Placing multi-modal, and multi-lingual Data in the Humanities Domain on the Map: the Mythotopia Geotagged Corpus

Voula Giouli, Anna Vacalopoulou, Nikolaos Sidiropoulos, Christina Flouda, Athanasios Doupas, Giorgos Giannopoulos, Nikos Bikakis, Vassilis Kaffes, Gregory Stainhaouer
ATHENA Research and Innovation Centre
{voula, avacalop, nsidir, cflouda, adoupas, giann, bikakis, vkaffes, stein}@athenarc.gr

Abstract

We present an infrastructure that comprises a multi-lingual and multi-modal corpus (i.e., a corpus of textual data supplemented with images and video) that belongs to the humanities domain along with a dedicated database (content management system) with advanced indexing, linking, and search functionalities. This infrastructure also includes a geotagging component and will be integrated into a platform aimed at defining personalized itineraries and providing, thus, a multi-faceted experience to visitors of Eastern Macedonia and Thrace in Northern Greece using mythology as a starting point.

Rationale & Scope

The Mythotopia corpus is being created as part of an infrastructure aimed at the development of an online platform offering a multifaceted view of the area using mythology as starting point. The mythological content is further enhanced with tangible and intangible elements that pertain to the domains of history, architecture, natural environment, culture, society, folklore, recreation, gastronomy, travel and tourism, leisure, and more (Vacalopoulou et al., 2021). The platform uses different types of data to facilitate search and retrieve functionalities based on several criteria, also offering the option of defining personalized itineraries in the area based on these criteria. Therefore, the problem of "multimodal location estimation" lies within the heart of the overall project.

Consequently, Points of Interest (POIs), i.e., geospatial entities that are (a) characterised by at least a name and a set of coordinates, and (b) correspond to a place of interest to end users, are deemed as core elements handled in the Mythotopia corpus. Places, facilities, artefacts, living entities (i.e., persons, plants, and animals), events, and even intangible cultural heritage items that may be placed on a map are deemed as POIs.

Corpus description

The corpus comprises three components:

- ✓ Sub-corpus of literary texts (LA, GRC, EL, EN) ➔ primary data + accompanying texts (EL, EN)
- ✓ The cultural component (images, videos) of artifacts coupled with texts (EL, EN) as accompanying material
- ✓ Sub-corpus in the domain of Travel, texts (EL, EN) and images as accompanying material.

Selected by experts based on certain criteria defined:

- ✓ relevance to the myths of the area;
- ✓ availability of primary data.

Textual data in the Travel domain are crated ad hoc.

Metadata were added to both primary data and accompanying material (texts, images). Annotations were integrated across the following pillars: (a) efficient documentation aimed at indexing and retrieval of the content; (b) interlinking of the various entries in the database; (c) placing certain entities on the map, and (d) modelling linguistic features of the textual data. More precisely:

- ✓ Documentation based on standards (TEI, CIDOC);
- ✓ Indexing based on vocabularies defined early in the project life-cycle;
- ✓ Linguistic annotation using dedicated tools, namely, ILSP pipeline (Papageorgiou et al, 2000) and UDPipe (Straka et al., 2019).
- ✓ Part of the textual data has been geotagged.

In total, the corpus currently comprises c. 130 images, and 450 (EL) texts that amount to 45K tokens.

Storing in a Database

The corpus is stored in a Database that was ultimately used for searching the data.

It serves as a content management system (CMS) and was built in-house.

Database functionalities:

- ✓ Documentation of texts, images and video data
- ✓ Interlinking of the data, i.e., texts with images, video, and accompanying texts
- ✓ Geotagging of the textual data referring to POIs.

The platform follows the database schema (depicted in Figure 1) and provides a menu containing the main data elements of the corpus: myths, reference texts, artistic representations, mythological figures (people), multimedia files and POIs. For each of these entities, the platform provides a common management mechanism.

To better account for quality assurance and security control, different accessibility rights are granted to users. At the lower level, annotators are allowed to access all corpus elements and can create, edit, and delete their own records. Editors, on the other hand, have the same access to system functionalities, but also the right to publish completed records.

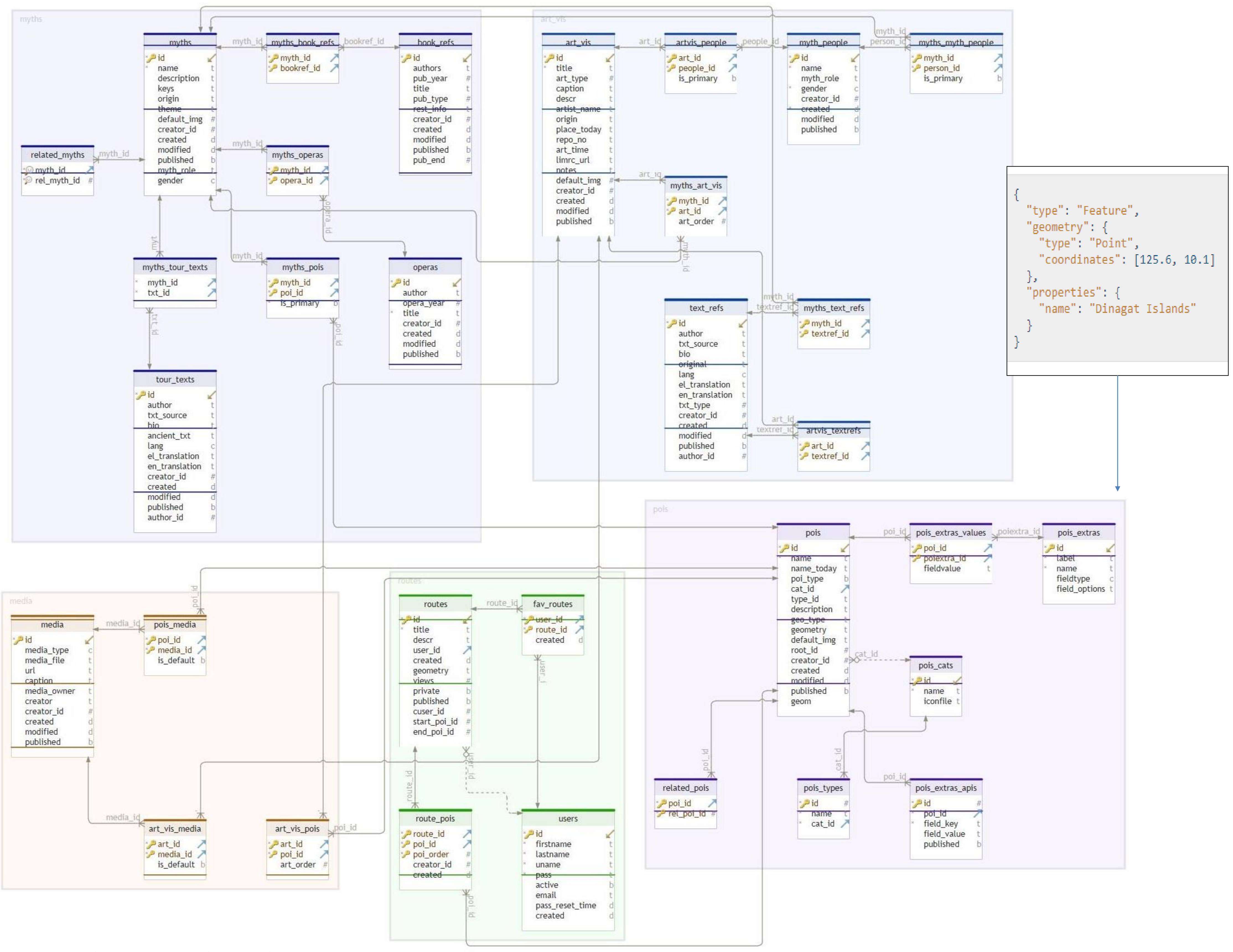
Geotagging the corpus

Besides documentation and linguistic annotation, the Mythotopia corpus was geotagged, i.e., textual data that pertain to the domain of Travel, that is, POI entries, are assigned geographical coordinates. Geottaging was performed manually via a dedicated functionality provided by our platform; Annotation

- ✓ interacting with the map and clicking on the corresponding place, or
- ✓ using the "Forward & Reverse Geocoding" API based on OpenStreetMap data. This API converts addresses into geocoordinates and vice versa. ➔ annotators can type the coordinates into the corresponding text field and the API returns the name of the place - if found.

After the marker shown on the map is saved, the geographic data are successfully stored in the database in GeoJson format, that is, a format that is used for encoding a variety of geographic data structures using JavaScript Object Notation (JSON).

By linking the POIs with the rest of the entries in the Database, this information is finally propagated to the rest of the data as well.



```
{
  "type": "Feature",
  "geometry": {
    "type": "Point",
    "coordinates": [125.6, 10.1]
  },
  "properties": {
    "name": "Dinagat Islands"
  }
}
```

Figure 1: The main Database schema

Creating routes

Sight-seeing scenarios based on users' preferences & restrictions

Input: the user provides a starting and an ending location + a set of POIs which they wish to visit

Output: the route generated comprises road segments/paths that need to be followed by the user, to visit all the desired POIs in a nearly optimal order

the system integrates several transportation networks extracted from OpenStreetMap (e.g., road networks, railways, boat routes) to generate routes based on the following criteria: (a) route optimization objective, (b) travel type, (c) visiting order mode, and (d) approximate route mode.

The routing functionality allows efficient route planning over multiple POIs (Kaffes et al., 2018) and involves the combination of shortest path algorithms, multi-criteria optimizations, approximation methods, and techniques for efficient route generation.

Conclusions

Mythotopia is a dataset that comprises a multilingual and multimedia corpus in the humanities domain. The corpus bears multi-layered and multi-faceted annotations; however, the novel feature is the fact that part of the textual data has been geotagged manually. The work is still in progress. Upon completion the corpus will be available via APOLLONIS, the Greek Infrastructure for Digital Arts, Humanities and Language Research and Innovation with open access.

Acknowledgements

We acknowledge support of this work by the project "Mythological Routes in Eastern Macedonia and Thrace" (MIS 5047101) which is implemented under the Action "Regional Excellence", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund)

References (selected)

Kaffes V., Belesiotis A., Skoutas D., Skiadopoulos S. (2018) Finding shortest keyword covering routes in road networks. In *Proceedings of the International Conference on Scientific and Statistical Database Management*.

Papageorgiou, H., Prokopidis, P., Giouli, V., and Piperidis, S. (2000). A unified tagging Architecture and its Application to Greek. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resource Association (ELRA).

Vacalopoulou, A., Mastrogianni, A., Michalopoulos, C., Tsiafaki, D., Michailidou, N., Mourthos, I. Botini, P., and Stainhaouer G. (2021). Mythological Itineraries Along the Western Silk Road: Finding Myths in Visits to Eastern Macedonia and Thrace Today. In *Silk Road Sustainable Tourism Development and Cultural Heritage*. The University of Thessaloniki and European Interdisciplinary Silk Road Tourism Centre.

Straka, M., Straková, J., Hajič, J. (2019): Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing. In: ArXiv.org *Computing Research Repository*, ISSN 2331-8422, 1904.02099.