Voula Giouli/Anna Vacalopoulou/
Nikos Sidiropoulos/Christina Flouda/Athanasios Doupas/
Gregory Stainhaouer

FROM MYTHOS TO LOGOS: A BILINGUAL THESAURUS TAILORED TO MEET USERS' NEEDS WITHIN THE ECOSYSTEM OF CULTURAL TOURISM

Abstract Thesauri have long been recognized as valuable structured resources aiding Information Retrieval systems. A thesaurus provides a precise and controlled vocabulary which serves to coordinate data indexing and retrieval. The paper presents a bilingual Greek and English specialized thesaurus that is being developed as the backbone of a platform aimed at enhancing and enriching the cultural experiences of visitors in Eastern Macedonia and Thrace, Greece. The cultural component of the intended platform comprises textual data, images of artifacts and living entities (animals and plants in the area), as well as audio and video. The thesaurus covers the domains of Archaeology, Literature, Mythology, and Travel; therefore, it can be viewed as a set of inter-linked thesauri. Where applicable, terms and names in the database are also geo-referenced.

Keywords Thesaurus; cultural heritage resources; content management platform; controlled vocabularies; bilingual resources

1. Introduction

Cultural tourism tends to be a learning-intensive experience, in the sense that visitors most often wish to

> learn, discover, experience, and consume the tangible and intangible cultural attractions and products in a tourism destination. These attractions/products relate to a set of distinctive material, intellectual, spiritual, and emotional features of a society that encompasses arts and architecture, historical and cultural heritage, culinary heritage, literature, music, creative industries and the living cultures with their lifestyles, value systems, beliefs, and traditions.[1]

In this regard, the project seeks to address the visitors' needs to learn and experience virtual itineraries in Eastern Macedonia and Thrace, Greece, by developing a platform that integrates cultural heritage content relative to this area. In the paper, we present work aimed at building a thesaurus that is tailored to meet the needs of indexing and retrieval of the cultural heritage content in the resulting platform.

The paper is structured as follows: In section 2 we will give a brief account of the project in terms of its scope, aims, and expected results. Then, the cultural content will be presented in section 3 along with the web interface that serves as a content management system. The focus will be on the description of the thesaurus developed, the design principles it abides to, as well as the twofold purpose it serves in the project and the final platform (section 4). We will then elaborate on the methodological approach taken for creating the thesaurus and

[1] This definition was adopted by the United Nations World Tourism Organization General Assembly, at its 22nd session (2017).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/
Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society.
Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

controlled vocabularies (section 5). In section 6 we will present background on thesauri and controlled vocabularies aimed at cultural heritage content management. Finally, our conclusions and plans for future research will be outlined in section 7.

2. Project aims and scope

Mythotopia (short for "Mythological routes in East Macedonia and Thrace") is currently a Research & Development project in progress, aimed at developing an online platform, which offers a multi-faceted view of Eastern Macedonia and Thrace, Northern Greece. This includes a wealth of information from several points of view including mythology, history, architecture, natural environment, culture, society, folklore, recreation, gastronomy, travel and tourism, and leisure. The primary aim of Mythotopia is to bring to light the cultural wealth of the region in the form of a bilingual Greek (EL) and English (EN) informative platform, in view of eventually contributing to its tourism development (Vacalopoulou et al. 2021). In order to achieve this aim, Mythotopia records, maps, and highlights local cultural and tourist attractions under the scope of mythology. Users of Mythotopia will be given the chance to combine elements of interest connected to the current natural, social, and cultural landscape with the rich mythological background of the area to create their preferred virtual routes. They can, then, use their selected routes and the extensively researched accompanying information as guides in actual visits on site. By combining these elements, Mythotopia attempts to join the past and the present to help create a complete tourist experience (ibid.).

3. The cultural tourism ecosystem: the content and web application

Project development is centred around two main pillars: (a) the cultural content relative to the areas of Eastern Macedonia and Thrace, which is multi-faceted and highly heterogenous in nature, and (b) the resulting platform, that is, a Graphical User Interface (GUI) for integrating, documenting, geo-tagging, storing, and retrieving data based on specific metadata elements. In between, the thesaurus serves as the backbone of the platform and helps indexing and retrieving the content.

3.1 The platform content

Following project specifications, the data types stored and, therefore, documented in the platform comprise: (a) myths that are relevant to the area of Eastern Macedonia and Thrace; (b) Points of Interest (POIs), that is, a variety of tangible and intangible elements that are characteristic of the area pinpointing its cultural, social, and environmental identity; and (c) audio-visual material.

To start with, ancient Greek myths related to the area are the primary data to be stored and documented in the platform in the form of narratives. These myths are centred around specific living entities, as for example deities, heroes, and mythological creatures; moreover, they are linked to specific localities (places), religious or ritual practices, and the culture of the time. Ancient myths have been made known via the seminal literary works of ancient Greek and Latin authors, and they have been also referred to in travel memoirs of the past.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/
Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society.
Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Therefore, a significant part of the mythological component consists of ancient Greek and Latin literary texts featuring myths or mythological figures and localities that are linked to the area. The literary texts are excerpts that were selected from established scholarly editions; they cover a variety of genres and a wide range of Ancient Greek and Latin literary production both in prose (historiography, myth-writing, biography, rhetoric, philosophy, and scientific texts, such as geographical works and ancient scholia) and in verse (epic, drama, elegy, epigram, lyric poetry, bucolic poetry, and didactic poetry). Moreover, texts that pertain to the genre of travel literature, that is, travel writings produced by authors who visited the area of Eastern Macedonia and Thrace both in ancient times (i. e., Pausanias, Strabo), and in Medieval and Modern times (i. e., Buondelmonti, Ciriaco d'Ancona) were also selected for inclusion. The Ancient Greek and Latin literary texts are stored in the original; consequently, their translations in EL and EN are also provided by the project participants. Additionally, accompanying textual material in the form of biographies of authors, general information about the mythological figures is also developed by the project participants and included in the raw data. Finally, the corpus comprises narrative texts in EL and their translations in EN depicting the myths that are of relevance to the broader area of interest. To date, a total of c. 300 texts has been collected or produced.

Along the same lines, visitors may gain valuable insights of the culture and myths related to the area by means of artifacts of all sorts. Thus, images of clay pots, sculptures, engravings, coins, various metal and glass objects, mosaics, sarcophagi, etc. portraying figures, or scenes of the myths, dating back to Greek and Roman antiquity, were also selected. In addition, the collection includes references to excerpts of classical operas depicting mythological scenes and stories. Finally, informative texts, accompanying images, and videos tailored to meet the needs of tourists visiting the area have also been included in the collection. In terms of content or subject matter, the textual and audio-visual content features primarily entities of the following types: living entities (i. e., animals, plants that are endemic in the area), geopolitical entities (i. e., cities, towns, villages, or minor settlements), geographical entities (e. g., mountains, rivers, beaches, lakes), facilities and archaeological sites, events, and activities, as well as intangible cultural elements (i. e., food and gastronomy, folklore, and cultural events of the area). In essence, these entities constitute the body of POIs of the resulting application.

Besides the various annotations applied to the data (Giouli et al. 2022), the development of a taxonomy of object types was of paramount importance in view of better accounting for the heterogeneous dataset stored in the project database. At the same time, according to the specifications of the project, the informative material that was initially prepared in Greek was to be translated to English. The controlled vocabulary that was developed also helped in the translation process, as will be shown in section 4.1 below. Therefore, the purpose of thesaurus creation was two-fold: on the one hand, it was aimed at documenting and indexing the primary data, facilitating retrieval over the database constructed; and on the other hand, it assisted translators towards ensuring consistency throughout the translation phase.

3.2 Collaboratively storing and managing data: the database and web application

A web platform was developed as the front-end of a database for collaboratively storing, documenting, and searching the cultural content. The database supports various modalities (text, images, audio, and video) in two languages: EL and EN. In addition, it was designed to meet the following requirements as a minimum: (a) user friendliness for both people who are responsible for data entry and the end users who will use the taxonomy for browsing and retrieving information; (b) extensibility, that is, the taxonomy should be open and easily modified or extended in view of incorporating new datatypes in the future; (c) functionality, that is, the thesaurus and the overall platform should provide certain functionalities for inter-linking entries, as well as managing the workflow from the initial entry of data to the final publishing of entries; and (d) scalability, the system must deal with (relatively) large amounts of data, still offering a friendly environment. Therefore, the platform was designed so as to meet a series of crucial requirements that support end outcome quality and user satisfaction.

Items stored in the platform (myths, artefacts, POIs) were indexed with respect to a location term provided in the thesaurus. In addition, one novel feature of the digital collection lays in the fact that part of the cultural content has also been geotagged manually. More precisely, POIs have been assigned geospatial metadata in the form of geographical coordinates using OpenStreetMap and Leaflet,[2] a light, open-source JavaScript library that works across all major desktop and mobile platforms.

The web application is designed to provide access to the main taxonomy objects in an order of importance as a part of a workflow from the main taxonomy to secondary ones. As a starting point, myths constitute the main or central taxonomy; therefore, they are managed first. All the other taxonomies (literature, media, POIs) are imported and connected to the main taxonomy. However, during the data entry period, a more myth-independent approach was also employed, and secondary – or less central - taxonomies were submitted to the database on their own, and then editors provided inter-connections. This proved to be faster and more efficient in creating and managing the final collections.

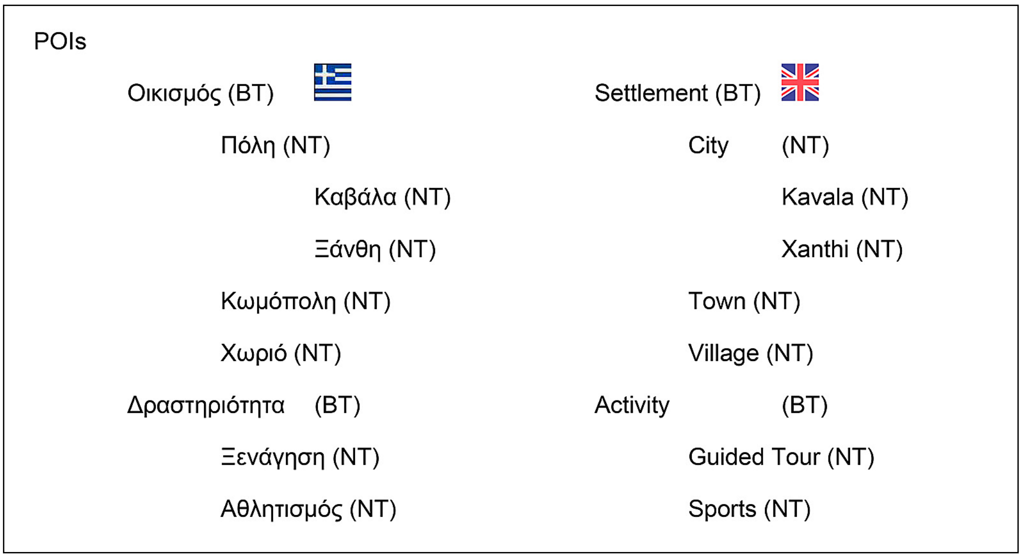
4. Thesaurus description

In this section, we will elaborate further on the thesaurus that supports this application. The major issue the project team had to cope with when defining the taxonomy, was the heterogenous nature of the dataset at hand. To overcome this challenge, we initially identified the subject fields or domains that different types of data fall into; each domain was treated as a separate thesaurus thereof with inter-dependencies where applicable. Thus, the following fields were identified: literature, archaeology, arts, and travel. The backbone of the thesaurus is an upper-level taxonomy that starts from the aforementioned fields and depicts a set of classes and subclasses that correspond to the concepts of the taxonomic schema. Ultimately, the concepts or classes are being populated with domain-specific terms.

Following standard specifications for thesauri creation (Aitchison/Gilchrist/Bawden 2000), a set of relations have also been defined that hold between concepts, between concepts and terms or between terms. Thus, three types of relations are foreseen: semantic, spatial, and associative ones. These ultimately create a semantic network that shows links and paths between terms and concepts. The standard semantic relations that have been defined, namely the hierarchical relation (between concepts) and the equivalence one (between terms), designate the vertical and the horizontal axes of the taxonomic schema respectively. The former deals with defining the hierarchical structure of the thesaurus. Concepts at the upmost level define the basic categories or classes of cultural data objects. A total of 37 classes have been

[2] https://leafletjs.com/.

defined in our schema; these are instantiated as the Broad Terms (BTs) of the taxonomic schema. The hierarchical relation in the thesaurus links these concepts with their subordinate ones, the so-called Narrow Terms (NTs), establishing thus the hierarchical structure in a thesaurus. In total, 144 subclasses have been stipulated, which are further grounded to the terms of the thesaurus.



POIs	
Οικισμός (BT) 🇬🇷	Settlement (BT) 🇬🇧
Πόλη (NT)	City (NT)
Καβάλα (NT)	Kavala (NT)
Ξάνθη (NT)	Xanthi (NT)
Κωμόπολη (NT)	Town (NT)
Χωριό (NT)	Village (NT)
Δραστηριότητα (BT)	Activity (BT)
Ξενάγηση (NT)	Guided Tour (NT)
Αθλητισμός (NT)	Sports (NT)

Fig. 1: Snippet of the hierarchical structure of the thesaurus

Concepts at the first two upmost levels are unambiguously grounded to terms in Greek and English. Moreover, guidelines elaborated in the framework of the project further specify the content of these concepts. For example, types of settlements are distinguished with respect to the number of inhabitants, whereas official records are provided to the people involved in the task of data entry and indexing. This is not the case with the concepts at the downmost level of the hierarchy. For these concepts, a preferred term is employed assuming the role of the descriptor. According to the specifications set, the descriptor is the unmarked form of a term, that is, one that pertains to general language, as opposed to dialectical or otherwise marked terms. Historical and geographical variations of terms are also encoded in the thesaurus as alternative terms or non-descriptors. The equivalence relation is then established between a descriptor and one or more non-descriptors. This relation is of particular importance for the data at hand and the final application since it is used in computing itineraries. The thesaurus encompasses more than 1820 terms, of which c. 1,400 are descriptors. Apart from terms, a set of 139 keywords or key phrases featuring myth-related concepts have been so far included in the thesaurus for better indexing the mythological narratives.

In addition, associative relations are also specified in the thesaurus. The generic relation RelatedTo (RT) is further specialised in a set of sub-relations, each one specifying the items related. Therefore, the relation RelatedTo-POI is used to link an item documented in the platform with one or more POIs in the database; the relation RelatedTo-image (RT-image) is used to link an item with an image, whereas RelatedTo-file and RelatedTo-author are used to link myths with literary and travel texts and corresponding authors respectively. Finally, spatial relations are also foreseen. More precisely, the relation Area links one of the regional units of the area with the POIs in the database. Where applicable, terms in the database are also geo-referenced. Figure 2 illustrates a snippet of the model for the description of POIs.

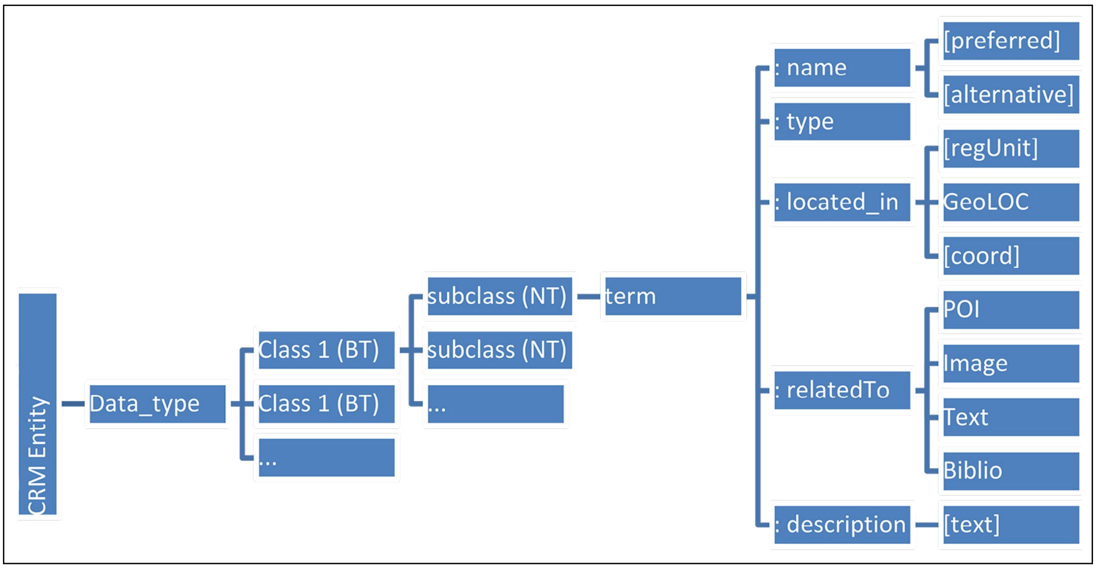


Fig. 2: A sample of the taxonomic schema with relations

We opted for establishing hierarchical and associative relationships between concepts rather than between terms, since this has been proved as a prerequisite for thesaurus interoperability (Dextre Clarke/Zeng 2012).

Apart from common nouns relating to concepts and terms, the thesaurus also includes proper nouns that is, names of mythological characters of the area of Eastern Macedonia and Thrace (heroes, gods, etc.), as well as names of artists, and place names (i. e., names of geopolitical and geographical entities). Historical and dialectal variants of toponyms are also included as terms; moreover, dialectal as well as scientific names of entities (i. e., animals and plants) are also retained.

4.1 Using controlled vocabularies to bridge languages and domains

To achieve the highest possible level of control, standardisation and homogeneity throughout the collection, detailed guidelines were elaborated specifying linguistic choices and presentation. These include choice and ordering of terms, selection of descriptors over non-descriptors, orthographic conventions, grammatical forms, capitalisation policy, abbreviations, and acronyms, as well as usage of punctuation marks.

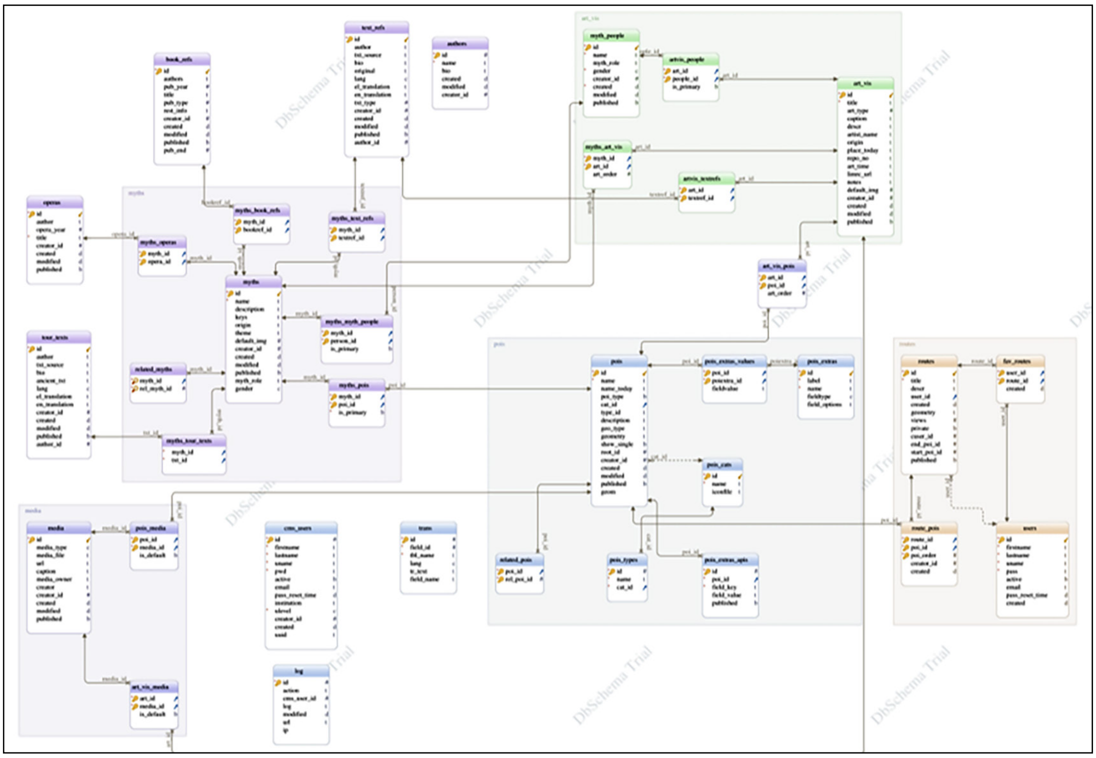


Fig. 3: Interlinking of entities springing from the central "myths" table in a small snapshot of the thesaurus architecture

Interlinking of data or entries in the platform via the relations of terms in the thesaurus results in bridging domains which seemed unrelated. For example, the myth of Voreas in the mythological component of the platform is linked to literary works (Literature) and images of Archaeological artifacts (Archaeology). By linking it to the item Pangeon (a mountain that is documented in the Travel domain), a bridge is established between the two domains (Travel and Literature). This is depicted in Fig. 3, in which the myth of Voreas belongs in the central "myths" table, which is linked through intermediate tables to "text_refs" (where the respective literary works are stored), to "art_vis" (where images of related artifacts are kept), to "pois" (populated by related POIs, in this case, Pangeon, a geographical element and the respective bridge, a transport element).

From another perspective, the controlled vocabularies were also consulted when translating the platform content from Greek to English in an attempt to achieve accuracy as well as consistency throughout the translated content. As a first step, the establishment of translational equivalents between the controlled vocabularies in the two languages was in order. This has not always been an easy task, especially in the case of translating specialized terms, as for example those referring to plant and animal species and sub-species. In this case, the translation of the descriptors was based on non-descriptors, namely, the scientific names in Latin.

5. Methodology for thesaurus construction

The main activities within the project life cycle can be outlined as follows: 1) identification and road mapping of the data types that will be stored in the platform (myths, POIs, audio-visual material) and the sources for acquiring them; 2) definition of the metadata elements that are appropriate for documenting each one of them; 3) selection of the appropriate concepts and terms that were relevant to the dataset, and 4) building the platform and deciding upon its functionalities taking into consideration not only the people who are responsible for data entry, but also the prospective users of the platform.

The project team opted for a hybrid approach to thesaurus creation: a top-down development phase was complemented by a bottom-up one. In our top-down approach, a list of seed terms that roughly correspond to our taxonomy was initially postulated. Using the textual data that was collected and produced within the project, we then populated the classes with instances; revisions and extensions of the initial schema were made where needed. For our bottom-up procedure to thesaurus creation, the textual data collected served as a corpus; terms were selected for inclusion from the data. The terms were mined from reliable sources that had been selected early in the project life cycle. Our work then involved defining the hierarchy and rest of relations, designating the preferred term or descriptors of the concepts and its variant term(s). Following standard methodologies, and previous best practices for thesaurus creation (Aitchison/Gilchrist/Bawden 2000), a set of guidelines were implemented to ensure the usage of controlled vocabularies throughout the lexicographic process.

The thesaurus broadly covers the domains of Archaeology, Literature, Mythology, and Travel. From another perspective, it can be viewed as a set of inter-linked thesauri. The main challenge that we had to deal with when constructing the thesaurus, however, was the high degree of fragmentation between the various domains and datatypes that we needed to include in the platform. In this respect, we tried to be inclusive and develop self-sustained thesauri that would be adequate for each domain, allowing for overlapping concepts/classes across domains if needed.

6. Background on thesauri and controlled vocabularies

Over the last decades, thesauri have been developed and utilized by digital content holders and providers, such as digital libraries, archives, and museums. These are mainly characterized by a considerable degree of specialization and granularity in distinguishing and layering concepts. For example, UNESCO Thesaurus[3] is a controlled and structured list of terms used indexing and retrieval of publications in seven subject fields or domains, namely education, culture, natural sciences, social and human sciences, communication, and information; each subject field is broken down further into micro-thesauri which allow users to gain a quick overview of the subject matter. With its first release dating back in 1977 in English, the thesaurus has over the years been extended to also include French, Spanish and Russian. Similarly, the CIDOC Conceptual Reference Model (CIDOC CRM)[4] is a theoretical and practical tool for information integration in the field of cultural heritage (Bekiari et al. 2021). The Getty Vocabularies contain structured terminology for art, architecture, decorative arts, archival materials, visual surrogates, art conservation, and bibliographic materials. Compliant with international standards, they provide authoritative information for cataloguers, researchers, and data providers. However, the decision for building our own taxonomy instead of adopting an existing one was essentially directed by the nature of our data. Being highly heterogenous, not only in form (texts, video, audio) but also in terms of the domain (or even the genre in the case of texts) they pertain to, they were difficult to organize.

[3] https://skos.um.es/unescothes/.

[4] https://www.cidoc-crm.org/Version/version-7.2.

7. Conclusions and future research

We have presented a thesaurus that has been elaborated in view of indexing and retrieval of cultural heritage content over a dedicated platform aimed at enhancing tourists' experiences in the area of Eastern Macedonia and Thrace. The core of the thesaurus is a collection of concepts represented by terms and interlinked by relationships. These concepts pertain to the domains of Archaeology, Literature, and Travel. Terms in the thesaurus have already been employed to index text, video and audio that is featured on the touristic platform. Future work has been already planned towards enriching the thesaurus with new terms. Moreover, to render our taxonomic schema compatible with existing digital libraries and other state-of-the-art conceptual models (i. e., CIDOC CRM), mappings from our schema to standardized ones are underway. The resulting platform and the thesaurus will be freely available and accessed over the web.

References

Aitchison, J./Gilchrist, A./Bawden, D. (2000): Thesaurus construction and use: a practical manual. 4th edition. London.

Bekiari, C./Bruseker, G./Doerr, M./Ore, C.-E./Stead, S./Velios, A. (2021): Definition of the CIDOC Conceptual Reference Model. Version 7.2.
https://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_7.2.pdf (last access_ 20-06-2022).

Dextre Clarke, S. G./Zeng, Marcia Lei (2012): From ISO 2788 to ISO 25964: the evolution of thesaurus standards towards interoperability and data modelling. Information Standards Quarterly 24, 1 (Winter), pp. 20–26.

Garshol, L. M. (2004): Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. In: Journal of Information Science 30 (4), pp. 378–391.

Giouli, V./Vacalopoulou, A./Sidiropoulos, N./Flouda, C./Doupas, A./Giannopoulos, G./Bikakis, N./ Kaffes, V./Stainhaouer, G. (2022): Placing multi-modal, and multi-lingual data in the humanities domain on the map: the Mythotopia Geo-tagged Corpus. In: Proceedings of the 13th Chapter of the International Language Resources and Evaluation Conference (LREC 2022), 20–25 June 2022, Marseille, France.

Nielsen, M. L. (2004): Thesaurus construction: key issues and selected readings. In Cataloging & Classification Quarterly 37 (3–4), pp. 57–74.

Ryan, C. (2014): Thesaurus construction guidelines: an introduction to thesauri and guidelines on their construction. Dublin. DOI: 10.3318/DRI.2014.1.

UNESCO and International Bureau of Education (IBE) (1984): IBE education thesaurus: a list of terms for indexing and retrieving documents and data in the field of education – with French and Spanish equivalents. Paris.

Vacalopoulou, A./Mastrogianni, A./Michalopoulos, C./Tsiafaki, D./Michailidou, N./Mourthos, I./ Botini, P./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.

Contact information

Voula Giouli
Institute for Language and Speech Processing, ATHENA RC
voula@athenarc.gr

Anna Vacalopoulou
Institute for Language and Speech Processing, ATHENA RC
avacalop@athenarc.gr

Nikos Sidiropoulos
Institute for Language and Speech Processing, ATHENA RC
nsidir@athenarc.gr

Christina Flouda
Institute for Language and Speech Processing, ATHENA RC
cflouda@athenarc.gr

Athanasios Doupas
Institute for Language and Speech Processing, ATHENA RC
adoupas@athenarc.gr

Gregory Stainhaouer
Institute for Language and Speech Processing, ATHENA RC
stein@athenarc.gr

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).