NFDI4Objects
Research Data Infrastructure
for the Material Remains of
Human History

TRAIL 3.2:

Poseidon 2.0: A data repository for genetic, bioarchaeological and archaeological data

Partner **Lead**: Matthias Renz (Kiel University), and Stephan Schiffels (Max Planck Institute for Evolutionary Anthropology)

Members: Stephan Schiffels (Max Planck Institute for Evolutionary Anthropology), Philipp Stockhammer (Max Planck Institute for Evolutionary Anthropology), Ben Krause-Kyora (Kiel University), Christoph Rinne (Kiel University), Toralf Kirsten (Mittweida University of Applied Sciences)

Contact Matthias Renz (Kiel University) / mr@informatik.uni-kiel.de

Summary

This TRAIL addresses the challenge of linking genetic data extracted from human remains together with archaeological context data and other analytic data such as from stable isotope analysis. Within the research data lifecycle, we address three aspects: discovery, integration and analysis. We start with the existing Poseidon framework hosted by MPI Leipzig (https://poseidon-framework.github.io/#/), which currently provides i) a minimal data format to package archaeogenetic data, ii) software to handle such packages and iii) an extensive public data repository including more than 5,000 ancient human genomes. Based on this framework, we aim to develop an extension which explicitly allows for cross-domain data linking to include analyses of stable isotopes and burial contexts (such as grave goods). To guide the development process, we will use several well-described sets of archaeological sites in Germany (Lech valley, Niedertiefenbach, Trebur and Lauchheim) for which cross-domain analytical information is either already published or soon to be published and contributed by the co-authors of this TRAIL. An important benchmark will be the ability of the extension to query these data with explicitly cross-domain research questions, by linking genetic and extended contextual information.

Description

By extending the existing framework of Poseidon, this TRAIL aims to develop a Data Service (DaS), together with a Software Application Service (SAS). Poseidon 2.0 combines a software tool with methods of cross-domain data fusion based on heterogeneous information networks. It will also facilitate development of a Discovery Service (DiS). We do this using data from dedicated archaeological sites, which has already been analysed (Knipper et al. 2017, PNAS; Mittnik et al. 2019, Science, Immel et al. 2021, Commun Biol, co-authors' work in preparation for publication). In total, this includes nearly 250 individuals with available cross-domain information such as genetics, stable isotopes, biological relationships and grave goods.

At present, the genetic data in Poseidon is stored in a dedicated standard format (PLINK), and the metadata as tab-separatedvalues (tsv) files, as well as Bibtex and Markdown. Both are human and machine readable. Poseidon package descriptions and metadata are compatible with the W3C-recommended CSV-on-the-web standard (CSVW). We will retain these formats.

The existing metadata format consists of a flat table which links every genetic sample to certain basic metadata like spatiotemporal coordinates. To enable cross-domain analyses we will extend this flat-table approach to a more hierarchical or graph-like layout that is able to link the entity layers involved (e.g. sites, graves, individuals and samples) with experimental or observational features.

Genome-wide data are very large (consisting typically of at least 1.2 million genetic markers for each sample). This is a challenge for the performance our software handling such datasets (e.g. merging, validating, extraction or computing summary statistics) and for version control and file serving in our public data repository.

Relevance

This TRAIL develops semantic links between data structures in order to integrate and describe the data together with its context information and the mutual relationships between the data entities. Based on this, we will create scalable discovery services (through our public repository) to support information retrieval and analysis tasks. The typical users of this extended framework will be scientists who generated the data in the first place and would like to make it available in an integrated and standardised way. Downstream users include scientists of all disciplines researching the human past. For example, by linking genetic data with information on grave goods, one can ask how status is shared among genetic relatives. And by linking genetics with stable isotope data like strontium, one can compare genetic ancestry with biogeographic origin. While such questions have been asked before for individual projects, we aim to make it possible to query published data routinely, bringing such datasets closer to the broad community of researchers.

In principle, our extension should not stop at isotopes and grave goods, but provide a general means to store and package analytic data for human remains and their associated burial contexts. As such, the methods to link the data semantically provide blueprints for other data repositories. The German Genome-Phenome Archive (GHGA) is an NFDI. GHGA is setting up a German "node" within the existing European Genome-Phenome Archive, which is Europe's largest existing repository for genotype data. We have contacted GHGA leadership and plan to "hook" our Poseidon repository into the GHGA.

All elements of FAIR will be addressed. As a package repository and standardise package format, Poseidon already makes archaeogenetic data more findable and accessible. Interoperability is enhanced with this project (by cross-linking domains). Analyses that use Poseidon and its tools are reproducible (all public packages are versioned and available from the server). One key mission of N4O is to makes data on objects interoperable. Semantic linking between experimental approaches to the same objects is a prime example of this, so the TRAIL will enable us to move forward with the entire initiative. This TRAIL will be conducted in cooperation with the SFB1266 TransformationsDimensionen – Mensch-Umwelt Wechselwirkungen in Prähistorischen und Archaischen Gesellschaften (Scales of Transformation: Human–Environmental Interaction in Prehistoric and Archaic Societies) and the ROOTS cluster of excellence, both at the Universität Kiel. In this way, the project will guarantee FDM, consolidation and semantic linking of (analytical/fieldwork) data.

Deliverables

Package format/schema: Existing package description and schema (https://github.com/poseidon-framework/poseidon2-schema) is extended with additional relations and appropriate identifier schemas to link the entities.

Package repository: server requests are added to the Poseidon-HTTP Server API (https://poseidon-framework.github.io/#/server) to obtain metadata on packages.

Software: command line software trident adapted to allow for the extended schemas.

Website/documentation: A landing page, server queries via a web interface and documentation for the extended version of Poseidon. This will be flexible and usable for all aspects of archaeological protocols, contributing to a knowledge commons based on the FAIR principles. The tool can easily be integrated into existing courses and trainee programmes, contributing to a broad competence framework.

Work plan

Topics	Months											
	1	2	3	4	5	6	7	8	9	10	11	12
Kick-off meeting: introduction of deliverables and responsibilities	■											
Clarification of the data, metadata and data structures in Poseidon and functionalities to be enabled Identification of the most appropriate data model for the heterogeneous information network	■	■	■									
After 3 months: 1st milestone meeting Discussion and evaluation of data model for the heterogeneous information network.												
Prototype implementation (proof of concept), development of data access and online service strategies				■	■	■	■	■	■			
After 9 months: 2nd milestone meeting Posidon users evaluate service functionality and quality												
Plan for providing services: StoS; DaS; DiS; web service for data access and analysis										■	■	■