Rapid Communication

Samantha S. Reiter*, Robert Staniuk, Jan Kolář, Jelena Bulatović, Helene Agerskov Rose, Natalia E. Ryabogina, Claudia Speciale, Nicoline Schjerven, Bettina Schulz Paulsson, Victor Yan Kin Lee, Elisabetta Canteri, Alice Revill, Fredrik Dahlberg, Serena Sabatini, Karin M. Frei, Fernando Racimo, Maria Ivanova-Bieg, Wolfgang Traylor, Emily J. Kate, Eve Derenne, Lea Frank, Jessie Woodbridge, Ralph Fyfe, Stephen Shennan, Kristian Kristiansen, Mark G. Thomas, Adrian Timpson

The BIAD Standards: Recommendations for Archaeological Data Publication and Insights From the Big Interdisciplinary Archaeological Database

https://doi.org/10.1515/opar-2024-0015
received April 29, 2024; accepted August 26, 2024

Abstract: This article presents a series of recommendations for the publication of archaeological data, to improve their usability. These 12 recommendations were formulated by archaeological data experts who mined thousands of publications for different data types (including funerary practices, accelerator mass spectrometry dating, stable isotopes, zooarchaeology, archaeobotany and pathologies) during the initial construction of the Big Interdisciplinary Archaeological Database (BIAD). We also include data harmonisation

*** Corresponding author: Samantha S. Reiter,** Environmental Archaeology and Materials Science, National Museum of Denmark, Kongens Lyngby, Denmark, e-mail: samantha.scott.reiter@natmus.dk
Robert Staniuk, Jan Kolář, Alice Revill, Stephen Shennan: Institute of Archaeology, University College London, London, United Kingdom
Jelena Bulatović, Helene Agerskov Rose, Natalia E. Ryabogina, Nicoline Schjerven, Bettina Schulz Paulsson, Fredrik Dahlberg, Serena Sabatini: Department of Historical Studies, University of Gothenburg, Gothenburg, Sweden
Claudia Speciale: Department of Historical Studies, University of Gothenburg, Gothenburg, Sweden; Archaeobotany Unit, Catalan Institute of Human Paleoecology and Social Evolution (IPHES), Tarragona, Spain
Victor Yan Kin Lee, Elisabetta Canteri, Fernando Racimo: Globe Institute, University of Copenhagen, Copenhagen, Denmark
Karin M. Frei: Environmental Archaeology and Materials Science, National Museum of Denmark, Kongens Lyngby, Denmark
Maria Ivanova-Bieg, Eve Derenne: Department of Pre- and Protohistoric Archaeology, Institute of Ancient Studies, Johannes Gutenberg University, Mainz, Germany
Wolfgang Traylor: Quantitative Biogeography, Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany
Emily J. Kate: Faculty of Life Sciences, University of Vienna, Vienna, Austria
Lea Frank: Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway
Jessie Woodbridge, Ralph Fyfe: School of Geography, Earth and Environmental Sciences, University of Plymouth, Plymouth, United Kingdom
Kristian Kristiansen: Department of Historical Studies, University of Gothenburg, Gothenburg, Sweden; Globe Institute, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark
Mark G. Thomas, Adrian Timpson: Research Department of Genetics, Evolution and Environment, University College London, London United Kingdom
ORCID: Samantha S. Reiter 0000-0001-8872-0640; Robert Staniuk 0000-0002-9941-1875; Jan Kolář 0000-0001-8013-6992; Jelena Bulatović 0000-0002-0672-067X; Helene Agerskov Rose 0000-0003-1061-3129; Natalia E. Ryabogina 0000-0003-1098-0121; Claudia Speciale 0000-0003-1527-9000; Nicoline Schjerven 0000-0002-2820-4422; Bettina Schulz Paulsson 0000-0003-4800-031X; Victor Yan Kin Lee 0009-0001-5598-9632; Elisabetta Canteri 0000-0001-9867-8247; Fredrik Dahlberg 0000-0002-0328-7604; Serena Sabatini 0000-0001-6404-1280; Karin M. Frei 0000-0001-5198-073X; Fernando Racimo 0000-0002-5025-2607; Wolfgang Traylor 0000-0002-4813-1072; Jessie Woodbridge 0000-0003-0756-3538; Ralph Fyfe 0000-0002-5676-008X; Stephen Shennan 0000-0001-6605-064X; Kristian Kristiansen 0000-0003-2423-308X; Mark G. Thomas 0000-0002-2452-981X; Adrian Timpson 0000-0003-0292-8729

vocabularies utilised for the integration of data from different recording systems. The case studies we cite to illustrate the recommendations are grounded in examples from the published literature and are presented in a problem/solution format. Though practically oriented towards the facilitation of efficient databasing, these recommendations – which we refer to as the BIAD Standards – are broadly applicable by those who want to extract scientific data from archaeological information, those who work with a specific region or theoretical focus and journal editors and manuscript authors. We anticipate that the use of the BIAD Standards will increase the usability, visibility, interoperability and longevity of published data and also increase the citations of those publications from which data were mined. The Standards will also help frame a unified foundation to support the continued integration of the natural sciences with archaeological research in the future.

Keywords: publishing, FAIR, data longevity, "big data", archaeology

1 Introduction

Archaeology is a broad discipline encompassing both a wide range of data types (from rock art to isotopes) and a wide range of inferential approaches (from purely interpretive to model-based statistical inference). All inferential approaches can benefit from large amounts of systematically collated data, ideally stored in relational databases. This together with the influence of related scientific disciplines (such as isotope studies, population genetics and geographical information systems or geographical information system [GIS]) have provided increasingly quantitative and computational components to archaeology.

The establishment of large and centrally organised databases in other fields (such as DNA and protein sequence databases in the early 1980s) together with journal requirements to deposit any new data into such databases for the purpose of accessibility and reproducibility has been a major factor in propelling those fields forward. We contend that similar data configurations would benefit archaeology in a comparable way and would facilitate greater interdisciplinarity. Furthermore, making published archaeological data available in accessible databases increases the visibility, usability and longevity of data and also increases citations of source publications. Lohr's *New York Times* article (2012) welcomed in the "Age of Big Data," though various archaeological scholars have noted the increasing value placed on and engagement with large datasets over the last 20 years (Cooper & Green, 2016; Kintigh et al., 2014; Kristiansen et al., 2014; Levy, 2004; Onsrud & Campbell, 2007).

Much of this general move towards "big data" analysis in archaeology is based upon the amalgamation of comparatively smaller topically, regionally, or chronologically focused data into larger databases covering broader geographic regions and longer time spans. This facilitates a different scale of analysis, enabling researchers to approach problems and questions that are beyond the scope of site-specific, fine-resolution research. The complexities of archaeological data in terms of the breadth and interrelationship of data types pose major databasing challenges compared to the relatively trivial task of storing strings of nucleotides or amino acids in DNA or protein databases, respectively. Essentially, for scientists to be able to analyse archaeological data as "big data," those data have to be assembled in such a way as to make them a single body of information (O'Malley & Soyer, 2012). As it says on the SEADDA homepage (Saving European Archaeology from the Digital Dark Age), "[t]he emerging research challenge for the next decade is optimising archaeological data for re-use[…]." We believe that the first step in meeting such challenges is through the publication of archaeological data in a standardised way. As it was through the construction of a relational database (Big Interdisciplinary Archaeological Database [BIAD]) that we as authors first became aware of the need for a standardised format for data publication in archaeology, we describe relational databases and their influence regarding our current suggestions for publication standards in greater detail in the following.

2 About Relational Databases

It is important to distinguish between an archaeological data repository and an archaeological relational database, as the two are often confounded (see Appendix 1 for a list of archaeological databases and

repositories). A data repository provides a single location for storing a large number of independent datasets. Typically, there is a low barrier to entry, requiring the data provider to give some basic metadata (information about the dataset, such as its publication source, size or entry date) to assist users in searching for relevant entries. These datasets then remain independent; each is structured as originally provided in accordance with the research interests that generated them. By analogy, a public library provides a data repository of books, where metadata such as title and author are required to search for its physical location, but the content of each book remains unrelated to the content of other books.

In contrast, an archaeological relational database stores all data within highly organised hierarchical tables connected in a consistent format, such that associated data from a single entity (like isotopic, genetic and cultural data from a burial) are necessarily connected. Constructing such a relational database requires considerable amounts of effort and expertise to mine, clean, harmonise and aggregate source data (Geser et al., 2022, p. 3; Ribeiro, 2019, p. 120), as well as a deep specialist understanding of the material and the analytical processes/techniques used to produce the data. Effort grows exponentially whenever legacy data are included; indeed it is often easier to mine data from primary literature than to transform old datasets (Gattiglia, 2015, p. 118). The expert labour required to properly datamine published literature is increasingly being mitigated by computer algorithms, such as in the digitisation and translation of legacy data still only available on paper and, more recently, in assisting image extraction and data organisation using machine learning (Klein et al., 2023). While repositories are increasingly common in archaeological data curation, archaeology has a "pressing need" for relational database infrastructures (Kintigh, 2005, p. 16).

In theory, amalgamating data into a single consistent relational database should simply require an *a priori* agreed data structure for any new archaeological primary data. In practice, the greatest availability of archaeological data comes from the existing published literature that has already been collected under a myriad of research interests and practices and, therefore, now requires large efforts to harmonise. To put this issue into perspective, although archaeologists have been collecting and collating data since at least the sixteenth–seventeenth centuries (Schnapp, 1996; Steibing, 1993, pp. 29–30; Trigger, 2008, p. 84), it was only in the nineteenth century that archaeologists began to standardise practices and recording methods (Fleming, 2020; Trigger, 2008, p. 124). This was often linked to the formation of national surveys of monuments and sites requiring protection (Cleere, 1989; Kristiansen, 1984). To date, there remains enormous variation regarding not only what information archaeologists consider worth examining but also how data have been recorded (Evans, 2013, p. 31). For example, a 2017 survey by the European Archaeological Council included a segment on how fieldwork documentation is archived and published. The survey showed that only 32% of the European member states reported *any* kind of formal documentary archiving system whatsoever (Innesti et al., 2018; Novák et al., 2023; Novák et al., 2024; Oniszczuk, 2021). Even when raw data are included in a publication, as is increasingly required by journals and enforced by editors, datasets can rapidly become unusable if crucial metadata (e.g. methodologies, standard and sample information) are not included, as has been suggested in the ecological sciences (Mitchener et al., 1997; Vines et al., 2014). When this is compounded by a transient workforce (where samples may be processed by researchers on fixed-term contracts), the project PI may well not even be able to re-use their own data within only a couple of years of publication if those researchers have transitioned away from the discipline/left the institution.

3 Continuing Towards Archaeological "Big Data"

Although integrating archaeological data and relating it is a challenge, archaeologists are in principle already well-suited to building connections, albeit often in an implicit manner (Howey et al., 2020; Newhard, 2013). Since the birth of the discipline, archaeologists have been searching the varied information they hold in their heads, collections, libraries, files, archives, etc., for correlations within the archaeological record. This is what Boyd and Crawford define as the key to "big data": "[...] [it] is less about data that is big than it is about a capacity to search, aggregate and cross-reference large data sets" (Boyd & Crawford, 2012, p. 663). In this

context, archaeologists address the problems of what might be called "broad data," whereas "big" data might be better attributed to large amounts of narrower data, such as DNA sequences.

Where "big data" approaches have been embraced, there has been a tendency to focus on larger amounts of data from a narrow research focus, rather than cross-disciplinary amalgamations of data (for exceptions, see e.g. Gerbault et al., 2011; Larson et al., 2014). This is already evident in the way that many data are presented and archived. For example, there persists a sharp division between (1) publications of or about databases and (2) publications that utilise databases. The former tend to focus on a single database or repository and are oriented towards specialist researchers (such as geneticists, isotope scientists or researchers who work with radiocarbon dates) or information managers (Mallick et al., 2023; Plomp et al., 2022; Saktura et al., 2023; Williams et al., 2018). The latter are generally made within traditional archaeological fora and tend to focus on the stories or trends that can be discovered from the data (often of a single type, or two data types at most) (Cooper & Green, 2016, p. 2).

4 Future-Proofing the Publication of Archaeological Data

Overarching guiding principles for those wishing to enhance the reusability and inter-relatability of their data holdings have existed for some time. These suggest that scientific data should be Findable, Accessible, Interoperable and Reusable (FAIR) (Wilkinson et al., 2016). Although archaeology and its related disciplines have already made great strides in complying with these principles in the publication of their data, in spite of over two decades' work on archaeological data integration initiatives, the main impact of many of these resources remains specific to the subdiscipline(s) involved and/or specific research project agendas rather than applicable to broader research (Atici et al., 2013; Cooper & Green, 2016; Spielmann & Kintigh, 2011).[1]

Part of the concern here involves being aware that data have inherent (sometimes invisible or difficult-to-discern) biases already integrated from the moment they were produced. As we are reminded by Gitelman, "raw data is [sic] an oxymoron" (2013). Data are not an entirely neutral resource; their manner of collection, collation, storage, harmonisation and original interpretation (classification or categorisation) introduce the possibility for inherent bias (Johnson, 2011). Integrating different kinds of data as well as integrating data recorded in various manners (each with separate biases) are two distinct challenges (Atici et al., 2013, p. 4). However, being aware of these challenges – a concept which has been referred to as the development of the "data gaze" (Beer, 2019, p. 6; Huggett, 2020) or "data journeys" (Leonelli, 2014) – is already a step towards finding a solution.

Archaeological data have been and always will form an incomplete, biased and messy record of the past, not only due to the diverse cultural, taphonomic, regional and practical influences of every archaeological tradition but also due to differences in publication and archiving practices (Huggett, 2020, p. 9; Huggett, 2023). The publication of archaeological data in accordance with a set of Standards will be a big step in helping to minimise bias as well as an aide in the eventual integration of those new data into a single body of information (e.g. through a relational database) which also has the advantage of facilitating the exploration, testing and exposure of inherent biases. In turn, this will naturally assist in minimising and adjusting for biases in downstream analysis (Sabatini & Kristiansen, in prep.). However, merely being aware of the inherent biases in data does not help us to ameliorate them.

Rather than examining those biases more closely (a subject to which we will return in future publications), our purpose with this article is to take a practical approach with the intent of concretely changing the future outlook of archaeological publication, especially – but not exclusively – as related to databasing. While the work of mining, cleaning and harmonising previously published data is ongoing (and will presumably continue for the foreseeable future), we can nevertheless take steps to unify and standardise the framework for

1 Further cross comparisons of such initiatives and relevant ontologies (such as the International Committee of Documentation's Conceptual Reference Model or CIDOC-CRM) with BIAD will be shortly forthcoming (Timpson et al., in prep.).

data reporting in new publications. Doing this will increase the visibility, usability and longevity of published data and also increase citations regardless of whether or not the authors of those publications themselves have an interest in quantitative data, models and/or methods.

Here, we propose a series of 12 publication recommendations and practical guidelines that help to ensure that archaeological data are intelligible and interoperable between different recording systems and archaeological traditions and also that they can eventually be efficiently incorporated into a relational structure. Publishing in accordance with the BIAD Standards has low costs in terms of labour (in its simplest form, it advocates for the inclusion of basic information, clearly presented and may at most involve additional information or supplementary files in publications) and high benefits. Over and beyond service to the discipline in terms of the interoperability of data produced, publishing in accordance with the BIAD Standards will result in an increase in citations and a longevity of the publications in which the data are first presented. In sum, by following the BIAD Standards, we can move beyond studying the archaeological material we curate towards a future-oriented curation of the knowledge we gather and will continue to gather in years to come.

5 Background

These Standards have been developed as a consequence of the data mining component of the 2021 European Research Council (ERC) Synergy Grant COREX Project ("From Correlations to Explanations: Towards a New European Prehistory") which began to build BIAD. COREX's overall goal is to explain the key processes that formed the genetic and cultural diversity of Europe north of the Mediterranean from France to the Urals between 6000 and 500 BCE. To facilitate such analyses, it became crucial to integrate large amounts of disparate data. To that end, COREX assembled a team of archaeologists, topical specialists (e.g. in accelerator mass spectrometry [AMS] dating, stable isotopes, zooarchaeology, archaeobotany and aDNA) and database experts to begin constructing the relational database that became BIAD. An initial dozen data experts were soon joined by others from different research projects (SUSTAIN and SEASCAPES). To date, the work is still ongoing. Our common goal is to facilitate a new scale for archaeological analysis.

Over the course of the past few years, BIAD data experts have had ample opportunities to compare different means of presenting data based on variations in e.g. language, region, nationality or educational/training system for published data relating particularly to prehistoric Europe (though we have worked with data from other chronological periods and geographic areas as well). Nevertheless, integration challenges such as those we encounter daily represent the bread and butter of data-driven international and interdisciplinary research, no matter the period or region. Finding a way through those challenges requires continual work on collecting heterogeneous datasets which undergo expert harmonisation. BIAD is currently driven by manual data collection, which is harmonised according to the standards outlined below.

There exists an important body of literature discussing the biases – and the inherent dangers – within the data produced by archaeological practice, which is essentially co-creative (Cooper & Green, 2016; Djindjian & Moscati, 2021; Gitelman, 2013; Jackson et al., 2007; Johnson, 2011; Kintigh et al., 2014; Leonelli, 2014, 2015; O'Malley & Soyer, 2012; Onsrud & Campbell, 2007), as well as how these may help to perpetuate specific narratives (for discussion within an American context, see e.g. Atalay, 2006; Thomas, 2001, which transferred to European discussions as well e.g. Biddick, 1993). We are currently in the process of preparing a manuscript (Sabatini & Kristiansen, in prep.) that examines some of these ethical concerns specifically in relation to databasing and the construction of archaeological infrastructures (i.a., legacy data, the nature of data and the implications of data categorisation processes). We hold that, as Kansa (2022) and Huvila (2018) have previously argued, while knowledge and information infrastructures such as that which we seek to cultivate through the use of the BIAD Standards may limit some approaches to archaeological inquiry, they nevertheless also facilitate new methods and support new opportunities. In short, publishing in a manner that is "big data-friendly" does not prevent us from continuing to do "small data" archaeology; it merely leaves the door open for other studies, more citations and potential collaborations (see Speciale et al., in prep.). For this reason, we firmly believe that the publication of archaeological information for the promotion of accessibility, interoperability and longevity of data nevertheless remains a worthy goal.

Although there have been previous recommendations for the presentation of data specifically for the promotion of interoperability (e.g. for inclusion in databases; see Cooper & Green, 2016; Roskams & Whyman, 2007), it has been our experience that these tend to be more theoretical than practical in scope. Over the course of our data collection for BIAD, we have all come across a series of issues that render such data coalition problematic. In the following, we present common problems encountered during data collection as well as proposed solutions that may help avoid or prevent them. These solutions count as our concrete recommendations for publishing and presenting new archaeological and related data to promote the inter-operability and longevity of the data and also facilitate efficient databasing for international, interdisciplinary research. Finally, we also present new vocabularies that we developed to facilitate the formalised categorisation (data harmonisation) across regional/national/linguistic/educational boundaries.

6 The BIAD Standards

We present the following issues and our proposed solutions addressing data publication and database entry, which we have grouped together into six overarching themes: (1) securely positioning data in space and time, (2) referencing data sources, (3) consistently harmonising data outcomes, (4) reporting results comprehensively, (5) describing data fully and (6) common vocabularies. These are further supplemented by a checklist which should ensure that any publication is in compliance with the BIAD Standards (see Appendix 2).

It is important to note that these recommendations are not meant as a criticism of current practices or directed towards any particular group of researchers, or journal and book editors. We have taken our examples almost exclusively from prehistoric European material, as this is the area of common expertise between all contributing authors. However, the Standards themselves apply equally well to the publication of archaeological data from any and all chronological periods and geographical areas.

7 Securely Positioning Data in Space and Time

7.1 Problem I: From Which Site/Part of Site/Excavation Does Material Derive?

As mentioned above, archaeology has a long disciplinary history. Many generations of archaeologists have held a sustained interest in individual sites. Multiple excavations in the same (or nearby) areas can lead to confusion regarding the origins of material from which new data have been obtained.

In addition, many archaeological sites have multiple names (including nicknames) and variations in spelling. Some countries may have multiple sites which share the same name. For example, a search of the Danish national sites and monuments record (*Fund og Fortidsminder*) for the site name “Ølby” produces 22 separate hits. If we provide the System Number (95447), the Heritage Protection Number (34275) and/or the Heritage Site Number (a system in which county, district, parish and sites are assigned specific registration numbers; in this case 020105-3), it is clear that we are referring to a place which has been registered under the site name “Nordhøi” (“høi” being an older Danish spelling variant of the modern “høj,” meaning “mound”) but which is also commonly referred to as both “Ølby” and “Ølbys Nordhøj.”

In many cases, archaeological literature does not contain any spatial coordinates or even, sometimes, a map clearly indicating where an archaeological site is located. When combined with ambiguity within the site name, this leads to situations when such scientific data cannot be integrated with other data, therefore making them unusable.

7.1.1 Solution I: Be as Specific About Site Name and Location as Possible

The ideal solution to eliminating confusion regarding the origin of the materials under discussion is to include a (detailed) map *alongside* written coordinates in an internationally recognised geographical coordinate

system (e.g. WGS84). Ideally, such locations should be precise to the fourth decimal place (e.g. accurate to tens instead of hundreds of metres). A dot on a map covering a whole region or even a sub-region is often not sufficient information for differentiating between sites, as the increasing intensity of research in certain areas means that repeated visits to a single area/site/tomb by different research teams and projects are likely to continue or even increase (e.g. Troy has had 24 separate campaigns over the last c. 150 years; UNESCO, 1998).

It is also important to note that a site name in a different alphabet (e.g. Cyrillic) is not cause for unease. Attempts at transliteration often cause more problems than they solve, as they create references to a site that exist in one language or country/ies, but not in others. Furthermore, different languages have different transliteration regulations. For example, Serbian Cyrillic is not problematic, as Serbia also uses the Latin alphabet (and Serbian diacritics generally follow the one-sound-per-letter rule). However, Bulgarian, Russian and Ukrainian transliteration is less standardised. The diversity of archaeologists themselves and their linguistic histories and habits multiplies a problem that would be solved by sticking to an original site name and spelling.

More generally, the translation of site names into English can generate confusion even if the alphabets used for translation are fairly similar. For example, when reported in English, Hungarian sites accompanied by toponyms are often translated, which creates an anglicised form that is difficult to trace in the original literature, e.g. “Kaposvár-Road 61, Site no. 1.” In cases where the differences are caused by a larger number of diacritic signs (like in Polish), the English translation of the site name is generally the original site without the diacritic signs, e.g. “Złota Grodzisko I; Zlota Grodzisko I.” We recommend reporting both site names, as they speed up site identification and – more importantly – prevent data duplication.

On a more regional level, challenges in site identification can be caused by incorporating results from excavation campaigns preceding and following the World Wars, when state border changes were accompanied by administrative and linguistic shifts. In such cases, sites already known under one name would receive a second name, e.g. “Otomani-Cetatea de Pământ” (present-day Romania) and “Ottomány-Földvár” (previously part of Hungary) or “Вертеба” (present-day Ukraine) and “Bilcze Złote Werteba” (previously part of Poland). Reporting all the names known for a site helps create a more reliable archive of associated data, effectively preventing it from becoming duplicated under two different names. Additional names can always be added as and when they become associated with a site.

Proper publication protocols include as many different names for a site as are known, as well as any reference numbers to local and/or national registries and catalogues associated with the site. Returning to the Danish example of “Ølby” above, the ideal publication of new data from this site will include not only a map of the location but also the national heritage site number (020105-3) and the site coordinates in a written format (WGS84 N55.4912, E12.1514).

There are also situations in which researchers are unwilling or legally unable to provide the exact coordinates of a site for fear of looting (American National Parks Service on “Looting and Vandalism”). For example, the American Archaeological Resource Protection Act of 1979 prohibits the public disclosure of “sensitive” information including the location of archaeological resources (16 U.S.C. §470hh). In such cases, we suggest identifying a site to a local region only (and to explicitly state that this is the cause for reticence in terms of site location).

Furthermore, it is crucial to situate the materials within a site correctly, especially when reporting sampling procedures from selected contexts. Rather than introducing a site in the “background” section of a publication and discussing only the most sensational parts of it, the key is to report the *actual* details about the parts of the site from which samples were obtained. This should be standard practice even if those contexts may not be as “exciting” as other parts of the site. If word count is an issue, we suggest that authors make use of supplementary data for in-depth context descriptions, as is becoming more and more common in the literature (e.g. Frei et al., 2019). An alternative solution may be to upload archaeological information/documentation to public repositories like Zenodo, which can provide a referenceable DOI (e.g. Carlson et al., 2020).

7.2 Problem II: Culture Is Used Differently (or Not at All)

We are currently in the midst of a shift in how archaeologists define and look at cultures. Those adhering to a more traditional approach still hold cultures to be a keystone of archaeological practice, while others view

them as an increasingly woolly, problematic or even unnecessary category (Riede et al., 2019; Ross, 2012). This affects database infrastructure insofar as the concept of culture is used differently (or not at all) within different time periods and regions. There are some areas for which the level of detail discernible in archaeological material does not allow for classification into a culture. In such cases, specialists distinguish only by period rather than by culture group. For example, this is standard for the Chalcolithic in the Urals, the Early Iron Age in the Volga region and the periods of the Early and Late Bronze Age in Scandinavia. In other cases, regional specialists may not define culture groups at all, either because this may be glaringly obvious to them (but not necessarily to others) or due to unwillingness to make use of a concept that they feel to be heuristically (or even ideologically) flawed.

7.2.1 Solution II: Data Should Always be Published in Association With a Specific Culture Group or Absolute Chronological Period and the Grounds for This Association Should be Made Explicit

As our understanding of cultures evolves, what we define as "diagnostic" features for a particular culture may change. To future-proof publications against later changes in cultural assignation, we suggest the best practice is to either (1) clearly assign culture *and also specify how that culture was identified* (e.g. the diagnostic traits mentioned above, such as characteristic ceramic types, jewellery, house structure or something entirely other, such as particular husbandry practices) or (2) not assign culture but provide a thorough overview of the documented cultural practices. A publication adhering to these standards might state, e.g. "this grave was identified as coming from the Nagyrév Culture due to the fact that the individual was cremated and that the grave included 14 vessels with the typical geometric patterns associated with this culture within it." Alternatively, an example without a cultural determination might contain the following information "inhumation A was the only example of a burial rite discovered on-site; individual 1/1984 was found inside a rectangular grave pit measuring 200 cm × 75 cm × 150 cm; the body was placed in a supine position with the head directed towards 180° (where true North is 0°); grave goods: none documented; skeletal abnormalities: none documented; post-depositional manipulation: none documented; skeletal completeness: not evaluated."

We are also aware that some academic traditions are moving away from cultural classifications on the whole (Riede et al., 2019; Ross, 2012). It has also been our experience that some data experts publish data on archaeological specimens by chronological periods (e.g. "Neolithic," "Roman" or "Migration Period") rather than by archaeological cultural association (e.g. "Funnel Beaker Culture"). While chronological associations may be more easily relatable than culture on an international stage, we note that the publication of new data that has been assigned a culture group only on the basis of radiocarbon dates or very broad-spectrum periods is extremely difficult to integrate with other data for further archaeological work. The reason for this is that broad archaeological periods are not universally applicable across space; for example at the same moment in time what is Iron Age in Central Europe is still considered to be Bronze Age in Scandinavia. Last but not least, synchronisation of archaeological periods is a challenging task in itself since, e.g. the Early Neolithic period designates extremely different chronological and cultural phenomena in Anatolia, France or Estonia.

7.3 Problem III: Grouping Quantifications of Data

It stands to reason that data from different branches of archaeometry come in different volumes. For example, a single cemetery may have sixty human skeletons across three phases from Late Neolithic and Early Bronze Age contexts and only a small amount of animal remains. In those cases, the easiest approach for the publication of the faunal data may be the publication of the animal remains by period (e.g. all animals from Late Neolithic contexts and all animals from Early Bronze Age contexts) rather than by phase (e.g. Late Neolithic phases 1, 2 and 3 and Early Bronze Age phases 1 and 2). Unfortunately, if such data are published grouped together by period (or even for the site as a whole), they cannot be later separated by the phases used

elsewhere in related publications. In addition, if excavations at the cemetery are taken up again and more materials are analysed at a later stage, grouping by period rather than by phase may become a hindrance to future research.

7.3.1 Solution III: Always Specify Original Chronological (Period/Phase) Associations of the Data Alongside Raw Data

Our experience in data processing suggests that, although it may be analytically more interesting to approach data on a period (vs. phase) basis, specifying the phases from which data stem (e.g. in the supplementary data) future proofs that data for a longer lifespan and greater usability. We recommend keeping the initial, most detailed phasing for primary data and amalgamating only later for analytical purposes. Therefore, the ideal publication should contain both non-amalgamated primary data and analytically grouped datasets. Wherever a specific chronology (e.g. Müller-Karpe, 1959) is used, this should be named and referenced to help future scholars easily adapt any potential adjustments to period associations at a later point.

8 Referencing Data Sources

8.1 Problem IV: Integrating Legacy Data for Re-analysis

A longer lifespan and usability of the data we collect are fundamental to justify the huge expense of collecting those data in the first place and to our ability to reuse them to move the discussion forward or to assess the plausibility of a new research angle. This is just as true for archaeometric as it is for archaeological data, especially considering the past few decades' scientific advances. What can be confusing for data integration in this regard are occasions in which new data are added to old data without clearly differentiating between the two. This is particularly problematic in the case of e.g. interim reports or syntheses which add new data to a body of extant work.

8.1.1 Solution IV: Always Provide a Clear Source and Reference, Also for Data That Have Already Been Published (Even – and Especially if – They Are Your Own!)

To avoid any confusion down the line, the best practice for the use of legacy data – even data that you and/or co-authors have produced – is to clearly label them as previous work (and provide a reference). This is equally as important in tables as it is in charts and graphs. For an example, please see the recent series of publications on isotopic values for Neolithic contexts in southwest Sweden (Blank et al., 2018a,b; Sjögren et al., 2016). In these publications, the sources of newly published results are labelled as "source: this study," while previous results are labelled as "source: [citation]."

Another way to ensure traceability and clarity in data publications is through the use of laboratory codes. In almost all cases, laboratory codes are unique identifiers. As such, laboratory codes are incredibly useful in identifying data that have previously been published. Like other unique identifiers in academia, the use of laboratory codes in publications may impact the success of a publication (Park et al., 2011).

Unique laboratory code designations (see a current list of radiocarbon laboratory codes) also have the additional benefit of ensuring traceability in terms of the initial production of the data. As we move forward in time and become increasingly aware of the importance of laboratory metadata (simultaneously measured standards, measured parts per million/pMC, δ^{13}C measured by AMS, etc.), having the name of the laboratory which produced the analyses makes them traceable. For this reason, we also recommend that the name of the laboratory that conducted the analyses be listed in the methods section of the publication. In cases in which an issue with a particular laboratory is identified *post-facto* (e.g. Meadows et al., 2015; Rose et al., 2019, Appendix

1), this will help with identification for future re-analysis or eventual data exclusion. Furthermore, if follow-on analyses are made based on a legacy sample, it is essential to include information on *both* the original sample designations as well as any new sample codes to ensure data continuity and transparency.

Unfortunately, the issue of unique identifiers, such as laboratory codes, being duplicated due to minor alterations is a common problem in large datasets. For instance, a search in the XRONOS database for the site of La Sente in France reveals duplicates such as GrN-32097, 3,760 ± 50 and GrA-32097, 3,760 ± 50, as well as GrA-4468, 3,805 ± 35 and GrA-44685, 3,805 ± 35. Although this may on the surface seem to be a "database" problem, we have it listed here as we find that these errors often stem from source publications (e.g. errors in transcription from lab reports to final publication). Even small discrepancies create duplicate entries in the database, complicating data analysis and interpretation. As the volume of available data grows beyond the capacity for manual detection, even seemingly minor errors can significantly hinder the efficient and accurate utilisation of the data. Large community efforts are required to correct and keep a record of such errors, as can be seen by the modern European Pollen Database (EPD). It is crucial to implement robust data management practices to prevent such issues and to ensure the integrity of the data from the very beginning.

Robust data management protocols also extend to referencing. By including adequate references for your publication, you ensure that the data and the reasonings behind your results are traceable by future researchers. Moreover, this rule of thumb should also be applied to specialist papers that focus on a particular aspect of archaeological data. Rather than giving a brief overview of a site only with regard to the eventual conclusions discussed by the specialist paper, it is crucial to provide clear references for further information about the contexts discussed.

This can sometimes be problematic with regard to e.g. grey literature/unpublished specialist reports (even if the authors have given their permission for their data to be used), as some journals will not accept citations from unpublished works or on occasions in which hundreds of datasets are used in a single paper. There are two ways forward in such instances. First, authors may be able to list such works as a separate category under "other sources" or "full dataset references" following their bibliography. Alternatively, such references may be mentioned in the supplementary materials (in which case they may additionally not count towards the official paper references).

9 Consistently Formatting and Harmonising Data Outcomes

9.1 Problem V: Data Are Not Easily Extracted

Although it may seem to be self-evident that the data used for quantitative research should be published alongside the inferences and conclusions drawn from them, this is not always the case. Even though we all nominally agree that papers that do not present the new data they discuss should not make it through peer review, the increasing trans- and interdisciplinarity of archaeology means that the person(s) who carried out the analyses are not always the first or main authors of the paper, and certain information can be left off or edited out.

9.1.1 Solution V: Present Data Clearly in an Easily Tabulated Format

Here we cannot emphasise enough the importance of supplementary data, especially for large quantities of information. Ideally, these should be presented in a downloadable format (spreadsheets in .csv or similar). Information that is presented in tables within the body of a manuscript is a clear second choice, as these must be converted (via commercial software) to a tabulated format for integration and further analysis.

Our experience shows that the less manipulation needed for further data use, the less chances there are for user error. In the same line, if data are divided across several sheets or tables between manuscript and

supplementary data, it is a great help if the presentation of those data is consequent and also follows a recognisable format (e.g. if there is a unique identifier – such as a laboratory code – which helps differentiate samples from each other, especially when multiple samples were taken from the same context/individual/material; see Section 8.1.1).

This is not to say that we advocate publishing in a manner that differs greatly from the standard means of presenting data for each (sub)discipline. In most cases, the rows and columns present in a data table that holds to the discipline's standards will be sufficient, so long as other points (see above and below) are also taken into account. For example, the contents and structure of an archaeobotany database will be the same as those listed in e.g. the ArboDat system used in more than 30 laboratories today (Kreuz & Schäfer, 2022). Another option may be to include the original laboratory report in the supplementary materials (while also ensuring that there is continuity between labelling of sample/laboratory numbers between the publication and the supplementary data).

The BIAD wiki (https://biadwiki.org/en/Structure) provides links to a series of templates (.csv files) which we have formatted specifically in response to disciplinary standards but which also encompass the variety of data presentation methods we have come across during our data harvesting thus far. These are also automatically updated whenever BIAD is altered. Information on e.g. the categories described within those spreadsheets is also freely accessible (and continually updated) via the BIAD wiki. Importantly, although the spreadsheets are specifically designed for uploads to BIAD, they include columns for all relevant data, ensure sufficient coverage for ease of upload for other databases as well, are open to download and use by other scholars in data publications and help ensure adherence to the BIAD Standards.

10 Describing Data Fully

10.1 Problem VI: Insufficient Context Information

We have come across many occasions in which authors have published, e.g. ^{14}C dates with site names, but without specifying either which contexts were present at the site or what was actually dated. Not knowing whether scientific data come from, e.g. a settlement or a burial site, or from human bone or charcoal hinders their future use.

On multiple occasions, we have been confused about the particular individual/grave/context from which published data originate. Identifying which individual/grave/context was sampled after the results are published is extremely time-consuming and can potentially lead to misidentification. We particularly wish to emphasise that reporting data on an individual identified by grave name/number/other ID and a dot on a map alone is not always sufficiently clear to ensure that said individual can be identified with confidence for future work (see Section 7.1.1).

We include here an example from a site in Scandinavia (which shall remain nameless for the purposes of this exercise):

> The site consists of a Neolithic megalithic tomb containing seven individuals. Of those seven, three have undergone archaeometric analysis. Comparison between the publication of the new results and the original site publication reveals that the original provides no individual-specific record for one of the individuals. It further reveals that the way of identifying those individuals has changed from the original publication: from grave letters followed by sequential numbers (e.g., individuals A1 and A2) to simple numbers (e.g., individual 1, 2, 3, etc.), which may have caused an individual to 'disappear' in the process. After a thorough examination of both publications, paying particular attention to which teeth were sampled, the source of this disappearance was identified: two individuals from the same grave, a subadult and an adult individual, were misnumbered during the switch to the new numbering system. Individual D1 (subadult) was numbered 4b, and D2 (adult) became 4a. However, the archaeometric publication (which followed the archaeological one) only referred to the subadult individual as 4a which, though correct, created the need for additional research for reference to the original publication.

10.1.1 Solution VI: All Context Information and Other Names/Specifications Should Be Included With Each Data Entry

Archaeological recording conventions are important for the later traceability of data. For this reason, we strongly suggest that all context information, including individual names or numbers, grave names or numbers, archaeological find contexts, museum accession numbers, box/shelving/unit numbers and archive acquisition codes or years accompany data to the publication stage (at least in supplementary data, if not in the main manuscript itself). In short, anything written on the box or find bag *must* accompany any data produced from the material in the publication. This may be crucial to future work with the material that was sampled and/or other material from the same site/find. If new naming conventions are used, these must be specified in the publication (e.g. Aner & Kersten, 1976).

For example, publications that list isotopic data (e.g. $^{87}Sr/^{86}Sr$, ^{18}O, ^{34}S, ^{13}C, ^{14}C and ^{15}N) and aDNA data values should specify the particular individuals from which the new data derive in relation to grave, individual and sample type as well as lab code (see Section 8.1.1). Again, as future studies may choose to go back and sample other tissues/individuals, this helps produce consequent metadata about what has been done (see Section 11.1.1). If this information is clearly set out from the outset, e.g. “This paper presents new data on individual 4b from XYZ site, previously referred to as ‘Individual D1’ [Smith 1990]”, a tremendous amount of time and effort has been saved and we both minimise the margin for error and the perpetuation of those errors across the long term.

11 Reporting Results Comprehensively

11.1 Problem VII: Failed Analyses (and the Causes Thereof) are Lost if Not Reported

We are not always successful in obtaining the desired results when analysing archaeological samples. Sometimes, the aDNA is too damaged or on other occasions, there is not enough collagen for ^{14}C dating, etc. The elevated sensitivity of modern methods and instruments can sometimes overcome hurdles presented by the material that were not possible even a decade ago. However, institutions are often reluctant to give sampling permissions if they have no guarantee that a subsequent destructive analysis will give results. To give the best chances for not only sampling permissions to be granted, but also for success in the laboratory, we need to have respect for the material and not lose the knowledge that has already been gained by a failed analysis.

11.1.1 Solution VII: Report Failed Analyses (and the Causes Thereof)

We readily acknowledge that publishing fully null results and failed studies is challenging and that the fact that the importance of such studies is consistently undervalued is an endemic failing across disciplines (Sindall & Barrington, 2020). Nevertheless, we argue that the failure to report a failed analysis is the true crime, not the failure of the analysis itself. Scholars should notify the curating institutions/archives of the results of their analyses – *even and especially if those analyses fail* – and should, furthermore, include that information in their final publication. For example, if a radiocarbon sample failed due to, e.g. insufficient collagen preservation, publishing that information will help future researchers orient their sampling strategy. In addition, it is extremely important to also include the type of material (especially in case of repeat samplings of different materials from a single individual) as well as sample identification and name of the facility doing the analysis with each data entry.

This is a question not only of respect for the material, but also of helping future scholars avoid the perpetuation of the same mistakes (e.g. by taking an insufficient sample size, or of a particular material type). Although it is important in all aspects of science, it is particularly noticeable in areas such as aDNA in which sample preparation procedures to establish the preservation of collagen are laborious, time-

consuming and expensive. In addition, this may also help other researchers (especially those in the earlier career phases) identify sites and regions that are rife with preservation issues in spite of the fact that they present excellent research questions. Just as the absence of proof is not proof of absence, so is the absence of results not the same thing as the absence of analysis. Finding a balance between destructive sampling and the necessary advance of science is an ethical challenge (Pálsdóttir et al., 2019, p. 4); reporting failed analyses, their causes and the analytical methods utilised in publications is a simple means of helping future scholars glean as much knowledge as possible from previous attempts.

It is also worthwhile mentioning that we advocate the publication of results which may seem odd insofar as they fall outside an established narrative. For example, although authors may be reluctant to publish ^{14}C dates that fall outside the realm of what is deemed probable for a particular sample at the time of publication, we suggest that these instead be included and noted as "unlikely for xyz reason." Different approaches (e.g. Bayesian chronological modelling) may be able to make sense and fit together data pieces which may seem odd. However, they can only do this *if those original data are published.*

11.2 Problem VIII: How Were Data Calibrated?

Primary data are often adjusted to e.g. a specific standard to make them applicable or comparable with other data. Although this problem is common to almost all data types, most scholars can relate to issues surrounding failure to report the raw uncalibrated values (and their errors) of radiocarbon dates (^{14}C years BP, i.e. before 1950). Different schools of thought, traditions and laboratories calculate dates in various ways. Different journals allow for different publication standards of radiocarbon dates; in extreme cases, only cal BCE/CE ranges are listed, but without information on uncalibrated ^{14}C ages or what ^{14}C calibration dataset, laboratory code or software was used to calibrate the dates. The problem here is that while it is easy to calibrate a raw radiocarbon age using a variety of different approaches and calibration curves, it is practically impossible to uncalibrate a calibrated date. This means that if only a calibrated date range is provided, any interpretation of that date is now tied to the calibration approach and curve used. Although guidelines on how these should be reported exist (Bayliss & Marshall, 2022; Millard, 2014; Polach & Stuiver, 1977), the discipline has for some time neglected to report radiocarbon dates according to conventions. Similarly, the listing of dates as years BP can also be problematic, as different calculations are used depending on how one defines the "present." For example, it has been suggested that luminescence ages standardise the use of AD 2000 as "present" as opposed to the AD 1950 "present" usually used in radiocarbon dating (Duller, 2011). Being inexplicit in such areas can lead to incomparability of the data and to errors in downstream analysis.

11.2.1 Solution VIII: Always (Also) Publish Uncalibrated Data

Averages (e.g. of $^{87}Sr/^{86}Sr$ for all males/females within a certain cemetery or culture group therein) and calibrated data (e.g. $\delta^{18}O$ values to VPDP or VSMOW standard) may be appropriate for many different kinds of discussions and interpretations of primary data. However, unless (and usually, even if) any treatment of the data is presented and explained alongside the primary data, the original work is lost (Roberts et al., 2018). For the long-term reusability of data and analyses, we need to ensure that the uncalibrated values, their errors, laboratory codes and detailed information on the sample material, sample species, archaeological context and laboratory pre-treatments are included in every publication (in supplementary material, if not in the body of the publication itself).

11.3 Problem IX: Publishing Without Mentioning Material (or Part Thereof)

Although we approve of the publication of e.g. .csv files for making data available (see Section 9.1.1), there are cases in which listing multiple results within a single file can lead to confusion. For example, although the

same tooth may have been used for e.g. aDNA, $^{87}Sr/^{86}Sr$, ^{13}C, ^{15}N and ^{18}O, it may not necessarily be the same *parts* of that tooth (e.g. the root dentine might go to aDNA, the enamel might go to $^{87}Sr/^{86}Sr$ and ^{18}O). Bone collagen is formed and remodelled over an individual's lifetime, but the rate at which this happens within a skeleton is not consistent and is affected by sex, age, health, physical activity, etc. (Fahy et al., 2017). Samples of different skeletal elements (bone and teeth) with different intrinsic bone collagen ages can, thus, provide input on different stages of an individual's life history (Chmielewski et al., 2020; Dury et al., 2021): the *pars petrosa* frequently sampled for aDNA and strontium analyses will reflect the early childhood years, whereas the long bones frequently sampled for ^{14}C dating will likely reflect the last few decades prior to death (Geyh, 2001; Hedges et al., 2007).

11.3.1 Solution IX: Be Specific About the Nature of the Sample for Each Analysis

The ideal solution to these issues is to only sample identifiable archaeological materials, be they bone, tooth, charred seeds or wood charcoal. One should also make every attempt to identify the species of the sample as well. If an indeterminate sample must be used, it is better to label that sample clearly as "indeterminate" than to not mention the sample material at all. Failing to specify the sample type/species forces subsequent data users to make guesses about the nature of the sample, potentially resulting in ambiguities. Furthermore, it is in many cases important that every effort is made to identify the part of the original entity that has been sampled (e.g. dentine or enamel for teeth or heartwood, sapwood or short-lived branch for wood (Bayliss, 1999; Bayliss & Marshall, 2022)), as that information is key to unlocking many other forms of data the same may hold e.g. information about the sample's inherent age (Bronk Ramsey, 2009). Therefore, it is also important to register stable isotope values for bones, if they are available. These could contain important information on a possible age-offset or reservoir effect and, thus, on the quality of the data. Whether or not the sample consists of a single entity (e.g. one cereal grain or bone fragment) or is a bulk sample (e.g. a group of cereal grains or several bone fragments) should also be reported.

11.4 Problem X: How Can Quantifiers be Numerically Quantified?

Quantifiers like "some," "a quantity of," "a pile" or "a group" are essentially meaningless once taken out of the context of the site or, at most, the region in which the finds were made.

11.4.1 Solution X: Find Quantification and Classification Should Be the Default

When we record finds from catalogues or excavation reports within a publication, it is crucial to not only classify the finds by type but also to quantify the number. To prepare the best presentation of data for integration in interdisciplinary, international research, we should always make every attempt to quantify numerically and to the greatest exactitude possible. In cases where something cannot be counted, some attempt should be made to estimate the quantity. Number ranges (e.g. 50–100 beads or 500–750 cereal grains) are preferable to orders of magnitude (e.g. "less than a hundred beads" or "several hundred cereal grains").

12 Common Vocabularies

Having a similar data quality across archaeological publications of different varieties is a prerequisite for creating any large interdisciplinary archaeological database. Different scholarly traditions do, however, publish high-quality standardised scientific data each in their own way(s), using different sets of practices that are

unfortunately not always directly relatable to each other. In the following, we detail strategic classification approaches (what we term “vocabularies”) to aid in the harmonisation of data relating to (1) osteological sex, (2) osteological age and (3) plant and animal taxa in a simple but concise way. These vocabularies represent both categories and methods that are universally applicable across different regional research traditions and, furthermore, can be applied flexibly in relation to the level of detail obtainable from the material.

12.1 Problem XI: Differences in Osteological Recording Methods

Some of the recording methods that vary most strongly in the circum-archaeological literature are those related to osteological estimations of the sex and age of human skeletal remains. Osteological age estimation methods rely on the use of reference collections of known age and sex, such as the William M. Bass collection. Many such collections vary in demographic composition and cannot take the full scale of modern human and (pre)historic variation into account. As a result, methods for sex and age estimation are continuously being adapted and improved upon in relation to increased awareness of a variety of different important factors that influence the human body, including e.g. socioeconomics (Cardoso, 2007). Unfortunately, the integration of this kind of data is particularly problematic, as the modes by which data are reported and classified depend largely on the training tradition of the expert who carried out the analysis rather than on the chrono-geographic characteristics of the material (cf. Sosna et al., 2010). For example, forensic anthropologists are more likely to give specific age ranges for each particular case, whereas more traditional osteoarchaeologists are more likely to classify individuals into ordinal groups whose categories vary between national training systems (Buckberry, 2000), no matter the time and region from which the bones derive. Such classification disparities can create significant bias when the resulting data are later integrated by other scholars.

12.1.1 Solution XIa: A Systemised Vocabulary for Recording Anthropological Age

To mitigate the issues brought on by publishing data via recording systems which are still evolving, BIAD has developed a tripartite system for age recording for inputting that information into the database. BIAD records age-at-death in three different (related) formats: (1) the published age range (where available) with a minimum and maximum age (e.g. min. 25 years, max 35 years); (2) the original published description (including original language; and (3) an interpreted BIAD category (Figure 1).

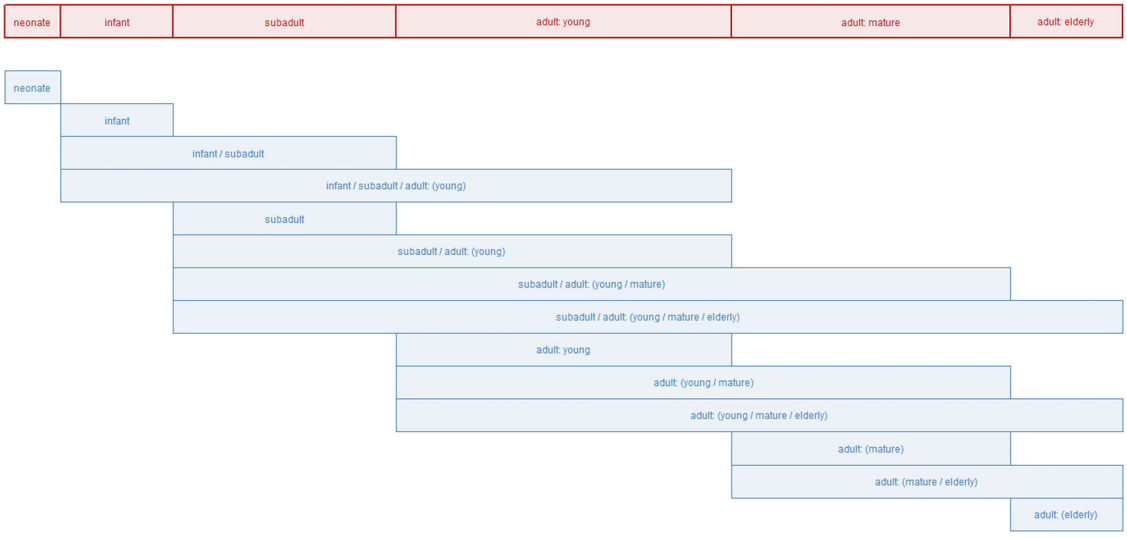

Figure 1: BIAD’s ordinal age categories and amalgamated overlapping age groups.

Both category names and their associated age ranges vary from tradition to tradition. Numerical age ranges are often omitted from publication, but where available can fall only partially within an age category, or straddle several. The BIAD approach considers six common classifications (Figure 1, red) that in principle are ordinal, cover all possible ages and are mutually exclusive (neonate; infant; subadult; adult: young; adult: mature; adult: elderly) and allow for any grouping of one or more of these ordinal categories provided they are sequential. Figure 1 (blue) illustrates the groupings that are currently used in BIAD, reflecting the variability in precision available in the literature. For example, one can differentiate a neonate from an infant osteologically despite the fact that both categories only cover the first few months of early life (Baker et al., 2005). Indeed, as yet, no data have been mined that require a grouping of both (but could be easily generated if required). By comparison, however, there is often substantial osteometric uncertainty distinguishing between "adult: young" and "adult: mature," and the combined category of "adult: (young/mature)" covers many mined data. This approach to harmonising anthropological age ranges enables us to analyse archaeological populations entering publications from diverse academic traditions, and therefore comparing data quality across time and space.

12.1.2 Solution XIb: A Standardised System for Recording Sex Estimates

One of the key issues for transferring published information about osteological sex estimations is a robust and universal expression of uncertainty. Our approach in BIAD is to quantify that uncertainty as a probability of being female, such that 1 represents the certainty of being female and 0 represents the certainty of being male (Table 1). Individuals for whom osteological sex was not estimated or for whom it was estimated as indeterminate based on the material available for analysis are given a score of 0.5. This has required some interpretation at the data mining stage to convert a plethora of uncertain descriptions to a probability, such as "maybe female," "probably male" and even more cryptic (albeit relatively common) assignments such as "male?" and "male??." These descriptions indicate a degree of subjective probability for the original authors. Nevertheless, we encourage assignments to be explicit about their probabilities, which are a consequence of various sources of uncertainty such as the preservation of the osteological material, recovery techniques or the methodology used for sex determination (White & Folkens, 2005).

The rise of molecular anthropology increasingly provides additional data on archaeological human remains which need to be documented systematically (Buonasera et al., 2020). In the notes section of the supplementary materials, authors should consider including the basis for any sex assessment, e.g. whether sex was determined via aDNA, peptide analysis, osteological assessment or if it was based on grave goods/archaeological context (as is more common in areas with poor general bone preservation).

Table 1: BIAD's numeric system for recording sex estimations

Original description	BIAD probability
Female	1
Probably female	0.75
♀ ?	0.75
Possibly female	0.625
♀ ??	0.625
Sex unknown/indeterminate	0.5
Possibly male	0.375
♂ ??	0.375
Probably male	0.25
♂ ?	0.25
Male	0

12.2 Problem XII: Recording Taxa

Just as archaeology has journeyed towards standardisation and formalisation of its codes of practice, so too have there been similar processes within other disciplines. The journey towards standardisation of biological nomenclature began almost 200 years ago (Nicolson, 1991). Today, faunal specialists and archaeobotanists habitually use binomial nomenclature (what many of us refer to as the "Latin name") of the species they identify. This is useful insofar as it can both remove doubt with regards to the specific *variety* of a plant or *subspecies* of an animal that has been identified as well as express uncertainty (e.g. when preservation is sufficient only for a seed to be identified to genus level). For example, the non-expert (including many archaeologists) would refer to examples of all species within the genus *Lavendula* as "lavender." However, there are at least 32 different species within that category, some of which are toxic and others that have medicinal applications (Lis-Balchin, 2012). Though layman's terms (common or colloquial names) of the taxa identified in archaeological and -related research are the most relatable for the greatest number of people, lack of specificity in terms of the species can lead to misidentification and problems with data harmonisation.

12.2.1 Solution XII: A New Vocabulary System for Specifying and Grouping Taxa and Specifying the Degree of Identification

The easiest and most accurate way to record botanical and zoological data from archaeological contexts is by means of the Latin names (binomial nomenclature) of the identified taxa. Therefore, the publication of tables with quantitative data and Latin names of taxa (rather than their common or colloquial names) is extremely important and allows us to avoid ambiguous interpretations during translation.

Built based on a system originally developed for the EUROEVOL Project, BIAD moves the recording of binomial nomenclature forward by defining a "TaxaList," a list of taxon codes based on the Latin system which facilitates rapid data harmonisation, integration and later querying. Each unique code usually consists of the first four letters of the genus and the first three letters of the species (capital letters for archaeobotanical, e.g. TRITDIC for *Triticum dicoccum*, and lowercase for archaeozoological data, e.g. ovisari for *Ovis aries*). Similar coding systems are occasionally applied in the supplementary materials of publications (e.g. Filipović & Obradović, 2013; Gaastra & Vander Linden, 2018; Manning, 2016; Orton et al., 2016).

However, as mentioned, there are also cases in which preservation is insufficient to support a precise identification at the species level. In such cases, plant/animal remains are coded only by the first four letters of the identified level and +SPE/spe is added (for genera, e.g. HORDSPE for *Hordeum* spp.; canispe for *Canis* sp.) or +IND (for families, e.g. CHENIND for Chenopodiaceae indeterminate; cyprind for Cyprinidae indeterminate) (Table 2).

However, plant seeds in particular are often grouped together on the basis of certain important features, because it is not possible to identify them precisely. In such cases, we specify a group via the last letter of the taxon code. In this way, LEGUINL refers to Leguminosae (Fabaceae) indeterminate large and LEGUINS refers to Leguminosae (Fabaceae) indeterminate small. This grouping is particularly important to mark the difference between naked or hulled cultivated cereals (TRITFTW for *Triticum* species free threshing wheat and TRITGLW for *Triticum* species glume wheat; HORSHUL for *Hordeum sativum* [*vulgare*] hulled barley and HORSNAK for *H. sativum* [*vulgare*] naked barley). Such combined assemblages for archaeobotanical remains that have been identified by groupings (e.g. glume wheat) or by some features of a species (e.g. naked barley) allow them to be included in subsequent analyses of dietary reconstructions, the study of agricultural practices, the exploration of exchange networks and the comparison of regional variations in economies.

13 Concluding Remarks

It has been our intent to present some practical solutions for solving many of the data integration and harmonisation hurdles we come across on a daily basis with the goal of providing guidelines for the

Table 2: Examples of BIAD's taxon code recording system

Taxon code	Full name of taxon	Species	Genus	Wild/domestic status	Family	Class	Kingdom
Canifam	*Canis familiaris*	*familiaris*	*Canis*	Domestic	Canidae	Mammalia	Animal
Canispe	*Canis* species	—	*Canis*	Wild/domestic	Canidae	Mammalia	Animal
Cyprind	Cyprinidae	—	—	Wild	Cyprinidae	Pisces	Animal
HORDSPE	*Hordeum* spp.	—	*Hordeum*	Wild/domestic	Poaceae (Gramineae)		Plant
TRITDIC	*Triticum dicoccum*	*diococcum*	*Triticum*	Domestic	Poaceae (Gramineae)		Plant
TRITGLW	Triticum species glume wheat	—	*Triticum*	Domestic	Poaceae (Gramineae)		Plant
CHENIND	Chenopodiaceae indeterminate	—	—	Wild	Chenopodiaceae		Plant

presentation of archaeological data to promote legibility and interoperability. These take the form of twelve practical standards for the publication of archaeological and related data for integration into large-scale interdisciplinary international research. These also include descriptions of common classification vocabularies and strategies we have developed to ease data harmonisation and integration for osteological age, osteological sex and bioarchaeological taxa recording. Throughout we have approached big archaeological data as made up of unique entities which are not inherently unbiased. Like Beer (2019) and Leonelli (2014), we are increasingly aware of how our data come to us, where we see them going and how best we can prepare them for the journey. To quote again from SEADD's homepage,

> [o]ver the last decade, innovation has centred on making archaeological data more interoperable, both to increase the discoverability of data through integrated cross-search, and to facilitate knowledge creation by combining data in new ways. The emerging research challenge of the next decade is optimising archaeological data for re-use, and defining what constitutes good practice around re-use.

The quantitative analyses of large datasets must handle the bias and uncertainty which come from data use and re-use on a day-to-day basis. While we have attempted to help lay the groundwork for tackling these challenges relative to data publication here, there still remains work to be done to assess how best to statistically adjust for certain types of biases, such as varying sampling density/absences or divergent data quality (Isaac et al., 2020). Correcting for data biases, explicitly accounting for uncertainty on all levels and exploring the relevance of absences promise to be fruitful research frontiers to further improve reporting standards and data integration, and which will be explored in a forthcoming publication (Sabatini & Kristiansen, in prep.).

Cultural heritage professionals are acutely aware of the transitory nature of what it is that we study. Archaeological material must survive the ages to make it to our desk or laboratory. While material analysis methods are becoming increasingly less invasive and/or demanding as science advances, most of our sampling is still destructive. While the new knowledge we obtain from such studies and analyses certainly furthers our understanding of the past within a specific topic, the advance of "big data" broadens its potential, so that a small study can be part of something much bigger.

By ensuring that our data are presented in an accessible and interoperable way, we give our new knowledge the best chances of being impactful and having a long "use-life" within the literature, thereby contributing to a wider understanding of the past. As we move towards the goal of increased research e-infrastructures in the cultural heritage domain, it is crucial to have metadata standards and archaeological dataset aggregation schemes in place so as to minimise the amount of work needed to make that data FAIR. Applying the suggestions and principles by both authors and journal and book editors alike such as are laid out above is the first step in this direction, as it formats data for publication in a clear and easily relatable way, be it for integration in a database or for the use of international or non-specialist colleagues who need to contextualise their own work. In so doing, we future-proof our research as best we can, make it more interoperable and also increase citations. Preparing data for publication in this way also has the benefit of facilitating not only databasing but also the development of services and programmes aimed at discovering, accessing and utilising such resources (Klein et al., 2023). If you look at it that way, in publishing archaeological and -related data thoroughly and in as open and inter-relatable a way as possible, we are curating not only the subject of our research but also the research itself.

Funding information: This article is an output of the Synergy project "COREX: From Correlations to Explanations: towards a new European prehistory" funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement no. 95138). We also acknowledge the ERC: CoG project "SUSTAIN: Sustainability of Agriculture in Neolithic Europe" (Grant no. GA 865515), The UKRI Arts and Humanities Research Council (AHRC) and the Austrian Science Foundation (FWF) project "Seascapes: Tracing the emergence and spread of maritime networks in the Central and Western Mediterranean in the 3rd millennium BC" project (Grant nos AH/T012803/1 and I05088, respectively) and the ERC-StG project NEOSEA (Grant No. 949424). We are grateful for the support of the Carlsberg Foundation Semper Ardens research grant (CF18-0005) "Tales of Bronze Age People" and the research grant "Tales of Bronze Age

Women" (CF-15 0878) both to KMF. The views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results and approved the final version of the manuscript. All authors contributed to the study conception and design. Investigation, conceptualisation and visualisation (including (material preparation, data collection and analysis) were performed by SSR, RS, JK, JB, HAR, NR, CS, NS, VYKL, EC, FD, MI-B, EJK, ED, LF and AT. Database structure and programming was done by RS, VYKL, EC, JK and AT. The first draft of the manuscript was written by SSR, RS, JK, JB, HAR, NR and CS. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Conflict of interest: Authors state no conflict of interest.

Data availability statement: All tables and resources discussed in this article are freely available from the BIAD wiki (https://biadwiki.org/). BIAD is also currently in the process of preparing an open-access version of the database for public release.

References

Aner, E., & Kersten, K. (1976). *Die Funde der älteren Bronzezeit des Nordischen Kreises in Dänemark, Schleswig-Holstein und Niedersachsen: Holbæk, Sorø und Præstø Amter*. The National Museum of Denmark, Karls Wachholtz.

Atalay, S. (2006). Guest editor's remarks: Decolonizing archaeology. *American Indian Quarterly*, 30(3/4), 269–279. Retrieved from http://www.jstor.org/stable/4139015.

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*, *20*(4), 663–681. doi: 10.1007/s10816-012-9132-9.

Baker, B. J., Dupras, T. L., & Tocheri, M. W. (2005). *The osteology of infants and children* (1 ed.). Texas A&M University Press.

Bayliss, A. (1999). On the taphonomy of charcoal samples for radiocarbon dating. In J. Evin (Ed.), *Actes du 3ème Congrès International 14C et Archéologie: Lyon 6-10 avril 1998* (pp. 51–56). Soc. Préhist. Française.

Bayliss, A., & Marshall, P. (2022). *Radiocarbon dating and chronological modelling: Guidelines and best practice*. Historic England. https://historicengland.org.uk/images-books/publications/radiocarbon-dating-chronological-modelling/.

Beer, D. (2019). *The data gaze: Capitalism, power and perception*. Sage Publications.

Biddick, K. (1993). Decolonizing the English past: Readings in medieval archaeology and history. *Journal of British Studies*, *32*(1), 1–23. doi: 10.1086/386018.

Blank, M., Sjögren, K. G., Knipper, C., Frei, K. M., & Storå, J. (2018a). Isotope values of the bioavailable strontium in inland southwestern Sweden – A baseline for mobility studies. *PLoS ONE*, *13*(10), e0204649.

Blank, M., Tornberg, A., & Knipper, C. (2018b). New perspectives on the late neolithic of south-western Sweden. An interdisciplinary investigation of the gallery grave falköping stad 5. *Open Archaeology*, *4*(1), 1–35. doi: 10.1515/opar-2018-0001.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679. doi: 10.1080/1369118X.2012.678878.

Bronk Ramsey, C. (2009). Dealing with outliers and offsets in radiocarbon dating. *Radiocarbon*, *51*(3), 1023–1045. doi: 10.1017/S0033822200034093.

Buckberry, J. (2000). *Missing, presumed buried? Bone diagenesis and the under-representation of Anglo-Saxon children*. Assemblage 5. http://hdl.handle.net/10454/676.

Buonasera, T., Eerkens, J., de Flamingh, A., Engbring, L., Yip, J., Li, H., Haas, R., DiGiuseppe, D., Grant, D., Salemi, M., Nijmeh, C., Arellano, M., Leventhal, A., Phinney, B. Byrd, B. F., Malhi, R. S., & Parker, G. (2020). A comparison of proteomic, genomic, and osteological methods of archaeological sex estimation. *Scientific Reports*, *10*(1), 11897. doi: 10.1038/s41598-020-68550-w.

Cardoso, H. F. V. (2007). Environmental effects on skeletal versus dental development: Using a documented subadult skeletal sample to test a basic assumption in human osteological research. *American Journal of Physical Anthropology*, *132*(2), 223–233. doi: 10.1002/ajpa.20482.

Carlson, D. F., Walsh, M. J., Tejsner, P., & Thomsen, S. (2020). *A 3-D model of The Bear Trap: A unique stone structure on the northwest tip of the Nuussuaq Peninsula, Greenland*. Zenodo. https://zenodo.org/records/4075144.

Chmielewski, T. J., Hałuszko, A., Goslar, T., Cheronet, O., Hajdu, T., Szeniczey, T., & Virag, C. (2020). Increase in C dating accuracy of prehistoric skeletal remains by optimised bone sampling: Chronometric studies on eneolithic burials from Mikulin 9 (Poland) and Urziceni-Vada Ret (Romania). *Geochronometria*, *47*(1), 196–208. doi: 10.2478/geochr-2020-0026.

Cleere, H. (1989). *Archaeological heritage management in the modern world*. Unwyn Hyman.

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.

Djindjian, F., & Moscati, P. (Eds.). (2021). Big data and archaeology. *Proceedings of the XVIII UISPP World Congress (4–9 June 2018, Paris, France)*. Archaeopress.

Duller, G. A. T. (2011). What date is it? Should there be an agreed datum for luminescence ages? *Ancient TL*, *29*(1), 1–3.

Dury, J. P. R., Lidén, K., Harris, A. J. T., & Eriksson, G. (2021). Dental wiggle matching: Radiocarbon modelling of sub-sampled archaeological human dentine. *Quaternary International*, *595*, 118–127. doi: 10.1016/j.quaint.2021.03.030

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. doi: 10.1179/1756750513Z.00000000023.

Fahy, G. E., Deter, C., Pitfield, R., Miszkiewicz, J. J., & Mahoney, P. (2017). Bone deep: Variation in stable isotope ratios and histomorphometric measurements of bone remodelling within adult humans. *Journal of Archaeological Science*, *87*, 10–16. doi: 10.1016/j.jas.2017.09.009.

Filipović, D., & Obradović, R. (2013). Archaeobotany at Neolithic Sites in Serbia: A critical overview of the methods and results. In N. Miladinović-Radmilović & S. Vitezović (Eds.), *Bioarchaeology in the Balkans balance and perspectives* (pp. 25–57). Sremska Mitrovica.

Fleming, D. (2020). The Internationalization and Institutionalization of Archaeology, or, How a Rich Man's Pastime Became an International Scientific Discipline, and What Happened Thereafter. *Bulletin of the History of Archaeology*, *30*(1), 1–12. doi: 10.5334/bha-628.

Frei, K. M., Bergerbrant, S., Sjögren, K.-G., Jørkov, M. L., Lynnerup, N., Harvig, L., Allentoft, M. E., Sikora, M., Price, T. D., Frei, R., & Kristiansen, K. (2019). Mapping human mobility during the third and second millennia BC in present-day Denmark. *PLOS ONE*, *14*(8), e0219850. doi: 10.1371/journal.pone.0219850.

Gattiglia, G. (2015). Think big about data: Archaeology and the big data challenge. *Archäologische Informationen*, *38*, 113–124. doi: 10.11588/ai.2015.1.26155.

Gerbault, P., Liebert, A., Itan, Y., Powell, A., Currat, M., Burger, J., Swallow, D. M., & Thomas, M. G. (2011). Evolution of lactase persistence: An example of human niche construction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1566), 863–877. doi: 10.1098/rstb.2010.0268.

Geser, G., Richards, J., Massara, F., & Wright, H. (2022). Data management policies and practices of digital archaeological repositories. *Internet Archaeology*, *59*(2), 1–49. doi: 10.11141/ia.59.2.

Geyh, M. A. (2001). Bomb radiocarbon dating of animal tissues and hair. *Radiocarbon*, *43*(2B), 723–730. doi: 10.1017/S0033822200041382.

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron*. MIT Press.

Gaastra, J. S., & Vander Linden, M. (2018). Farming data: Testing climatic and palaeoenvironmental effect on Neolithic Adriatic stock-breeding and hunting through zooarchaeological meta-analysis. *The Holocene*, *28*(7), 1181–1196. doi: 10.1177/0959683618761543.

Hedges, R. E. M., Clement, J. G., Thomas, C. D. L., & O'Connell, T. C. (2007). Collagen turnover in the adult femoral mid-shaft: Modeled from anthropogenic radiocarbon tracer measurements. *American Journal of Physical Anthropology*, *133*(2), 808–816. doi: 10.1002/ajpa.20598.

Howey, M. C. L., Sullivan, F. B., Burg, M. B., & Palace, M. W. (2020). Remotely sensed big data and iterative approaches to cultural feature detection and past landscape process analysis. *Journal of Field Archaeology*, *45*(sup1), S27–S38. doi: 10.1080/00934690.2020.1713435.

Huggett, J. (2020). Is big digital data different? Towards a new archaeological paradigm. *Journal of Field Archaeology*, *45*(sup1), S8–S17. doi: 10.1080/00934690.2020.1713281.

Huggett, J. (2023). Deconstructing the digital infrastructures supporting archaeological knowledge. *Current Swedish Archaeology*, *31*, 11–38. doi: 10.37718/CSA.2023.01.

Huvila, I. (2018). Ecology of archaeological information work. In I. Huvila (Ed.), *Archaeology and archaeological information in the digital society* (pp. 121–141). Routledge.

Innesti, A., Barbero, M., Linz, F., de Vries, M., Mauritz, S., Wauters, P., Chrzanowski, P., Jakimowicz, K., Bartz, K., Tenge, E., Graux, H., Osimo, D., Ypma, P., & Hillebrand, A. European Commission, Directorate-General for Communications Networks. (2018). *Study to support the review of Directive 2003/98/EC on the re-use of public sector information – Final report*. Publications Office. doi: 10.2759/373622.

Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R. Simmonds, E. G., & O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67. doi: 10.1016/j.tree.2019.08.006.

Jackson, S. J., Edwards, P. N., Bowker, G. C., & Knobel, C. P. (2007). Understanding infrastructure: History, heuristics and cyberinfrastructure policy. *First Monday*, *12*(6). doi: 10.5210/fm.v12i6.1904.

Johnson, M. H. (2011). On the nature of empiricism in archaeology. *Journal of the Royal Anthropological Institute*, *17*(4), 764–787. doi: 10.1111/j.1467-9655.2011.01718.x.

Kansa, E. (2022). On infrastructure, accountability, and governance in digital archaeology. In K. Garstki (Ed.), *Critical Archaeology in the Digital Age: Proceedings of the 12th IEMA Visiting Scholar's Conference* (pp. 141–152). Cotsen Institute of Archaeology Press.

Kintigh, K. W. (2005). The promise and challenge of archaeological data integration. *Anthropology News*, *46*(7), 16.

Kintigh, K. W., Altschul, J. H., Beaudry, M. C., Drennan, R. D., Kinzig, A. P., Kohler, T. A., Limp, W. F., Maschner, H. D. G., Michener, W. K., Pauketat, T. R., Peregrine, P., Sabloff, J. A., Wilkinson, T. J., Wright, H. T., & Zeder, M. A. (2014). Grand challenges for archaeology. *American Antiquity*, *79*(1), 5–24. Retrieved from http://www.jstor.org/stable/24712724.

Klein, K., Wohde, A., Gorelik, A. V., Heyd, V., Diekmann, Y., & Brami, M. (2023). AutArch: An AI-assisted workflow for object detection and automated recording in archaeological catalogues. *ArXiv*, 2311.17978. doi: 10.48550/arXiv.2311.1797.

Kreuz, A., & Schäfer, E. (2022). *Archaeobotanical database programme ArboDat*. GFZpublic. Retrieved from https://gfzpublic.gfz-potsdam.de/rest/items/item_5012429_5/component/file_5012431/content?download=true.

Kristiansen, K. (1984). Dansk arkæologi – Fortid og fremtid. *Fortid Og Nutid*, *164*, 164–205. https://tidsskrift.dk/fortidognutid/issue/view/6795/921.

Kristiansen, K., González-Ruibal, A., Chilton, E., & Niklasson, E. (2014). Towards a new paradigm? The third science revolution and its possible consequences in archaeology. *Current Swedish Archaeology*, *22*, 11–34. Retrieved from https://www.academia.edu/10100372/TOWARDS_A_NEW_PARADIGM_The_Third_Science_Revolution_and_its_Possible_Consequences_in_Archaeology.

Larson, G., Piperno, D. R., Allaby, R. G., Purugganan, M. D., Andersson, L., Arroyo-Kalin, M., Barton, L., Climer Vigueira, C., Denham, T., Dobney, K. Doust, A. N., Gepts, P., Thomas, M., Gilbert, P., Gremillion, K. J., Lucas, L., Lukens, L., Marshall, F. B., Olsen, & Fuller, D. Q. (2014). Current perspectives and the future of domestication studies. *PNAS*, *111*(17), 6139–6146. doi: 10.1073/pnas.1323964111.

Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. *Big Data & Society*, *1*(1), 2053951714534395. doi: 10.1177/2053951714534395.

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, *82*(5), 810–821. doi: 10.1086/684083.

Levy, T. E. (2004). Editorial. *Near Eastern Archaeology*, *77*(3), 1–2. doi: 10.5615/neareastarch.77.3.fm.

Lis-Balchin, M. T. (2012). 17 – Lavender. In K. V. Peter (Ed.), *Handbook of herbs and spices* (2nd ed., pp. 329–347). Woodhead Publishing. doi: 10.1533/9780857095688.329.

Lohr, S. (2012, November 2). *The age of big data*. The New York Times. Retrieved from https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html.

Mallick, S., Micco, A., Mah, M., Ringbauer, H., Lazaridis, I., Olalde, I., Patterson, N., & Reich, D. (2023). The allen ancient DNA resource (AADR): A curated compendium of ancient human genomes. *BioRxiv*, 2023.04.06.535797. doi: 10.1101/2023.04.06.535797.

Manning, K. (2016). The cultural evolution of Neolithic Europe. EUROEVOL dataset 2: Zooarchaeological data. *Journal of Open Archaeology Data*, *5*, 1–5. doi: 10.5334/joad.41.

Meadows, J., Hüls, M., & Schneider, R. (2015). Accuracy and reproducibility of 14C measurements at the Leibniz-Labor, Kiel: A first response to Lull et al., "when 14C dates fall beyond the limits of uncertainty: An assessment of anomalies in Western mediterranean bronze age 14C series." *Radiocarbon*, *57*(5), 1041–1047. doi: 10.2458/azu_rc.57.18569.

Millard, A. R. (2014). Conventions for reporting radiocarbon determinations. *Radiocarbon*, *56*(2), 555–559. doi: 10.2458/56.17455.

Mitchener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications*, *7*, 330–342. doi: 10.2307/2269427.

Müller-Karpe, H. (1959). Beiträge zur Chronologie der Urnfelderzeit nördlich und südlich der Alpen. *Römisch-Germanische Forsungen*, Band 22. Berlin, pp. 334.

Newhard, J. (2013). *Archaeology, humanities and data science* [College of Charleston Blog]. Retrieved April, 10, 2023, from The ArchaeoInformant website: https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html.

Nicolson, D. H. (1991). A history of botanical nomenclature. *Annals of the Missouri Botanical Garden*, *78*(1), 33–56. doi: 10.2307/2399589.

Novák, D., Oniszczuk, A., & Gumbert, B. (2023). Digital archaeological archiving policies and practice in Europe: The EAC call for action. *Internet Archaeology*, *63*, 1–21. doi: 10.11141/ia.63.7.

Novák, D., Oniszczuk, A., Tsang, C., Gumbert, B., de Langhe, K., & Watson, J. (2024). *Revisiting the Valletta Convention for the Digital Age: Position statement on archiving primary archaeological data (EAC Guidelines No. 6)*. European Archaeological Council. Retrieved from European Archaeological Council website: https://zenodo.org/records/10695890.

O'Malley, M. A., & Soyer, O. S. (2012). The roles of integration in molecular systems biology. *Data-Driven Research in the Biological and Biomedical Sciences*, *43*(1), 58–68. doi: 10.1016/j.shpsc.2011.10.006.

Oniszczuk, A., Tsang, C., Brown, D. H., Novák, D., & de Langhe, K. (2021). *Guidance on selection in archaeological archiving (EAC Guidelines No. 3)*. Brussels: European Archaeological Council. Retrieved from European Archaeological Council website: doi: 10.5281/zenodo.10671360.

Onsrud, H., & Campbell, J. (2007). Big opportunities in access to "small science" data. *Data Science Journal*, *6*, 1–9. doi: 10.2481/dsj.6.OD58.

Orton, D., Gaastra, J., & Vander Linden, M. (2016). Between the Danube and the deep blue sea: Zooarchaeological meta-analysis reveals variability in the spread and development of Neolithic farming across the Western Balkans. *Open Quaternary*, *2*, 1–6. doi: 10.5334/oq.28.

Pálsdóttir, A. H., Bläuer, A., Rannamäe, E., Boessenkool, S., & Hallsson, J. H. (2019). Not a limitless resource: Ethics and guidelines for destructive sampling of archaeofaunal remains. *Royal Society Open Science*, *6*(10), 191059. doi: 10.1098/rsos.191059.

Park, S., Zo, H., Ciganek, A. P., & Lim, G. G. (2011). Examining success factors in the adoption of digital object identifier systems. *Electronic Commerce Research and Applications*, *10*(6), 626–636. doi: 10.1016/j.elerap.2011.05.004.

Plomp, E., Stantis, C., James, H. F., Cheung, C., Snoeck, C., Kootker, L., Kharobi, A., Borges, C., Moreiras Reynaga, D. K., Pospieszny, L., Fulminante, F., Stevens, R., Alaica, A. K., Becker, A., de Rochefort, X., & Salesse, K. (2022). The IsoArcH initiative: Working towards an open and collaborative isotope data culture in bioarchaeology. *Data in Brief*, *45*, 108595. doi: 10.1016/j.dib.2022.108595.

Polach, H. A., & Stuiver, M. (1977). Discussion reporting of 14C data. *Radiocarbon*, *19*(3), 355–363. doi: 10.1017/S0033822200003672.

Ribeiro, A. (2019). Science, data, and case-studies under the third science revolution: Some theoretical considerations. *Current Swedish Archaeology*, *27*(1), 115–132. doi: 10.37718/CSA.2019.06.

Riede, F., Hoggard, C., & Shennan, S. (2019). Reconciling material cultures in archaeology with genetic data requires robust cultural evolutionary taxonomies. *Palgrave Communications*, *5*(1), 55. doi: 10.1057/s41599-019-0260-7.

Roberts, P., Fernandes, R., Craig, O. E., Larsen, T., Lucquin, A., Swift, J., & Zech, J. (2018). Calling all archaeologists: Guidelines for terminology, methodology, data handling, and reporting when undertaking and reviewing stable isotope applications in archaeology. *Rapid Communications in Mass Spectrometry*, *32*(5), 361–372. doi: 10.1002/rcm.8044.

Rose, H. A., Boudin, M., Hamann, C., Huels, M., Meadows, J., & Palstra, S. W. L. (2019). Radiocarbon dating cremated bone: A case study comparing laboratory methods. *Radiocarbon*, *61*(5), 1581–1591. doi: 10.1017/RDC.2019.70.

Roskams, S., & Whyman, M. (2007). Categorizing the past: Lessons from the archaeological resource assessment for Yorkshire. *Internet Archaeology*, *23*(23). doi: 10.11141/ia.23.2.

Ross, D. E. (2012). Transnational artifacts: Grappling with fluid material origins and identities in archaeological interpretations of culture change. *Journal of Anthropological Archaeology*, *31*(1), 38–48. doi: 10.1016/j.jaa.2011.10.001.

Sabatini, S., & Kristiansen, K. (in prep.). “Ethics of databases”.

Saktura, W. M., Rehn, E., Linnenlucke, L., Munack, H., Wood, R., Petchey, F., Codilean, A. T., Jacobs, Z., Cohen, T. J., Williams, A. N., & Ulm, S. (2023). SahulArch: A geochronological database for the archaeology of Sahul. *Australian Archaeology*, *89*(1), 1–13. doi: 10.1080/03122417.2022.2159751.

Schnapp, A. (1996). *The discovery of the past: The origins of archaeology*. Harry N. Abrams.

Sindall, R. C., & Barrington, D. J. (2020). Fail fast, fail forward, fail openly: The need to share failures in development. *Journal of Trial and Error*, *1*(1). doi: 10.36850/ed2.

Sjögren, K.-G., Price, T. D., & Kristiansen, K. (2016). Diet and mobility in the Corded Ware of Central Europe. *PLOS ONE*, *11*(5), e0155083. doi: 10.1371/journal.pone.0155083.

Sosna, D., Sládek, V., & Galeta, P. (2010). Investigating mortuary sites: The search for synergy. *Anthropologie*, *XLVIII*, 33–40.

Speciale, C., Allué, E., Riabogina, N., Timpson, A., (in prep.). Wood databases in archaeology: structures and perspectives (provisional title).

Spielmann, K., & Kintigh, K. (2011). The digital archaeological record: The potentials of archaeozoological data integration through tDAR. *The SAA Archaeological Record*, *11*(1), 22–25.

Steibing, W. H. (1993). *Uncovering the past: A history of archaeology*. Oxford University Press.

Thomas, D. H. (2001). *Skull wars: Kennewick Man, archaeology, and the battle for Native American identity*. Basic Books.

Timpson, A., Blanz, M., Bulatović, J., Canteri, E., Cramp, L., Dahlberg, F., Dankov, G., Davy, T., Derenne, E., Frank, L., Fyfe, R., Ivanova-Bieg, M., Kate, E. J., Kjær, K. Kolář, J., Kristiansen, K., Lee, V. Y. K., Manning, K. M., Paulsson, B. S., … Thomas, M. G. (in prep.). BIAD 2025 Report.

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

UNESCO. (1998, December 2). Archaeological Site of Troy. https://whc.unesco.org/en/list/849/#:~:text=24%20excavation%20campaigns%2C%20spread%20over,portions%20of%20five%20defensive%20bastions.

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J. S., Renaut, S., & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, *24*(1), 94–97. doi: 10.1016/j.cub.2013.11.014.

White, T. D., & Folkens, P. A. (2005). *The human bone manual*. Elsevier Academic.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J-W. Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. doi: 10.1038/sdata.2016.18.

Williams, J. W., Grimm, E. C., Blois, J. L., Charles, D. F., Davis, E. B., Goring, S. J., Graham, R. W., Smith, A. J., Anderson, M., Arroyo-Cabrales, J., Ashworth, A. C., Betancourt, J. L., Bills, B. W., Booth, R. K., Buckland, P. I., Curry, B. B., Giesecke, T., Jackson, S., Latorre, C., & Takahara, H. (2018). The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. *Quaternary Research*, *89*(1), 156–177. Cambridge Core. doi: 10.1017/qua.2017.105.

Appendix 1

Table A1: Examples of major large databases used in archaeology and its related disciplines

Name	Type	URL
Allen Ancient DNA Resource (AADR)	Database	https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data
Archaeobotanical Database of Eastern Mediterranean And Near Eastern Sites (ADEMNES)	Database	https://www.ademnes.de/
Archaeology Data Service (ADS)	Data repository	https://archaeologydataservice.ac.uk/
ArkeoGIS	Data repository	http://www.arkeogis.org
Comprehensive Archaeological and Archaegenomic Database and Online GIS (Anthropology GIS)	Database	https://www.anthropology-gis.com/
African Pollen Database (APD)	Database	https://africanpollendatabase.ipsl.fr/#/home
Arches open source data management platform	Data repository	https://www.archesproject.org/
ARIADNEplus	Data repository	https://ariadne-infrastructure.eu/
Bugs Coleopteran Ecology Package (BugsCEP)	Database	https://www.bugscep.com/
Canadian Archaeological Radiocarbon Database (CARD)	Database	https://www.canadianarchaeology.ca/
CBAB: Cremation Bronze Age Burials	Database	https://cbab.acdh.oeaw.ac.at/
Cronica (Repository preliminary archaeological reports in Romania)	Database	
Data Archiving and Networked Services (DANS-EASY)	Data repository	https://dans.knaw.nl/en/
Dietary Isotope Baseline for the Ancient North (DIANA)	Database	https://www.oasisnorth.org/diana.html
Diatom Paleoliminology Data Cooperative (DPDC)	Database	https://diatom.ansp.org/dpdc/Information.aspx
The Digital Archaeological Record (tDAR)	Data repository	https://core.tdar.org/
EPD	Database	https://epdweblog.org/
FAUNMAP	Database	https://ucmp.berkeley.edu/faunmap/about/index.html
Human Relations Area Files (eHRAF)	Data repository	https://hraf.yale.edu/
Hungarian National Museum Archaeology Database	Data repository	https://archeodatabase.hnm.hu/en
Forschungsdatenarchiv der Archäologie und Altertumswissenschaften [Archive of research data from Archaeology and Ancient Sciences] (IANUS)	Database	http://datenportal.ianus-fdz.de/pages/collectionView.jsp?dipId=1650048
Isotopic Database for Archaeology (IsoArcH)	Database	https://isoarch.eu/
Latin American Pollen Database	Database	https://www.latinamericapollendb.com/
Mappa open data (MOD)	Data repository	http://mappaproject.arch.unipi.it/mod/Index.php
North American Non-Marine Ostracode Database Project (NANODe*)	Database	https://www.personal.kent.edu/~alisonjs/nanode/index.htm

(Continued)

Table A1: *Continued*

Name	Type	URL
Radiocarbon dates from Late Mesolithic/Early Neolithic transition in the Southern European Atlantic Coast (NeoNetAtl)	Database	https://digitallib.unipi.it/it/raccolta/The-NeoNetAtl-dataset/
Neotoma	Database	https://www.neotomadb.org/
North American Packrat Midden Database	Database	https://geochange.er.usgs.gov/midden/
ORAU database of 14C samples (ORAU)	Database	https://c14.arch.ox.ac.uk/database/db.php
Ostracode Metadatabase of Environmental and Geographical Attributes (OMEGA)	Database	https://www.gbif.org/dataset/a779af82-1422-4b00-9e7f-8e1c1f07bea2
The Paleobiology Database	Database	https://paleobiodb.org/#/
Pandora	Data repository	https://pandora.earth/
PANGAEA data publisher for earth and environmental science	Data repository	https://www.pangaea.de/
People 3000 Radiocarbon (p3k14C)	Database	https://www.p3k14c.org/
National Archaeological Repertory (RAN)	Database	http://ran.cimec.ro/sel.asp?lang=EN
Standard Cross-Cultural Sample (SCCS)	Database	https://hraf.yale.edu/resources/reference/sccs-cases-in-ehraf/#:~:text=The%20Standard%20Cross%2DCultural%20Sample,pinpointed%20in%20time%20and%20space
Strategic Environmental Archaeology (SEAD)	Database	https://www.sead.se/
Sheshat Global History Databank	Database	https://seshatdatabank.info/data/
Svenskt Hällristnings Forsknings Arkiv [Swedish Rock Art Research Archive] (SHFA)	Database	https://shfa.dh.gu.se/
THANADOS The Anthropological and Archaeological Database of Sepultures	Database	https://thanados.net/
XRONOS – Open Chronometric data for archaeology	Database	https://xronos.ch/

*Now discontinued (as of 2021).

Appendix 2 Checklist for Publishing to BIAD Standards

- Are different site names/spellings mentioned clearly?
- Are the coordinates of the site provided in as detailed a manner as possible alongside the name of the coordinate system used?
- Is any cultural identification accompanied by a description of the criteria used to define this culture?
- Are raw data published alongside original chronological period/phase?
- Are datasets as well as data in charts and graphs issued from previous publications referred to as such (e.g. with appropriate references)?
- Are data provided (e.g. in supplementary materials)?
- Are said data within an easily tabulated format (e.g. .csv file)?
- Have you reported on failed analyses and the reasons for failure?
- Are any calibrations explained?
- Are uncalibrated results disclosed?
- Is the exact material for each sample clearly detailed?
- Have finds been clearly quantified?
- Have osteological age/sex and taxa been recorded clearly and with vocabularies that are commonly applicable?