey data, or how these data might be re-used by a scholar who was not part of the original project, yet their study offers an example of distributed, collaborative and version controlled management of data in archaeological field projects.

Conclusion

Preparation of archaeological data for further analysis, curation and re-use is tied to data quality, which in turn, is integral to research design and an archaeologist's intended use of archaeological data. Geospatial technologies such as GIS are routinely used in archaeology to manage, store and analyse large amounts of digital archaeological data. However, these spatial databases are known to have poor error management, a situation that can result in error propagation that impacts on subsequent analysis and the final result. The widespread use of GIS in archaeology therefore can constrain a broader assessment of archaeological methods and the appropriateness of data in terms of interpretation and re-use. Furthermore, processing of data and analysis within computational pipelines is rarely documented and shared, a situation that limits what is known on procedures and transformations and the overall quality of archaeological data. Without clear documentation of how data were processed, archaeologists impose black boxes within the archaeological workflow that prevent examination of how data were transformed from acquisition to their final presentation and publication. This situation is especially problematic when point-and-click software is uncritically utilized. Archaeologists must become ready to

dismantle black boxes at a moment when greater amounts of 'born digital' archaeological data, are being generated.

Recent interests in the preparation of archaeological data and the quality of those data are influenced by the growing use of digital and geospatial tools and technologies, and the rapid growth of communication tools that facilitate exchange and sharing of data between scholars, institutions and non-specialists. Archaeologists are accumulating large amounts of data through real-time digital documentation in the field, that are paperless and these efforts are thought to minimize redundancy and human-introduced errors in the recording of archaeological sites and archaeological data. These efforts can also potentially shorten the time interval between data acquisition, processing, analysis and presentation and publication.

A growing awareness of a digital data-rich environment is encouraging archaeologists to think in terms of data-intensive methods and big data whilst highlighting that greater efforts are needed to document and report how decisions were made on cleaning, analysing and publishing data. This situation presents challenges and opportunities for archaeologists. Calls for Open Science in archaeology reflect these tensions, and offer a way forward in terms of promoting the generation of scripted workflows, version control for data management and collaborative research. Recognition that the interests and needs of social groups differ in terms of ownership of the past are attuning archaeologists to the role of institutional practices in data quality issues.

Greater and more stringent control over metadata is enabling archaeologists in documenting their data creation methods, sampling techniques and contextual information that facilitates the re-use of digital archaeological data. Metadata typically include authorship information, basic project and site descriptions, keywords, chronological ranges and geographical coverage (e.g. bounding box coordinates). The data publisher Open Context, for example, has shown leadership in preparing digital data for re-use, including data cleaning, such as performing basic checks on received data to correct data entry errors and inconsistencies in classification fields, as well as more involved transformations to translate code books and reconcile them with tabular information (Kansa et al., 2014, p. 60). We do not yet have sufficient information on how

metadata are being used beyond search, browse and filtering for specific records. Nonetheless, recent developments show that archaeologists are aware of data quality issues and are actively taking steps to communicate the level of confidence they have in their analysis and interpretation of archaeological data.

The apparent democratization of archaeological site information has renewed concerns over privacy and the security of sensitive locational information. Publishing archaeological information presents significant challenges and opportunities. Conventional wisdom is that archaeological data collected in the field contain sensitive locational information, and that sharing locations of archaeological and historical sites can facilitate, if not result in the destruction of those sites through looting. Looting and illegal trafficking of archaeological artefacts and human bones is an issue observed in many places (Brodie, Doole & Renfrew 2001; Huffer & Graham, 2017). These concerns are often heightened in national contexts where tensions over ownership of the past exist between archaeologists and local communities or ethnic and linguistic minorities. Yet recent developments in geovisual analytics demonstrate that scholars can meaningfully analyse data even when they contain sensitive location information (Andrienko et al., 2007). Archaeologists are now putting greater efforts into examining how to share sensitive archaeological information, and making explicit scenarios in which such efforts are inappropriate. These efforts are reflected in conference sessions at the 2018 Society for American Archaeology meetings, such as the 'Futures and Challenges in Government Digital Archaeology' symposium organized by Jolene Smith, and a forum, 'Keeping Our Secrets: Sharing and Protecting Sensitive Resource Information in the Era of Open Data', that was chaired by David Gadsby and Anne Vawser. The ethos of 'openness' is encouraging archaeologists to better understand possibilities in and potential implications of publishing archaeological data on the Web.

The re-use of digital archaeological data require scholarly efforts in cleaning and better documentation of these procedures and transformations. This scholarship is being promoted to facilitate deeper engagement with archaeological methods, which in turn, can open new forms of

research in archaeology. Archaeologists are increasingly extracting geographical information from historical documents and repurposing these data for spatial analysis (Murrieta-Flores & Gregory, 2015). Employing sophisticated techniques such as Natural Language Processing, archaeologists draw out place-names in historical texts and incorporate them into GIS software. Tools such as geoparsers that automate annotation of texts, and geo-reference place-names (e.g. create pairs of coordinates) are now being developed for specific corpora. Platforms such as ORBIS (2018), a geospatial network model of the Roman World developed at Stanford University and the Pelagios Commons (2018), an online community that enables linked open data on historical places, are highlighting the range and scope of interdisciplinary scholarship. These efforts often emphasize collaborative code development and code sharing.

Growing numbers of archaeologists are employing programming languages such as R and Python in documenting their research processes. Scripted workflows and code sharing is facilitated by Web-based platforms such as GitHub and Jupyter notebooks. For example, the Open Digital Archaeology Textbook and Environment (Graham et al., 2018), an open-access digital textbook makes extensive use of Jupyter notebooks to share code and data for teaching purposes (https://o-date.github.io/support/notebooks-toc/). Most crucially, the digital environment offers scholars and learners a platform to read and experiment with code writing. Notebooks of particular interest include one on spatial analysis developed by Rachel Optiz (https://mybinder.org/v2/gh/ropitz/spatialarchaeology/master), as well as one on processing public data such as Light Detection and Ranging (LiDAR) that are published by local and national institutions. These notebooks greatly extend the potential and possibilities for scripted workflows in spatial analysis, within and without traditional GIS software.

Greater attention is now given to archaeological site records as historical documents, and the history of site discovery as a way to assess data quality. In this context, archaeologists are making greater effort to employ version control systems that log changes in files and potentially reduce the likelihood of errors going undetected through the archaeological workflow. Because all files in versioning systems are authored, it is possible to organize and manage multi-authored

projects on these platforms. As such, documentation of changes to a digital object offers a history of that object, and a way to track error and its propagation through the archaeological workflow. Implementing good documentation practices into the archaeological workflow can enable better data quality. Yet version control systems present barriers in terms of implementation; expertise and institution support and resources are necessary prerequisites. Nonetheless, version control systems offer great potential for archaeologists who manage site inventory information that changes and where archaeological data are managed by experts who do not have access to the original data collectors and their contextual knowledge. These challenges underscore the need for better data management techniques in archaeology more broadly.

With more stringent control over metadata, there is enormous scope for data-intensive methods in archaeology. Federal funding agencies are placing greater emphasis on data management plans for funded projects and these developments are creating an environment in which archaeological data are being more closely scrutinised for sharing on Web-based platforms. As a result, greater amounts of better documented data are available for re-use in archaeology, which in turn, can facilitate a better understanding of the human past. More fundamentally, these efforts are creating opportunities for new forms of research in archaeology that can promote collaboration with anthropologists, historians, cognitive scientists, geographers and computer scientists, which in turn, can have broader implications in the social sciences and humanities.

Open Refine: kml cleaning file [Gupta-openRefine-cleaning.txt]

Figure 2.1: Gupta, Neha. (2018, November 1). Sources and types of errors in digital archaeological data. Zenodo. http://doi.org/10.5281/zenodo.2556803

Figure 2.2: Gupta, Neha. (2018, November 1). Digital Archaeological Workflow. Zenodo. http://doi.org/10.5281/zenodo.2556805

References

Allison, P. (2008). Dealing with legacy data: An introduction. Internet Archaeology, 24. http://dx.doi.org/10.11141/ia.24.8.

Andrienko, G., Andrienko, N., Jankowski, P., Kraak, M-J., Keim, D., MacEachren, A. M., & Wrobel, S. (2007). Geovisual analytics for spatial decision support: Setting the research agenda. International Journal of Geographical Information Science, 21(8), 839–857.

Atici, L., Kansa, S. W., Lev-Tov, J., & Kansa, E. C. (2013). Other people's data: A demonstration of the imperative of publishing primary data. Journal of Archaeological Method and Theory, 19, 1–19.

Austin, A. (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.

Australian National Data Service (ANDS). (2018). Data versioning. ANDS. Retrieved October 2018, from www.ands.org.au/working-with-data/data-management/data-versioning

Averett, E. W., Counts, D. B., & Gordon, J. (2016). Introduction. In D. B. Counts, E. W. Averett, & J. Gordon (Eds.), Mobilizing the past for a digital future: The potential of digital archaeology. Retrieved from http://dc.uwm.edu/arthist_mobilizingthepast/

Bampton, M., & Mosher, R. (2001). A GIS driven regional database of archaeological resources for research and CRM in Casco Bay, Maine. Bar International Series, 931, 139–142.

Banning, E. B., Hawkins, A. L., Stewart, S. T., Hitchings, P., & Edwards, S. (2017). Quality assurance in archaeological survey. Journal of Archaeological Method and Theory, 24(2), 466–488. https://doi.org/10.1007/s10816-016-9274-2.

Bevan, A., Crema, E., Li, X., & Palmisano, A. (2013). Intensities, interactions, and uncertainties: Some new approaches to archaeological distributions. In A. Bevan & M. W. Lake (Eds.), Computational approaches to archaeological spaces (pp. 27–52). Walnut Creek, CA: Left Coast Press.

Bevan, A., & Lake, M. W. (2013). Computational approaches to archaeological spaces. Walnut Creek, CA: Left Coast Press.

Brodie, N., Doole, J., & Renfrew, C. (Eds.). (2001). Trade in illicit antiquities: The destruction of the world's archaeological heritage. Cambridge: McDonald Institute for Archaeological Research.

Burg, M. B, Peeters, H., & Lovis, W. A. (Eds.). (2016). Uncertainty and sensitivity analysis in archaeological computational modeling. Switzerland: Springer.

Chrisman, N. (2006). Development in the treatment of spatial data quality. In R. Devillers & R. Jeansoulin (Eds.), Fundamentals of spatial data quality (pp. 22–30). Newport Beach, CA: ISTE.

Cooper, A., & Green, C. (2016). Embracing the complexities of “Big data” in archaeology: The case of the English landscape and identities project. Journal of Archaeological Method and Theory, 23(1), 271–304. doi:10.1007/s10816-015-9240-4

Costa, S., Beck, A., Bevan, A. H., & Ogden, J. (2013). Defining and advocating open data in archaeology. In G. Earl, T. Sly, A. Chrysanthi, P. Murrieta-Flores, C. Papadopoulos, I. Romanowska, & D. Wheatley (Eds.), Archaeology in the Digital Era: Papers from the 40th annual conference of computer applications and quantitative methods in archaeology, Southampton, 26–29 March, 2012 (pp. 449–456). Amsterdam: University Press.

Crema, E. (2012). Modelling temporal uncertainty in archaeological analysis. Journal of Archaeological Method and Theory, 19, 440–461.

Devillers, R., & Jeansoulin, R. (2006). Spatial data quality: Concepts. In R. Devillers & R. Jeansoulin (Eds.), Fundamentals of spatial data quality (pp. 31–42). Newport Beach, CA: ISTE.

Dibble, H. L., & McPherron, S. P. (1988). On the computerization of archaeological projects. Journal of Field Archaeology, 15(4), 431–440. https://doi.org/10.2307/530045.

Dunnell, R. C., Teltser, P., & Vercruysse, R. (1986). Efficient error reduction in large data sets. Advances in Computer Archaeology, 3, 22–39.

Evans, T. N. L. (2013). Holes in the archaeological record? A comparison of national event databases for the historic environment in England. The Historic Environment: Policy & Practice, 4(1), 19–34. https://doi.org/10.1179/1756750513Z.00000000023.

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). Journal of the Medical Library Association : JMLA, 105(2), 203–206. doi:10.5195/jmla.2017.88.

Gobalet, K. W. (2001). A critique of faunal analysis: Inconsistency among experts in blind tests. Journal of Archaeological Science, 28(4), 377–386.

Google Developers. (2018). What is KML? Retrieved March 2018, from https://developers.google.com/kml/

Graham, S. Open lab notebook. Retrieved March 20, 2018, from https://electricarchaeology.ca/

Graham, S., Gupta, N., Smith, J., Angourakis, A., Carter, M., & Compton, B. (2018). The open digital archaeology textbook environment. Retrieved from https://o-date.github.io/draft/book/

Green, C. (2011). It’s about time: Temporality and intra-site GIS. In E. Jerem, F. Redő, & V. Szeverényi (Eds.), On the ooad to reconstructing the past: Computer applications and quantitative methods in archaeology (CAA): Proceedings of the 36th international conference, Budapest, April 2–6, 2008 (pp. 206–211). Budapest: Archaeolingua.

Gupta, N., & Devillers, R. (2017). Geographic visualization in archaeology. Journal of Archaeological Method and Theory, 24(3), 852–885.

Gupta, N., Nicholas, R., & Blair, S. (n.d). Post-colonial and indigenous perspectives in digital archaeology. In E. Watrall & L. Goldstein (Eds.), Digital heritage and archaeology in practice. University Press of Florida. Retrieved from http://dhainpractice.anthropology.msu.edu/

Heilen, M. P., Nagle, C. L., & Altschul, J. H. (2008). An assessment of archaeological data quality: A report submitted in partial fulfillment of legacy resource management program project to develop analytical tools for characterizing, visualizing, and evaluating archaeological data quality systematically for communities of practice within the department of defense. Department of Defense Legacy Resource Management Program, Technical Report 08–65, Statistical Research Inc., Tuscon, AZ.

Hodder, I., & Orton, C. (1976). Spatial analysis in archaeology. New York: Cambridge University Press.

Huffer, D., & Graham, S. (2017). The insta-dead: The rhetoric of the human remains trade on Instagram. Internet Archaeology, 45. https://doi.org/10.11141/ia.45.5

Huggett, J. (2015). Digital haystacks: Open data and the transformation of archaeological knowledge, In A. T. Wilson & B. Edwards (Eds.), Open source archaeology: Ethics and practice (pp. 6–29). Walter de Gruyter GmbH & Co KG.

Hunter, G. J., & Beard, K. (1992). Understanding error in spatial databases. Australian Surveyor, 37(2), 108–119. https://doi.org/10.1080/00050326.1992.10438784

Kansa, E. C. (2011). Introduction: New directions for the digital past. In E. C. Kansa, S. W. Kansa, & E. Watrall (Eds.), Archaeology 2.0: New approaches to communication and collaboration (pp. 1–25). Los Angeles, CA: Cotsen Institute of Archaeology Press.

Kansa, E. C. (2014). The need to humanize open science. In S. Moore (Ed.), Issues in open research data (pp. 31–58). Ubiquity Press. doi:10.5334/ban.c.

Kansa, E. C., Kansa, S. W., & Arbuckle, B. (2014). Publishing and pushing: Mixing models for communicating research data in archaeology. International Journal of Digital Curation, 9(1), 57–70. https://doi.org/10.2218/ijdc.v9i1.301

Kintigh, K. (2006). The promise and challenge of archaeological data integration. American Antiquity, 71(3), 567–578.

Kintigh, K., Altschul, J. H., Kinzig, A. P., Limp, W. F., Michener, W. K., Sabloff, J. A., . . . Lynch, C. A. (2015). Cultural dynamics, deep time, and data: Planning cyberinfrastructure investments for archaeology. Advances in Archaeological Practice, 3(1), 1–15.

Kohl, P. L., & Fawcett, C. (Eds.). (1995). Nationalism, politics and the practice of archaeology. Cambridge: Cambridge University Press.

Kohl, P. L., Kozelsky, M., & Ben-Yehuda, N. (Eds.). (2007). Selective remembrances: Archaeology in the construction, commemoration and consecration of national pasts. Chicago: The University of Chicago Press.

Kolar, J., Macek, M., Tkáč, P., & Szabó, P. (2015). Spatio-temporal modelling as a way to reconstruct patterns of past human activities. Archaeometry, 58(3), 513–528. doi:10.1111/arcm.12182

Leeper, T. J. (2015). Collecting thoughts about data versioning: Contribute to Leeper/data-versioning development by creating an account on GitHub. Retrieved October 2018, from https://github.com/leeper/data-versioning

Leroux, A., Lemonsu, A., Bélair, S., & Mailhot, J. (2009). Automated urban land use and land cover classification for mesoscale atmospheric modeling over canadian cities. Geomatica, 63(1), 13–24.

Levy, T. E., & Smith, N. G. (2007). On-site GIS digital archaeology: GIS-based excavation recording in southern Jordan. In T. E. Levy (Ed.), Crossing Jordan: North American contributions to the archaeology of Jordan (pp. 47–58). Oakville, CT: Equinox Publishing.

Llobera, M. (2007). Reconstructing visual landscapes. World Archaeology, 39(1), 51–69. http://doi.org/10.1080/00438240601136496

Marwick, B. (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.

Marwick, B. (2018). Using R and related tools for reproducible research in archaeology. In J. Kitzes, D. Turek, & F. Deniz (Eds.), The practice of reproducible research: Case studies and lessons from the data-intensive sciences. Oakland, CA: University of California Press. Retrieved from www.practicereproducibleresearch.org/case-studies/benmarwick.html

Marwick, B., d'Alpoim Guedes, J., Barton, C. M., Bates, L. A., Baxter, M., Bevan, A., . . . Wren, C. D. (2017). Open science in archaeology. The SAA Archaeological Record, 17(4), 8–14.

McCoy, M. D. (2017). Geospatial big data and archaeology: Prospects and problems too great to ignore. Journal of Archaeological Science, 84, 74–94.

Meskell, L. (2005). Archaeology under fire: Nationalism, politics and heritage in the eastern Mediterranean and Middle East. London: Routledge.

Molloy, J. C. (2011). The open knowledge foundation: Open data means better science. PLoS Biology, 9(12), e1001195. doi:10.1371/journal.pbio.1001195

Murrieta-Flores, P., & Gregory, I. (2015). Further frontiers in GIS: Extending spatial analysis to textual sources in archaeology. Open Archaeology, 1(1), 166–175.

National Science Foundation. (2017). Dissemination and sharing of research results. Retrieved February 2018, from www.nsf.gov/bfa/dias/policy/dmp.jsp

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., & Buck, S. (2015). Promoting an open research culture. Science, 348(6242), 1422–1425. doi:10.1126/science.aab2374

ORBIS: The stanford geospatial network model of the roman world. (2015). Stanford University Libraries. Retrieved October 2018, from http://orbis.stanford.edu/

Osborne, J. W. (2013). Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data. Retrieved from http://srmo.sagepub.com/view/best-practices-in-data-cleaning/SAGE.xml

Pelagios Commons. (2018). Linking the places of our past. Retrieved October 2018, from http://commons.pelagios.org/

Plewe, B. (2002). The nature of uncertainty in historical geographic information. Transactions in GIS, 6(4), 431–456.

Porter, S. T., Roussel, M., & Soressi, M. (2016). A simple photogrammetry rig for the reliable creation of 3D artifact models in the field lithic examples from the Early Upper Paleolithic sequence of Les Cottés (France). Advances in Archaeological Practice, 4(1), 71–86.

Rabinowitz, A. (2014). It's about time: Historical periodization and linked ancient world data. ISAW Papers, 7(22). Retrieved March 2018, from http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/rabinowitz/

Rahm, E., & Hai Do, H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3–13.

Roebroeks, W., Gaudzinski-Windheuser, S., Baales, M., & Kahlke, R.-D. (2017). Uneven data quality and the earliest occupation of Europe: The case of untermassfeld (Germany). Journal of Paleolithic Archaeology, 1(1), 5–31. https://doi.org/10.1007/s41982-017-0003-5

Roosevelt, C. H., Cobb, P., Moss, E., Olson, B. R., & Ünlüsoy, S. (2015). Excavation is destruction digitization: Advances in archaeological practice. Journal of Field Archaeology, 40(3), 325–346. https://doi.org/10.1179/2042458215Y.0000000004

Silberman, N. A. (1989). Between past and present: Archaeology, ideology, and nationalism in the modern Middle East. New York: Holt. Retrieved from http://hdl.handle.net/2027/heb.02303.0001.001

Sitara, M., & Vouligea, E. (2014). Open access to archeological data and the Greek law. In A. Sideridis, Z. Kardasiadou, C. Yialouris, & V. Zorkadis (Eds.), E-democracy, security, privacy and trust in a digital world. e-Democracy 2013. Communications in Computer and Information Science, 441. Cham: Springer.

Snow, D. R., Gahegan, M., Giles, C. L., Hirth, K. G., Milner, G. R., Mitra, P., & Wang, J. Z. (2006). Cybertools and archaeology. Science, 311(5763), 958–959.

Strupler, N., & Wilkinson, T. C. (2017). Reproducibility in the field: Transparency, version control and collaboration on the project panormos survey. Open Archaeology, 3(1). https://doi.org/10.1515/opar-2017-0019

Thompson, P. A., Matloff, N., Fu, A., & Shin, A. (2017, August). Having your cake and eating it too: Scripted workflows for image manipulation. ArXiv:1709.07406 [Eess]. Retrieved from http://arxiv.org/abs/1709.07406

Tri-Agency Statement of Principles on Digital Data Management. (2016). Retrieved February 2018, from www.science.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html?OpenDocument

Trigger, B. (2006). A history of archaeological thought (2nd ed.). New York: Cambridge University Press.

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities. PLoS Medicine, 2(10), e267. https://doi.org/10.1371/journal.pmed.0020267

Van der Linden, M., & Webley, L. (2012). Introduction: Development-led archaeology in northwest Europe: Frameworks, practices and outcomes. In L. Webley, M. Van der Linden, C. Haselgrove, & R. Bradley (Eds.), Development-led archaeology in Northwest Europe proceedings of a round table at the University of Leicester 19th–21st November 2009 (pp. 1–8). Oxford: Oxbow.

Vincent, M. L., Kuester, F., & Levy, T. E. (2014). OpenDig: Contextualizing the past from the field to the web. Mediterranean Archaeology and Archaeometry, 14(4), 109–116.

Wells, J. (2011). Four states of mississippian data: Best practices at work integrating information from four SHPO databases in a GIS-structured archaeological Atlas. Society for American archaeology e-symposium. Retrieved from http://visiblepast.net/see/americas/four-states-of-mississippian-data-best-practices-at-work-integrating-information-from-four-shpo-databases-in-a-gis-structured-archaeological-atlas/

Wheatley, D., & Gillings, M. (2002). Spatial technology and archaeology: The archaeological applications of GIS. London: Taylor & Francis.

Wheaton, J. M., Garrard, C., Whitehead, K., & Volk, C. J. (2012). A simple, interactive GIS tool for transforming assumed total station surveys to real world coordinates: The CHaMP transformation tool. Computers & Geosciences, 42, 28–36.

Willems, W. J. H., & Brandt, R. (2004). Dutch archaeology quality standard. Den Haag: Rijksinspectie voor de Archeologie.

Wilshusen, R. H., Heilen, M., Catts, W., de Dufour, K., & Jones, B. (2016). Archaeological survey data quality, durability, and use in the United States. Advances in Archaeological Practice, 4(2), 106–117. https://doi.org/10.7183/2326-3768.4.2.106

Wylie, A. (2002). Thinking from things: Essays in the philosophy of archaeology. Berkeley, CA: University of California Press.

Yulmetova, M. (2018). Python script: Transformation of local coordinates to global coordinates. Retrieved March 2018, from https://github.com/MariaYulmetova88/Transferring-local-coordinates-to-UTM-using-the-GPS-coordinates

Zoghlami, A., de Runz, C., Akdag, H., & Pargny, D. (2012). Through a fuzzy spatiotemporal information system for handling excavation data. In J. Gensel, D. Josselin, & D. Vandenbroucke (Eds.), Bridging the geographic information sciences: International AGILE'2012 Conference, Avignon (France), April 24–27, 2012 (pp. 179–196). New York: Springer.