Poseidon - A toolbox for archaeogenetic data management

Clemens Schmid, Max Planck Institute for the Science of Human History (MPI-SHH), schmid@shh.mpg.de
Ayshin Ghalichi, Max Planck Institute for the Science of Human History (MPI-SHH), ghalichi@shh.mpg.de
Wolfgang Haak, Max Planck Institute for the Science of Human History (MPI-SHH), haak@shh.mpg.de
Stephan Schiffels, Max Planck Institute for the Science of Human History (MPI-SHH), schiffels@shh.mpg.de

The recent increase in openly available ancient human DNA samples demands for new software solutions to store, distribute and analyse both genomic as well as archaeological context data. In this paper we present Poseidon, a computational framework including an open data format, software, and a public online repository to enable convenient, reliable, and FAIR access to genotype data from all around the world.

Archaeogenetics has become a fast accelerating field, with new data coming out faster than people can co-analyse (Orlando et al. 2021). If one considers samples currently being processed in the world's largest laboratories, we are now quickly approaching genome-wide data for 10,000 ancient human individuals. In addition, emergent fields such as ancient metagenomics and paleo-proteomics are adding complexity to a data landscape that already hosts traditional archaeological data with contributions from long-established non-genetic technologies like radiocarbon dating and stable isotope analyses.

Data from genetic analyses in academic papers is usually shared by releasing raw sequencing output into public repositories like the European Nucleotide Archive (ENA). Archaeological context information for individual samples is documented in supplementary tables attached to the publications. So while most data in the field is technically accessible, it does not satisfy the FAIR principles of open data: Findability, Accessibility, Interoperability, and Reproducibility (Wilkinson et al. 2016)

Beyond ethical concerns, this raises a number of concrete practical issues:

- Intermediate data such as genotypes or metagenomic profiles are often not released at all, making it hard to reproduce specific results.
- The connection between individuals, contextual information and genetic data is challenging to maintain across very different repositories and sources.
- Meta-analyses spanning datasets require enormous amounts of work on data collection and curation.
- Incrementally produced data, for example by adding new data to previously published individuals, cannot be easily connected to the same individuals.

To mitigate some of these problems, we propose a package data format which bundles the big genotype data in industry-standard formats (EIGENSTRAT, PLINK) with a flat context data file: The .janno file. For each sample, this human- and machine-readable .tsv file format stores metadata (e.g. publication, keywords) together with information about spatiotemporal origin (e.g. coordinates, radiocarbon dates) and genetic data preparation context (e.g. shotgun sequencing vs. target enrichment) as well as quality markers (e.g. number of autosomal SNPs on the 1240k array). Genotype data and .janno file are supplemented with a BibTeX file for the relevant citations plus optional metadata files (README, CHANGELOG) to make the package complete.

The main workhorse for operations on these Poseidon packages is the command-line software trident. It is written in the functional programming language Haskell, builds on strong type-safety and clean interfaces, and provides modules for package creation, inspection, validation and analysis. trident handles both context data as well as the large genotype data files – the latter via stream processing. Sample entities are internally represented with specific data types, which ensures strict parsing constraints to maintain structural correctness and machine readability for all data in Poseidon. trident also serves as a command line client to download already available packages from a central online repository we host and maintain. We implemented the respective webserver relying on the same Haskell infrastructure as trident.

For integration of Poseidon packages with an R data analysis pipeline, we provide the poseidonR package to load .janno files into a tidyverse-compatible, tibble-derived S3 object. poseidonR also provides functions for bulk radiocarbon date calibration and age sampling on these janno objects.

In the spirit of the session, our presentation will highlight multiple aspects of our learning experience when developing Poseidon. This includes…

- the value we saw in a clear file format definition for the .janno file
- the advantages of a strictly typed programming language like Haskell for parsing operations
- the challenges of command line interfaces as the true lingua franca of scientific computing (Prasad et al. 2020)
- the necessity for code review, unit tests and manual alpha testing

Poseidon is a community project. Code and data are open. All documentation for the data format as well as the software tools are available online. Both for software development and for (context) data keeping we rely on version control with Git/Github.

References

Orlando, Ludovic, Robin Allaby, Pontus Skoglund, Clio Der Sarkissian, Philipp W. Stockhammer, María C. Ávila-Arcos, Qiaomei Fu, et al. 2021. "Ancient DNA Analysis." Nature Reviews Methods Primers 1 (1): 14. https://doi.org/10.1038/s43586-020-00011-0.

Wilkinson, Mark D., Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR

Guiding Principles for Scientific Data Management and Stewardship." Scientific Data 3 (March): 160018. https://doi.org/10.1038/sdata.2016.18.

Aanand, Prasad, Firshman Ben, Tashian Carl, and Parish Eva. 2020. "Command Line Interface Guidelines." December 23, 2020. https://clig.dev.

CHRONOLOG: a tool for computer-assisted chronological research

Eythan Levy, Tel Aviv University, eythan.levy@gmail.com
Gilles Geeraerts, Université Libre de Bruxelles, gigeerae@ulb.ac.be
Frédéric Pluquet, Haute École Louvain en Hainaut, fredericpluquet@gmail.com

Abstract unavailable.

Digital Ecosystems in Archaeological Science: A History and Taxonomy of R packages in Archaeology

Ben Marwick, University of Washington, bmarwick@uw.edu

Introduction

The digital ecosystem of a research community is important to understand because it reveals what is possible, what is common, and what unrealised opportunities exist for future work. The recent growth in the use of R in archaeological science is of especial interest because it is a free, open source, and highly extensible programming language that any researcher can contribute to by writing a package and sharing it for others to use. In this presentation we present the results of a study of how archaeologists use R packages, and how archaeologists write them for other researchers to use. We ask: what are the patterns in R package use and production among archaeologists, and what packages are yet to be developed that might satisfy the specific needs of archaeological scientists?

Methods and materials

We have two data sources for this study: (1) a list of packages written by archaeologists, and (2) a list of 150+ scholarly publications that include R code. Both datasets are openly curated at https://github.com/benmarwick/ctv-archaeology To answer our question about how packages are used, we will extract all the packages referenced in the 150+ papers that include R code. We will classify these packages according to keywords in their title to identify the types of packages that are often used by archaeologists, and how they associate with journals and archaeological topics. We will study co-citation patterns to understand what packages tend to be used together, and for what types of data. To answer our question about archaeologists produce packages, we will use our dataset of packages authored by archaeologists and analyse data on dates of first publication, usage data from CRAN downloads and citations, and co-citation patterns. We will also examine software engineering attributes, such as the presence/absence of vignettes, tests, continuous integration, dependencies, etc. to understand how archaeologists adhere to modern package development conventions ,

Results

The analysis is underway and results are not available yet.

Discussion

Our results will be the first empirical study to examine the frequency of R package use for the reported analyses of archaeological data in publications. We will also report on patterns of R package production by archaeologists. We expect that our results will be useful for archaeologists looking to begin using R but feel overwhelmed by the 10,000+ packages currently on CRAN. Our results will show the most commonly used packages, which will guide archaeologists towards packages that will probably be a good choice for their data analysis needs, assuming a strong frequency-based bias in software choice. In looking at how archaeologists product R packages, we will reveal where current effort has been concentrated, and recommend where future effort be directed. We will also report on the best practices currently in use by archaeologists writing R packages, to inspire and motivate future package authors to produce high quality software that will be used and cited by the research community

References

Joo, Rocio, Matthew E. Boone, Thomas A. Clay, Samantha C. Patrick, Susana Clusella-Trullas, and Mathieu Basille. “Navigating through the r packages for movement.” Journal of Animal Ecology 89, no. 1 (2020): 248-267.
Lai, Jiangshan, Christopher J. Lortie, Robert A. Muenchen, Jian Yang, and Keping Ma. “Evaluating the popularity of R in ecology.” Ecosphere 10, no. 1 (2019): e02567.
Li, Kai, and Erjia Yan. “Co-mention network of R packages: Scientific impact and clustering structure.” Journal of Informetrics 12, no. 1 (2018): 87-100.
Lortie, Christopher J., Jenna Braun, Alessandro Filazzola, and Florencia Miguel. “A checklist for choosing between R packages in ecology and evolution.” Ecology and evolution 10, no. 3 (2020): 1098-1105.

outlineR: An R package to derive outline shapes from (multiple) artefacts on JPEG images

David Matzig, Aarhus University, david.matzig@cas.au.dk
Felix Riede, Aarhus University, f.riede@cas.au.dk

Geometric morphometric methods (GMM) in archaeology are experiencing a sharp increase in application and popularity since the last decade or so and seem to be more popular now than ever. In general, they constitute a major advance vis-à-vis earlier qualitative descriptions, typological assessment, or linear measurements of artefacts. GMM approaches can be divided into methods that use landmarks, and those that use trigonometric descriptions of whole outlines. The bulk of archaeological applications of GMM have so far relied on landmark-based approaches, although a surge of recent studies is demonstrating the utility of whole-outline approaches using so-called elliptical Fourier analysis (EFA; Kuhl and Giardina 1982) and cognate approaches. Landmark approaches offer a straightforward way of delineating homologous structures, but their application also incurs a significant loss of shape information. In addition, the a priori identification of homologous landmarks on artefacts can be difficult and inherently subjective unless unambiguous theoretical expectations are available. Therefore, outline approaches offer an alternative, robust and information-rich way of capturing artefact shape data. Accurate artefact outlines can also be extracted efficiently from widely available legacy data, especially from artefact line drawings. There currently exist various standalone software applications as well as some R packages for the extraction and analysis of landmarks and whole-outlines. However, the extraction step always involves a considerable amount of manual processing and manual tracking of either the landmarks or whole-outlines, which proves to be the definite bottleneck of many studies.

In this paper we introduce the R package outlineR that allows for a fast and efficient extraction of whole-outlines from multiple artefacts on images, ready to be analysed in the Momocs (Bonhomme et al. 2014) environment. We give insight to the workflow and how it compares to existing methods of whole-outline extraction thus showing the advantages and savings in time when using outlineR for the digitization of large amounts of legacy data, such as artefact photographies or drawings. Finally, we present a case study using a large dataset of Late Neolithic/Early Bronze Age projectile points from Northwestern Europe extracted using the outlineR package to showcase the possibilities of whole-outline GMM regarding the creation of typologies and inference of chronological information.

References

Bonhomme, V., Picq, S., Gaucherel, C., & Claude, J. (2014). Momocs : Outline Analysis Using R. Journal of Statistical Software, 56(13). https://doi.org/10.18637/jss.v056.i13

Kuhl, F. P., & Giardina, C. R. (1982). Elliptic Fourier features of a closed contour. Computer Graphics and Image Processing, 18(3), 236–258. https://doi.org/10.1016/0146-664X(82)90034-X

An open-source approach for the vulnerability assessment of archaeological deposits using GPR data in QGIS environment

Philip Fayad, Alma-sistemi, pkf@alma-sistemi.com
Matteo Serpetti, Alma-sistemi, mse@alma-sistemi.com
Stefano De Angeli, University of Tuscia (UNITUS), deangeli@unitus.it

Introduction

A specific tool for GPR data processing has been designed in the context of the project RESEARCH (REmote SEnsing techniques for ARCHaeology – H2020-MSCA-RISE- 2018 n. 823987). The project addresses risk assessment procedures for archaeological sites threatened by environmental pressures, as land-use change, land movement and soil erosion, and the creation of a Web-GIS Platform able to automatically perform the risk assessment procedures (www.re-se-arch.eu). The methodology adopted by the project required data about the depth of the archaeological deposit and the distance from the ground of its most superficial layer, in order to precisely evaluate the vulnerability of buried archaeological features to threats acting on soil (in particular soil erosion). To do so, an innovative method of GPR data processing has been designed, that can automatically recognize subsurface features depth. The challenge was to implement in a GIS environment an automatic open-source tool that can replace the manual interpretation of archaeological features detected with GPR investigation by automatically working on pixel values. The study illustrates the specific methodology adopted to automate the GPR data processing procedure. The GIS-based tool was tested in the case study of the Roman town of Falerii Novi.

Methods and Materials

RESEARCH risk assessment procedure give prominence to unexcavated archaeological heritage, focusing on main buried features, such as structures (which can be more easily identified through geophysical survey) and stratigraphy. The burial depth of structures and possibly intact stratigraphy is the principle on which vulnerability values are assigned, since the more a feature is far from the soil surface, the less it is impacted by pressures acting on it, such as agricultural activities and soil erosion. The goal was to identify an interface that represents the superficial extent of the undamaged layer affected by agricultural activities, and to assign the vulnerability values on small scale, in order to more precisely evaluate the risk for the preservation of the archaeological deposit. In our study, we translated the vulnerability assessment methodology proposed by RESEARCH into specific processing steps creating in this way a GIS workflow. The whole chain of operations was wrapped into a single process, a single algorithm, in order to make it convenient to execute it later with different sets of inputs, thus saving time and effort. The algorithm resumes the entire process of pixel-based vulnerability mapping of archaeological deposits, starting from processing raster GPR time-slices to produce the most superficial layer of the archaeological deposit, up to finally assigning vulnerability values to each pixel. The tool was developed inside QGIS platform, the most popular geographic information system (GIS), Open Source, licensed under the GNU General Public License, which is part of the Open-Source Geospatial Foundation project (OSGeo). The high-resolution GPR data concerning the site of the roman town of Falerii Novi, recently published in open-access, were used for the first testing of this procedure (Millett et al. 2019; Verdonck et al. 2020).

Results

This automated procedure allows to produce good results in a short time, with no need for manual intervention of the user, all within QGIS environment. Usually, manual interpretation of detected features mostly concerns structures, overlooking stratifications. This algorithm allows for a very detailed representation of the archaeological deposit considered in its wholeness. It also allows for the production of detailed vulnerability maps, where vulnerability values are assigned to pixel-size areas in color scale. This level of detail ensures project RESEARCH a more reliable risk assessment, where risk is calculated on very small scale, therefore with more accuracy.

Discussion

The paper presents the specific procedure designed for producing the upper layer of archaeological deposits detected with GPR, within QGIS environment, highlighting the opportunities and obstacles of this type of procedure and approach. Further development of the tool will be also discussed, such as the possibility to fully automate the procedure for archaeological features recognition and structures representation.

References

REmote SEnsing techniques for ARCHaeology (RESEARCH). Accessed March 1, 2021. https://www.re-se-arch.eu.

Verdonck, Lieven, Alessandro Launaro, Frank Vermeulen, and Martin Millett. 2020. "Ground-Penetrating Radar Survey at Falerii Novi: A New Approach to the Study of Roman Cities." Antiquity 94 (375): 705–23. https://doi.org/10.15184/aqy.2020.82.

Martin Millett, Lieven Verdonck, Ninetta Leone, Alessandro Launaro. 2019. "Beneath the surface of Roman Republican cities" (data set). York: Archaeology Data Service. https://doi.org/10.5284/1052663.

Managing and analysing pictorial documentation with GIS and graphs

Craig Alexander, independent researcher, craiga304@gmail.com
Jose Pozo, independent researcher, josmpozo@gmail.com
Thomas Huet, LabEx ARCHIMEDE, ANR-11-LABX-0032-01, thomashuet7@gmail.com

Since its creation, archaeology has largely focused on material culture (personal ornaments, pottery, burials, settlements, etc.) with an ongoing special interest shown in the 'bel objet'. Recovered artefacts and the relations they share (spatial, temporal, typological, etc.) are documented in words, numbers, but above all pictures (drawings, photographs, 3D models, etc.). As a result, archaeology has a long tradition of technical drawing (artefact drawings, x-y ground plans, stratigraphic or elevation cross-sections, etc.). Graph theory and network analysis allow one to overcome the main difficulties arising from non-georeferenced and unscaled drawings by modelling qualitative relationships when the quantitative measurements are poor. For example, for a freehand 20th century historical map of an Urnfield necropolis, the spatial distribution of goods within the burials and the spatial relationship between the burial goods can be model with a spatialized network. Besides these spatialized networks, many other types of graphs are used (implicitly or otherwise) in archaeology. The two main ones are probably the directed acyclic graph (DAG) used to model stratigraphy (i.e, Harris matrices) and the tree-like structures modelling hierarchical relations between categorical variables (i.e, typology). Graph theory is multi-paradigm (e.g, can be used for spatial and temporal modelling), multi-scalar, and graph drawing is a well-known heuristic to communicate results. This type of modelling is currently used for linked open data (LDO, e.g. JSON-LD format).

We believe that pictorial documentation within a spatial database and with graph-based methods is one of the priorities for IT development in archaeology (package creation, shared conceptual models, best practices dissemination, open software, etc.). The recent R package 'iconr' employs spatialized networks to model prehistoric graphical content at the scale of the decorated support (pottery, wall, statues-menhirs, etc.) and favours GIS data entry. Currently, the package CRAN version allows the user to tag graphical units (GUs) with attributes and to filter the whole iconographical content when two or more chronological layers of GUs are present (superimposition, diachronic structure of the representations, etc.). The next version of the package will focus on functions to create DAGs and tree-like structures.

After an introduction to the importance of management of pictorial documentation with spatial databases and graph theory (network analysis), we will provide a critical review of DAGs and tree-like R packages available in archaeology. We will then briefly present the 'iconr' package and the planned developments in future versions.

Open archaeology: a survey of collaborative software engineering in archaeological research

Zachary Batist, University of Toronto, z.batist@mail.utoronto.ca Joe Roe, University of Bern, joe@joeroe.io

Archaeologists increasingly rely on specialised digital tools and computer code to conduct their research. Scientific programming languages such as R and Python are used to write, modify, comment on, review, share and reuse scripted analyses, enabling more advanced manipulation of digital data (Schmidt and Marwick 2020). This usually consists of custom scripts and project files that parse and transform archaeological data using generic functions for data manipulation, statistical analysis, and visualization. These are drawn from a range of 'off-the-shelf' packages, such as: the tidyverse, ROpenSci, or NumPy data analysis ecosystems; database management systems and visualisation engines; or modelling frameworks and specialised software from other disciplines and industries. Collectively, these tools support a broad set of applications common across research settings. However, as archaeological workflows are rarely explicitly accounted for in their design, they can also impose limits and force archaeologists to adapt tools in ways unanticipated by their developers.

As digital methods have become increasingly central to archaeological research, it is therefore becoming more common for archaeologists to develop software explicitly targeted at their use cases. These tools fundamentally differ from analysis scripts in that they are designed for reuse by multiple analysts for multiple applications within a given problem domain. They provide generic functions that can handle mutable inputs, rather than custom procedures tailored to a specific dataset and analysis, based on the designers' assumptions about what data structures, processes and desire outputs are shared by the tool's intended users. As such, building these tools evokes a distinct set of skills, shifting the developer from the role of analyst to that of a 'research software engineer' (Baxter et al. 2012). Although the intersection of archaeology and software engineering is not new, these contemporary projects are distinguished by their adoption of practices from open source software development. In particular, widespread use of the git version control system, associated web-based source code management platforms such as GitHub and GitLab, is a relatively recent trend, opening up a new set of workflows for collaboration between multiple developers.

In this paper, we survey the state of the art in archaeological software engineering, documenting the wide range of general-purpose digital tools currently in development. Using open-archaeo (https://open-archaeo.info/), a curated list of 300+ active open source archaeological software packages, augmented with data collected from GitHub's API, we seek to identify emerging norms in software development and collaboration, focusing on three key questions:

1. What types of open source projects are have been developed by archaeologists over the last 5–10 years?

2. To what extent to these projects leverage the collaborative features of git/GitHub?
3. Does collaboration in software development mirror, or differ from, collaborative practices in archaeological research more broadly?

We find that collaborative open source software development in archaeology, measured both in the number of projects and discrete contributions tracked in git repositories, has seen a rapid and sustained increase beginning around 2015 (see figure). This growth is seen across a range of languages and categories of tools, but is strongest in standalone web apps and R packages. In terms of collaboration, our analysis shows an uneven use of git and GitHub's extended features, beyond their basic usage as a version control system and repository host. The vast majority of repositories have 1–3 contributors, with only a few distinguished by an active and diverse developer base. Similarly, collaborative features such as GitHub "issues" are used in only a minority of repositories. However, a network analysis of repository contributors may point to some nascent communities of practice.

We highlight areas in which archaeologists are either pooling resources for common goals or working independently and in a redundant manner, factors that may contribute to either enthusiastic upkeep or abandonment of software development projects, how various means of communication and contribution are valued, and how GitHub is leveraged for either one-way or discursive means of engaging with relevant stakeholders. We consider these aspects of collaborative software development in relation to common structures, practices and challenges that bind archaeologists together as a distinct community, and draw comparisons with potentially conflicting underlying assumptions, attitudes and processes accounted for and encouraged by the infrastructures that archaeological software developers have come to rely upon. We demonstrate how archaeological software engineering is beginning to foster new kinds of collaborative commitments while also being rooted in established archaeological sociotechnical structures.

References

Baxter, Rob, N Chue Hong, Dirk Gorissen, James Hetherington, and Ilian Todorov. 2012. "The Research Software Engineer." In Digital Research 2012, Oxford.

Schmidt, Sophie C, and Ben Marwick. 2020. "Tool-Driven Revolutions in Archaeological Science." Journal of Computer Applications in Archaeology 3 (1): 18–32. https://doi.org/10.5334/jcaa.29.