SEAD - The Strategic Environmental Archaeology Database Inter-linking Multiproxy Environmental Data with Archaeological Investigations and Ecology

Philip Iain Buckland
Umeå University, Sweden

Abstract:
The volume of data on past environmental and climate changes, as well as human interactions with these, has long since passed the level where it is manageable outside of large scale database systems. The Strategic Environmental Archaeology Database project aims to not only store and disseminate such data, but also provide tools for querying and analysing them, whilst maintaining a close connection with the archaeological and ecological data that are essential for their comprehensive interpretation. Large scale, geographically and chronologically unrestricted databases provide us with essentially unlimited scope for putting individual sites into a broader context and applying locally collated data to the investigation of earth system level changes. By providing integrated access to data from a variety of proxies, including plant macrofossils, pollen, insects and geochemistry, along with dating evidence, more complex questions can be answered where any single proxy would not be able to provide comprehensive answers.

Keywords:
Environmental Archaeology, Database, Archive, Software, Palaeoecology

1. Introduction

SEAD, the Strategic Environmental Archaeology Database, is an open access archive for environmental archaeology and Quaternary science data. It aims to provide easy access to the raw data from a variety of investigation types where, primarily, biological or physical proxy data have been used to study the past. As an environmental archaeology database, the focus naturally leans towards the human and cultural aspects of prehistory, but the "natural", Quaternary science, background data essential in the interpretation of these phenomena are considered fundamental to the long-term usefulness of the system. Modern calibration and environmental survey data (e.g. insect pitfall traps, vegetation surveys) are also within the scope of the system. These are limited in number at the moment, but are scheduled to play an increasingly more important role in quantitative landscape and archaeological reconstructions, as science demands more empirically supported interpretations. This paper discusses a number of data ingestion and archiving related aspects of the project, the initial development phase of which runs from 2008 to 2013. It is hoped that the information here will prove useful to those considering undertaking similar ventures, or considering entering their own data into an existing archive such as SEAD. Readers interested in more applied aspects of the project are kindly directed towards Buckland et al. 2010.

Corresponding author: phil.buckland@arke.umu.se

2. Project Outline

The primary objective of SEAD is to provide a research infrastructure for environmental archaeology and palaeoenvironmental science. Specifically, systems and pathways for 1) data collation, storage and management; 2) integration, access and dissemination; 3) analysis and visualisation and 4) networking and support. These aims are to be achieved through the long-term financing of a stable resource based on expert domain science and technology. SEAD is a bottom up, research driven project, in that it originates from an expressed need from primary researchers and is designed and implemented by the same, in collaboration with technology experts. An appreciation of the need for large scale databases is essentially nothing new (e.g. Maurer et al. 2000; see earlier CAA volumes for archaeological examples), and the SEAD initiative stems from the Environmental Archaeology Lab's observation

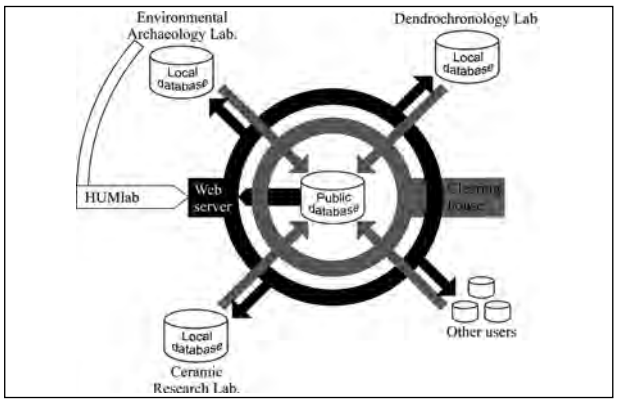

***Figure 1**. SEAD project structure and overview data flow model. Development is led by the Environmental Archaeology Lab (MAL), in collaboration with HUMlab, who in turn run the web server. Data submitted by partners and users are vetted by the clearing house before publication in the public database.*

that the large amount of raw data accumulating its archives (paper and digital) were no longer manageable without a systematic data storage and management system. In common with much archaeological research and consultancy data, these data are also largely the results of publicly funded projects, and as such subject to open access requirements. An online database system is the ideal solution for public dissemination of digital data, and with the addition of data-entry and quality control systems can provide a trusted resource for archiving research data.

Whilst these are worthy goals, the SEAD team believe that the addition of advanced data interrogation and analysis tools are essential if we are to realise the vision of transdisciplinary science, especially with respect to the use of data from multiple institutions, regions and disciplines. These tools may not only save research users time when aggregating sample data from multiple sites, over multiple chronologies, or synthesising habitat data for species, but also inspire and empower them to interrogate or analyse the data in new ways. The team firmly believe that breakthrough science will be enabled by systems such as SEAD.

2.1 Project organisation

The database and associated software is a cooperative venture between the Environmental Archaeology Lab. (http://www.idesam.umu.se/english/mal/) and HUMlab (http://www.humlab.umu.se/english/) at Umeå University, Sweden. It is funded by the Swedish Research Council, with co-funding from the Department of Historical, Philosophical & Religious Studies at Umeå University. Closely related research and development projects have been funded by Umeå University Faculty of Arts, and systems analysis and data entry of ceramics and dendrochronological data have been funded by the Nation Laboratories for Ceramic Research and Dendrochronology respectively at Lund University, Sweden (Figs 1 and 3). The project is advised by national and international reference groups, and close collaboration with the Neotoma Palaeoecology database ensures the optimal use of funds, by sharing ideas, software tools and data. As well as being an open access database, the project is run on open development principles, in that those interested in collaborating, both in terms of software development and data entry, are actively encouraged.

2.2 Data collation, storage and management

Data are entered into SEAD through a downloadable software application which includes an internal copy of the database (see http://www.sead.se for download). This copy can be synchronised with the online master at any time to provide local access to the latest datasets, and thus ensure that newly entered data can be analysed with respect to the latest data before publication. Users may of course wish to analyse their data with respect to a snapshot of the database, providing a stable base for quantification. Once ready for publication, the data can be submitted to the SEAD clearing house (one or more experts in the relevant field) for quality control (Fig. 1). This includes an assessment of the completeness of the data, with special attention being paid to the often neglected meta- and spatial location data, and a dialog with the submitter for any potentially inconsistent or extreme values. As archaeology is a science where it is most often the anomalies that are interesting, expert knowledge is essential for differentiating unusual data from errors.

2.3 Transparency

A fundamental concept in the SEAD project is that of data and analysis transparency; that research

results and interpretations should be traceable to the raw data underpinning them and that methods should be thoroughly described and reproducible. All datasets are uniquely identified (DOI is supported, but has yet to be implemented), permanent and version logged. The user can identify exactly which data were used to undertake an analysis, even when the datasets used have been subsequently revised. This also facilitates the retesting of earlier hypotheses with updated data, and the online interface supports saved viewstates which can be used to save, share and retrace analysis pathways. Datasets and selected metadata are linked to an internal bibliography, and the system is capable of providing citations at multiple levels, so that the nature of the publication's content can be assessed (e.g. publication of primary data; reanalysis of existing data; summary site interpretations). Virtual constituent databases (or master datasets) are used to help retain the identity of datasets ingested from other sources, or projects wishing to maintain their individual profiles, whilst taking advantage of access to the larger archive and analysis system, without the substantial costs involved in creating and managing it.

Transparency is also essential in connection with the storage of modern reference data. At the time of writing, this includes mainly insect ecology, distribution and climate data from the BugsCEP database (Buckland & Buckland 2006), but the aim is to expand the database, either internally or through linking, to cover other proxies and fields such as botany and ethnography. The methodology of environmental reconstruction, and the interpretation of archaeological deposits, ranges from subjective descriptions, based on the expert knowledge of an individual researcher, to complex mathematical models which are difficult to understand for those without expert knowledge. Giving users access to the raw source data allows them to see where interpretations come from and even assess the viability of reconstructions based on their understanding of the models and the description of the methods (Fig. 2).

2.4 System and database structure

SEAD is built around a PostgreSQL relational database, with online and offline components written in Java, Javascipt and PHP. The highly normalised

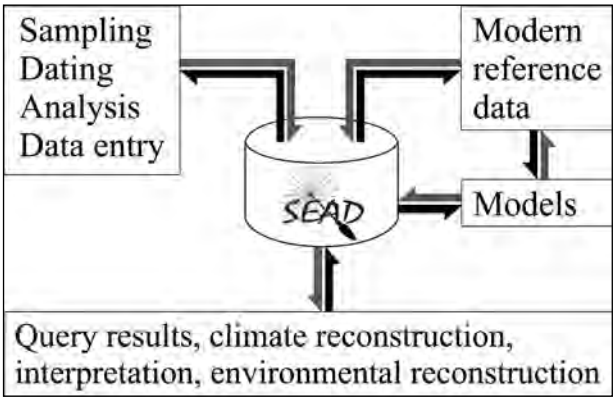

***Figure 2**. Schematic of transparent database/software based use of SEAD, showing the flow of data (light arrows) and transparency of process (dark arrows) from models to results, and vice versa.*

relational model, though perhaps not the most flexible system (when compared to object orientated or more abstracting database models), provides a level of detail and analytical power difficult to realise through more abstract models without a more considerable programming overhead. The structure is extensible, in that data tables are able to accept any form of continuous, discrete or categorised values or measurements and that any type of sample or feature metadata can be included. The entry of metadata types is regulated by the clearing house to prevent duplication or ambiguity. As the database implements a standard minimum divisible unit, that of a single analysis of a single sample, further tables can easily be added in order to extend the scope into the territory of new data forms, such as spectroscopic data, or linked to external databases. The potential for linking to external databases is also facilitated by permitting the storage of multiple identifiers for objects common to archaeological research (e.g. sites, features, contexts, samples etc.). This means that the same sample can be referred to by lab number, museum number or field number as appropriate.

The database stores comprehensive details and references for all methods (e.g. proxy analysis methods, pre-treatments, measurements, coordinate systems, processing methods), and care is taken to assign method descriptions to any area of metadata which would require a qualified description for quality assurance. Methods may also be grouped, allowing for multiple methods for each analysis type, such as the use of different photospectrometers

in phosphate measurement. Whilst the details of many methods may be considered irrelevant for many users, they are essential for the assessment of different methods, and providing background data for improving existing and developing new methods. The system is also capable of recording uncertainty values where relevant and available. To paraphrase an anonymous attendee of another conference, "true transparency, is not achieved by providing the data, but by providing enough information to enable the assessment of the quality of the data".

The latest database design, modelled using MicroOlap (http://www.microolap.com/), can be found on the project website at http://www.sead.se/database/.

2.5 Data scope

Whilst SEAD's structure technically allows it to cover an essentially unlimited scope of methods in environmental archaeology, the current scope of the project (Table 1) is confined by its budget. This is in no way an indication of future limits, integration of the ceramics and dendrochronological components having begun after the project's first three years. Other laboratories and research groups are actively encouraged to contact the SEAD team with respect to further expanding the scope of the database and analysis tools. In the long term, large scale, sustainable and queryable research archives can only be realised through community based data entry and development. The project is particularly interested in integrating archaeological and Quaternary science data, the latter dominated by studies of "natural" environmental change not directly related to archaeological investigations, and more often than not conducted from the point of view of the natural sciences. As well as being interesting in their own right, these studies provide data and information essential for the study of past human activities, not only in terms of the background climate and environment, but also for describing the immediate history and surroundings of archaeological sites.

SEAD has no chronological limits, meaning not only that any fossil data may be entered, but also any modern data. Currently, this extends only to insect survey data, which will be used as proof of concept for the integration of fossil and modern data, building on work undertaken in the Bugs database project (Buckland 2007). This will improve the efficiency with which researchers can interpret sites, but also opens up possibilities for refining palaeoenvironmental reconstructions. With time, similar data will be added for plants and geoarchaeology in order to build up a powerful reference database. Similar projects are well established for landscape reconstruction through pollen (Mitchell 2011), although they are not yet represented by integrated databases.

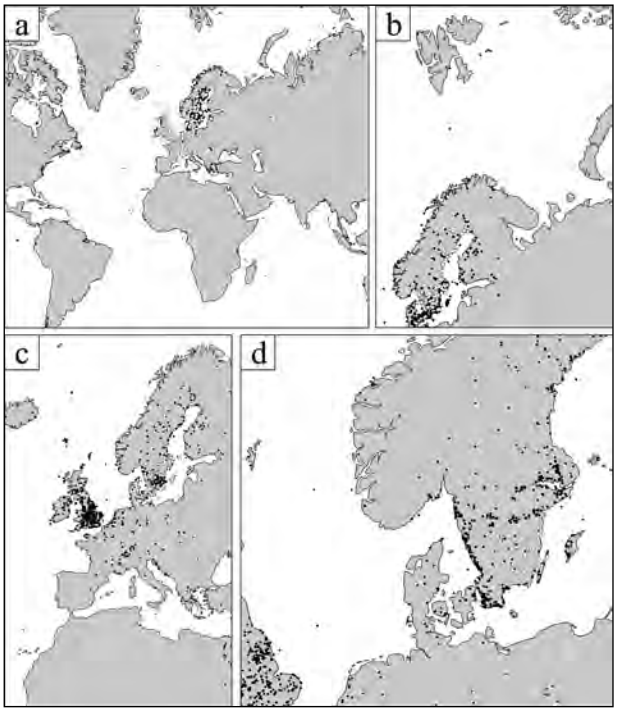

***Figure 3**. Four maps showing the geographical location of sites currently being ingested into SEAD. The white points in each map indicate the sites originating from one of the initial four master datasets: a) BugsCEP; b) Environmental Archaeology Lab, Umeå; c) National Laboratory for Ceramic Research, Lund; and d) National Laboratory for Dendrochronology, Lund (Småland test data). Note that coordinate errors may still be present as all the data have yet to be checked.*

2.6 Geographical scope

The majority of data currently included in SEAD's data entry schedule are European, reflecting the research scope and history of the labs involved in the initial development. Figure 3 shows the contribution of each of the data sources to the current

Proxy data sources	
Biological proxies	Raw counts of insects/arthropods, plant macrofossils, pollen, molluscs
Geoarchaeology	Soil chemistry (pH, phosphates) and physical properties (conductivity, organic content, colour)
Ceramics	Thin section quantification and properties (e.g. tempering material, inclusions, firing temperature, vessel chara cteristics)
Dendrochronology	Dates and support data, tree species, building history, dated object description and location (e.g. church, west tower, third beam from roof)
Dating evidence	
Scope	14C and other radiometric methods, dendrochronology, archaeological typology dates, period classifications, calendar dates and ranges, tephras
Chronological extent	Theoretically unlimited, but current range from 2.4 MyBP to present day.
Bibliographic data	
References	May be linked to site, sample group, sample and dataset levels as well as to methods, ecological codes and more.
pdf files	For references where available and not restricted by copyright
Modern reference data	
Abstracted text	Insect habitats and distributions, abstracted from trusted sources, with citations
Coded descriptors or classifications	Insect ecology, in-house system and Koch (1989-1992). Used for quantitative habitat reconstruction/visualisation (Buckland 2007)
Climate	Beetle Mutual Climatic Range (MCR) temperature reference data
Location data	
Coordinates	Three dimensional at site, sample group and sample levels (latitude, longitude, altitude and project survey grids). Capacity for national grid based storage. See section 2.6 and figure 3 for current geographical extent
Depth	Multiple types of depth recorded as positive or negative numbers, e.g. depth from lake/soil surface, depth from datum line or reference level (especially useful for stratigraphic sequences)
Archaeological, geological and sampling data	
Descriptive metadata	Site, feature and sample metadata to allow correlation between environmental and archaeology or lithology datasets. Sample names (e.g. field label, lab number, museum number). Descriptive information for all objects
Sample dimensions	Capacity for multiple measurements (volume, size, weight) at multiple stages of analysis (initial sample, analysed subsample, residue etc.). Position in sample group
Images	Scans, photographs, plans etc. for multiple levels in the site hierarchy and reference data (e.g. seeds, insects)

***Table 1**. Outline of the initial data scope of SEAD. A full list of metadata will be provided in the database documentation.*

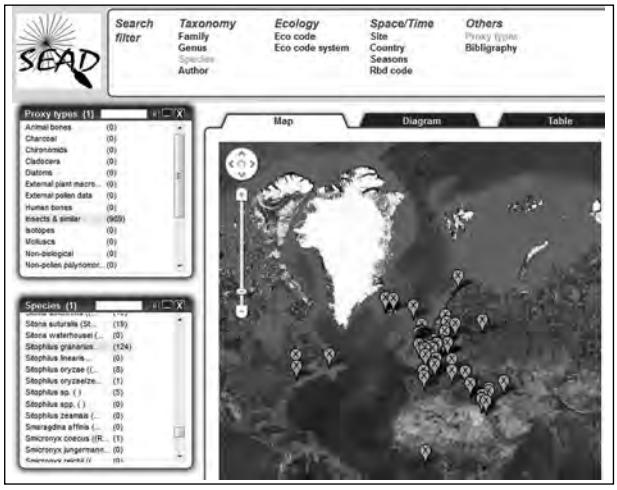

***Figure 4**. QSEAD, the online facetted browser for the SEAD database. Facets are classified by theme and activated by clicking on their names along the top of the screen. Selected filters appear on the left, collapsible to for an easier overview. Records may be multi-selected within each filter, the results cascading to subsequent filters and the results displaying after a short delay, dependent on the complexity of the query and the amount of data involved. The example uses Google Maps to show the location of the 127 sites which make up the databased fossil record of the grain weevil, Sitophilus granarius (L.). Map imagery ©2012 NASA, TerraMetrics. Insect data from the BugsCEP dataset (Buckland & Buckland 2006) in which citations for original sources can be found.*

geographical scope of the database. At the time of writing, the dendrochronological data consists only of a small proof of concept digitalisation, roughly 1.5 % of the Dendrochronology lab's archived data (Meissner et al. 2012).

3. Interfaces and Analysis Tools

The SEAD online interface, referred to as QSEAD after its aesthetic and technical origins in the QVIZ inter-archive project (www.qviz.eu; Palm 2008, 2009), is based around a facetted browser system using selectable, rearrangeable and cascading filters (Fig. 4). Each filter passes its results to the next one on the list, and can be rearranged should the logic of the research question require it. Figure 5 shows an example of the type of (palaeo)biogeographical and archaeological research which is made considerably easier by the system. The inclusion of relevant archaeobotanical and geological data in the system will increase the scope for such studies, both through

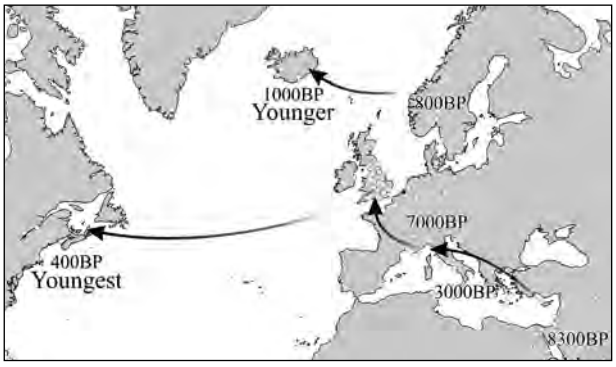

***Figure 5**. An example research application of the results exported from Figure 4. The spread of points and dates shows the state of (databased) knowledge on the palaeobiogeography of the species and highlights some obvious gaps, such as the relatively late dates from Greece and southern Scandinavia. Arrows show the probable spread of the grain weevil with agriculture and transport of store products. See Plarre (2010) and references for a more comprehensive study of the species.*

the investigation of parallel lines of evidence and the capacity to undertake the synthetic analysis of multiple proxies.

The system is currently able to show results as a list of datasets in tabular form as well as on maps, and customisable reporting functions are in the pipeline. Data can be exported (csv, xml), as a composite dataset of aggregated samples where the query includes multiple sites. Later, the ability to transfer the same dataset to more advanced analysis functions within the online software will be provided. A climate reconstruction explorer is under development at the moment, which will allow the faunal implications of insect based climate reconstructions to be investigated, summary statistics generated, and the relative importance of species in an assemblage assessed. Further tools are planned, including more flexible versions of the abundance and habitat diagrams and statistics provided in BugsCEP (see Buckland 2007), which can be applied to any data of appropriate type.

The engine behind QSEAD is a flexible, semi-automated, abstracting query system based on a network graph of the database (Palm 2009). Whilst this works well for many queries and up to a certain level of database complexity, additional web services are under development to provide database subsets which can be chained together for more

complex queries. These services will also allow for easier integration with other databases and external software.

4. Data Entry and Ingestion

The four initial data sources are described briefly below with respect to their scope, quality and specific problems relating to their ingestion. In SEAD terminology, these are referred to as master datasets, denoting their nature as virtual constituent databases. There is no limit to either the number of datasets stored under a master dataset or the volume and scope of its data. Although each of the master datasets are continually being added to, the import of legacy data is the main focus below.

4.1. The BugsCEP master dataset

BugsCEP is a downloadable, Microsoft Access and Visual Basic for Applications based insect ecology database and software package which contains the large part of the Quaternary fossil insect record (http://www.bugscep.com, Buckland & Buckland 2006; Buckland 2007). It contains over 1050 sites (Fig. 3a), most of which include quantified abundance data in the form of minimum numbers of individuals of each taxon found in each sample. The database includes ecology, distribution and climate reference data, as well as a comprehensive bibliography. The software incorporates a number of advanced data interrogation and analysis tools. Bugs has its origins in the 1980s (Sadler et al. 1992) and was presented in a previous incarnation at the 2001 CAA meeting in Gotland, Sweden (Buckland & Buckland 2002).

Having existed in some form for over 30 years, there have been various versions and improvements. Each stage has refined the data, parsing text and memo fields, and added to the scope of the project. The database is partially normalised, and includes a number of compromises which enabled the more rapid development of interfaces. These compromises have occasionally resulted in a lack of structural referential integrity, which in combination with insufficient checks on data entry have allowed inconsistent data to be entered, and important data to be omitted. However, the ingestion of Bugs into SEAD has comprised of the relatively simple task of devising SQL schema for mapping the two database structures, with a set of integrity catches which report to the data manager. This report can be used to correct any problems in BugsCEP or SEAD before publication, either using the interfaces or via backend hacking. Although simple in nature, this task requires a large amount of manual checking given the more than 10 500 taxa and 144 000 fossil records in the database. As always, the most vital resource in this undertaking is the knowledge of experts in the field. The assessment of data quality cannot be undertaken by inexperienced technicians or students, and training is therefore essential for long-term sustainability.

The BugsCEP software provides a number of analysis tools which could be usefully applied to other proxy data sources, and it is the author's intention, given time and funds, to reproduce all of these features online. Some of this will be achieved through the SEAD project, but until it is fully realised, and for the benefit of current users, the two systems must run in parallel. This has required a synchronisation system based on the SQL transactions which can cope with alterations as well as additions and deletions. Bugs unfortunately has a more limited and free text based approach to metadata, the automated linking of archaeological databases being outside of its scope. Sites are, by practice, uniquely named on data entry. The BugsCEP unique site identifier has no external relevance and is not seen by the user, and sites are unique on the basis of their name, location, date and references combined.

The primary linking attribute of the Bugs species reference data is that of the unique identifier for an insect taxon, the name alone being insufficient due to differences between international and the numerous national taxonomies. Where Bugs can only hold a single identifier for each taxon, SEAD can store multiple parallel taxonomies and thus record and export occurrences in whatever taxonomic system the user chooses. The database stores a comprehensive and referenced synonym list, but interrelating these taxonomies for data retrieval may be far from simple, and more problems will be encountered where pollen are concerned (e.g. Birks & Birks 2000). The approach of this project is, therefore, based on the concept of getting the data in and sorting out the details later, a principle which works best in the presence of

comprehensive metadata recording and consistency of data entry. Ontologies are being developed parallel to this process, rather than in advance, due to the complexity of the data and large number of unknowns.

4.2 The Environmental Archaeology Lab. (MAL), Umeå, master dataset

Research and consultancy in environmental archaeology has existed in some form since the 1970's at Umeå University. Consequently, MAL (Miljöarkeologiska laboratoriet) has amassed a considerable amount of data (Fig. 3b), much of it analogue and very little of it in database ready form. The lab's activities have encompassed a number of methods, but most frequently the analysis of plant macrofossils, pollen, geoarchaeology and insects (the latter covered by BugsCEP above). Although much of the recent data (from at least the year 2000) are stored digitally, far from all datasets are in easily ingestible spreadsheet or database form (Buckland et al. 2006). Even the spreadsheets, which should theoretically be based on a small number of templates, have had columns moved, renamed and new ones added due to the lack of a strict data handling regime. Occasionally, the wrong data have been entered under a heading, with an explanatory note in another column. These factors are a severe hindrance to the automated import of data, and have necessitated the development of data ingestion tools which allow spreadsheet columns to be checked and mapped to the appropriate analysis methods or metadata tables. Fully automated ingestion is out of the question due to the unpredictable nature of the variation.

A primary goal of ingesting the MAL dataset, and one of the pillars of the SEAD project, is that of making so called grey literature publicly available. These reports, although always available on request, are often only included as appendices in archaeological reports and would be more conveniently available online. SEAD will make the raw data which underpin or are included in the reports available online and queryable, with the reports available as pdf files. Metadata will be ascribed, national antiquities numbers checked, and English summaries provided for Swedish texts. The short reports presenting null results will also be included as a guide to future studies. This work is extremely time consuming, and requires qualified staff capable of undertaking detective work in the reports and paper and digital archives (which span five generations of Macintosh). A considerable part of the project budget goes to such assignments, a fact that is sometimes forgotten when planning or reviewing large scale database projects.

It may be prudent to mention at this point that, from the contract side of archaeology, it is the scientific literature that may be considered as "grey" in that it is not accessible without membership of a university or through considerable fees. The publication of data from these publications, if not the papers themselves, in open archives is a step towards a more democratic dissemination of scientific data where those not fortunate enough to have a university post may make use of them.

If SEAD is to achieve the vision of its designers, then all sample and analysis data should be cross-queryable – that is to say that it should be possible to identify the commonality of samples based on their location, chronology and properties, between projects. This is no simple task, and the data detective must deal with obscure or badly recorded sample descriptions and identifiers, missing or incorrect dating information, partial data, missing metadata and even the occasional bit of cryptography (older ecologists have been known to code their identification data). Where multiple proxies have been used on an excavation their results may be published in separate reports, sometimes with little indication of having come from the same samples. Part of the solution has been the use of standardised spreadsheets for the intermediate storage of data whilst problems are sorted out. An alternative exists where pollen data are concerned, in that Erik Grimm and the Neotoma project's (Neotoma 2009) Tilia software is able to read older Tilia files and export them as XML or spreadsheets. These files can then be easily imported into SEAD, leaving the main issue as the finding of metadata associated with each sample. Unfortunately, such metadata are rarely stored together and in unfortunate cases common sample names were not used for all analyses. In such cases, it is only by the interrogation of staff involved in the projects that an acceptable level of metadata can be extracted (a process that is familiarly referred to as "chaining senior researchers to their desks" within the SEAD project).

4.3 The National Laboratory for Ceramic Research, Lund, master dataset

At the inception of SEAD it was never conceived that ceramics data would form part of its scope. After discussions with researchers at Lund, however, it was clear that certain commonalities existed between the environmental data and aspects of the ceramics data. In particular, it was decided that the ingestion of already partially digitised data from the analysis of ceramic thin sections would be highly desirable, and SEAD was adapted to accept them (Fig. 3c). A number of the recorded properties of thin sections are essentially geochemical or can be recorded as the presence/absence of specific components. The facility to cross-query the ceramics and macrofossil data of multiple sites holds great potential for archaeological research. Among other things, it could expedite contributions to the study of plants as foods and medicines, crop processing, agricultural development, diets and rituals by facilitating the integration of different streams of evidence for the same activities.

The ingestion of the ceramic data serves to illustrate an interesting problem in the assimilation of legacy data: that of recording absence. In this master dataset, representing data accumulated by several researchers over the past 30 years, but also including some data from the 1950's, there are five different absence recording methods: blank/null, "no", "no information", "missing" and "undefined". For the sake of simplicity, it has been decided to assume that all of these mean the absence of whatever was being recorded, although as the authors of earlier investigations are inactive we may never know the truth. There is always a danger of losing resolution when digitising or migrating between database systems. To ensure data transparency the intermediate spreadsheets, which contain this variety, will be preserved in case the variation has meaning. The general problem of recording absences is perhaps one of standardisation, and common to many databases. Problems occur more often when merging from different data sources, where internal standardisation may not have been consistently applied to each source. We can, however, at least hope to try to maintain that null is a state and not a value.

4.4 The National Laboratory for Dendrochronology, Lund, master dataset

With the exception of a few dates, dendrochronological data are the latest edition to SEAD (Fig. 3d), and a test dataset consisting of material from the Swedish county of Småland was entered in early 2012 (Meissner et al. 2012). Although data from a number of archaeological projects (35) were entered, using intermediate spreadsheets, the majority of the data relate to building heritage management (71 projects). The latter orientation was also an extension of the initial mandate of SEAD, and required the expansion of generic description tables to enable the attribution of multiple types (e.g. building type, building purpose, building structural form etc). Although dendrochronology is often considered primarily as dating evidence for the context in which wood is found, there is a large amount of measured and descriptive data behind every date. For SEAD to provide anything but a date retrieval system for these data, it was therefore essential that it be able to store these supporting data. This was achieved by a minor modification of spatial data tables, so that they could cater for a descriptive location within a structure; and the addition of a specific dendrochronological support data table. The existing structure provides for the storage of tree species, and allows the retrieval of all analyses performed on a particular species, be they macrofossil, pollen, ^{14}C or dendrochronological.

Entering the remainder of the archived dendrochronological data is estimated to take the equivalent of four years (Meissner et al. 2012), a monumental task planned to start in 2013. In addition to the above, the Dendrochronology Lab. has also undertaken the analysis of wood in wrecks and living trees. With time, these datasets will also be entered into SEAD and provide an extremely useful continuity between ancient and modern measurements, land and sea.

5. External-linking

The BugsCEP project has shown the advantages of being able to rapidly summarise the ecology of fossil insects found in samples, with over 80 publications citing use of the latest version alone (see http://bugscep.com/publications.html). Bugs has been managed from the environmental

archaeology point of view, with an understanding of the importance to archaeology of Quaternary science data. Although the latter may be obvious to some, there are many archaeologists who believe that only samples taken on-site may reveal useful archaeological information. This is a misconception which must be overcome if archaeology is to progress and collaborate with other palaeoenvironmental sciences. By doing so, it may increase its contribution to the body of accepted knowledge on an integrated picture of past human activities and environmental and climate change. As ecologists become more interested in palaeobiogeography, perhaps with the explosion of molecular studies of potential origins and refugia (e.g. Schönwetter et al. 2005), there is an indication that they themselves may themselves become more interested in palaeoenvironmental databases (see Brewer et al. 2012), especially when the DNA, fossil record and morphology may disagree.

Within Sweden, the standard cultural heritage identifier is the RAÄ (Riksantikvarieämbetet) number, the equivalent of the British SMR number. The combination of parish name and RAÄ number are supposedly unique, although mistakes have occurred and this cannot be guaranteed. Discussions are underway on how to resolve this and similar issues and enable reliable linking between SEAD and the Swedish National Heritage Board's databases. The latter are undergoing a large scale redesign and there is good scope for improved low level accessibility (Lars Lunquist, RAÄ, pers comm.). Recently, the Swedish National Data Service (SND) has released excavation data in GIS form (http://snd.gu.se/sv/node/564), an important step in creating online archaeological databases with a level of detail sufficient for complex multi-site queries.

An important issue yet to be fully dealt with is the representation and visualisation of sites without spatial data or dates. These will not show up on maps or timelines respectively without special attention, and may contain useful data which should not be omitted from a query. A similar problem is posed by sites/objects with multiple locations, such as an Egyptian mummy stored at the British museum.

6. Concluding Remarks

The creation of large scale research data infrastructures is essential for scientific advancement in an increasingly data-rich world. Open access to published data, as well data citation, encourages data reuse and large scale syntheses which would otherwise be extremely difficult. Providing easier access to previous results effectively bootstraps research within and across disciplines (Mooney & Newton 2012). Such initiatives may also help to satisfy public and political demands for greater research transparency, the importance of which was made particularly clear during the "climategate" affair, where lack of transparency was used as one of numerous arguments against the viability of evidence for global warming (Maibach et al. 2012). The examples provided in this paper highlight the fact that although there are considerable advantages to integrated data sharing, the construction and maintenance of such infrastructures is far from simple. It not only requires long-term funding and commitment from a core research community, preferably distributed over multiple research groups, but perhaps more importantly a persistent user base. SEAD is still at an early stage and it remains to be seen as to whether it will achieve the required momentum for long term success.

Expert systems have been discussed earlier in the history of the Bugs project (Buckland et al. 1997), the implications being of a software system that could provide explanations for new datasets based on the existing database. This is still a goal for many palaeoecologists, even if, in the absence of artificial intelligence, the preferred term may now be decision support systems. Ultimately the system should be able to provide a list of statistically similar sites or samples to those selected, along with multiple environmental reconstruction scenarios and other tools to aid interpretation. There is however, always a cost based limit to the amount of "intelligence" that can be built into a system, after which the designers must rely on the intelligence of the user. Even with advanced ecology based, quantitative environmental reconstruction, a detailed knowledge of environmental science and the ecology of the organisms/proxies used for reconstruction is required, if the reconstruction is to be reliably interpreted. It is never a case of push a button and get an answer, but always push the right button and get something that will help provide potential answers to the research questions.

Acknowledgements

The SEAD project is a team effort, in cooperation with the Lund laboratories, and the work of all those involved should be acknowledged. Thanks in particular to Fredrik Palm and Erik J. Erikson for help with the conference presentation. Roger Engelmark should also be mentioned for pressing the burgeoning idea of the database home with his colleagues. I am also grateful to the Swedish Research Council and its external reviewers who provided useful comments and permitted SEAD to be funded. As always, the database would be nothing without the data contributors (see the database for full citations).

References

Birks, H.H., and H.J. B. Birks. 2000. "Future uses of pollen analysis must include plant macrofossils." *Journal of Biogeography* 27(1): 31-35.

Brewer, S., S.T. Jackson, and J.W. Williams. 2012. "Paleoecoinformatics: applying geohistorical data to ecological questions." *Trends in Ecology & Evolution* 27(2): 104-112. doi: 10.1046/j.1365-2699.2000.00375.x.

Buckland, P.I., Y. Zhuo, D. and P.C. Buckland. 1997. "Towards an Expert System in Palaeoentomology." In *Studies in Quaternary Entomology - An Inordinate Fondness for Insects*, edited by A. C. Ashworth, P. C. Buckland, and J. P. Sadler, 71-77. Chichester: John Wiley & Sons Ltd. http://bugscep.com/phil/publications/bucklandetal1997_expert.pdf.

Buckland, P.I., and P.C. Buckland. 2002. "How can a database full of Bugs help reconstruct the climate?" In *Archaeological Informatics - Pushing the Envelope - CAA 2001 - Computer Applications and Quantitative Methods in Archaeology, Proceedings of the 29th Conference, Gotland, April 2001,* edited by G. Burenhult and J. Arvidsson, 453-461. Oxford: Archaeopress. http://bugscep.com/phil/publications/buckland&buckland2002_caa.pdf.

Buckland, P.I., and P.C. Buckland. 2006. *BugsCEP Coleopteran Ecology Package*. Boulder: IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series # 2006-116. NOAA/NCDC Paleoclimatology Program. Accessed 24 June 2012. http://www.ncdc.noaa.gov/paleo/insect.html and http://www.bugscep.com.

Buckland, P.I., J. Olofsson, and R. Engelmark. 2006. *SEAD – Strategic Environmental Archaeology Database, planning report*. MAL Reports 2006-31. http://bugscep.com/phil/publications/bucklandetal2006_SEADplanning.pdf.

Buckland, P.I. 2007. "The Development and Implementation of Software for Palaeoenvironmental and Palaeoclimatological Research: The Bugs Coleopteran Ecology Package (BugsCEP)." PhD diss., University of Umeå, Sweden. http://www.diva-portal.org/umu/abstract.xsql?dbid=1105.

Buckland, P.I., E.J. Eriksson, J. Linderholm, K. Viklund, R. Engelmark, F. Palm, P. Svensson, P.C. Buckland, E. Panagiotakopulu, and J. Olofsson. 2010. "Integrating Human Dimensions of Arctic Palaeoenvironmental Science: SEAD - The Strategic Environmental Archaeology Database." *Journal of Archaeological Science* 38(2): 345-351. doi:10.1016/j.jas.2010.09.011.

Koch, K. 1989-92. *Die Käfer Mitteleuropas. Ökologie, 1-3*. Krefeld: Goecke & Evers.

NEOTOMA. 2009. "The Neotoma multiproxy palaeoecology database." Accessed 30 May 2012. http://www.neotomadb.org/.

Maibach, E., A. Leiserowitz, S. Cobb, M. Shank, K.M. Cobb and J. Gulledge. 2012. "The legacy of climategate: undermining or revitalizing climate science and policy?" *WIREs Clim Change* 3: 289–295.

Maurer, S.M., R.B. Firestone and C.R. Scriver. 2000. "Science's neglected legacy." *Nature* 405: 117-120. doi:10.1038/35012169.

Meissner, K., P.I. Buckland, D. Hammarlund, and H. Linderson. 2012. "Pilotprojekt 'Dendro-databas' i SEAD." MAL rapports nr. 2012-23. Lund: Umeå universitet & Lunds universitet. http://www.sead.se/files/pilotprojekt_dendro_2012.pdf.

Mitchell, F.J.G. 2011. "Exploring vegetation in the fourth dimension." *Trends in Ecology & Evolution* 26(1): 45-52. doi:10.1016/j.tree.2010.10.007.

Mooney, H., and M.P. Newton. 2012. "The Anatomy of a Data Citation: Discovery, Reuse, and Credit." *Journal of Librarianship and Scholarly Communication* 1(1): eP1035. http://dx.doi.org/10.7710/2162-3309.1035.