DOI 10.13131/UNIPI/2785-0668/A1CC-XT56

Philip Buckland*/§, Mattias Sjölander*, Johan von Boer**, Roger Mähler**, Johan Linderholm*

* Environmental Archaeology Lab, Department of Historical, Philosophical and Religious Studies, Umeå University; 90187 Umeå, Sweden.

** Humlab, Umeå University; 90187 Umeå, Sweden.

§ Corresponding author: philip.buckland@umu.se

THE INTRICATE DETAILS OF USING RESEARCH DATABASES AND REPOSITORIES FOR ENVIRONMENTAL ARCHAEOLOGY DATA

Abstract: Environmental archaeology is a complex mix of empirical analysis and qualitative interpretation. It is increasingly data science oriented, and databases and online resources are becoming increasingly important in large scale synthesis research on changes in climate, environments and human activities. Research funders, journals and universities place much emphasis on the use of data repositories to ensure transparency and reusability in the research process. Although these are important, researchers themselves, however, may have more use for research databases which are oriented more towards advanced querying and exploratory data analysis than conforming to archiving standards. This paper explores the pros and cons of these different approaches. It also discusses and problematizes some key concepts in research data management, including the definitions of data and metadata, along with the FAIR principles. Research examples are provided from a broad field of environmental archaeology and palaeoecology. In contrast to most publications, the developer's perspective is also included, and a worked example using the Strategic Environmental Archaeology Database (SEAD) to investigate fossil beetle data demonstrates the implementation of some of this in the real world. This example may be followed online using the SEAD browser, and all described data downloaded from there. After providing both encouragement and warnings on the use of digital resources for synthesis research, some suggestions are made for moving forward.

Keywords: Palaeoecology, geoarchaeology, research data infrastructure, linking data and metadata, interdisciplinary research.

1. Introduction

Environmental archaeology interprets the past through the analysis of empirical evidence gathered from archaeological sites and the landscapes around them. To do this, an understanding of the ecological, ethnographic or otherwise implications of the constituent parts of the evidence is needed. For biological organisms (palaeoecology), this is often called modern reference data, and may include, among many other things, the known habitat and distribution of plants and insects, or the cultural use of domestic or wild animals. For stone tools, the equivalent could include the provenance of utilised rock, the ethnographic evidence or archaeological theories, based on wear patterns or co-occurrence of finds etc. used to interpret regional distributions of site types. Both the fossil and modern components of this research process can be understood in terms of the data, metadata (data about data) and information (= processed data) collated when interpreting a fossil assemblage (fig. 1). Prior to the 1990s (e.g. Sadler et al., 1992), an

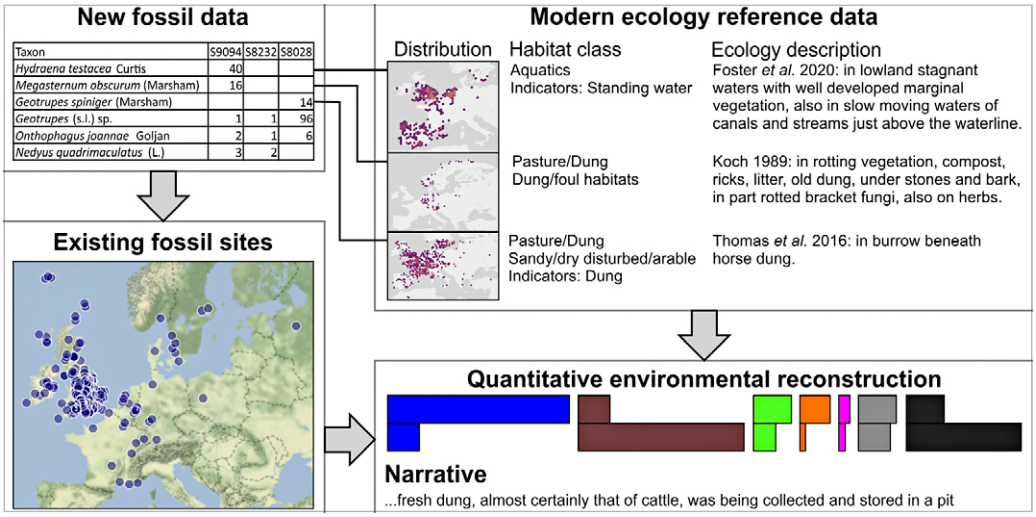

fig. 1. Environmental archaeology uses modern reference data, metadata and information to create quantitative and qualitative reconstructions of the past. New data contribute to the number of fossil sites which can help interpret future data. (Fossil list and narrative: Buckland & Buckland 2019; habitat class data: Buckland & Buckland 2006; maps captured from SEAD browser and GBIF; ecology descriptions: Foster et al. 2020, Koch 1989 and Thomas et al. 2016).

extensive literature review would have been the primary method for obtaining modern reference data, but the collation of massive amounts of this information into databases has considerably accelerated this task. Access to data on species distributions and ecology, and the facility to link these to other data on climate, vegetation and soils etc. has given the process of quantitative environmental reconstruction, where the data permit it, the potential to be considerably more systematic and transparent. It enables more empirically sound interpretations from which the essential narrative component of publications can be authored (Buckland et al., 2018a). The use of large amounts of open and linked data are driving the data science revolution in archaeology (Schmidt & Marwick, 2020), and helping environmental archaeology especially gain more prominence in debates on nature conservation, biodiversity, climate change and sustainability (Murphy & Fuller, 2017; Pilotto et al., 2021; Torben et al., 2020).

The study of any site with macrofossils involves an overview of other sites from which the newly identified species are already known. Previously, considerable library and archive skills would have been needed, and much of the archaeological 'grey literature' would only be discoverable through contacts in the museum and contract archaeology sectors. This task has also been made considerably easier through databases, open archives and data portals. These vary considerably in purpose and features, ranging from providing access to reports and comprehensive data from individual sites in sustainable formats (e.g. text files; Richards et al., 2021), to online access to more limited data from every site from which a species is known (Buckland et al., 2018b; Williams et al., 2018). More recently, Application Programming Interfaces (API's) have started to emerge for providing computational access to such aggregated data through GIS and statistical packages (Uhen et al., 2021).

Access to these resources is provided with the unwritten expectation that any newly created data be deposited in an appropriate database. There is a strong positive feedback loop in the deposition of data into open access resources – science will move forward not only through the accumulation of knowledge, but by providing better background data with which to interpret new sites. Despite this, there is still considerable variability in data sharing practices (Tenopir et al., 2020), and although few areas of environmental archaeology have been subject to systematic overview, those that have appear to require considerable improvement (Lodwick, 2019).

2. Data sources, documentation and transparency

Environmental archaeology relies to a large degree on methods adapted or derived from other areas of science (Reitz & Shackley, 2012). It is a pragmatic field which readily adopts techniques developed in other domains to analyse landscapes, sites, materials or data. The core of the field is something that is continually evolving, as new methods are developed (e.g. spectroscopy, ancient DNA, neutron methods) or as new possibilities arise for investigating old or new materials. Some of the more well-known methods are archaeobotany, palynology, entomology, osteology, and elements of geoarchaeology such as soil chemistry and magnetic properties analysis. The standard methods for studying most fossil organisms create similar raw data: (integer) counts (or estimates) of 'taxa' (identifications at different taxonomic levels, such as genus and species) per sample. In some cases, the abundance of different elements (e.g. bones, plant parts) is recorded, along with information on the developmental stage, articulation or modification of these. Other methods may produce measurements on a continuous scale, such as phosphate amounts or degree of magnetic susceptibility, or be recorded as relative amounts, such as percentage organic content.

Data are entered into databases in a number of manners, including manually through a data entry interface, transcription from analogue recording sheets, or import or ingestion from files created by other software (e.g. Tilia; Grimm, 1993), measurement instruments (spectrophotometers, magnetic susceptibility meters, balances) or scanning devices (cameras, laser scanners). Unfortunately, information on these chains of data entry is often left unrecorded in databases, despite their potential usefulness in identifying the source of errors. The Neotoma (Williams et al., 2018) and SEAD (Buckland et al., 2018b) databases do, however, include the capacity for storing this information, and these and other more established databases (e.g. Arbodat; Kreuz & Schäfer, 2002) will have gone through several phases of data entry mode in their lifetimes. The term 'paradata' is becoming more commonly adopted for describing such data provenance information (Kansa et al., 2020), although for many it is still considered part of the essential metadata describing the analysis process.

Whilst some metadata may be embedded in the output of analysis equipment or cameras, most are entered manually into databases, either directly upon inception or transcribed from analogue records such as sample bags or context sheets. To ensure reliable, reproducible and inter-comparable analyses, any lab needs to keep track of these data and metadata from when a sample enters its regime, to when the results are published and data archived (fig. 2). (Ideally the sample should be submitted to the lab with full metadata on sampling methods, purpose etc.) Doing so also helps ensure the transparency and future reuse of the results, as well as enabling the auditing of sample stores and projects. This is easier said than done, and it is not uncommon for the full details of analysis chains to get lost or confused towards the end of larger projects, when most database entry is undertaken. The digitalization of research and analysis processes (e.g. digital recording and sample processing logs) has helped improve this situation somewhat, but archaeology has much to learn from the growing use of electronic lab notebooks in the biological and physical sciences (see Kanza et al., 2017 for an overview). There has nevertheless been a move from a situation where even different analyses of the same sample could be difficult to relate to each other, to a more collected and consistent recording of information. However, even with digital help, the reliability of any information system is dependent on the individuals involved, and the consistency of their workflows. These workflows most often break down under stress and high workloads, and thus large projects are not always the best recorded. Such details, which could potentially be useful in evaluating the reliability of data during reuse, are rarely included in metadata, paradata or publications, and most often relegated to excavation diaries and conference anecdotes.

Methodological traditions may create gaps in information chains that could be considered unacceptable in other domains. For example, fossil insect parts are recorded individually during the identification process but then summarised as Minimum Numbers of Individuals (MNI)

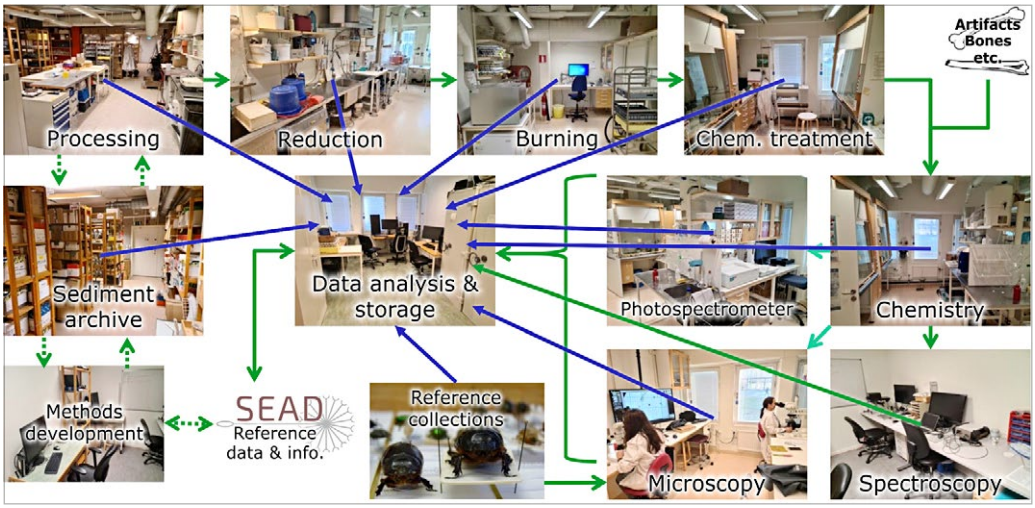

fig. 2. A generalised model of data and metadata sources in an environmental archaeology lab. Green arrows: flow of materials and information in the analysis process; blue arrows: the storage of data and metadata. Field collection data and metadata are missing here, but essential for the interpretation of the data. The model is based on, and includes pictures from, the Environmental Archaeology Lab in Umeå, Sweden.

for interpretation and publication (so 5 heads, 8 left elytra and 10 right elytra of one species become 10 MNI). The recording sheets with the raw data are at best filed in paper archives and almost never published. Praxis and experience dictates that there is no systematic bias in this method, but without the raw data there is no actual way of testing this (although Neotoma does include some North American insect data at this level). Similarly, the raw pre and post treatment weights of samples used to calculate organic content are rarely published. Even if the raw data are stored on lab servers, the more useful derived percentage values are more often archived and made available through public databases. These examples serve to illustrate potential transparency issues where the reputation of the lab or individual undertaking the analyses may factor into a user's evaluation of research results or the reliability of data for reuse. Systematic standardization and certification of labs is rare in environmental archaeology, and mostly limited to larger facilities for radiocarbon dating or where services such as environmental chemistry require them. Periodic workshops to compare results and interpretations are a pragmatic alternative, but these also require time, funding and consensus across research groups. The inclusion of more information on methods and processes in databases, and not just publications, as paradata, would most likely be beneficial for the transparency of environmental archaeology in the long term.

It is now widely acknowledged that to ensure research transparency, both data and metadata must be made available along with published results (Marwick et al., 2017). This has been recognized not only within the research community (Tenopir et al., 2020), but also by policy makers and funding agencies (Fecher et al., 2015). In the long term, research transparency relies on two major components: 1) the openness and sustainability of data, and 2) the openness and sustainability of information on the analysis methods applied to the data. The latter includes not only the methods applied in the field or laboratory, but also the software and code used to process and analyse the data. Just as access to raw data is meaningless without the contextual information on its production, the published results are almost meaningless without the methods used to analyse the data (Marwick et al., 2017). This is especially the case when datasets are to be aggregated or compared between multiple sites (see Heginbotham et al., 2010, for an illustration with XRF data).

3. Data and/or metadata?

The boundary between data and metadata can become fuzzy in environmental archaeology. Habitat descriptions for organisms may be considered data to an ecologist (or biodiversity database), but metadata to an archaeologist (or archaeology database) when using them to interpret fossil assemblages. Whilst the FAIR principles were designed to be discipline independent (https://www.go-fair.org/fair-principles/; Wilkinson et al., 2016), a number of them may be somewhat confusing to implement in *interdisciplinary* research. This includes the use of "domain-relevant community standards" (R1.3) when reconstructing past landscapes, where reference data and metadata are collated from multiple domains then subject to a variety of statistical and summarization methods. For environmental archaeology as a field, cross-references would be needed at the very least between several ontologies defining standard terms for habitats, landscapes (modern and past), ecology etc, and when using data from different proxies. There is, for example, a considerable variety of modern landscape classification systems available (Simensen et al. 2018), and their archaeological relevance varies between geographical regions, site types and chronological periods. In the case of fossil insects, standardised habitat definitions for modern landscapes may not be entirely relevant for interpreting data on prehistoric landscapes, and their use must proceed with caution, especially in terms of differentiating between anthropocentric definitions and the immediate conditions the insects require (Pilotto et al., 2021).

While community standards are still developing, such issues may delay the uptake of FAIR, and have practical implications for the implementation and evaluation of repositories (see below). They are, however, potentially solvable when proxy specific database communities (e.g. Arbodat) interact with large communities such as ARIADNE, Neotoma and PAGES (Kohler et al. 2018). Indeed, recent initiatives to provide empirically derived, archaeologically useful, standardized classifications for land use (Morrison et al. 2021) and transparent definitions of chronological and cultural periods (Rabinowitz et al. 2016) represent significant progress. Nevertheless, users of these systems should be reminded that they are tools for interrogating and understanding the past, rather than devices which provide exact answers. This applies as much to an archaeologist collating regional vegetation reconstructions as a palaeoecologist assessing continental scale human impact.

4. A question of scale and aggregation

Different types of research question require different levels of data aggregation (fig. 3). For macrofossils, the lowest common denominator in a database is often a value denoting the number of fossils of a particular taxon found in a sample. Collectively, all the finds in a sample make up an assemblage that is considered representative of the sampled layer or excavation context. Dates and interpretations are usually convergent on the context level, and this is thus often the highest resolution at which inter-site comparisons can be undertaken. For regional or global studies, the sample data are usually binned (c.f. aggregated) into fixed time slices (e.g. 500 years) according to some principle on how to relate sample dates to the bins (e.g. intersection, partial overlap, radiocarbon probabilities; see Carleton and Groucutt, 2021, for an exposé of the latter). Data aggregated in this way are increasingly used to study past biodiversity (Barnosky et al., 2017), with recent examples including plants (Giesecke et al., 2019) and mammals (Andermann et al., 2020). Other studies have promoted the use of the fossil record for understanding modern biodiversity, particularly in terms of conservation in different environments (Kiessling et al., 2019) and in connection with threatened species (Pilotto et al., 2021). In biodiversity investigations, the objects of study are the species themselves, whereas for the large part of research in environmental archaeology, the species are proxies for other information. Through the use of modern reference data, species names are substituted by their environmental or climatic implications (e.g. temperature tolerances,

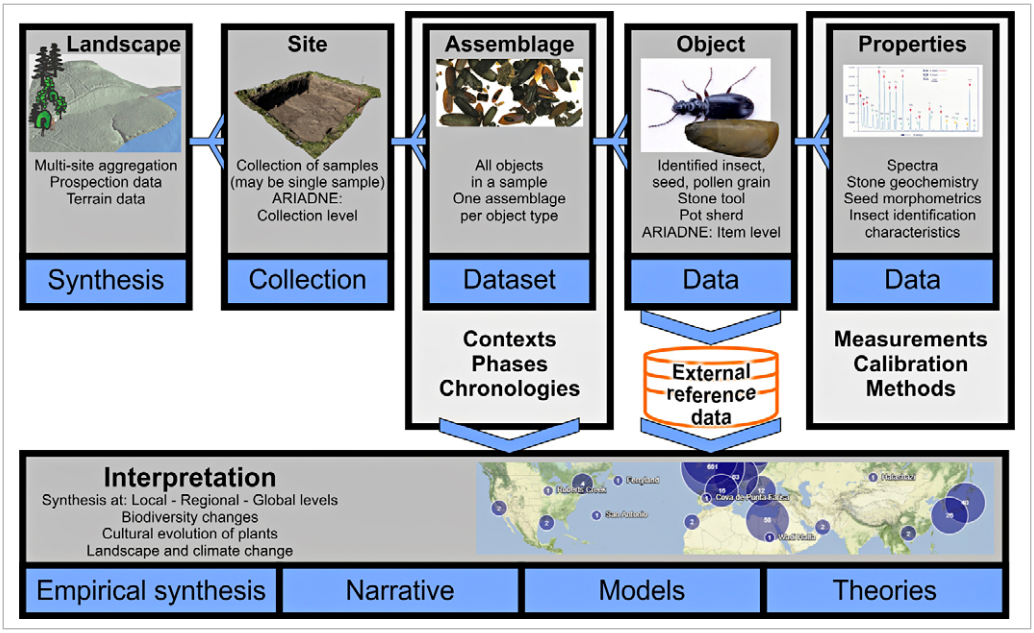

fig. 3. Data and aggregation levels, with increasing resolution from left to right. Light grey boxes show important metadata components at levels most relevant in an environmental archaeology database (other metadata is obtained from excavation databases and national site registers). Other 'Properties' could include ancient DNA, radiocarbon dates etc. (3D images a courtesy of Nicolo Dell'Unto and https://www.darklab.lu.se/digital-collections).

habitats) and summarised (quantitatively or qualitatively; fig. 1) to build models of changes in climate and landscape over space and time.

In artefact terms, a parallel could be the use of different stone tool types as proxies for understanding spatio-temporal changes in prehistoric cultures. Manninen et al. (2021) use an array of Mesolithic stone tool data to demonstrate cultural complexity not visible in ancient DNA (aDNA) results. The latter are, as yet, not linked to archaeological databases in any useful way, and it is thus very difficult to directly compare aDNA based results with any other form of archaeological evidence over wider areas. From a database point of view, aDNA could be considered as properties of the studied objects (e.g. bones, insects, although extracting aDNA from Quaternary fossil insects is proving difficult; Simpson et al., 2020). Whether the properties of an object are important will largely depend on the research questions – for example, the material properties of stone tools are essential in provenance studies (Sciuto et al., 2019), and the relative size of different seeds are used in crop diversification and cultivation studies (Karg, 2018). Manninen et al. (2021) cite a lack of integration of aDNA based research with more traditional archaeological methods and interpretations. The qualitative contextual aspects of the studied object are always essential for the interpretation of any empirical data. They thus need to be carefully considered and either included in or linked to environmental archaeology databases.

In practice, it is often a combination of analysis methods, reference data and aggregation levels that is needed to answer more complex, spatio-temporal research questions. Changes in the distribution or biodiversity of one organism type may act as a proxy or corroboration for the interpretation of another (e.g. Woodbridge et al., 2021; Schweiger & Svenning 2018). Multi-proxy data integration is not a simple process, and the tendency to ignore the intricacies, and variable reliability of dating evidence, can lead to unreliable conclusions (Price et al., 2018). With this perspective in mind, an example follows of the use of the SEAD online browser in an interdisciplinary context.

4.1 Facetted browsing and the advantage of item level data

The past presence of grazing animals in the landscape can be tracked, to an extent, by using indicators of their dung as a proxy. Dung beetles have been used in this respect (Schweiger & Svenning, 2018), as well as for helping to assess evidence for the impact of large herbivores in mid Holocene woodlands (Whitehouse & Smith, 2010). The genus *Aphodius* (s.l.) is a common taxonomic group of dung beetles, and thus a reasonable starting point for investigating such questions. The SEAD online browser allows the user to easily find all sites (datasets) which contain any species of the genus *Aphodius* (fig. 4). Selecting the genus in the appropriate filter sends a query to the database via an API which utilizes the database's relational structure. As the minimum common denominator in SEAD is the number of individuals of each species found in each sample, the system can easily retrieve and display points for all sites with samples containing *Aphodius* fossils.

SEAD has inherited ecological reference data from the BugsCEP database (Buckland & Buckland, 2006) for over 5000 species of insect, collated and updated over more than 30 years, by several individuals (Buckland & Sjölander, *in press*). These data can also be accessed in the SEAD browser, and the filters fulfil both visualization and querying functions, allowing for basic exploratory data analysis. In fig. 4, the 'Eco code' filter shows the variety of habitats preferred by all of the species in the genus *Aphodius* that have been found fossil. Of the results shown, only about 15% of the samples include species that can be considered as reliable indicators of dung, the genus *Aphodius* including many species which are also indicative of other environments.

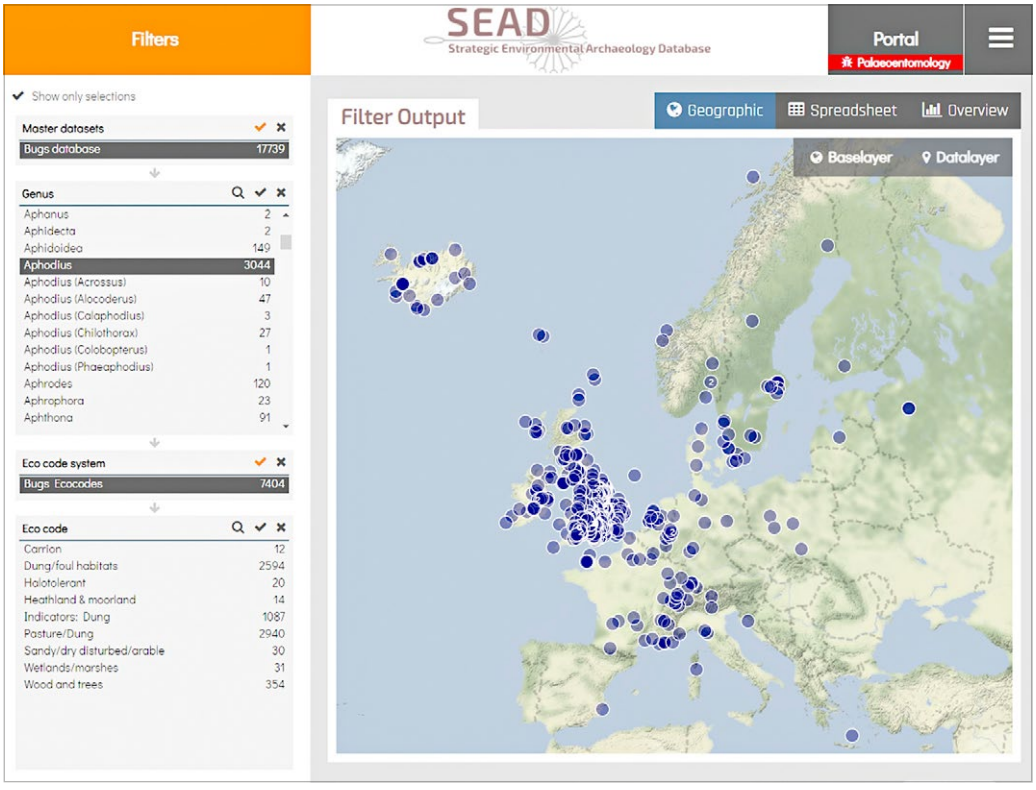

fig. 4. The SEAD online browser showing sites where species of the beetle genus *Aphodius* have been found in the Bugs virtual constituent database. The map updates as items are selected in the filters on the left. Filters are shown for 'Genus', 'Eco code system' and 'Eco code', but others can be activated to influence the search results. https://browser.sead.se/.

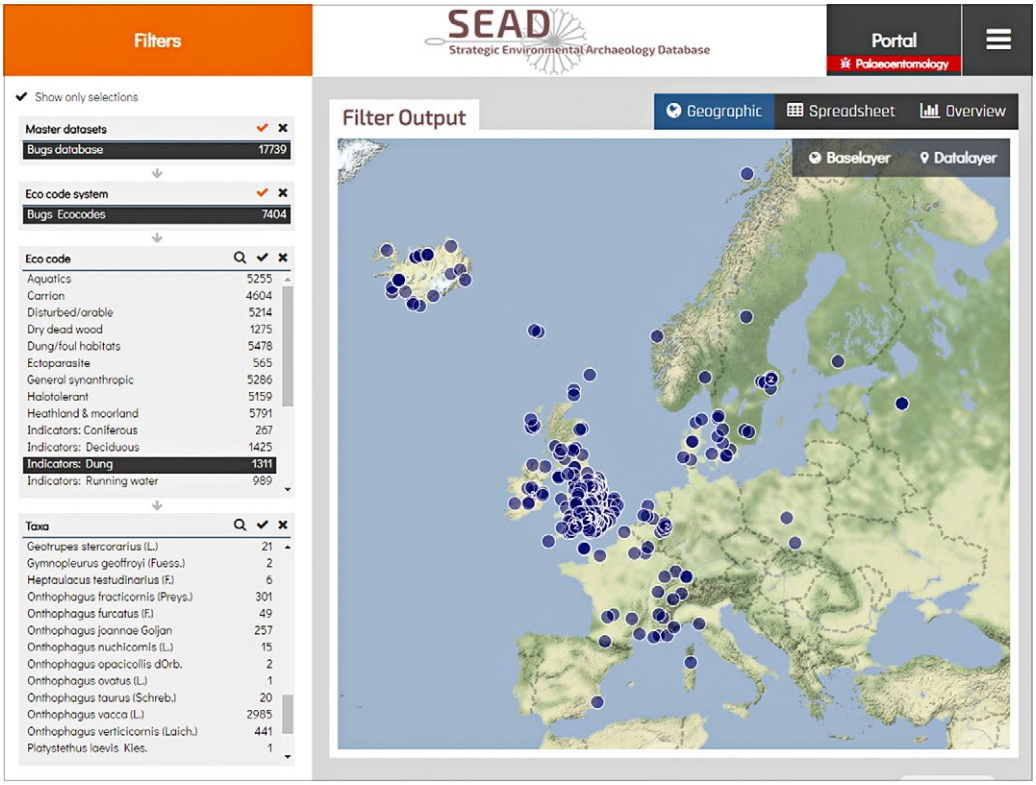

fig. 5. The SEAD online browser showing only sites which have samples that include beetle species which are definite indicators of dung. Note the new filters and selections relative to fig. 4. https://browser.sead.se/.

The query clearly needs to be narrowed if these other landscape elements are of no interest. SEAD's facetted browser system allows filters to be rearranged, with the upper ones filtering those below. Removing the Genus filter, selecting 'Indicators: Dung' as 'Eco code' and adding a 'Taxa' filter both narrows the search to only explicit (obligate) indications of dung, but also expands the range of species retrieved by including other genera than *Aphodius* (fig. 5). There are now 1311 samples in the results set, and SEAD's API will allow these data to be exported for further analysis in the users own software. The landing pages for individual sites, which have permanent web addresses, can also be viewed by clicking on the map points.

Deep diving into the individual datasets selected by this query reveals considerable variation in their size and metadata completeness. The BugsCEP database always contains information on site location and type, along with the type of data (partial or full quantification/presence only) and source publication references. Other references for reinterpretations or reuse of the data are also included, and it is sometimes difficult to identify the original data source. Metadata on sampling methods is intermixed with site descriptions, and not always present, sometimes being alluded to in the name of datasets (e.g. "bulk samples from..."). At the moment, dating information is poorly exposed in the SEAD browser, but inspection of the downloadable version of BugsCEP (https://www.bugscep.com/downloads.html) shows that the usual messy nature of archaeological dating is reflected in the insect datasets, with many types of dating and many samples only indirectly dated. Although the name of the entomologist identifying the specimens is almost always present (but not currently shown in the SEAD browser), the quality and extent of general information on the interpretation of the site varies considerably. The latter is very useful for further narrowing down the selection of

relevant sites, or assessing the relevance or reliability of the data, but requires time consuming browsing. Such text-based data, including the ecological descriptions of species beyond simple habitat classifications, are important for understanding the true implications of any synthesis research, but are often neglected in big data (macroecological) analyses (see for example Price et al., 2018, and Alexander's 2016 criticism of Sandom et al., 2014).

The fossil data in the BugsCEP dataset have been continually added to, through a variety of desktop software interfaces, since its inception in the late 1980s (Buckland, 2007; Sadler et al., 1992). Most of the datasets in BugsCEP can be generally described as having been collected for fossil insect analysis (either for archaeology or Quaternary science) by the circular fact that they are in a fossil insect database. Further details on the true project aims are only available in publications, grant applications or project descriptions, and for much of contract archaeology these may be inaccessible or lost. It is entirely possible that BugsCEP contains datasets that were collected for other purposes – e.g. a hydrology study of a peat bog, where insect fragments were noticed and then further analyses undertaken, or a curious find in a museum collection. In most cases this will have little meaning, but the difference between research projects and consultancy projects with more limited budgets may be reflected in limited identification and/or quantification in the latter. This is an important detail for any study where different taxonomic levels (genus, species etc.) are to be quantified (as, for example, a biodiversity measure). Such details become even more important, and problematic, when linking between databases, especially as compromises must usually be made to ensure compatibility between systems (e.g. Uhen et al., 2021). The Swedish Biodiversity Data Infrastructure (SBDI; https://biodiversitydata.se/) links data from SEAD to contemporary biodiversity data sources, including the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/), using the scientific names of species and other taxa. Due to the variability in the resolution of fossil identifications, only about 75% of the names in SEAD have direct equivalents in the Swedish species names database (Dyntaxa) or GBIF. A study of taxonomic biodiversity in the fossil record could therefore miss a quarter of the available data if it was initiated from the GBIF side of the connection rather than the SEAD side.

5. The developer's perspective

A flexible database design allows for more flexibility when designing for user interaction. Users need not be restricted by predetermined categories of metadata, and can be allowed to create queries based on inventive combinations of metadata and analysis results (raw data). However, with the exception of those skilled in working with APIs and backend databases, users will still need to interact with an interface designed with a specific purpose in mind, and thus limited by the filters and search tools provided. It is up to the database provider to weigh up the costs and benefits of what can be implemented online, or in desktop software, within the available project budget. This often entails limiting the search functions and exposed data to the minimum required by the main target user group. Research databases also have a tendency to focus on the data at the expense of detailed information on the analysis methods, which are rarely digitised in detail, and most often only available in the papers providing the data. This is not a problem unique to environmental archaeology, however, and even the field of artificial intelligence appears to have a significant lack of transparency in terms of published code (Hutson, 2018).

Designing a user interface along the lines of SEAD's browser requires developers to obtain knowledge of the relevant research domains, and insights into the perspectives of the potential users. Most research fields have established community standards for software tools and visuals, some of which may be complex or even archaic (such as the tendency for palynologists to hang on to older versions of the Tilia software; Grimm, 1993). These might contrast with current interface design trends (such as responsive, multi-device websites, information minimalism, smart content loading and videos) and the developer's ideas on designing an

intuitive user interface that is easy to understand and use for a layman. Here it is vital to understand that an expert interface for domain scientists is not necessarily compatible with one for the general public. Understanding how research workflows may differ from those of a layman or non-expert, when navigating the same data, is essential. Similarly, the way in which various data formats are used in different circumstances must be considered (e.g. the toolchain process from data creation and digitization, exploratory data analysis, data export and subsequent import into a desktop statistical package). For example, the majority of research users are more familiar with the Microsoft Excel format than GeoJSON, even though the latter is probably easier to work with in GIS software.

Expanding the domain knowledge of developers to take these things into account should be seen as an integrated part of any research database development project. This requires a considerable investment in time and funding, as well as what for the research PI may be unusual forms of tasks prioritization and team management. It may be as difficult for a researcher to prioritize programming tasks as for a developer to prioritize different analysis methods. Similarly, the nuances of distinct systems development skills may be as opaque to the researcher as the differences between apparently similar research domains are to the developer. A mutual dialogue and flexibility is thus essential for matching priorities with competences and expected outcomes.

In this respect, in-house development teams at universities and research institutes may have a considerable advantage over external contractors when developing research databases. Researchers tend to ascribe to the "build it once and never change anything" design philosophy, which is somewhat at odds with the current trend towards continual updates and software as a service. As with any user group, researchers are often interested in, and influenced by, both the aesthetics and performance of the tools they use. Progress indicators, for example, are essential for preventing frustration and the repeated pressing of keys which may impact performance. In contrast to commercial software, performance tends to take a backseat to functionality and reliability in a research context, and time spent on fixing bugs or adding features may be more appreciated than more efficient code or query optimisation. The adage 'Hardware is Cheap, Programmers are Expensive' is highly relevant even in the context of research databases, where resources are limited. However program code should be thoroughly documented through inline comments and manuals, which should be made as transparent and understandable as the data being stored. Doing so will help ensure that subsequent development teams can save time by improving on, and possibly reusing, parts of any software developed in a research context.

6. Data repositories and active research databases

The plurality of evidence, data types and research orientations in environmental archaeology requires a flexible approach to the storage and provision of data. This is reflected in the different approaches observable in data infrastructures which can handle the subject's data. A general division can be made between repositories, designed with long-term sustainability and making data and metadata findable and accessible in mind, and research databases designed for continued research (tab. 1). Whilst both types of system may facilitate the reuse of data outside of its original collection purpose, they tend to embrace different goals and principles. Where a repository will view the deposited research as being "finalized", research databases are often designed to enable further research, without necessarily archiving the data for future reuse in its original form. Most repositories, such as the Swedish National Data Centre (snd.gu.se), PANGAEA (www.pangaea.de) or the Arctic Data Center (arcticdata.io), store, or make available, data as individual files (e.g. xls/csv) with accompanying higher level metadata (e.g. txt, xml) that includes critical information on site locations, analysis method, researchers, etc. Permanent links to datasets or sites are essential for the transparency and reproducibility of research results. This is easier with static files than a continually updated

	Repositories	Research data infrastructures
System structure	• File based (hierarchical files, folders) • Generic context • Generic (fits many domains)	• Database based (often relational) • Domain specific context • Often has similarities with structure of research data levels (fig. 3)
Visible data format (user perspective)	• Defined by user according to repository guidelines • Often understandable by users	• Defined by system • Unintelligible to most users
Data	• Single files, references between files but not relationships • Frozen snapshot, "packaged" data • Downloadable or offline (data must be prepared for user)	• Distributed through related tables • May be versioned & updated • Online (data are ready to use)
Export data formats	• Same as input	• Usually domain specific but flexible
Metadata	• Single files describing data • Generic description (often files)	• Distributed through related tables
Research use	• Simple re-use or re-assessment of single or limited numbers of datasets • Reuse can be limited based on data provenance	• Complex, multi-site synthesis or analyses of many datasets (but single dataset approaches also possible) • Enables more advanced re-use based on affordances given by infrastructure
Primary aims	• Advancement of science through data re-use • Long term preservation, one project/file/dataset at a time	• Advancement of science through advanced analyses • High availability • Multi-dataset
FAIR Findability	• Usually good, but sometimes let down by poor search facilities	• Variable, but good if online systems can expose lower level metadata or data
FAIR Accessibility	• Good for single datasets (one at a time)	• Good for single or multiple datasets (cross-analysis)
FAIR Interoperability	• Generally not good from a research perspective	• Better, although the FAIR definition may not be entirely useful for practical purposes
FAIR Reusability	• Generally good for individual datasets but susceptible to file content problems	• Generally good for individual datasets and multi-dataset analyses
Documentation requirements	• Strict	• Often minimal
Adherence to data and metadata standards	• Often required	• Variable, sometimes problematic to implement
Ease of access	• Often poor websites for research use • Websites usually well maintained	• Often good user interfaces, either online or as desktop software • Desktop software can become out of date and unusable
Ease of data submission	• Variable but generally well developed and supported data entry systems • General requirements dictate file or data formats – reformatting often required OR General requirements allow for a diversity of data formats • Custom ingestion of larger datasets requires more resources	• Highly variable and often dependent on key individuals or data stewards • Proxy specific traditions guide data entry formats – data entry simple for experts in field • Ad-hoc/custom ingestion of larger datasets common
Data entry flexibility	• Metadata requirements often strict and ontology guided • Data/metadata fidelity sometimes lost when matching to lowest common denominator	• Metadata requirements often flexible (e.g. not linked to standards or ontologies) • Data fidelity usually retained, emphasis on preserving raw data
Research power	• Limited due to lack of relevant links between datasets (e.g. species, methods)	• Good
Aggregation tools	• Most often not present or limited to file collections after geographical search	• Variable, but often accessible through the database backend (if permitted) or API
Origins	• Usually national or institutional initiatives (but there are many domain orientated repositories)	• Usually researcher initiated and driven
Sustainability	• Good. Often backed by large organisations	• Variable. Generally smaller grants and rely on the enthusiasm of creators and users

tab. 1. Some generalised differences between research archives and databases from a multiproxy perspective. The list is not comprehensive and there is much variation. FAIRness comments relate to definitions at https://www.go-fair.org/fair-principles/.

database, where versioning must be maintained so as to expose the history of changes to any dataset. As new research results are added to a database or repository, the same query will provide different results at different points in time. Effective citation methods are thus also essential, and ideally tools for re-asking old questions on both new and old datasets. For full scientific reproducibility, the actual queries (which are in essence parts of the research methods and can be equated with the code used in statistical analyses) should be provided with permanent links as well. The inclusion of code and queries in repositories like GitHub (github.com) to accompany publications is an important step forward in this respect.

In repositories, datasets can often be found on the basis of predetermined categories (or ontologies) of metadata or keywords, but the data themselves are generally buried within files and not exposed to the search engine. Repositories often cover a wider range of research areas and data than research databases, and there is considerable diversity between and within repositories as to how a dataset is defined (site, study of many sites, data from three different seasons of sampling, etc.). Linking between datasets in a repository or archive is rare, and selection or aggregation of data from multiple datasets is generally only possible at site or higher metadata levels, if at all. This makes multisite, multi-regional analysis and synthesis, as exemplified above using SEAD, difficult as it requires the interoperability of data at the object or item level (fig. 3).

Where repositories need to be somewhat simplified in their (visible) structure in order to accommodate a wider user base, or adhere to certification standards, research orientated data infrastructures may have the freedom to be more flexible (Buckland & Sjölander, *in press*). The underlying database is often designed to facilitate complex and new research angles, with an emphasis on being used as a 'research tool' rather than a 'data discovery tool'. To enable queries on data at a high resolution (fig. 3), it is necessary to store the data (and metadata) in the database itself, rather than in the single files required by most archives. The FAIR principles appear to have been designed with the latter in mind ("The metadata and the dataset they describe are usually separate files"; https://www.go-fair.org/fair-principles/f3-metadata-clearly-explicitly-include-identifier-data-describe/), along with simple scientific domains, and not interdisciplinary research. FAIR principle F3 ("Metadata clearly and explicitly include the identifier of the data they describe") is particularly problematic, and arguably unnecessary, in a relational database where the separation of metadata and data into 'files' is not a relevant concept. For such a system to obtain a good FAIR score, an API and landing pages must be created which de-normalise the data into a format that FAIR evaluation tools can understand. This may lead to resources which could be considered as more usefully spent on improving data quality or interface functionality, being diverted to adapting API output to pass FAIR evaluations. However, increasing a system's FAIR evaluation score should not be underestimated in terms of its benefits for the findability of data and securing sustainable infrastructure funding.

7 Conclusions

There is no single optimal solution model for a research data infrastructure, and environmental archaeology, as an interdisciplinary science, provides extra challenges over more traditional domain orientated systems. The very act of making data openly available may contribute to scientific progress, but there are a number of ways in which this can be made more useful than simply by making data available for reuse. Tools for searching and aggregating data, not just metadata, as well as enabling the same through (external) reference data (ecology, cultural use etc.) vastly improve the usage potential for any database. The increasing prevalence and importance of synthesis research (Altschul et al. 2017) is not only a testament to this, but also a proof of concept for even more interdisciplinary use of archaeological data in the future. To ensure this, we need to enable a continual dialogue between researchers, archivists and developers and ensure the mutual transference of knowledge, ideas and creativity between

these groups. Finally, the FAIR principles were born into, or created by, a linked digital world, and several of the principles naturally assume that data exist in an online environment. It may be useful to reflect over the fact that digital metadata is older than the World Wide Web, and that different types of metadata are important for different types of use (see table in Strebel, et al., 1994, p. 43). Archaeologists are inherently aware of the past, and still tend towards the use of desktop, offline systems for most of their analysis.

List of references

Alexander, K. N. A. (2016). Europe's wood pastures-rich in saproxylics but threatened by ill-conceived EU instruments. *Bulletin de la Société belge d'Entomologie*, 152, 168-173.

Altschul, J. H., Kintigh, K. W., Klein, T. H., Doelle, W. H., Hays-Gilpin, K. A., Herr, S. A., ... & Sabloff, J. A. (2017). Opinion: Fostering synthesis in archaeology to advance science and benefit society. *Proceedings of the National Academy of Sciences, 114*(42), 10999-11002.

Andermann, T., Faurby, S., Turvey, S. T., Antonelli, A. & Silvestro, D. (2020). The past and future human impact on mammalian diversity. *Science advances, 6*(36), eabb2313.

Barnosky, A.D., Hadly, E.A., Gonzalez, P., Head, J., Polly, P.D., Lawing, A.M., Eronen, J.T., Ackerly, D.D., Alex, K., Biber, E., Blois, J., Brashares, J., Ceballos, G., Davis, E., Dietl, G.P., Dirzo, R., Doremus, H., Fortelius, M., Greene, H.W., ... Zhang, Z. (2017). Merging paleobiology with conservation biology to guide the future of terrestrial ecosystems. *Science 355*. https://doi.org/10.1126/science.aah4787

Buckland, P. (2007). The development and implementation of software for palaeoenvironmental and palaeoclimatological research: the Bugs Coleopteran Ecology Package (BugsCEP). Doctoral dissertation, Arkeologi och samiska studier, Umeå University. *Archaeology and Environment*, *23*. http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-1105

Buckland, P.I. & Buckland, P.C. (2006). *BugsCEP Coleopteran Ecology Package*. IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series # 2006-116. NOAA/NCDC Paleoclimatology Program, Boulder CO, USA. URL:http://www.ncdc.noaa.gov/paleo/insect.html or http://www.bugscep.com

Buckland, P.I., Palsson, G. & Dell'Unto, N. (2018a). To tree, or not to tree? On the Empirical Basis for Having Past Landscapes to Experience. *Digital Humanities Quarterly 12(3)*. http://www.digitalhumanities.org/dhq/vol/12/3/000383/000383.html

Buckland, P. I. & Buckland, P. C. (2019). When a Waterhole is Full of Dung: An Illustration of the Importance of Environmental Evidence for Refining Archaeological Interpretation of Excavated Features. *Archaeometry, 61*(4), 977-990. https://doi.org/10.1111/arcm.12461

Buckland, P. I. & Sjölander, M. (*in press*). Approaches to Research Data Infrastructure for Archaeological Science. In Watrall, E. & Goldstein, L. (Eds.) *Digital Heritage and Archaeology in Practice*. University Press Florida.

Buckland P.I., Sjölander M., Eriksson E.J. (2018b) Strategic Environmental Archaeology Database (SEAD). In: Smith C. (eds) *Encyclopedia of Global Archaeology*. Springer, Cham. https://doi.org/10.1007/978-3-319-51726-1_833-2

Carleton, W. C., & Groucutt, H. S. (2021). Sum things are not what they seem: Problems with point-wise interpretations and quantitative analyses of proxies based on aggregated radiocarbon dates. *The Holocene*, *31*(4), 630-643.

Fecher B., Friesike, S. & Hebing, M. (2015). What Drives Academic Data Sharing? *PLoS ONE 10*(2), e0118053. https://doi.org/10.1371/journal.pone.0118053

Foster, G. N., Bilton, D. T., Hammond, M. & Nelson, B. H. (2020). *Atlas of water beetles of Britain and Ireland – smaller families of Polyphaga*. FSC Publications, Telford.

Giesecke, T., Wolters, S., van Leeuwen, J. F., van der Knaap, P. W., Leydet, M., & Brewer, S. (2019). Postglacial change of the floristic diversity gradient in Europe. *Nature communications*, *10*(1), 1-7. https://doi.org/10.1038/s41467-019-13233-y

Grimm, E.C., 1993. *TILIA v2.0 (computer software)*. Illinois State Museum, Research and Collections Center, Springfield, IL.

Heginbotham, A., Bezur, A., Bouchard, M., Davis, J. M., Eremin, K., Frantz, J. H., Glinsman, L., Hayek, L-A. C., Hook, D., Kantarelou, V., Karydas, A. G., Lee, L., Mass, J., Matsen, C., McCarthy, B., McGath, M., Shugar, A., Sirois, J., Smith, D., & Speakman, R. J. (2010). An evaluation of inter-laboratory reproducibility for quantitative XRF of historic copper alloys. In *Metal 2010: International Conference on Metal Conservation, Interim Meeting of the International Council of Museums Committee for Conservation Metal Working Group, October 11-15, 2010*, Charleston, South Carolina, USA. Clemson University.

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science, 359*(6377), 725-726. https://doi.org/10.1126/science.359.6377.725

Kansa, S. W., Atici, L., Kansa, E. C., & Meadow, R. H. (2020). Archaeological analysis in the information age: Guidelines for maximizing the reach, comprehensiveness, and longevity of data. *Advances in Archaeological Practice*, *8*(1), 40-52. https://doi.org/10.1017/aap.2019.36

Kanza, S., Willoughby, C., Gibbins, N., Whitby, R., Frey, J. G., Erjavec, J., Zupančič, K. Hren, M. & Kovač, K. (2017). Electronic lab notebooks: can they replace paper? *Journal of cheminformatics*, *9*(1), 1-15. https://doi.org/10.1186/s13321-017-0221-3

Karg, S. (2018). New research on the cultural history of the useful plant *Linum usitatissimum* L. (flax), a resource for food and textiles for 8,000 years. *Veget Hist Archaeobot, 20***,** 507-508. https://doi.org/10.1007/s00334-011-0326-y

Kiessling, W., Raja, N. B., Roden, V. J., Turvey, S. T., & Saupe, E. E. (2019). Addressing priority questions of conservation science with palaeontological data. *Philosophical Transactions of the Royal Society B*, *374*(1788), https://doi.org/10.1098/rstb.2019.0222

Koch, K. (1989). Die Käfer Mitteleuropas. Ökologie, 1. Goecke & Evers, Krefeld.

Kohler, T. A., Buckland, P. I., Kintigh, K. W., Bocinsky, R. K., Brin, A., Gillreath-Brown, A., Ludäscher, B., McPhillips, T.M., Opitz, R., & Terstriep, J. (2018). Paleodata for and from archaeology. PAGES Magazine, 26(2), 68-69. https://doi.org/10.22498/pages.26.2.68

Kreuz, A. & Schäfer, E. (2002). A new archaeobotanical database programme. *Vegetation History and Archaeobotany* 11, 2002, 177-179.

Lodwick, L. (2019). Sowing the Seeds of Future Research: Data Sharing, Citation and Reuse in Archaeobotany. *Open Quaternary, 5*(7), 1-15. https://doi.org/10.5334/oq.62

Manninen, M. A., Damlien, H., Kleppe, J. I., Knutsson, K., Murashkin, A., Niemi, A. R., Rosenvinge, C. S. & Persson, P. (2021). First encounters in the north: cultural diversity and gene flow in Early Mesolithic Scandinavia. *Antiquity*, *95*(380), 310-328. https://doi.org/10.15184/aqy.2020.252

Marwick, B., d'Alpoim Guedes, J., Barton, C. M., Bates, L. A., Baxter, M., Bevan, A., Bollwerk, E. A., Bocinsky, R. K., Brughmans, T., Carter, A. K., Conrad, C., Contreras, D. A., Costa, S., Crema, E. R., Daggett, A., Davies, B., Drake, B. L., Dye, T. S., France, P., ... Wren, C. D. (2017). Open science in archaeology. *SAA Archaeological Record, 17*(4), 8-14.

Morrison, K. D., Hammer, E., Boles, O., Madella, M., Whitehouse, N., Gaillard, M-J., Bates, J., Vander Linden, M., Merlo, S., Yao, A., Popova, L., Hill, A. C., Antolin, F., Bauer, A., Biagetti, S., Bishop, R. R., Buckland, P., Cruz, P., Dreslerová, D., ... Zanon, M. (2021). Mapping past human land use using archaeological data: A new classification for global land use synthesis and data harmonization. *PLOS ONE*, *16*(4). https://doi.org/10.1371/journal.pone.0246662

Murphy, C., & Fuller, D. Q. (2017). The future is long-term: Past and current directions in environmental archaeology. *General Anthropology*, *24*(1), 1-10. https://doi.org/10.1111/gena.12020

Pilotto, F., Dynesius, M., Lemdahl, G., Buckland, P. C. & Buckland, P. I. (2021). The European palaeoecological record of Swedish red-listed beetles. *Biological Conservation 260,* 109203. https://doi.org/10.1016/j.biocon.2021.109203

Price, G. J., Louys, J., Faith, J. T., Lorenzen, E., & Westaway, M. C. (2018). Big data little help in megafauna mysteries. *Nature, 558*(7708), 23-25. https://doi.org/10.1038/d41586-018-05330-7

Rabinowitz, A., Shaw, R., Buchanan, S., Golden, P., & Kansa, E. (2016). Making sense of the ways we make sense of the past: The PeriodO project. *Bulletin of the Institute of Classical Studies*, *59*(2), 42-55. https://doi.org/10.1111/j.2041-5370.2016.12037.x

Reitz, E., & Shackley, M. (2012). *Environmental archaeology.* Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-3339-2

Richards, J.D., Jakobsson, U., Novák, D., Štular, B. & Wright, H. (2021). Digital Archiving in Archaeology: The State of the Art. Introduction. *Internet Archaeology, 58*. https://doi.org/10.11141/ia.58.23

Simensen, T., Halvorsen, R., & Erikstad, L. (2018). Methods for landscape characterisation and mapping: A systematic review. *Land use policy*, *75*, 557-569. https://doi.org/10.1016/j.landusepol.2018.04.022

Sadler, J.P., Buckland, P.C., Rains, M. (1992). BUGS: an entomological database. *Antenna, 16*, 158-166.

Sandom, C. J., Ejrnæs, R., Hansen, M. D., & Svenning, J. C. (2014). High herbivore density associated with vegetation diversity in interglacial ecosystems. *PNAS*, *111*(11), 4162-4167. https://doi.org/10.1073/pnas.1311014111

Schmidt, S. C., & Marwick, B. (2020). Tool-driven revolutions in Archaeological Science. *Journal of Computer applications in archaeology*, *3*(1). http://doi.org/10.5334/jcaa.29

Schweiger, A. H., & Svenning, J. C. (2018). Down-sizing of dung beetle assemblages over the last 53 000 years is consistent with a dominant effect of megafauna losses. *Oikos*, *127*(9), 1243-1250. https://doi.org/10.1111/oik.04995

Sciuto, C., Geladi, P., La Rosa, L., Linderholm, J., & Thyrel, M. (2019). Hyperspectral imaging for characterization of lithic raw materials: the case of a Mesolithic dwelling in northern Sweden. *Lithic Technology*, *44*(1), 22-35. https://doi.org/10.1080/01977261.2018.1543105

Simpson, A., Fernández-Domínguez, E., Panagiotakopulu, E., & Clapham, A. (2020). Ancient DNA preservation, genetic diversity and biogeography: A study of houseflies from Roman Qasr Ibrim, lower Nubia, Egypt. *Journal of Archaeological Science*, *120*, https://doi.org/10.1016/j.jas.2020.105180

Strebel, D. E., Meeson, B. W., & Frithsen, J. B. (1994). Metadata Standards and Concepts for Interdisciplinary Scientific Data Systems. In R.D. Melton, D.M. DeVaney and J.C. French (Eds.) *The Role of Metadata in Managing Large Environmental Science Datasets* (pp. 41-48). Richland, WA, USA: Pacific Northwest Laboratory.

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R. & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PloS one, 15*(3). https://doi.org/10.1371/journal.pone.0229003

Thomas, J. R. A., Eccles, T. & Bowestead, S. (2016). *The Coleoptera of the sandhills of South Lancashire.* Iver, Raven Entomological & Natural History Society / Pemberley Books

Torben, R. C. & Sandweiss, D. H. (2020). Archaeology, climate, and Global Change in the Age of Humans. *PNAS 117*(15): 8250-8253. https://doi.org/10.1073/pnas.2003612117

Uhen, M.D., Buckland, P.I., Goring, S.J., Jenkins, J.P. & Williams, J.W. (2021). The EarthLife Consortium API: an extensible, open-source service for accessing fossil data and taxonomies from multiple community paleodata resources. *Frontiers of Biogeography*. https://doi.org/10.21425/F5FBG50711

Whitehouse, N. J., & Smith, D. (2010). How fragmented was the British Holocene wildwood? Perspectives on the "Vera" grazing debate from the fossil beetle record. *Quaternary Science Reviews, 29*(3-4), 539-553. https://doi.org/10.1016/j.quascirev.2009.10.010

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*(1), 1-9. https://doi.org/10.1038/sdata.2016.18

Williams, J. W., Grimm, E. C., Blois, J. L., Charles, D. F., Davis, E. B., Goring, S. J., ... & Takahara, H. (2018). The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. Quaternary Research, *89*(1), 156-177. https://doi.org/10.1017/qua.2017.105

Woodbridge, J., Fyfe, R., Smith, D., Pelling, R., de Vareilles, A., Batchelor, R., Bevan, A., & Davies, A. L. (2021). What drives biodiversity patterns? Using long-term multidisciplinary data to discern centennial-scale change. *Journal of Ecology*, *109*(3), 1396-1410. https://doi.org/10.1111/1365-2745.13565