DIGITAL CLASSICS ONLINE

Band 12,2 (2026)

ISSN: 2364-7957

DIGITAL CLASSICS ONLINE

Bd. 12,2 (2026)

Table of Contents

DIGITAL CLASSICS ONLINE

Bd. 12,2 (2026)

Table of Contents

DIGITAL CLASSICS ONLINE

Nomina Omina: Ancient Greek and Latin Proper Names in the Age of Artificial Intelligence – Introduction

Monica Berti

Abstract: This paper is the introduction to the volume *Nomina Omina. Detecting and Preserving Ancient Greek and Latin Proper Names in the Age of Artificial Intelligence*, which collects the proceedings of a workshop held at Leipzig University in June 2024, thanks to the support of the German Research Foundation (Deutsche Forschungsgemeinschaft). This introduction outlines the goals and contributions of the workshop and the volume.

Nomina Omina – The Workshop

This volume is a collection of papers published as proceedings of the international workshop entitled *Nomina Omina. Detecting and Preserving Ancient Greek and Latin Proper Names in the Age of Artificial Intelligence*.[1] The workshop was held at Leipzig University from June 27 to 29, 2024, thanks to the generous support of the German Research Foundation (Deutsche Forschungsgemeinschaft – DFG) and as part of the project *Text-based Extraction, Analysis, and Annotation of Ancient Greek References to Authors and Works*.[2]

The goal of the workshop was to discuss the current state of data and research concerning proper names in ancient Greek and Latin, with a focus on the computational extraction of personal names, geographic names, and proper names related to authors and works cited in ancient sources. This kind of data is essential to studies in disciplines such as linguistics, philology, and historiography, and, in fact, traditional scholarship has consistently published repertories, lexica, and indices to collect it. The advancement of Digital Classics is showing the importance of this kind of data not only for traditional purposes, but also for computational ones, given that proper names function as anchors in the text to structure unstructured data.[3] This is one of the reasons why the so-called Named Entity Recognition (NER) technique is gaining growing interest in the Classics community, with projects devoted to extracting and annotating named entities (i.e., proper names) in ancient sources.[4]

The challenge with historical languages is due to their complexities, the fact that they are no longer spoken, and their high inflection, as in ancient Greek and Latin. As a matter of fact, data for proper names in the original language is still scarce, considering also that traditional print dictionaries, from which digital data is digitized, are not rich in proper names. Research in recent years has been focusing on this aspect to expand corpora of proper names, providing data in the original languages and not

1 The cover image was generated using Envato AI and licensed via Envato.

2 DFG Project No. 434173983. On this project, see Berti (2023b) and Berti (2024a).

3 See Berti (2019b) and Berti (2021), 398–414 with bibliography.

4 Berti (2019a), Berti (2019b), Beersmans et al. (2023), Beersmans et al. (2024), Palladino / Yousef (2024), Berti (2025b), Berti (2025c), Berti (2026a), Berti (2026b), Berti (2026c). Cf. also Ripoll Alberola et al. (2025).

just in modern contemporary ones. The goal is to enrich datasets with inflected forms and their lemmata, and annotate and represent this data according to the best recommendations of the Linked Open Data (LOD) principles.[5] Moreover, current new possibilities offered by Large Language Models (LLM) and Artificial Intelligence (AI) request scholars to increase the number of data and control their quality in order to feed and train machines, so that in the future it will be possible to parse more texts and get better results.[6]

The workshop was structured in four sessions (see fig. 1) to introduce the concepts of Named Entities (NEs) and Linked Open Data (LOD) (1), and to deal specifically with Onomastics and Prosopography (2), Geography (3), and Authors and Works (4). International specialists currently working on these topics were invited to Leipzig to discuss various aspects of these themes.[7]

Fig. 1: *Nomina Omina* Workshop – the program.

5 On LOD for the Ancient World, see Cayless (2019) and Middle (2024).

6 Berti (2025c) and Berti (2026a).

7 Links to online resources, projects, and corpora cited in the following paragraphs can be found at the end of the paper in the section listing online resources.

After the introduction to the workshop, the first session (Named Entities and Linked Open Data for the Ancient World) was devoted to the current state of research for Named Entity Recognition (NER) of ancient Greek and Latin, on linguistic annotation as a foundational method to generate data for extracting named entities, and on digital projects born with the digitization of print publications and the annotation of proper names. Andrea Beyer (Humboldt-Universität zu Berlin) described methods developed at the DFG project *Daidalos* for Named Entity Recognition (NER) in ancient Greek and Latin, while Carina Geldhauser (Munich Center for Machine Learning – MCML) showed examples for *nomina sacra*; Giuseppe G. A. Celano (Universität Leipzig) presented the current state of linguistic data for ancient Greek in the DFG project *Opera Graeca Adnotata* (*OGA*); finally, Adam Gitner (Bayerische Akademie der Wissenschaften) introduced the work done at the *Thesaurus Linguae Latinae* (*TLL*) project to extract and annotate proper names and bibliographic citations from the *Index Librorum*, and Irine Darchia (Ivane Javakhishvili Tbilisi State University) described Georgian projects to extract data from epigraphic, lexicographic, and encyclopedic material.

The second session (Onomastics and Prosopography) included three projects on different kinds of personal names. Matilde Garré (Université de Paris 1 Panthéon–Sorbonne) presented the work done for linguistic annotations of personal names in the project *LGPN-Ling*, which is a semantical complement to the *Lexicon of Greek Personal Names* (*LGPN*) Database, while Yanne Broux (KU Leuven) and Sylvain Lebreton (Université Toulouse – Jean Jaurès) introduced projects working on divine onomastics, such as *Trismegistos People*, *Trismegistos God*, and the *Mapping Ancient Polytheisms* (*MAP*) database.

The third session (Geography) was devoted to projects and methods for extracting toponyms from ancient sources. Pietro Zaccaria (KU Leuven) and Monica Berti (Universität Leipzig) presented the *TM Paradoxography* project of *Trismegistos* to extract place names from paradoxographical texts, and Chiara Palladino (Furman University) discussed geographical entities in ancient texts.

The fourth and last session (Authors and Works) had a focus on the individuation and extraction of proper names related to names of authors and forms referring to their works in ancient sources, which can be defined as bibliographic references in past texts. Margherita Fantoli (KU Leuven) and Annette von Stockhausen (Berlin-Brandenburgische Akademie der Wissenschaften) discussed this kind of entities in the two projects *NIKAW* (*Networks of Ideas and Knowledge in the Ancient World*) and *Patristic Text Archive* (*PTA*); Andrea Balbo and Elisa Della Calce (Università degli Studi di Torino) presented the research currently done to extract proper names from Latin sources on China dated between Hellenistic times and the 19th century in the project *SERICA*, with a focus on the *Confucius Sinarum Philosophus*; finally, Ivan Matijašić (Università Ca' Foscari Venezia) discussed proper names in Greek fragmentary historiography with examples from Attidography.

The *Linked Ancient Greek and Latin* (*LAGL*) Project

The selection of the aforementioned topics and sections could have been expanded to include other aspects of Named Entity Recognition and the analysis of proper names. However, their choice was driven by two main reasons: 1) current needs and discussions in the community of Digital Classicists to improve NER and increase data for ancient Greek and Latin; 2) the project I am currently leading, which is called *Linked Ancient Greek and Latin* (*LAGL*) and which is extracting named entities from ancient Greek and Latin sources, and selecting those related to bibliographic references.[8]

Considering the still limited availability of digital data for ancient Greek personal names, the *LAGL* project was established to develop a workflow for extracting these forms from ancient sources, lemmatizing them, and labelling them according to the NE categories employed in computational linguist-

8 https://www.lagl.org (last access 23.09.2025).

ics, such as PER and PERderiv for personal names and derivates (e.g., Θεμιστοκλῆς and Πυθαγορικός), LOC and LOCderiv for place names and derivates (e.g., Ἁλικαρνασσός and Μυτιληναῖος), ORG and ORGderiv for organizations and derivates like festivals and schools (e.g., Παναθήναια and Στωϊκός), and OTH for other entities such as currencies, months, work titles, etc. (e.g., Ἱστορίαι and Ὀκτώβριος).[9] These broad categories allow us to generate training data that can be used to train computational models for NER and extract semi-automatically proper names from ancient sources. For example, the form Λαοδίκειαν can be extracted, lemmatized as Λαοδίκεια, and labeled as a location (LOC). Through the lemma, it is possible to individuate other inflected forms, like for example Λαοδικείας, that can be further disambiguated in context as Laodicea in Syria or Laodicea on the Lycus in Asia Minor.[10]

A second part of the project is the *LAGL Catalog*, which collects linguistic forms about author names and their works cited in ancient sources.[11] This analysis is related to citation detection and fragmentary literature, even if the focus is not on the content of citations (i.e., the so called fragments), but on those elements that signal the presence of quotations, like names of authors, titles and descriptions of their works, and indirect references to them, such as unnamed authors and works.[12] For this kind of analysis, NER is particularly helpful, because named entities function in the text as anchors to individuate and disambiguate further elements of bibliographic references. This type of research is rooted in my extensive experience with fragmentary texts and citation analysis, which reveal the numerous linguistic forms concealed behind modern scholarly labels like *fragmenta*, *testimonia*, *reliquiae*, etc.[13]

These linguistic forms are important to explore the ancient language of proper names and bibliographic references, advance our knowledge of citation practices in the ancient world, and enrich digital data in the original language. An example among the many others that I could mention is Duris (Δοῦρις: PER), the ancient Greek historian and tyrant of the island of Samos, who is connected by ancient sources to a great variety of named entities. Duris, in fact, is not only labelled as Samius (Σάμιος: LOCderiv) and tyrant of Samos (Σάμου: LOC), but also as a pupil of Theophrastus of Eresus (Θεοφράστου Ἐρεσίου: PER|LOCderiv), brother of the author Lynceus of Samos (Λυγκεὺς Σάμιος: PER|LOCderiv), descendant of Alcibiades (Ἀλκιβιάδου: PER), and probably relative of the Kaios (Καῖος: PER) victorious in the boy's boxing contest and commemorated with a statue in Olympia (Ὀλυμπίαι: LOC) made by a certain Hippias (Ἱππίου: PER).[14] The name of Duris is cited in lists of ancient Greek and Latin sources with authors like Phylarchos (Φύλαρχον: PER), Polybios (Πολύβιον: PER), Psaon (Ψάωνα: PER), Demetrios of Kallatis (Καλλατιανὸν Δημήτριον: LOCderiv|PER), Hieronymos (Ἱερώνυμόν: PER), Antigonos (Ἀντίγονον: PER), Herakleides (Ἡρακλείδην: PER), Hegesianax (Ἡγησιάνακτα: PER), Hellanicus (Ἑλλάνικοί: PER), Onesicritus (Onesicrito: PER), Clitarchus (Clitarcho: PER), Ctesia (Ctesia: PER), Philistus (Philisto: PER), and many others.[15] Moreover, references to Duris' works show his heterogeneous interests that ranged from Homer (Προβλημάτων Ὁμηρικῶν: OTH|PERderiv), to tragedians like Euripides (Εὐριπίδου: PER) and Sophocles

9 Berti (2019b), Berti (2023b), Berti (2025b).

10 On the two ancient cities, see the gazetteer entries in Pleiades at https://pleiades.stoa.org/places/668290 (last access 23.09.2025) and https://pleiades.stoa.org/places/638955 (last access 23.09.2025). For examples of sources mentioning the two places, see *Suda*, alpha 3398 (s.v. Ἀπολινάριος) and gamma 450 (s.v. Γρηγόριος).

11 https://catalog.lagl.org (last access 23.09.2025).

12 Berti (2013), Berti (2018), and Berti (2025a).

13 Berti (2021), Berti (2023a), and Berti (2024b) with examples, bibliography, and projects related to this topic.

14 For the ancient sources on these assertions and commentaries on them, see FGrHist (= BNJ) 76 and BNP s.v. Duris (3).

15 See FGrHist (= BNJ) 76.

(Σοφοκλέους: PER), and to historical works on Samos (Σαμίων Ἧροι: LOCderiv|OTH), Agathocles (Ἀγαθοκλέα: PER), and Macedonia (Μακεδονικά: LOCderiv).[16]

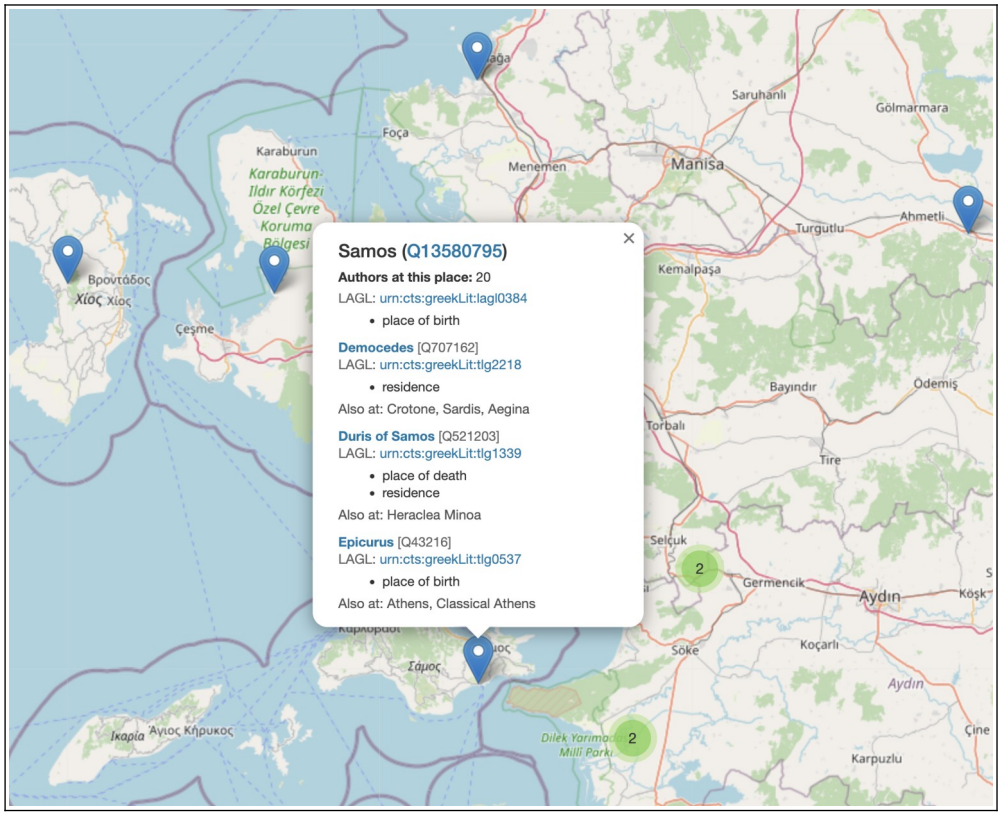


Fig. 2: *LAGL Catalog* – experimental map with *Wikidata* properties.

The *LAGL* project is collecting linguistic annotations like those mentioned in the previous paragraph to create a dataset of named entities in ancient Greek and Latin, and further disambiguate author names and work titles. To accomplish these goals, the project started with the semi-automatic annotation of the texts of Athenaeus of Naucratis (*Deipnosophists*), Valerius Harpocration (*Lexicon of the Ten Orators*), and the *Suda*, because they are very rich in named entities for ancient Greek, which is a language where data is still very scarce.[17] In order to get an estimate, the project currently contains the following occurrences of inflected proper names in ancient Greek: 23,583 personal names (PER) and 837 derivates (PERderiv), 3,916 place names (LOC) and 6,649 derivates (LOCderiv), 299 organization names (ORG) and 59 derivates (ORGderiv), and 2,721 other entities not classifiable in the previous categories (OTH).[18] As far as authors and works are concerned, the current coverage of the *LAGL Catalog* is of 1,242 authors and 3,228 works cited in the above mentioned sources.[19] The project will expand to other sources in order to increase the number of named entities in ancient Greek and add those in Latin.[20]

16 For a complete collection of sources citing Duris' works and commentaries on them, see FGrHist (= BNJ) 76. The Named Entity labels listed in this paragraph are examples of inflected Greek and Latin forms as they occur in ancient sources that are extracted and disambiguated in the context.

17 On the reasons for starting with these sources, see Berti (2026a).

18 These numbers refer to occurrences of inflected forms not lemmatized (e.g., every occurrence of Κτησιφῶντος). The annotation of the *Suda* is in progress, and data is now available for the entries of the letters alpha, beta, and gamma: see https://www.lagl.org/tools/suda/ (last access 23.09.2025). For the meaning of these NE labels, see above in the paper.

19 These numbers refer to the total number of authors and works with a unique identifier in the *LAGL Catalog* under which linguistic annotations referring to them are collected. These authors and their annotations can be found at https://catalog.lagl.org/ (last access 23.09.2025).

20 Berti (2024a) and Berti (2026a).

LAGL annotations of authors and works are enriched with CTS URNs and *Wikidata* IDs that function as unique identifiers to connect the annotated forms with metadata and external resources.[21] For example, annotations referred to Duris of Samos are collected under urn:cts:greekLit:tlg1339, which is the unique identifier of this author in the *Perseus Catalog* that preserves the four-digit number tlg1339 of the *Thesaurus Linguae Graecae* (*TLG*).[22] The *Wikidata* ID Q521203 accompanies the CTS URN urn:cts:greekLit:tlg1339 in the *LAGL Catalog* for Duris of Samos. The corresponding *Wikidata* item also includes a LAGL author ID property (P12869) that collects LAGL CTS URNs, linking the *LAGL Catalog* to metadata of the *Wikidata* knowledge base, and vice versa, linking *Wikidata* to the linguistic annotations of the *LAGL* project.[23]

Metadata deriving from *Wikidata* can be extracted and used for many possible views and analyses. For example, locations currently connected with *LAGL* authors in *Wikidata* permit the creation of maps. Fig. 2 shows an experimental map with Duris as one of the twenty authors currently collected in the *LAGL Catalog* and listed under ancient Samos (Q13580795 with *Pleiades* ID 599925), because ancient sources connect this place to these authors in assertions referring to their place of birth or intellectual activity.[24]

Nomina Omina – The Proceedings

These proceedings follow the structure and order of the workshop, excluding four papers that were not submitted to the editor, and adding three new ones by speakers who attended the workshop in person at Leipzig University and have an interest in its topics.

The paper by Anna Clara Maniero Azzolini (University of London) presents prosopographical data extracted from inscriptions of the Roman period from the ancient city of Altinum in Veneto, Italy, and its insertion into *Wikidata*. Camillo Carlo Pellizzari di San Girolamo (Scuola Normale Superiore Pisa) discusses named entities in the Wikibase *Hypotheseis*, which is a relational database of Greek rhetorical exercises written in Greek from the Hellenistic to the Byzantine age. Finally, Farnoosh Shamsian (Universität Leipzig) and Monica Berti (Universität Leipzig) present the translation alignment, annotation, and analysis of named entities in the trilingual inscription (Greek, Middle Persian, and Parthian) of Shapur I at Ka'ba-ye Zartošt (ŠKZ) located in Naqsh-e Rustam, Fars province, Iran.

These three new articles represent a significant addition to the topics covered in the workshop, as they provide further data from epigraphic sources and ancient languages other than Greek and Latin, demonstrating the potential and importance of Digital Classics methods in addressing the new challenges that await us in the face of the fascinating and important technological revolution of Artificial Intelligence.

---

21 On *Wikidata* and its use in the humanities, see Vrandečić et al. (2023) and Zhao (2023). On the interactions between the LAGL Project and *Wikidata*, see Berti (2025c).

22 See https://catalog.perseus.org/catalog/urn:cts:greekLit:tlg1339 (last access 23.09.2025). On the *Perseus Catalog* and the use of CTS URNs, see Babeu (2019).

23 See https://www.wikidata.org/wiki/Q521203 (last access 23.09.2025).

24 See https://www.wikidata.org/wiki/Q13580795 (last access 23.09.2025) and https://pleiades.stoa.org/places/599925 (last access 23.09.2025). The experimental map – available at https://catalog.lagl.org/ (last access 23.09.2025) – shows all *Wikidata* items with a *LAGL* author ID (P12869) and their associated *Wikidata* location properties: place of birth (P19), place of death (P20), country of citizenship (P27), ancestral home (P66), place of burial (P119), residence (P551), location of formation (P740), work location (P937). This map is updated regularly to correct possible mistakes and inconsistencies, and keep track of changes in *Wikidata*. For this map and other *Wikidata*-related resources in the LAGL Project, see Berti (2025c).

In conclusion to this introduction, I express my sincere thanks to the contributors, who accepted the invitation to come to Leipzig for the workshop and submitted their articles, providing stimulating ideas for the future of data on the ancient world and for our disciplines. I am also very grateful to the German Research Foundation (Deutsche Forschungsgemeinschaft) for supporting this publication, to Leipzig University for facilitating the organization of the workshop, to Alexander Plate for working on the layout of this publication, and to the Editorial Team of Digital Classics Online for hosting these proceedings as a special issue of the journal.

List of Abbreviations

BNJ I. Worthington (ed.), Brill's New Jacoby, Leiden 2006–.

BNP H. Schneider / M. Landfester / H. Cancik (eds.), Brill's New Pauly, Leiden 1996–.

FGrHist F. Jacoby (ed.), Die Fragmente der griechischen Historiker, I–III, Leiden 1923–1958.

Sources

Online sources

Altinum: a Wikidata Project for Digital Epigraphy: https://www.wikidata.org/wiki/User:Anna_Clara_Maniero_Azzolini/Altinum (last access 23.09.2025).

Daidalos-Projekt: NLP in der Klassischen Philologie: https://daidalos-projekt.de (last access 23.09.2025).

Hypotheseis: https://hypotheseis.wikibase.cloud/ (last access 23.09.2025).

LGPN – The Lexicon of Greek Personal Names: https://www.lgpn.ox.ac.uk/ (last access 23.09.2025).

LGPN-Ling. Etymology and semantic of ancient Greek Personal Names: https://lgpn-ling.huma-num.fr/ (last access 23.09.2025).

Linked Ancient Greek and Latin (LAGL) Project: https://www.lagl.org/ (last access 23.09.2025).

Linked Ancient Greek and Latin (LAGL) Catalog: https://catalog.lagl.org/ (last access 23.09.2025).

Mapping Ancient Polytheisms (MAP): https://base-map-polytheisms.huma-num.fr (last access 23.09.2025).

NIKAW (Networks of Ideas and Knowledge in the Ancient World): https://research.kuleuven.be/portal/en/project/3H220323 (last access 23.09.2025).

Opera Graeca Adnotata (OGA): https://varro.informatik.uni-leipzig.de/oga/en/ (last access 23.09.2025).

Patristic Text Archive (PTA) – An Open Access Archive of Ancient Christian Texts: https://pta.bbaw.de/ (last access 23.09.2025).

Perseus Catalog: https://catalog.perseus.org (last access 23.09.2025).

Pleiades Gazetteer: https://pleiades.stoa.org/ (last access 23.09.2025).

SERICA Project: https://serica.unipi.it (last access 23.09.2025).

Thesaurus Linguae Graecae (TLG): https://stephanus.tlg.uci.edu/ (last access 23.09.2025).

Thesaurus Linguae Latinae (TLL): https://thesaurus.badw.de/ (last access 23.09.2025).

Trismegistos God: https://www.trismegistos.org/god/ (last access 23.09.2025).

Trismegistos Paradoxography: https://www.trismegistos.org/paradoxography/ (last access 23.09.2025).

Trismegistos People: https://www.trismegistos.org/ref/ (last access 23.09.2025).

Word-level Alignment and Named Entities in the Trilingual Inscription at Ka'ba-ye Zartošt (ŠKZ): https://zenodo.org/records/15050878 (last access 23.09.2025).

References

Babeu (2019): A. Babeu, The Perseus Catalog: of FRBR, Finding Aids, Linked Data, and Open Greek and Latin, in: M. Berti (ed.), Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution, Berlin 2019, 53–72, https://doi.org/10.1515/9783110599572-005 (last access 15.04.2026).

Beersmans et al. (2023): M. Beersmans / E. de Graaf / T. Van de Cruys / M. Fantoli, Training and Evaluation of Named Entity Recognition Models for Classical Latin, in: Proceedings of the Ancient Language Processing Workshop, Varna (Bulgaria) 2023, 1–12, https://aclanthology.org/2023.alp-1.1/ (last access 23.09.2025).

Beersmans et al. (2024): M. Beersmans / A. Keersmaekers / E. de Graaf / T. Van de Cruys / M. Depauw / M. Fantoli, "Gotta catch 'em all!": Retrieving people in Ancient Greek texts combining transformer models and domain knowledge, in: Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), 152–164, https://aclanthology.org/2024.ml4al-1.16/ (last access 23.09.2025).

Berti (2013): M. Berti, Collecting Quotations by Topic: Degrees of Preservation and Transtextual Relations among Genres, Ancient Society 43 (2013), 269–288.

Berti (2018): M. Berti, Annotating Text Reuse within the Context: the Leipzig Open Fragmentary Texts Series (LOFTS), in: U. Tischer / A Forst / U. Gärtner (eds.), Text, Kontext, Kontextualisierung. Moderne Kontextkonzepte und antike Literatur, Hildesheim 2018, 223–234.

Berti (2019a): M. Berti (ed.), Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution, Berlin 2019, https://doi.org/10.1515/9783110599572 (last access 15.04.2026).

Berti (2019b): M. Berti: Named Entity Annotation for Ancient Greek with INCEpTION, in: K. Simov / M. Eskevich (eds.), Proceedings of CLARIN Annual Conference 2019, Leipzig 2019, 1–4.

Berti (2021): M. Berti, Digital Editions of Historical Fragmentary Texts, Heidelberg 2021, https://doi.org/10.11588/propylaeum.898 (last access 15.04.2026).

Berti (2023a): M. Berti, Ancient Greek Historians in the Digital Age, in: M. Althage / M. Dröge / T. Hiltmann / C. Prinz (eds.), Digitale Methoden in der geschichtswissenschaftlichen Praxis: Fachliche Transformationen und ihre epistemologischen Konsequenzen: Konferenzbeiträge der Digital History 2023, Berlin 2023, https://doi.org/10.5281/zenodo.8322062 (last access 15.04.2026).

Berti (2023b): M. Berti, Named Entity Recognition for a Text-Based Catalog of Ancient Greek Authors and Works, in: A. Baillot / W. Scholger / T. Tasovac / G. Vogeler (eds.), Digital Humanities 2023: Book of Abstracts, University of Graz: Austrian Centre for Digital Humanities 2023, 557–558, https://doi.org/10.5281/zenodo.8108058 (last access 15.04.2026).

Berti (2024a): M. Berti, Digital Canons and Catalogs of Fragmentary Literature, in: F. Neuerburg / T. Tsiampokalos / P. Wozniczka (eds.), Fragmente einer fragmentierten Welt. Zur Problematik des Umgangs mit Fragmenten in der gegenwärtigen klassisch-philologischen Forschung, Berlin 2024, 217–236, https://doi.org/10.1515/9783111508788-009 (last access 23.09.2025).

Berti (2024b): M. Berti, Digital Practice for Studying the Indirect Transmission of Classical Authors and Works, in: V. Mastellari / F. Favi (eds.), Treasuries of Literature: Anthologies, Lexica, Scholia, and the Indirect Tradition of Classical Texts in the Greek World, Berlin 2024, 189–205, https://doi.org/10.1515/9783111386010-010 (last access 15.04.2026).

Berti (2025a): M. Berti, Canoni e cataloghi collaborativi per una filologia sostenibile in ambiente digitale, in: S. Cannavale / V. Casapulla / M. Verstraete / M. Vitali-Rosati (eds.), Orizzonti della filologia digitale. L'Antologia Greca per ripensare formati, paradigmi e collaborazione, Napoli 2025, 131–152.

Berti (2025b): M. Berti, Digital Catalogs of Ancient Greek Authors and Works through Papyrological Data, in: N. Reggiani (ed.), Digital Papyrology III. The Digital Critical Edition of Greek Papyri: Issues, Projects, and Perspectives, Berlin 2025, 89–106, https://doi.org/10.1515/9783111070162 (last access 15.04.2026).

Berti (2025c): M. Berti, Linked Ancient Greek and Latin (LAGL) and Wikidata: Structuring and Reusing Data of Classical Literature (discussion paper): in Journal of Open Humanities Data 11/72 (2025), 1–12, https://doi.org/10.5334/johd.423 (last access 15.04.2026).

Berti (2026a): M. Berti, Traces of Lost Libraries in the Works of Alexandrian and Byzantine Scholars: Bibliographic open data from antiquity to Wikimedia, in: G. Bodard / V. Vitale (eds.), Open Data in Ancient and Byzantine Studies, London 2026 (forthcoming).

Berti (2026b): M. Berti, Annotating the Ancient World. Critical Annotations and Digital Editions, in: P. d'Hoine / D. Kohler / W. Decock (eds.), Charting the Future of Historical Humanities, Turnhout 2026, 21–46.

Berti (2026c): M. Berti, Digitale Philologie: in V. Schulz / A. Schwab (eds.), Moderne Zugänge zu antiken Texten. Theorien, Konzepte, Methoden und ihre Anwendung, Stuttgart 2026 (forthcoming).

Cayless (2019): H. A. Cayless, Sustaining Linked Ancient World Data, in: M. Berti (ed.), Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution, Berlin 2019, 35–50, https://doi.org/10.1515/9783110599572-004 (last access 15.04.2026).

Middle (2024): S. Middle, Linked Ancient World Data: Implementation, Advantages, and Barriers, Digital Classics Online 10/1 (2024), 16–49, https://doi.org/10.11588/dco.2024.10.104105 (last access 15.04.2026).

Palladino / Yousef (2024): C. Palladino / T. Yousef, Development of Robust NER Models and Named Entity Tagsets for Ancient Greek, in: Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024, Torino 2024, 89–97, https://aclanthology.org/2024.lt4hala-1.11/ (last access 23.09.2025).

Ripoll Alberola et al. (2025): L. Ripoll Alberola / L. D'Addario / M. Burghardt / M. Berti / M. Depauw, Tracing Antiquity: References to Greco-Roman Authors in Modern Academic Discourse, in: Digital Humanities 2025 (accepted long paper forthcoming).

Vrandečić et al. (2023): D. Vrandečić / L. Pintscher / M. Krötzsch, Wikidata: The Making Of, in: Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), New York (NY) 2023, 615–624, https://doi.org/10.1145/3543873.3585579 (last access 15.04.2026).

Zhao (2023): F. Zhao, A Systematic Review of Wikidata in Digital Humanities Projects, in: Digital Scholarship in the Humanities 38/2 (2023), 852–874, https://doi.org/10.1093/llc/fqac083 (last access 15.04.2026).

Figure References

Fig. 1: *Nomina Omina* Workshop – The Program

Fig. 2: *LAGL Catalog* – Experimental Map with *Wikidata* Properties

Author Contact Information[25]

PD Dr. Monica Berti
Universität Leipzig
Lehrstuhl für Alte Geschichte
Beethovenstraße 15
04107 Leipzig
E-mail: monica.berti@uni-leipzig.de

25

DIGITAL CLASSICS ONLINE

Naming Latin Texts in the *Thesaurus linguae Latinae* Index of Sources

Adam Gitner

Abstract: This paper discusses a specific category of names in the *Thesaurus linguae Latinae:* names of ancient authors and works used as source citations. These names cover nearly all known ancient Latin texts until about 600 CE, including ca. 1,035 authors or text groups and ca. 2,880 identifiable works. They were most recently printed as the *Index librorum scriptorum inscriptionum ex quibus exempla afferuntur* (1990) and, in partly updated form, appear online. Currently efforts are underway to transform this into an open-access database as part of a larger digital transformation of the project. The *Index* includes information about the dates of authors and works, authority control data, modern editions, the existence of ancient translations into or out of Greek, sample citations, and other metadata. The paper discusses theoretical and practical issues involving these names and their future data structure.*

1. Introduction

If you open any page of the *Thesaurus linguae Latinae* (*TLL*),[1] the most comprehensive dictionary of ancient Latin, one of the first things you will notice is a densely printed sequence of mostly abbreviated source citations. For instance, the following passage comes from the article *omen*, written by Frits Oomes and published in 1974, illustrating omens involving names (*nomina*):[2]

* I would especially like to thank Josine Schrickx, Nora Götze, and the rest of my colleagues on the *TLL* Digital Workgroup: Massimo Cè, Oliver von Criegern, Roberta Marchionni, and Martin Shedd. The views expressed are my own, as are the errors and omissions. I would also like to thank Manfred Flieger and Cornelis van Leijenhorst for advice and details about the history and use of the *Index* and Yannick Anné and Tim Denecker at Brepols for assistance with word counts of Latin literature.

1 The *TLL* is available online in open access excluding the three most recent years of publication: http://tll-open.badw.de/ (last access 04.04.2025). The subscriber-only De Gruyter platform offers the latest publications and the ability to do a full-text search: https://tll.degruyter.com/ (last access 04.04.2025). On using the *TLL*, see n. 4.; on its history, see e.g. Bögel et al. (1996), Corbeill (2007), Flow (2015), and the literature listed at https://thesaurus.badw.de/ueber-den-tll/literaturhinweise.html (last access 04.04.2025).

2 It is impossible to consult the article *nomen* since the *N* volume is not yet complete. The first *N* fascicle appeared in 2011 (volume XI, 1 fasc. 1: *n – navalis*) and work has progressed to *netura*. For the publication history of the *TLL*, see https://thesaurus.badw.de/ueber-den-tll/publikationen.html (last access 04.04.2025).

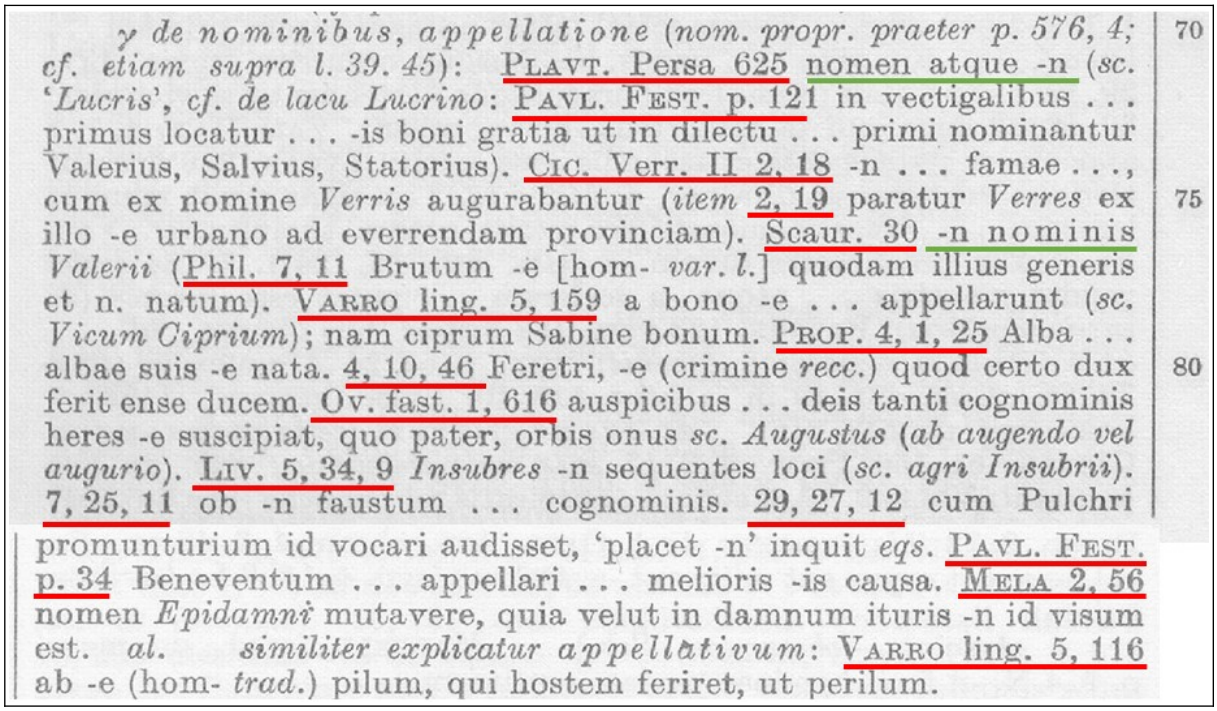

γ de nominibus, appellatione (nom. propr. praeter p. 576, 4; cf. etiam supra l. 39. 45): PLAVT. Persa 625 nomen atque -n (sc. 'Lucris', cf. de lacu Lucrino: PAVL. FEST. p. 121 in vectigalibus . . . primus locatur . . . -is boni gratia ut in dilectu . . . primi nominantur Valerius, Salvius, Statorius). CIC. Verr. II 2, 18 -n . . . famae . . ., cum ex nomine *Verris* augurabantur (*item* 2, 19 paratur *Verres* ex illo -e urbano ad everrendam provinciam). Scaur. 30 -n nominis *Valerii* (Phil. 7, 11 Brutum -e [hom- *var. l.*] quodam illius generis et n. natum). VARRO ling. 5, 159 a bono -e . . . appellarunt (sc. *Vicum Ciprium*); nam ciprum Sabine bonum. PROP. 4, 1, 25 Alba . . . albae suis -e nata. 4, 10, 46 Feretri, -e (crimine *recc.*) quod certo dux ferit ense ducem. Ov. fast. 1, 616 auspicibus . . . deis tanti cognominis heres -e suscipiat, quo pater, orbis onus sc. *Augustus* (*ab augendo vel augurio*). LIV. 5, 34, 9 *Insubres* -n sequentes loci (sc. *agri Insubrii*). 7, 25, 11 ob -n faustum . . . cognominis. 29, 27, 12 cum Pulchri promunturium id vocari audisset, 'placet -n' inquit *eqs.* PAVL. FEST. p. 34 Beneventum . . . appellari . . . melioris -is causa. MELA 2, 56 nomen *Epidamni* mutavere, quia velut in damnum ituris -n id visum est. *al.* *similiter explicatur appellativum*: VARRO ling. 5, 116 ab -e (hom- *trad.*) pilum, qui hostem feriret, ut perilum.

Fig. 1: TLL, s.v. *omen*, IX 2, 575, 70–576, 5 (Oomes [1974]).

The source citations (*Zitierweisen*) have all been underlined in red, such as "PLAVT. Persa 625", "PAVL. FEST. p. 121", "PROP. 4, 1, 25".[3] Since *TLL* articles are arranged in strict chronological order, the first citation in this section, from Plautus's *Persa*, is also the oldest surviving attestation of the word *omen* in the relevant sense. Strikingly it already illustrates the jingle *nomen atque omen* 'name and omen', which is part of the title of this volume. Though the conventions of the *TLL* may seem intimidating at first glance, with patience one can extract much more information about the meaning, range, and distribution of the word *omen*.[4]

As an onomastic convention, the use of standardized source references is essential for modern lexicography. Besides saving space, it provides the necessary scholarly precision so that the reader knows exactly which part of which edition is under discussion. This is especially important for ancient texts, which have been subject to proliferating and incompatible systems of pagination, capitulation, and numeration since the advent of printing. For instance, for the poet Propertius ("PROP."), modern editors disagree not just about how to number individual poems, but how many books have been transmitted, and where the divisions lie. Similarly, when a text is cited by page number, as in "PAVL. FEST. p. 121", how does the reader know which edition is referenced?

Answering these questions is the job of a meta-lexicographic reference work that has developed alongside the *TLL*: The *Index librorum scriptorum inscriptionum ex quibus exempla afferuntur* or here *Index* for short. In its latest incarnation, the *Index* exists as an open access HTML website in tabular format.[5] The website incorporates hundreds of *addenda* and *corrigenda* that have been adopted since the *Index* was last printed in 1990.[6] Since the *TLL* aims to cover *all* Latin produced before 600 in any medium, the *Index* provides, in effect, a database of nearly all known Latin texts within this time period. Currently it has nearly 5,000 entries, five times as many as the list of abbreviations in the *Oxford Latin*

3 When a following citation shares elements with the preceding citation, the identical elements are usually not repeated. For instance, "Scaur. 30" appears instead of "CIC. Scaur. 30" because the previous citation is also from Cicero.

4 A good place to start is the specimen article provided on the *TLL* website: https://thesaurus.badw.de/en/hilfsmittel-fuer-benutzer/specimen-article.html (last access 04.04.2025). The canonical account is the *Praemonenda* (Keudel [1990]).

5 https://thesaurus.badw.de/tll-digital/index/a.html (last access 22.04.2025).

6 Krömer / van Leijenhorst (1990); for the *addenda* see n. 32.

Dictionary, whose coverage only extends to around 200. But the *Index* is much more than just a list of abbreviations: it provides datings considered authoritative by the *TLL*, prosopographic information, concordances of various kinds, bibliographies of scholarly editions, and even information about translation from and into Greek. These features of the *Index* are often overlooked, and have scholarly value far beyond the lexicographic use for which it is primarily designed.

Currently the *Index* is undergoing a digital make-over as part of a more profound digital transformation of the *TLL*. Perhaps the most significant goal is to release all new lexicographic articles in open-access XML starting in 2026, which requires a more thorough digitalization of many components of its lexicographic workflow. As part of this, the *Index* data is being migrated to a new database that can facilitate the production of XML and handle complex queries.

Because of this digital transformation it is an opportune moment to take stock of the current *Index* and related onomastic conventions. First, I will discuss the underlying syntax and semantics of *TLL* source citations. These conventions deserve to be more widely known and adopted by classicists because of their consistency and breadth of coverage. Yet, as those of us working on transforming the *Index* have discovered, the conventions are more complex than they seem, raising difficult questions about the nature of authorship and ancient titulature. Explaining how these work in detail will be helpful to users of the *TLL* in general, and it will also facilitate the greater formalization of these names in future digital formats, such as Canonical Text Services. Second, I will provide an overview of the *Index* as a reference work for the names of ancient Latin sources, emphasizing the scope and variety of information available there. By way of conclusion, I will sketch some possible digital futures for the *Index* and its onomastic conventions.

2. The Syntax of *TLL* Source Citations

2.1 General Remarks

The practice of abbreviating Latin lexicographic sources can be traced back to antiquity,[7] and was elaborated by medieval lexicographers, such as Papias in the eleventh century.[8] But there is a crucial difference between the sporadic and inconsistent use of abbreviations one finds in ancient glossography – for instance variation among *Virg., Virgl.,* and *Virgilius* depending on how much space is at the end of a line – and the imposition of a *coherent system* of source abbreviation that is unambiguous and internally consistent. The increasing formalization and precision in citation practice since Robert Estienne's original *Thesaurus linguae Latinae,* first printed in 1531, reflects the growing professionalization of classical scholarship as well as the greater complexity of available Latin source material.

The modern *TLL* has developed these conventions over time into a comprehensive ontology for citing, as far as possible, all known ancient Latin texts produced before 600. The conventions strive to be as intuitive and self-explanatory as possible, so that a reader does not have to spend too much time decoding an unfamiliar abbreviation. Instances such as "PLAVT. Persa 625" (Plautus, *Persa*, line 625) and "AVG. in euang. Ioh. 10, 10" (Augustine, *in Iohannis euangelium tractatus,* tractate 10, chapter 10) are probably decipherable by most classicists even if they have never read these texts before. However, the pattern implicit in these straightforward cases has been extended over time – more by

7 For instance, in the medieval manuscripts of Arusianus Messius's *Exempla elocutionum*, there is inconsistent abbreviation of the names of Cicero, Sallust, Terence, Virgil, and their works; the most recent editor (Di Stefano 2011, p. XCVII) has preserved these on the grounds that it goes back to fourth-century practice. On the possibly related practice of *sigla personarum* in Latin dramatic texts, see e.g. Andrieu 1940.

8 On source marks in the eighth-century *Liber glossarum* see Grondeux (2015), in Papias, see Grondeux (2023), and in medieval usage more broadly, Grabowski (2024), 207–220.

improvisation than by any abstract rules – to cover some extremely complex sources, such as different witnesses of the *Vetus Latina*, e.g. "VET. LAT. I Cor. 4, 13 (cod. 75. *sim. al. et* VVLG.)",[9] and different recensions of the Latin Oribasius, e.g. "ORIBAS. eup. 2, 1 *Λ* 24 La p. 485".[10] Moreover, non- and para-literary sources, which have become available in greater abundance and diversity since the nineteenth century, require different paradigms of citation that do not center individual authorship. In the *TLL,* these texts may be cited by their place of publication (e.g. "CIL I^2 3191") or various combinations of genre and publication (e. g. "TAB. devot. Audollent 233, 32" for a curse tablet edited by Audollent 1904). To give one amusing example, there is even an abbreviation for inscriptions preserved on ancient gaming pieces, so called *tabulae lusoriae* ("INSCR. tess. lusor.").[11]

The brief, canonical account in the *TLL Praemonenda* does not offer terminology for describing all the elements of complex citations, nor does it say anything about how abbreviations are assigned.[12] To some extent the lack of strict codification is actually an advantage because it creates room for the improvisation that is sometimes necessary when citing complex sources and allows tolerance for the minor internal inconsistencies that have inevitably developed over time. What I offer here are some personal observations based on my experience as a lexicographer working on the dictionary and based on collective discussions in the *TLL* Digital Workgroup, but they are not meant to be an official or exhaustive account.

While the *TLL* system tries to be as comprehensive as possible, no single system is going to be adequate or appropriate for every scholarly task. Very often the form of a source citation is a compromise choice among conflicting criteria, such as brevity versus clarity or the latest scholarship versus backwards compatibility. Specialists in particular fields will sometimes need greater precision. For instance, a papyrologist may want Mertens-Pack abbreviations or *TM* identification numbers,[13] while a scholar of ancient grammar may need to distinguish the titles and authors of works that the *TLL* lumps together.[14] Ideally the *TLL* adapts to changing practices in these specialty fields, but every proposed change must be carefully weighed up against internal considerations, including the needs of non-specialist readers.

9 Translated into discursive prose this means: the *Old Latin* Bible text of the First Epistle to the Corinthians, chapter 4, verse 13, as found in codex number 75 (numbered according to Gryson 1999–2004), which is similar to the text in some other *Old Latin* witnesses and Jerome's *Vulgata*.

10 Translated into discursive prose: the Latin translation of Oribasius's *Euporista* book 2, chapter 1, under the Greek letter *lambda* entry number 24, in the La recension, which is found on page 485 of Molinier's edition (1873–1876).

11 These *tesserae lusoriae* sometimes preserve lexicographically relevant occurrences of relatively rare words, such as *nugator* 'trifler' at INSCR. tess. lusor. (cil x 8070) 13; see Baratta (2019).

12 Keudel (1990), 12 (under the subtitle *ad exempla afferenda et explicanda*). The account there is effectively bipartite, distinguishing between *nomina testium* (e.g. "CIC.") and *opera locorumque nomina* (e.g. "Verr. II 2, 156").

13 For some proposals about citing Latin literary papyri see Scappaticcio (2019). The Mertens-Pack^3 conventions are managed by *CEDOPAL* (http://www.cedopalmp3.uliege.be/ [last access 21.04.2025]); for *TM* stable identifiers, see https://www.trismegistos.org/about_how_to_cite.php (last access 21.04.2025).

14 For Roman grammarians, the most up-to-date list of sources, abbreviations, and editions is the *Index Grammaticorum Latinorum*, managed by Michela Rosellini and Elena Spangenberg Yanes as part of the Latin Grammarians Collection: https://latingrammarianscollection.uniroma1.it/site/index-grammaticorum-latinorum-intro (last access 21.04.2025).

2.2 *Tria nomina*?

Like Roman names, *TLL* citations have two obligatory elements, but can expand up to three or even four elements. The most general identifier is placed on the left and as one moves rightward one encounters increasingly specific hierarchies. The structure is best illustrated with some labeled examples:

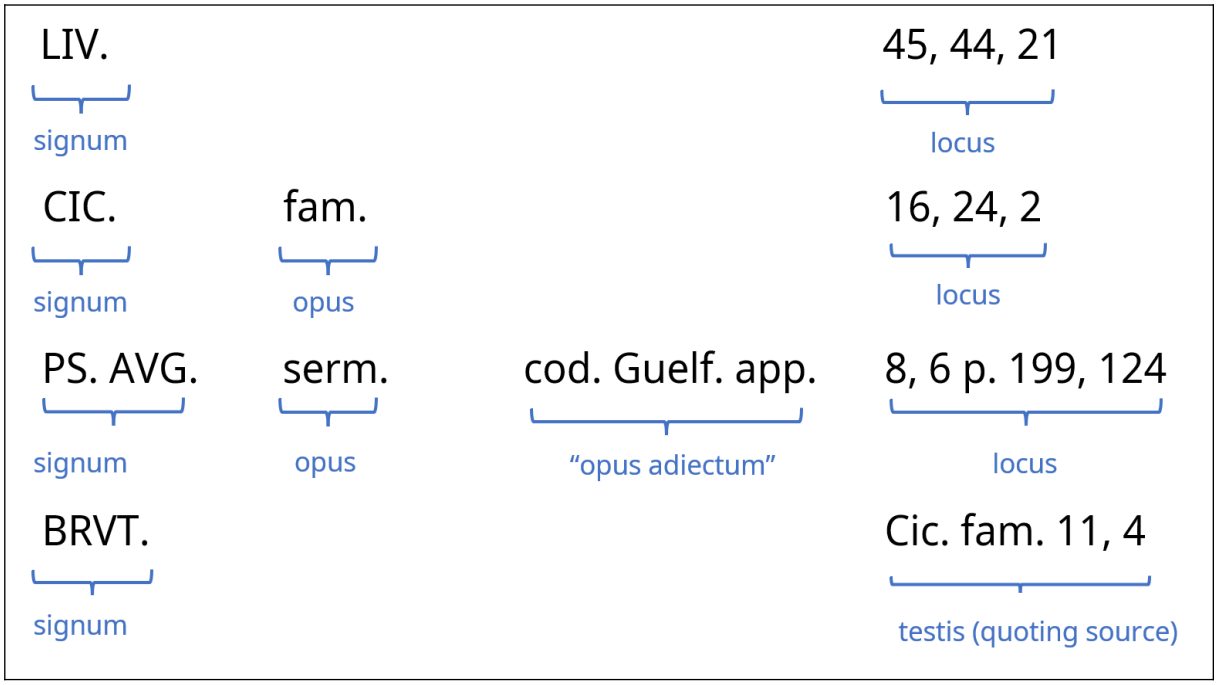


Fig. 2: Sample *TLL* Citations.

2.3 *Signum*

The first obligatory element is a word written in all capital letters here (or small caps in print). In the absence of an official title, I call this the *signum* in the sense 'nickname', as used in late antique onomastic formulae for various unofficial names.[15] In a *TLL* article, capitalization and Roman type are only used when the cited passage contains the lemma word; all other citations are written in sentence case and italicized.[16]

At first glance it is tempting to consider this element the 'author', and indeed in most of these examples this is so: "LIV." seems to refer to the historian T. Livius and "CIC." to Cicero. But the situation is more complex in the case of "PS. AVG." for 'Pseudo-Augustine', which, as used in the *Index*, is a cover term for several anonymous authors whose works have been intentionally or unintentionally confused with those of Augustine. In fact, if we look at "LIV." in the *Index,* we will see it is also used to identify later historical works *derived from* Livy, but not written by him: namely, the so-called Oxyrhynchus Epitome ("LIV. epit. Oxyrh."). Moreover, even where authorship can be confidently assigned, it is not always given prominence in first position: for instance, Augustus's *res gestae* are labeled "R. GEST. div. Aug." rather than "AVG. r. gest.".

Consequently, it is safest to regard the *signum* in the broadest sense as designating a group of works (or, when there is no further subdivision, a single work). In some cases these are defined by common authorship, but often the relationship is extremely slippery. In addition to the simple case of an au-

15 For discussion of the 'detached *signum*' (e.g. INSCR. Ann. Épigr. 1914 n. 63 *Aelius Saturninus signo Calvus*) see Kajanto (1966), 57–75.

16 Keudel (1990), 12. Furthermore, when writing Latin titles in sentence case, the *TLL* usually only capitalizes proper names: e. g. the works cited as "RHET. Her." and "VIR. ill.". The capitalization conventions can be inferred from the *notarum explicatio* column in the *Index*.

thor's name, there are at least five additional typologies among the ca. 1,374 unique *signa* in the *Index.* These include:

1. a fictive or pseudonymous author (e.g. "CANDID.", "PS. ALEX.")
2. an original Greek author of a work translated into Latin (e.g. "GALEN.")
3. a textual genre (e.g. "DECRET." for official decrees, "EPIST." for letters)
4. a modern edition or corpus (e.g. "GLOSS." for Goetz's six-volume *Corpus glossariorum Latinorum*)
5. a writing material or medium (e.g. "TAB. cer." for wax tablets, "PAP." for papyri)

While the *signum* is ordinarily capitalized, and can consist of several space-separated words (e.g. "CAEL. AVR." for Caelius Aurelianus), there are some exceptions where an additional word that belongs to the *signum* is not actually capitalized (e.g. "SCIP. mai." and "SCIP. min." for Scipio Africanus and Scipio Aemilianus).[17] Another peculiarity is that the current *signa* include several instances of homonymy. For instance, there are currently three different fragmentarily preserved authors named Brutus, who are all historically distinct (and distinguished by the *Index*) but notated by the same *signum* "BRVT.".[18] For human readers this is not a big difficulty, since the rest of the citation and *Index* provide disambiguation, but for machine processing this could be tricky.

2.4 *Locus*

The second obligatory element is the *locus*, indicating a precise location within a text. The notation depends on the way the work is structured in the standard edition. Ideally the text has been divided into logical units, such as books and chapters, that are independent of the pagination in a specific edition, but this is not always the case. When the sections are inconveniently long, sometimes the *TLL* also adds a page number (or page and line number) for convenience. Inscriptions and papyri may have columns, line numbers, and other complications. There may also be named elements in the textual hierarchy, such as a prologue ("prol.") or preface ("praef.").

Sometimes the *locus* is replaced by a cross-reference to a quoting source, labelled in fig. 2 as *testis* (e.g. "BRVT. Cic. fam. 11, 4"). The cross-reference takes the form of a complete citation, with its own *signum*, *opus* (if necessary), and *locus,* but all written in sentence case. This is a form of simple recursion. In effect it allows the *TLL* to attach a different *signum* to parts of a text or corpus that belong to different authors or textual typologies, such as embedded quotations, reported *senatus consulta*, and the like. The new *signum* usually emphasizes a feature of the embedded text that is important for its chronological placement or linguistic register.[19] The citation of fragments by their embedded location, rather than by a fragment number in a standard collection (when available) also helps to future-proof the *TLL*.[20] Fragment numbers have a tendency to change over time and not all collections are equally accessible.

17 By contrast, no distinction is made between the Elder and the Younger Seneca (both "SEN.") nor between the Elder and the Younger Pliny (both "PLIN."). Other cases where the *signum* arguably includes uncapitalized text: "BREV. expos.", "CVRIO avus" and "CVRIO pater", "OP. imperf. in Matth.", "ORAT. imp.", "PERVIG. Ven.", "RHET. Her.", "REG. urb.", "SERV. auct.", "SYMM. pater", "VIR. ill.", "VIRG. gramm.".

18 M. Iunius Brutus (Caesar's assassin), D. Iunius Brutus Albinus, and the jurisconsult M. Iunius Brutus.

19 Nevertheless, the *TLL* does not provide a distinct *signum* for every possible instance where an embedded text potentially has a different author or textual typology from its surrounding context. Much discretion is left to individual authors and editors.

20 On the complexity of citing fragmentary texts see Berti (2024).

2.5 *Opus*

In between the *signum* and the *locus* there is often an obligatory middle element, labelled here *opus*. It provides an important specification to the *signum*, typically identifying a single work, such as Virgil's *Aeneid* ("VERG. Aen."), as opposed to others by the same author or in the same work-group (e.g. "VERG. georg."). Where an original ancient title is unknown or impractical to use, the *opus* may be an abbreviation describing the genre (e.g. "serm." for *sermones*) or the published corpus where the relevant text is found (e.g. "gramm." for Keil's *Grammatici latini*).[21] Thus the *opus* element is more than just a 'work' in the strict sense. When the *signum* on its own refers unambiguously to a single work (e.g. "LIV." for Livy's *Histories* or "AETNA" for the Pseudo-Virgilian *Aetna*), there is no need to specify an *opus*.

The *opus* element is not typographically distinguished from the *locus*, nor does the current *Index* consistently identify where it starts and ends (§3.2). This sometimes creates ambiguity, which is not a problem at present, but will need to be resolved on a case-by-case basis if the citation syntax is implemented more formally. For instance, should an *appendix* ("app.") or *supplementum* ("suppl.") be regarded as a sub-section of an *opus*, and hence part of the *locus*, or a separate work entirely?[22] This may hinge on historical factors, such as whether the *appendix* was added by the author themself or is an entirely separate work that became attached in transmission.

2.6 *'Opus adiectum'*

Because of these and other textual complications, the *TLL* currently acknowledges a further specification to the *opus*, here labelled *opus adiectum.* This element is signalled in the current *Index* by a second level of indentation, as discussed below (§3.2). In the figure above, "PS. AVG. serm. cod. Guelf. app.". refers to a corpus of Pseudo-Augustinian sermons transmitted in a ninth-century manuscript (Wolfenbüttel, Herzog August Bibliothek, Cod. Guelf. 12 Weiss.).[23] While it would be possible to consider "cod. Guelf. app." part of the *locus*, this would efface the common features that set these sermons apart from other Pseudo-Augustinian sermons, such as their shared transmission and publication history. The label "*opus adiectum*" tentatively offered here is an awkward solution. It would be more accurate to say that "PS. AVG." and "serm." designate nested work-groups, and "cod. Guelf. app." functions as the specific *opus*.

2.7 Parenthetical *Addenda*

A final peculiarity to mention is the presence of obligatory parentheses or other *addenda* within a source citation. Typically these provide an alternative form of citation that may be superfluous, but nevertheless helps orient the reader. For instance, in citing the *Acta fratrum Arvalium,* the CIL number is always provided (e.g. "ACT. Arv. a. 105 (CIL VI 2075) II 7") and in citing the poems of Ausonius, the older Souchay numeration is given for reference (e.g. "AVSON. 27, 24 (417 S.), 124"). These obligatory parentheses, which interrupt the source citation, are different from an optional, explanatory parenthesis that sometimes follows the citation.

---

21 E.g. the treatise referred to by "CONSENT. gramm. V 385, 19" actually has the ancient title *De nomine et verbo*; in "LIV. ANDR. trag. 5" the *trag.* refers to Ribbeck's (1897) collection of fragments, instead of using the name of the tragedy, *Aegisthus*.

22 Some appendixes attached to literary corpora that are (or contain) arguably separate, anonymous compositions: e.g. "AVSON. app.", "CLAVD. carm. min. app.", "OPTAT. app.", "PHAEDR. app.". The *appendix* to Augustus's *res gestae* cannot have been written by Augustus since it was composed after his death ("R. GEST. div. Aug. app.").

23 Online facsimile and description available here: https://diglib.hab.de/?db=mss&list=ms&id=12-weiss&catalog=Butzmann (last access 04.04.2025). More specifically, the *opus adiectum* refers to a subset of sermons in this collection published by Morin in an appendix (1917).

3. The *TLL Index* of Sources as a Reference Work

To make sense of an unfamiliar source citation a reader should turn first to the *Index* online.[24] This is a master list of nearly all source citations currently in use, mapping abbreviations to full names and titles, as well as dates, editions, and other metadata. I say "nearly all" because in some cases, mainly for non-literary sources, only some representative citations are given rather than a complete listing of all surviving texts. Because it continues to be written in Latin, like the *TLL* itself,[25] explaining how it works in English should make it more accessible.

3.1 History

The *Index* is a continually evolving document, illustrating changes to *TLL* practice over its 130-year history, as well as advances in Latin scholarship. Its history has determined some its most important characteristics as a repertoire, as well as some constraints. Because of the need to maintain backwards compatibility, every change to existing citations must be weighed carefully against the harm of long-term inconsistency, and a high evidentiary bar must be met before adopting new scholarly findings.

The earliest list of source abbreviations was a short document reproduced by hectograph for internal use.[26] Meanwhile, readers of the earliest fascicles had to rely on a single page of *notanda* for guidance.[27] The need for a more comprehensive list of sources was finally met in 1904 with the publication of the first edition of the *Index* as a supplementary volume to the *TLL* under Friedrich Vollmer as General Editor.[28] This publication established the columnar layout and the main categories of data that remain today. This first edition of the *Index* runs to 109 pages and contains approximately 3,815 entries. Three short supplements appeared over the years, which were eventually combined together in 1958 into a separate, thirteen-page *Supplementum,* incorporating additional changes.[29]

The second edition from 1990 involved a systematic revision of every entry and a thorough search through the literature to discover missing items.[30] Dietfried Krömer was responsible for the general conception and the sections A to O, while Cornelis van Leijenhorst was enlisted as co-author to complete the sections from P to Z, as well as inscriptions and papyri. Colleagues and subject experts were frequently consulted, especially H. J. Frede on Christian material. Significant changes and scholarly progress are especially apparent in the presentation of late antique texts, some of which appear for the first time or with more accurate information, while other works long accepted as ancient have been rejected.[31]

The current HTML presentation of the *Index* incorporates numerous *addenda quaedam indici*, most recently updated in 2022 and downloadable separately.[32] Both the website, which went online in 2017

24 See n. 5.

25 For a defense of Latin in the *TLL* see Marchionni (2015).

26 Vollmer (1904); see Bögel et al. (1996), 54–55, 146.

27 The *notanda* appeared on the last page (xiv) of the *praefatio* that was enclosed with vol. I, fascicle 1 (1900), and reprinted at the beginning of vol. III (1907–1912). Though the *notanda* expand only a few abbreviations used in citations, they nevertheless formulate an important principle that source abbreviations should be formulated generously.

28 Bögel et al. (1996), 145–153. For internal and external contributors, including H. Dessau and L. Traube, see Vollmer (1904), first unnumbered page.

29 The three supplements appeared in vol. III, iv–v (1906), vol. V.1, iv–v (1910) and in a double-leaved printing in November 1936 (presumably included with fascicle VII.1.3); see Bögel et al. (1996), 154 n. 2.

30 Krömer / van Leijenhorst (1990), v–vii.

31 On the changes to late antique material see especially Krömer (2003) and the review by Doignon (1993).

32 The *addenda,* chosen by the *TLL* Editors, are available as a separate PDF file: https://thesaurus.badw.de/fileadmin/user_upload/Files/TLL/addenda.pdf (last access 04.04.2025). These include newly

with support from the IT Department of the Bavarian Academy, and the *addenda* are maintained by Josine Schrickx, current General Editor, and Marijke Ottink. The electronic version also includes its own directions for use.[33] Meanwhile the on-going work transforming the *Index* into a database has been overseen by the *TLL* Digital Workgroup, coordinated by Nora Götze.

3.2 Using the Online *Index*

To understand the contents of the *Index*, we can turn to a section from the beginning of the HTML table:

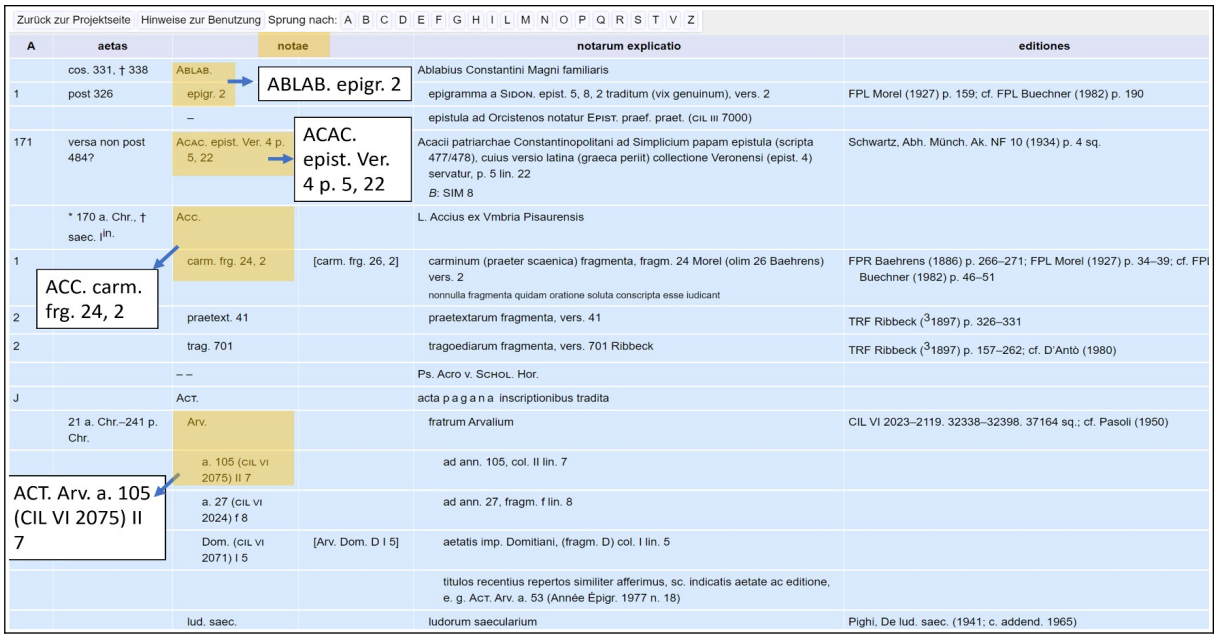


Fig. 3: *Index* online, from ABLAB. to ACT. lud.saec.

For most users, the most important column is the one labeled *notae*, providing sample abbreviations, some of which are highlighted in yellow, such as "ABLAB. epigr. 2", "ACAC. epist. Ver. 4 p. 5, 22", "ACC. carm. frg. 24, 2", "ACT. Arv. a. 105 (CIL VI 2075) II 7". As the illustration shows, a complete citation is formed from combining the text of the capitalized *signum* with all the text in the relevant indented lines below it until one reaches a *locus*.

It is important to note the presence of two levels of indentation within the *notae* column distinguishing between the *opus* and *opus adiectum* elements. Where there is a single indentation, this indicates the beginning of an *opus* element (e.g. before "epigr.", "carm. frg." and "Arv."). Where there is a double indentation, this either marks the beginning of an *opus adiectum*, such as before "a. 105 (CIL VI 2075) II 7", or emphasizes different types of *loci* belonging to the same *opus*. Nevertheless, some citations are compressed on a single line, such as "ACAC. epist. Ver. 4 p. 5, 22", and in these cases the division among the elements is ambiguous.

The sample citation *always* includes a complete *locus* so that it is clear which numeration or system of reference the *TLL* follows. This is extremely important in order to avoid confusion among the many competing systems of capitulation sometimes present within a single edition. The sample *locus* is usually the *last citable section of the work* since it is at the end of a work, rather than the beginning, where differences in capitulation are most evident.

published ancient Latin works and changes to the citation format of previously known works; new editions of known texts are not added unless they prompt the *TLL* to change its citation format.

33 https://thesaurus.badw.de/en/tll-digital/index/directions-for-use.html (last access 04.04.2025).

The full expansion of each element in the abbreviation is found in the adjacent column, labeled *notarum explicatio.* This column also contains additional information that may be relevant to understanding and finding the text in question. For instance, the citation "ACAC. epist. Ver. 4 p. 5, 22" is expanded as follows:

> *Acacii patriarchae Constantinopolitani ad Simplicium papam epistula (scripta 477/478), cuius versio latina (graeca periit) collectione Veronensi (epist. 4) servatur, p. 5 lin. 22.*
>
> Letter by Acacius, Patriarch of Constantinople to Pope Simplicius (written in 477–478), of which a Latin translation (the Greek has been lost) is preserved in the *collectio Veronensi* (letter 4), page 5, line 22.

Each component of the abbreviation ("epist. Ver.") is expanded, and the numbers in the *locus* ("4 p. 5, 22") are glossed. Furthermore, important contextual details in this case are provided about the text's addressee, its status as a translation from a lost Greek original, and its original date of composition (which differs from the date of its translation, indicated in the *aetas* column).

Incidentally, the use of the genitive case for the proper name, *Acacii,* in the *notarum explicatio* is significant. Whereas the nominative in apposition is regularly used to indicate genuine authorship (e.g. "Ablabius", "L. Accius"), the genitive usually indicates a more complex relationship, such as authorship of the original Greek text or pseudonymity.

3.3. Metatextual Information in the *Index*

In the figure below I have identified six kinds of metatextual data that are consistently or occasionally present:

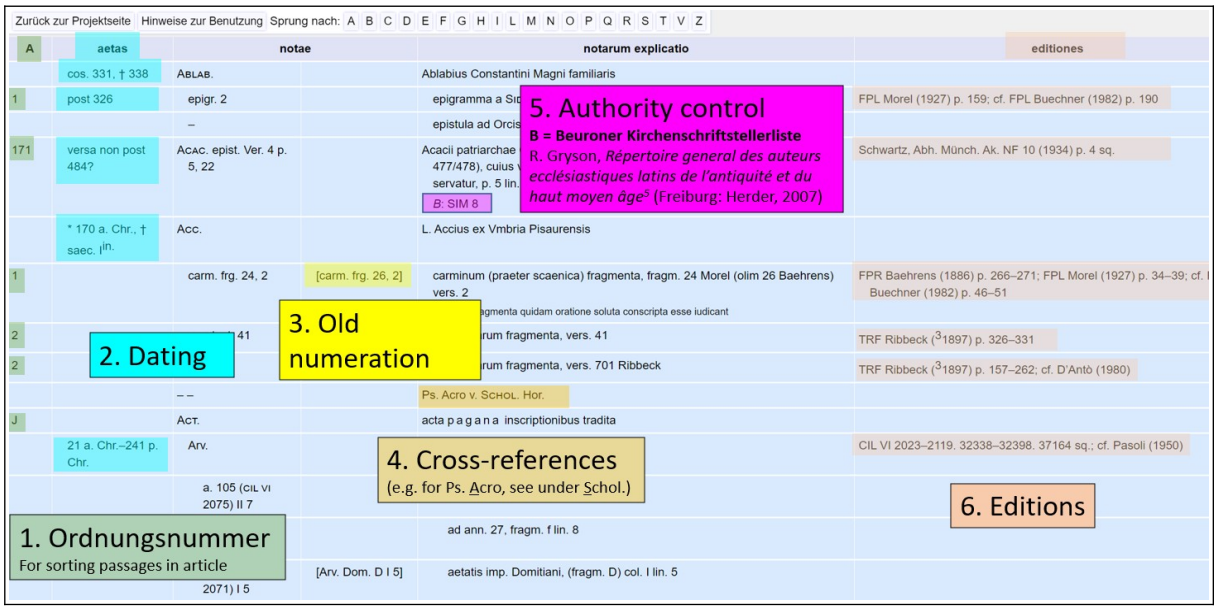


Fig. 4: Metatextual data in the Index.

Starting on the left, the first category is the *Ordnungsnummer,* found in the first column (labeled "A"). This number, which is important for ordering passages in a *TLL* article, provides an approximate relative chronological position for each author or text, starting with 1 for fragments of early Roman poetry and now ending at 226 for some of the latest, sixth- or seventh-century cited texts (e.g. "CHILP. hymn. Medard."). The numbers also correspond to the shelfmark in the *TLL* Library.

Second, the date of the author or work is found in the second column (*aetas*) where available. As even this short sample shows, the kind of dating and its precision vary considerably based on the available

information.[34] For instance, the author Ablabius ("ABLAB.") is identified only by the date of his consulship (331) and death (338), but the epigram Sidonius attributes to him is dated more precisely "after 326".[35]

The third category are out-of-date source abbreviations and numeration, used in older volumes of the *TLL* but now superseded. An out-of-date abbreviation always appears between square brackets ("[carm. frg. 26, 2]"). The third column (the lefthand side of *notae*) contains *all* abbreviations, whether current or not. When an out-of-date abbreviation appears in the third column, the current abbreviation appears in the fourth column (the righthand side of *notae*); when a current abbreviation appears in the third column, an out-of-date equivalent may be listed in the fourth column.[36] As the *notarum explicatio* makes clear, the fragments of Accius ("ACC.") were originally numbered according to Baehrens's edition, but now according to Morel.

Cross-references of various kinds constitute the fourth kind of metadata. For instance, a reader might expect to find the Pseudo-Acronian scholia to Horace listed under 'PS. ACRO'. Accordingly the name 'PS. ACRO' is inserted alphabetically in the *notarum explicatio* column between 'ACC.' and 'ACT.', but the reader is advised to find the text under 'SCHOL. Hor.'. The presence of a cross-reference is indicated in the *nota* column by double dashes ('--'), when it refers to the *signum* level, and by a single dash ('-'), when it refers to an *opus*. There are about 820 cross-references in the current *Index*.

The fifth kind is authority control information (*Normdaten*). This has become extremely important in information science as a way to guarantee that names, especially used across platforms in linked open data, refer persistently to the same person or object.[37] The second edition of the *Index* already provided a rudimentary form of this: whenever an italicized letter "*B"* appears in the *notarum explicatio* column, it is followed by the abbreviation for the author in the Beuron *Kirchenschriftstellerliste.*[38] This repertoire explicates the standard source abbreviations used in the monumental Beuron critical edition of the *Vetus Latina*, and as such often provides more detailed information about late antique Christian sources. In turn, the *Kirchenschriftstellerliste* provides the corresponding *TLL* abbreviations where possible. This system of mutual cross-reference is a form of authority control that guarantees long-term compatibility, and it also helps the reader find additional information about the relevant author or work.

The last kind of metadata, and for classicists perhaps the most valuable, is the list of reference editions provided in the rightmost column (*editiones*). The list is up-to-date until 1990; after 1990 only sporadic additions have been made, principally of editions that have altered the *TLL* source abbreviation or *locus*. Accordingly, many significant, post-1990 editions are currently absent from this list, but they are expected to be added soon. The most recent edition before the "cf." (if there is a "cf.") typically provides the capitulation and numeration of citations; other important editions may be added after a "cf.". Altogether there are over 3,100 unique editions listed in this column.

The use of multiple editions is essential for *TLL* practice: while it is necessary to use a single reference edition for capitulation, the text as quoted in the *TLL* articles is meant to be the most plausible text, regardless of whether it is found in an edition or not. This sometimes means making a choice among tex-

34 Krömer (2003) discusses some issues relating to dating.

35 The more precise dating is presumably based on the historical situation described by Sidonius. Note also that the *Index* specifies parenthetically that the poem is *vix genuinum.* This lampoon on the emperor, posted up on the palace wall, seems unlikely to have been written by a consular, despite Sidonius's report. Nevertheless, the *signum* remains "ABLAB." rather than "PS. ABLAB.".

36 On this innovation introduced by the *Index* second edition, see Krömer / van Leijenhorst (1990), v.

37 On the exemplary use of authority control in the *Perseus Catalog*, see Babeu (2019).

38 The most recent, fifth edition: Gryson (2007). H. J. Frede, who worked on the third and fourth editions (1981 and 1995 respectively), was in close contact with the editors of the *TLL Index* while both works were under revision.

tual variants, citing conjectures published elsewhere, or even suggesting a new emendation where necessary.

3.4 *Graeca*

One would not expect to consult an index of Latin sources for information about Greek texts, but there is a surprising amount of information in the *TLL Index* for Hellenists.[39] First, there are a few Greek-language texts scattered within the *Index* as primary sources occasionally cited by the *TLL*: notably, the sixth-century lexicographer Hesychius ("HESYCH. lex.") and Iohannes Lydus ("LYD."), a sixth-century scholar who preserves many etymologies and discussions of Latin words.[40] Secondly, and more significantly, there are over two hundred Greek texts referenced in the *editiones* column whenever an ancient Latin text was either translated from Greek or possibly translated into Greek. These are preceded by the abbreviation "*gr.*" when the Greek text is known to be source or "*cf. gr.*" when the Greek is potentially later or the precise relationship has not been sufficiently established. The systematic collection and evaluation of relevant Greek texts was one of the achievements of the second edition of the *Index*, and it offers a tantalizing possibility for future research into ancient bilingualism or even large-scale digital language alignment of Greek and Latin texts.

3.5 Visualizing the *Index*

Finally, to give an idea of the scale of the ancient Latin corpus depicted by the *Index*, I provide some graphic representations of the number of citeable works across time. By "citable work", I simply mean a unique combination of *signum, opus,* and *opus adiectum.* Altogether I count 2,880 of these, divided out over ca. 1,035 *signa.*

The results need to be treated with caution. First, a 'citable work' varies significantly in scale: for instance, the entire *CIL* (*Corpus inscriptionum Latinarum*) counts as one citable work, as does a two-line epigram attributed to Ablabius ("ABLAB. epigr."). Second, the *Index* does not always provide a date, and where it does, it may be approximate and/or uncertain.[41] To convert these dates into centuries I have treated uncertain cases as certain and converted any Arabic number or Roman numeral into a corresponding century (using the lower of two alternatives where provided). Third, some of the entries in the *Index* provide only a few sample works rather than an exhaustive listing of all known works.[42] Thus the number of works in the *Index* is less than the actual number of works cited in the *TLL*.

39 On Hellenists using the *TLL* see Vogt (1995).

40 Though not mentioned in the *Index*, Polybius ("POLYB.") is cited in the article *praefectus* (X 2, 624, 20; van Leijenhorst 1985).

41 There are about 453 citable without dates, either on their own or derivable from their parent *signum*. These can mostly be assumed to be late.

42 For instance, under "DECRET. decur." only some sample decurional decrees are given; listing all such texts would take up too much space. Other such *Mustereinträge* can be found by searching the *Index* for *similiter*.

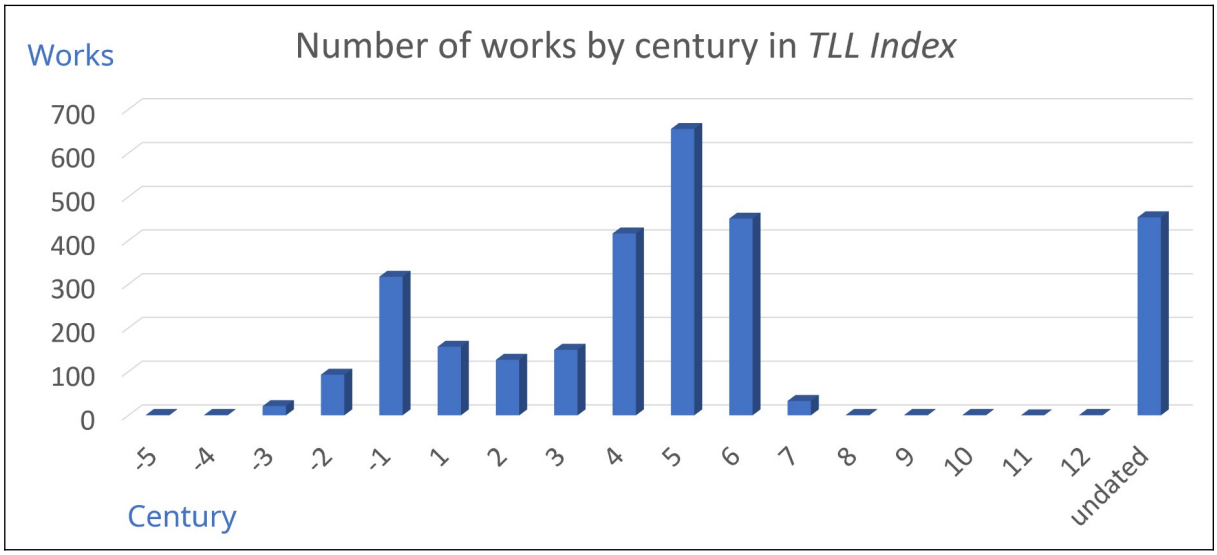


Fig. 5: Citable works in the *TLL Index* by century.

It is no surprise to see an abundance of material for late antiquity: if everything before 200 CE is considered 'classical', there are nearly 2.5 times as many works that survive from late antiquity. What did surprise me was to see that the contents did not drop off completely after 600 CE: the *Index* in fact includes 33 works from the seventh century, and one each from the eighth (PAVL. FEST.), ninth (BASILIC.), tenth (CONSTANT. PORPH.), and twelfth centuries (*Osbern*.). I encourage readers who are curious to look these up in the *Index*.

A more effective representation would involve word count. Since the *Index* does not include any word counts, I have relied on the figures generously provided by Brepols' *Library of Latin Texts* (*LLT*). Their database includes *TLL* source abbreviations wherever possible, which makes alignment easier. But not all the texts in the *Index* are in *LLT*, especially inscriptions and other non-literary documents. Moreover, even when a text is in both corpora, it is often difficult to align them automatically since each database may have made different choices about authenticity, text division, and what constitutes an independent work.

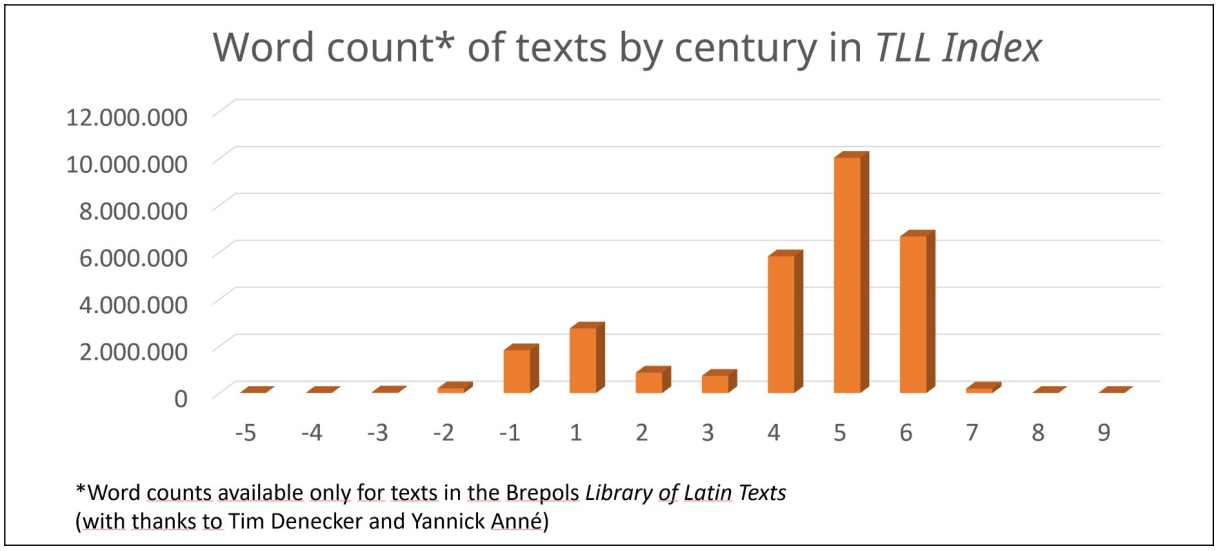


Fig. 6: Word count of citable works in the *TLL* by century.

The resulting distribution remains somewhat similar, with approximately three times as many words from late antiquity as from the classical period. The most dramatic difference is the relative drop in the first century BCE. This certainly has to do with the number of fragmentary texts from this period: very many citable works but with a very low word-count. Given the inconsistencies in the data, these fig-

ures should be regarded as an *amuse-bouche* for what can be derived from future digitization rather than a reliable representation of the *TLL* textual corpus.

4. Future Prospects

The *Index* is currently being migrated to a new relational database as part of a thorough digitalization initiative within the *TLL* that aims to increase access and functionality. Currently work on the *Index* has focused on updating the digital architecture rather than making changes to the content. It is still too early to share details about implementation. The most urgent priority is integrating the *Index* with a new digital workflow for producing *TLL* articles in XML so that lexicographers can quickly cite the correct source abbreviations and ensure their accuracy. Nevertheless, this transformation also opens the prospect for significant improvements for external users and the linked open data community at large.

By way of conclusion I will sketch here some of the main areas of future work as well as some open questions and desiderata. Not all of these features will be possible to implement in the near future, or perhaps at all, but discussing these issues in this public forum can generate awareness and excitement among the larger scholarly public and solicit feedback from interested parties at a formative moment of development.

4.1 Searchability and Complex Queries

A central desideratum is improving the searchability of the *Index* online. Currently a full-text search of the HTML table is the only way of querying the data, which is unwieldy when looking for short snippets of text. In a new relational data structure, it will be possible to perform targeted queries for text in specific elements: for instance, to find all *signa* containing the string "AVG" or to identify all texts certainly (or possibly) composed after the year 410. Indeed it should be possible to combine these queries into a complex search in order to generate a list of Augustine's works composed after 410.

For these searches to work effectively, some of the orthographic conventions of the *TLL* must be taken into account automatically: for instance, to recognize that an upper-case "V" in "AVG" can correspond to a lower-case "u" or "v" depending on the context. More generally, the natural-language dates provided in the *aetas* column for each author and work will need to be converted into a machine readable format, such as the Extended Date Time Format (EDTF) specification of the Library of Congress.

Furthermore, it will be necessary for some ambiguous elements in current source abbreviations to be clearly sorted into the relevant categories (*signum, opus, opus adiectum*, and *locus,* assuming the analysis proposed above is viable). As we have seen (§3.2), the presentation of the sources in the *Index* does not always make clear how to analyze a compound citation such as "ACAC. epist. Ver. 4 p. 5, 22". Fortunately, the Digital Workgroup has already gone through the entire *Index* and made preliminary assignments, but more internal testing and proofreading will be necessary to guarantee consistency across the data.

Once the *Index* in its current state has been successfully migrated to a new database, it will be possible to consider adding additional categories of information, such as tags to identify the genres of individual works, along the lines of the *TLG Canon*, or even information about the modality of transmission, which is currently implicit in some but not all of the source abbreviations (e.g. the *signum* "PAP." or the abbreviation "frg." in the *locus*). It would also be desirable for all this data to be accessible via an Application Programming Interface (API) to facilitate re-use.

4.2 Linked Open Data

As the most complete and up-to-date repository of information about ancient Latin texts, the *Index* can play an important role as a digital clearing-house, supplementing and refining the data already made available by trailblazers in Classics linked open data, such as *Perseus*, *Trismegistos*, and *Wikidata*. On the one hand, this would require linking *Index* entries outward with the relevant records in external databases. On the other hand, it would require making *Index* entries externally linkable by assigning persistent identifiers to all entries at the *signum* and *opus* levels.

An open question remains the form such persistent identifiers should take. The *TLG Canon* provides a useful model: in their case this is a bipartite structure consisting of a four-digit 'author' ID and three-digit 'work' ID (e.g. 0012.001 represents Homer's *Iliad* and 3123.001 represents Theodorus Daphnopates's *Epistulae*). The author IDs also encode some chronological information: IDs in the range 1–2000 are early/classical, the 3000 range is mostly for Byzantine authors.[43] Since *TLL* names are occasionally tripartite (when there is an *opus adiectum*), this complication would have to be met either by flattening out the hierarchy, and thereby losing information about textual affinities, or by tolerating three positions. Furthermore, the identifiers must be flexible enough to accommodate the possibility that authorship attributions and datings may change in future scholarship.

In addition to linking to other Latin digital catalogs, the most important targets for linkage will be open access Latin editions. This will facilitate quick access to relevant texts and could eventually provide the basis for translating *TLL* source citations articles automatically into hyperlinks. Currently there is a growing amount of ancient Latin material available online, thanks to collections such as *Corpus Corporum*,[44] *digilibLT*,[45] the *Digital Latin Library*,[46] and the *Perseus Digital Library*.[47] But in order to be in the public domain, many of the texts are not the most up-to-date. Meanwhile, through a pilot project led by Massimo Cè, a former *TLL* researcher, it was possible to map approximately 30% of the editions referenced in the *Index* to PDFs in the public domain (i.e. out of European copyright or available in open access). It remains a great desideratum to see more high-quality editions in open access, ideally with a critical apparatus, or at least with an XML structure compatible with the *TLL* source citations.

4.3 Source Citations via Canonical Text Services (CTS)

A standardized, machine readable format for *TLL* source citations is necessary not just for composing future *TLL* articles in XML, but to encourage wider adoption of these reference standards, and in the future to create links from citations to digital texts.

For encoding a complete citation, including the *locus,* the Canonical Text Services (CTS) model developed by the *Homer Multitext* project provides an attractive solution.[48] It has already been adapted to the citation of Latin texts in the *Perseus Catalog* using the Packhard Humanities Institute (PHI) and Stoa identifiers.[49] A CTS reference is a Universal Resource Name (URN) with the following basic structure: urn:cts:namespace:work:part. For instance, using the greekLit namespace,[50] a citation of the first line of the *Iliad* can be represented as: urn:cts:greekLit:tlg0012.tlg001:1.1. CTS allows for the in-

43 Pantelia (2022), xxiii n. 4.

44 https://mlat.uzh.ch/ (last access 23.04.2025).

45 https://digiliblt.uniupo.it/ (last access 23.04.2025).

46 https://digitallatin.org/ (last access 23.04.2025).

47 https://www.perseus.tufts.edu/ (last access 23.04.2025).

48 Blackwell / Smith (2019).

49 https://catalog.perseus.org/ (last access 22.04.2025); see Babeu (2019), 55.

50 Maintained by the Center for Hellenic Studies: https://github.com/chs-tg/greekLit (last access 23.04.2025).

clusion of an optional 'version', such as the representation of the text in a specific manuscript or edition, as well as the citation of the 'part' according to complex textual hierarchies, such as '1.proposition.20' in Euclid's *Elements*.

For the most part CTS syntax can be easily adapted to *TLL* source citations. Indeed, one can even provide users with a choice between numerical identifiers or human-readable names, such as the following two representations of VERG. Aen. 12, 952:

urn:cts:tll:sign04691.op0003:12.952
urn:cts:tll:Verg.Aen:12.952

Nevertheless, there are potential conflicts and limitations of CTS syntax for which the *TLL* would have to find custom solutions. First, on the *TLL* side, there is the problem of homonymous *signa* discussed above (§2.3). Ideally these should be distinguished in a machine readable way. Second, some of the more complex *TLL* source citations discussed above (§2.4) require recursion or the coordination of multiple citations. Some new notation might have to be developed to represent these relationships. For instance, to cite a letter within Cicero's epistolary corpus as having been written by Brutus (fig. 2), one might imagine the following syntax: "urn:cts:tll:Brut||urn:cts:tll:Cic.fam.11.4". Here the double pipes (||) represent the relationship 'is a description, synonym, or specification for'.

An even more complex case is the Latin Oribasius (e.g. "ORIBAS. eup. 2, 1 *Λ* 24 La p. 485"). Here there are three different layers ('versions' or 'manifestations'?) that can and sometimes must be distinguished after the work level: an individual recension of a work, a manuscript containing a recension, and a modern edition reporting the manuscript or recension. Moreover, the *TLL* cites in parallel both the logical capitulation of the work and its pagination in the only available edition. Perhaps it would have to look something like:

urn:cts:tll:Oribas.eup.La.La:2.1.Λ.24||urn:cts:tll:Oribas.eup.Molinier:pag.485

The URN before the pipes is a citation of the manuscript La (Laon Ms. 424, s. ix),[51] which belongs to the eponymous recension La; the URN after the pipes reports the same text according to Molinier's edition (1873–1876). Two citations are needed here because Molinier's pagination – still necessary for readers to find the text and compare it with that of other recensions – is not fully coterminous with or nestable within the logical capitulation.

More generally, citations in the *TLL* do something subtly different than what canonical citations are meant to do. The 'reference' edition is used as a *basis* for the capitulation and text, but the lexicographer ultimately quotes the text using all available evidence as they believe it to be correct (see §3.3). In other words, the lexicographer tries to print the Latin Oribasius, not Molinier's version of it. Accordingly, perhaps there needs to be a flag (a big exclamation point?) included as part of an URN, so that the user knows that the edition is referenced, but not necessarily quoted verbatim.

4.4 Envoi

After over 130 years of work, the *TLL* prepares to enter a new phase of digital life. The source citations and *Index* are only a small part of this, but they sit at a crucial meeting place where article production interacts with texts, names of works are encoded and decoded, dates are assigned, and external users come with queries of their own. There is still much work to do, and many of the developments will be gradual. Nevertheless, it will be possible, not only because of the dedication of the *TLL* team, but in part thanks to the digital trailblazers who have shown the way.

51 https://bibliotheque-numerique.ville-laon.fr/idviewer/1468/7 (last access 22.04.2025).

List of Abbreviations

OLD: P. G. W. Glare (ed.), Oxford Latin Dictionary, 2nd edition, Oxford 2012.

TLG: Thesaurus linguae Graecae. http://stephanus.tlg.uci.edu (last access 04.04.2025)

TLL: Thesaurus linguae Latinae, Leipzig / Berlin 1900–. https://tll-open.badw.de/ (last access 04.04.2025)

References

Andrieu (1940): J. Andrieu, Étude critique sur les sigles de personnages et les rubriques de scène dans les anciennes éditions de Térence, Paris 1940.

Audollent (1904): A. M. H. Audollent, Defixionum tabellae quotquot innotuerunt, tam in Graecis Orientis quam in totius Occidentis partibus praeter Atticas in Corpore inscriptionum atticarum editas, Paris 1904.

Babeu (2019): A. Babeu, The Perseus Catalog: of FRBR, Finding Aids, Linked Data, and Open Greek and Latin, in: M. Berti (ed.), Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution, Berlin / Boston 2019, 53–72.

Baratta (2019): G. Baratta, Benest, malest: Archeologia di un gioco tardo-repubblicano, Barcelona 2019.

Berti (2024): M. Berti, Digital Canons and Catalogs of Fragmentary Literature, in: F. Neuerburg et al. (eds.), Fragmente einer fragmentierten Welt: Zur Problematik des Umgangs mit Fragmenten in der gegenwärtigen klassisch-philologischen Forschung, Berlin / Boston 2024, 217–236.

Blackwell / Smith (2019): C. W. Blackwell / N. Smith, The CITE Architecture: a Conceptual and Practical Overview, in: M. Berti (ed.), Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution, Berlin / Boston 2019, 73–93.

Bögel et al. (1996): D. Krömer / M. Flieger (ed.), Thesaurus-Geschichten: Beiträge zu einer Historia Thesauri linguae Latinae von Theodor Bögel, Stuttgart / Leipzig 1996.

Corbeill (2007): A. Corbeill, “Going Forward”: A Diachronic Analysis of the Thesaurus linguae Latinae, American Journal of Philology 128 (2007), 469–496.

Di Stefano (2011): A. Di Stefano (ed.), Arusiani Messi: Exempla elocutionum, Hildesheim 2011.

Doignon (1993): J. Doignon, Review of Krömer / van Leijenhorst (1990), Revue d’Études Augustiniennes et Patristiques 39 (1993), 225–226.

Flow (2015): C. Flow, Thesaurus Matters: Frames for the Study of Latin Lexicography, in: C. Stray and G. Whitaker (ed.), Classics in Practice: Studies in the History of Scholarship, BICS Supplement 128, London 2015, 33–73.

Flury (1995): P. Flury, Vom Tintenfaß zum Computer, in: D. Krömer (ed.), Wie die Blätter am Baum, so wechseln die Wörter: 100 Jahre Thesaurus linguae Latinae, Stuttgart 1995, 29–56.

Grabowski (2024): A. Grabowski, The Craft of History: Turning History into a Discipline in the Twelfth and Thirteenth Centuries, Turnhout 2024.

Grondeux (2015): A. Grondeux, Le traitement des ‘autorités’ dans le Liber glossarum (s. VIII), Eruditio antiqua. revue électronique de l’érudition gréco-latine 7 (2015), 71–95.

Grondeux (2023): A. Grondeux, Comment définir un ‘dictionnaire latin’? Du Liber glossarum (VIIe s.) à l’Elementarium de Papias (XIe s.), Histoire Épistémologie Langage 45/2 (2023), 15–33.

Gryson (1999–2004): R. Gryson, Altlateinische Handschriften: Manuscrits vieux latins: répertoire descriptif, Freiburg 1999–2004.

Gryson (2007): R. Gryson, Répertoire general des auteurs ecclésiastiques latins de l'antiquité et du haut moyen âge, 5th edition, Freiburg 2007.

Kajanto (1966): I. Kajanto, Supernomina: A Study in Latin Epigraphy, Helsinki 1966.

Keudel (1990): U. Keudel, Thesaurus linguae Latinae: Praemonenda de rationibus et usu operis, Leipzig 1990, https://tll.degruyter.com/help (last access 04.04.2025).

Krömer (2003): D. Krömer, Don't Trust the Label: Zur Datierung lateinischer Texte, in: H. Solin / M. Leiwo / H. Halla-aho (ed.), Latin vulgaire – latin tardif: Actes du Vie colloque international sur le latin vulgaire et tardif, Hildesheim et al. 2003, 183–189.

Krömer / van Leijenhorst (1990): D. Krömer / C. G. van Leijenhorst, Thesaurus linguae Latinae: Index librorum scriptorum inscriptionum ex quibus exempla afferuntur, 2nd edition, Teubner 1990.

Marchionni (2015): R. Marchionni, Latein als Sprache des Thesaurus linguae Latinae, Akademie Aktuell 53 (2015), 54–57, https://badw.de/fileadmin/pub/akademieAktuell/2015/53/0215_13_Marchionni_V05.pdf (last access 24.04.2025).

Molinier (1873–1876): A. Molinier (ed.) in: U. C. Bussemaker / C. Daremberg (ed.), Œuvres d'Oribase, voll. 5 and 6, Paris 1873–1876.

Morin (1917): G. Morin (ed.), Sancti Aureli Augustini tractatus sive sermones inediti ex codice Guelferbytano 4096, Kempten / Munich 1917.

Oomes (1968): F. Oomes, Nomina ominosa, in: Lemmata: Donum natalicium Guilelmo Ehlers sexagenario a sodalibus Thesauri linguae Latinae oblatum, München 1968, 221–231.

Pantelia (2022): M. C. Pantelia, Thesaurus Linguae Graecae: A Bibliographic Guide to the Canon of Greek Authors and Works, Oakland (CA) 2022.

Ribbeck (1897): O. Ribbeck (ed.), Scaenicae Romanorum poesis fragmenta, third edition, Leipzig 1897.

Scappaticcio (2019): M. C. Scappaticcio, Testi latini su papiro e lessicografia. In margine ad un contributo possibile al Thesaurus Linguae Latinae, BStudLat 49 (2019), 689–698.

Vollmer (1904): F. Vollmer, Thesaurus linguae Latinae: Index librorum scriptorum inscriptionum ex quibus exempla adferuntur, Leipzig 1904.

Vogt (1995): E. Vogt, Ein Gräzist benutzt den Thesaurus, in: D. Krömer (ed.), Wie die Blätter am Baum, so wechseln die Wörter: 100 Jahre Thesaurus linguae latinae, Stuttgart / Leipzig 1995, 99–109.

Author Contact Information[52]

Dr. Adam Gitner
Thesaurus linguae Latinae / Bayerische Akademie der Wissenschaften
Alfons-Goppel-Str. 11
80539 München
E-mail: agitner@thesaurus.badw.de

52 The rights pertaining to content, text, graphics, and images, unless otherwise noted, are reserved by the author. This contribution is licensed under CC BY-SA 4.0.

DIGITAL CLASSICS ONLINE

Daidalos: NER for Literary Studies on Latin and Ancient Greek Texts

Andrea Beyer

Abstract: Literary texts offer a wealth of unstructured data that can be harnessed for data-driven text analysis through Natural Language Processing (NLP). Named Entity Recognition and Classification (NER) is a crucial initial step in this process, enabling the automatic identification of entities such as persons, organizations, locations, and dates. However, NER faces significant challenges, particularly with historical texts in low-resource languages like Latin and Ancient Greek, due to limited annotated corpora and the dynamic nature of language. This paper explores the evolution of NER from simple extraction to semantics-aware entity disambiguation and linking, highlighting the importance of multi-layer annotation systems to enhance data quality and model accuracy. The interdisciplinary *Daidalos* project aims to bridge the gap between Digital Humanities and Classical Studies by providing an NLP infrastructure that supports various data-driven research methods, among others NER. One of the project's case studies demonstrates the potential of NER in Classical literary studies; this is accompanied by proposals on other NER related literary research questions, e.g. on authorship attribution and stereotyping. Additionally, the paper offers some thoughts about teaching NER, presenting a framework to assess the required level of digital literacies when working on a specific research question. Finally, it discusses the implications of generative AI and Large Language Models (LLM) on NER and NLP in Classics, emphasizing the challenges for independent research posed by the high costs and limited transparency of LLMs.

Introduction[1]

Literary texts provide a lot of unstructured data that needs to be extracted and structured, so that a computer can support a data-driven text analysis (Natural Language Processing, NLP). Therefore, automatic taggers search, retrieve and explore the information in a given text corpus and annotate the text accordingly. One of the first and most crucial processing steps in doing so is the extraction technique Named Entity (NE)[2] Recognition and Classification (NER for short). This method enables the automatic identification of entities in texts, entities being for the most part categorised as follows: person, organisation, location and date. This is an essential part of question answering, media monitoring, and opinion mining, however it also helps with machine translation, text summarisation, and text classification.[3] It might additionally be useful for subsequent Information Extraction (IE) tasks like Rela-

1 Acknowledgements: This work is part of a project funded by the German Research Foundation (project no. 518919950) and led by Andrea Beyer, Malte Dreyer, and Anke Lüdeling.

2 Ehrmann et al. (2021), 5: "[…] named entities correspond to different types of lexical units, mostly proper names and definite descriptions, which, in a given discourse and application context, autonomously refer to a predefined set of entities of interest. There is no strict definition of named entities, but only a set of linguistic and application-related criteria which, eventually, compose a heterogeneous set of units." For further discussion see Ehrmann (2008).

3 Ehrmann et al. (2021), 2.

tion Extraction and Entity Linking[4] as well as for Topic Classification, Event Timelines, Network Analysis and various other analysis techniques.[5] Thus, NER has also undergone an evolution by shifting its focus from the mere extraction of information, i.e. the detection and classification of NE, to a more semantics-aware viewpoint, i.e. entity disambiguation and linking, "which can support the cross-linking of multilingual and heterogeneous collections based on authority files and knowledge bases".[6]

Essentially, NER is a sequence-labelling task that is enriched with features at three levels: the morphological level (words, e.g. *Roma, Zeus, Galli*), contextual level (close context or sentences, e.g. *C. Iulius Caesar*, *consulibus M. Tullio Cicerone et C. Antonio Hybrida*), and text level (document, e.g. *C. Plinius Traiano Imperatori*). Developed NER Systems are evaluated in terms of Precision (P), Recall (R) and F-measure (F-score, the harmonic mean of P and R) as well as more fine-grained evaluation metrics.[7] Additionally, the accuracy of NER results is dependent on which resource methods have been used, as well as the inherent challenges that certain research areas may pose, such in the case of Latin and Ancient Greek; here, difficulties may emerge due to the significant changes in the use of these languages over time, as well as the long transmission history of Latin and Ancient Greek texts.

Resource Types

In order to develop NER systems four types of resources exist:[8]

- Typologies: Typologies define a semantic framework for the entities under consideration. They constitute the source of annotation guidelines.

- Lexicons and knowledge bases: On the one hand, information about NE can be of lexical nature given verbatim in a textual unit. By using look-up procedures in lexicons, the NE might be extracted. On the other hand, information about NE can be encyclopaedic in nature, i.e. non-linguistic information on referred entities. For extracting this kind of information, knowledge bases like *Wikipedia* or *Wikidata* are widely used.

- Word embeddings and language models: Word embeddings are dense, low-dimensional vectors that are acquired from the distribution of words in continuous texts and represent the meaning of words. Their ability to be obtained self-supervised, i.e. from unlabelled data, is a major benefit that makes the shift from feature engineering to feature learning possible.

- Corpora: These can consist of labelled and/or unlabelled textual data. Unlabelled texts are used to train language models and embeddings, while labelled corpora are utilised as a learning base or as a point of reference for evaluation purposes.[9]

Low resources are one major problem for research communities working with historical texts and languages, because these communities are rather small and underfinanced. Therefore, they lack the power to significantly improve the unsatisfying situation.

4 Feng et al. (2018), 4071.

5 Chastang et al. (2021).

6 Ehrmann et al. (2021), 3. Other NE-related specific research directions are temporal information processing and geoparsing: Ehrmann et al. (2021), 5.

7 Ehrmann et al. (2021), 6–7.

8 Ehrmann et al. (2021), 7–8.

9 E.g. Latin NER with literary texts (1st century BC – 2nd century AD), 7.175 NE: Person, Location, Group. See Erdmann et al. (2016).

Methods

The engineering of NER systems is performed by four families of algorithms:[10]

	Description	Example	Constraints
Lexicon-based approaches	NE are detected by comparing a dictionary with the list of words in the selected corpus. These look-up procedures work better for historical data, because the lexicons do not require constant updating and careful maintenance to stay accurate and effective.	*Trismegistos*, Domain: state (papyri, 4C–1C BC, languages: Egyptian, Ancient Greek, Latin, cf. Broux / Depauw [2015]); *Pleiades*; *Lexicon of Greek Personal Names*; *Prosopographia Imperii Romani.*	Resources might not be available (digital, open access) nor well-maintained (updated databases). The complexity of NE is a problem in itself related to the ambiguity of names (e.g. father and son: same name, different person), spelling variation (e.g. *Caius* or *Gaius*), abbreviations (e.g. *C./ G.*, *SPQR*), patronyms (e.g. *Pelopides* – "descendants of Pelops"), and metonyms (e.g. *Tonans* – Jupiter). It is almost impossible to include all potential NEs. Chastang et al. (2021), 9: "Training a dictionary-based recognition model against a list of names can lead to a high ratio of recognition for a particular corpus, but the model is often not robust when applied to unseen texts or different types of data."
Rule-based approaches	NE rules are manually crafted by a developer or linguist on the basis of regularities (patterns) observed in the data, e.g. derivates for ancestry (*-ides*). Rule-based approaches have the advantage of not requiring training data and of being easily interpretable. In contrast, their design needs time and expertise and is thus costly. Chastang et al. (2021), 9: "[…] rule-based models […] show a valid global recall, but a slight tendency to poor precision-ratio on unseen texts."	"*ruLe-based lOCation nAmed-entity recognition method for Latin tExt*" (*LOCALE*); for Ancient Greek no example.	Morphological analysis (stemming, lemmatisation) might be useful for applying the rules, but this comes with a certain percentage of errors.

10 Chastang et al. (2021); Ehrmann et al. (2021), 8–10.

Machine-learning (ML) based approaches	These approaches are also called feature-based, because they use labelled data and learn from it to recognise patterns for applying them on unlabelled data. Due to their capacity to take into account the neighbouring tokens, conditional random fields (CRF) proved particularly well-suited for NER tagging and became the standard for feature-based NER systems.	*Classical Language Tool Kit* (*CLTK*); *CRF*. Domain: news (medieval charters, 10C-13C, cf. Aguilar et al. [2016]); Domain: literature (Classical texts, 1C BC-2C, cf. Erdmann et al. [2016], Beersmans et al. [2023]; Domain: literature (Herodotus, cf. Palladino et al. [2020]).	The ML algorithms require a lot of labelled data which is usually scarce. Besides, in order to resolve ambiguities, a deep understanding of the context (sentences, paragraphs, document) is necessary, something which is a challenge for a rather simple ML algorithm.
Deep learning (DL) approaches	DL systems use artificial neural networks with multiple processing layers, which is why they are called neural-based approaches. On the basis of word embeddings, DL models the semantic and syntactic relationship between various words by learning representations of the given data (corpus) with multiple levels of abstraction. The key benefit of neural networks is their ability to automatically learn input representations instead of relying on manually elaborated features – whether or not the input is topic-specific or rather general.	*Latin BERT* (trained on 640 million tokens spanning 22 centuries, cf. Bamman / Burns [2020]);LatinCy; *Grε(BERT|T)A*, *PHIL(BERT|T)A; AG_BERT_hypopt_reduced_NER; grc-ner-bert; flair_grc_bert_ner*	Deep learning models "frequently operate as opaque 'black boxes' with limited transparency in their decision-making processes" (Sankarapu et al. [2024], 1).

Tab. 1: Methods of Named Entity Recognition (NER).

Challenges of NE Recognition and Classification

Generally, NER on historical documents, particularly those that are written in low-resource languages like Latin and Ancient Greek, faces considerable challenges due to a lack of resources, in particular the documentation on high-quality annotated datasets.[11] The lack of resources for Latin and Ancient Greek summarised by Burns (2019), is still valid:

> "[…] named entity recognition (NER), or the systematic tagging of words in texts by category (so, Roma as a "location" or Σωκράτης as a "person") is not well-supported by standalone tools. With respect to Greek and Latin, a lack of annotated texts and robust language models underlies the problem."[12]

11 Beersmans et al. (2023).

12 Burns (2019), 168.

Even though the AI-driven development in the last few years has brought some advances relating to the models and evaluation procedures, the overall trend has not changed, particularly concerning annotated corpora.

Apart from the quality of data and methods there are challenges inherent to language corpora which contain texts from different genres, places, and eras. Although the Latin and Ancient Greek texts belong mostly to literature and use standardised, formal language, the dynamics of language might cause some troubles extracting and labelling NE correctly. These errors can be attributed to problems related to normalisation, genre-specifics, ambiguity, and multi-word expressions:

- Normalisation: Without normalising NE to a common standard, circumstances such as spelling variations and slightly different naming conventions may produce multiple results for the same NE, e.g. **C.** Iulius Caesar and **G.** Iulius Caesar, *par**v**um* and *par**u**um*.

- Genre specifics: Poetic texts in particular provide stylistic or rhetorical expressions like paraphrases, metaphors and metonymies, which obfuscate the NE for a non-expert like a standard automatic NER-tagger, e.g. *urbs* instead of Rome, *tonans* instead of Jupiter, *Dis* instead of Pluto.

- Ambiguity: Some of the naming conventions of antiquity like homonyms between father and son are very challenging for automatic NER taggers because they do not 'understand' the concept of one name and two referents, i.e. they were not trained to distinguish multiple homonymous referents. Similarly, this applies to adjectives, e.g. *romanus* meaning Roman (adjective, no NE) or Roman (noun, NE: person). In both examples, a NER tagger is prone to make mistakes in the classification process, e.g. identifying too many or too few entities.

- Multi-word expressions: Occurrences of multi-word expressions ('composed entities') come with different challenges. Firstly, they consist of more than one word with different linguistic construction possibilities like *vallum Hadriani*, *consulibus M. Tullio Cicerone et C. Antonio Hybrida*, *P. Lentulus Sura*, *P. et Ser. Sullae Ser. Filii*. Secondly, they might occur as discontinuous text spans like *rebus Sancti Vincentii Matiscensis*. Thirdly, they might be nested into each other or might overlap like *Guillelmus de Sancti Stephano de Ponte*. In every case they are highly complex and therefore mostly not recognised correctly, i.e. the NE will probably be identified only partially and often not correctly classified.

Approaches to enhance NER results

As mentioned above, to gain better results it is necessary to increase the quantity of data for training and evaluation purposes as well as to improve the quality of this data. By providing more annotated texts of different genres and eras the issue of quantity could be addressed, but without more sophisticated annotations the aforementioned challenges would be almost the same. Thus, the quality of data and consequently of the models needs to be enhanced. Accordingly, it is suggested to enhance the annotations by introducing a so-called multi-layer annotation system; this allows researchers to make explicit annotations in cases of uncertainty and to reveal the distinction in complexity between manual and automatic annotation through one layer of automatic annotation and multiple layers for manual ones.[13] Most notable is the latter functionality, which offers insights into the contrast between a broader automatic approach and a more subtle manual approach. Based on these multi-layer annotations, NER taggers using a DL approach could learn to recognise patterns of fuzzy multi-word expressions, extract them in their entirety as one NE, and classify the NE accurately.

13 Chastang et al. (2021).

Application of NER in Classics

As part of the Humanities and *inter alia* concerned with literary studies, modern Classical Studies has intersections with the Digital Humanities (DH) and with the Computational Literary Studies (CLS). Nowadays, digital research and teaching methods enrich the established methodology and offer new perspectives on the interaction with Latin and Ancient Greek texts. This encompasses all main areas of German Classical Studies: editions, translations, commentaries, interpretations, didactics. However, there is still a huge knowledge gap between the people who develop new resources and tools and the people who might use them for their research and teaching. This gap is two-sided: On the one hand, 'traditional' researchers and teachers lack a sufficient level of digital literacies[14] for fully exploiting the provided tools and methods. On the other hand, specialists for computational methods are often not sufficiently acquainted with the domain-specific needs and research interests – occasionally, they are not even familiar with the Classical languages or literary studies. Thus, matching the research interests and goals of both perspectives sometimes seems impossible. This marks an exciting starting point for the *Daidalos* project, which aims to develop an NLP research infrastructure for different competency levels.

Daidalos: Key Goals and Features

The interdisciplinary *Daidalos* project[15] intends to bring computational and traditional approaches closer together by providing an NLP infrastructure which offers a software-as-a-service for NLP methods like NER, word embeddings, or sentiment analysis. Additionally, *Daidalos* includes working in research tandems, offering material for further education, and consulting researchers or research groups on third-party funding or more generally on research questions appropriate to digital methods. Therefore, *Daidalos* has the following goals and features:

- Bridging the Gap between DH and CLS on one side and German (!) Classical Studies on the other side by working in so-called research tandems for a better understanding of what literary researchers need and want.[16]
- No-code and low-code access for various data-driven research methods and data visualisations.[17]
- Reuse of existing resources and literature overview on application possibilities.
- Systematic Evaluation Framework for NLP Models and Datasets in Latin and Ancient Greek: SEFLAG.[18]

14 This term sums up different technology-driven literacies, i.e. digital literacy, data literacy, and AI literacy. A domain-specific and case-sensitive theoretical model and framework has been designed as part of the *Daidalos* project. See Beyer (2026).

15 https://daidalos-projekt.de (last access 11.07.2025). The *Daidalos* Project (2023–2026) is funded by the German Research Foundation. Project number: 518919950. *Daidalos* is on GitLab (https://scm.cms.hu-berlin.de/daidalos/daidalos-platform [last access 11.07.2025]) and Zenodo (https://zenodo.org/communities/daidalos/ [last access 11.07.2025]).

16 Beyer / Schulz (2023a) and (2024).

17 Beyer / Schulz (2023b).

18 Schulz / Deichsler (2024).

Feature	Already implemented	Planned
Corpus of Latin and Ancient Greek texts	*Perseus* via *Perseids API*	Integration of other digital corpora like *LASLA*, *MQDQ*
NLP methods and visualisation	Pre-processing, word embeddings, NER, sentiment analysis	Part-of-Speech, topic modelling, fine-tuning of parameters
Workshop material	(Hosted on GitLab:) Jupyter Notebooks on Markdown/JN, data preprocessing, word embeddings, sentiment analysis, NER	Integration via JupyterLite for low-code access and increased flexibility during research
SEFLAG	NER, lemmatisation, dependency parsing	Expanding, web integration
Databases	Small documentation on resources	Research literature, AI tools
User-friendly functionality		Text upload, result download
Identity Access Management		Settings, own corpus

Tab. 2: Overview on features of the *Daidalos* NLP research infrastructure.

The *Daidalos* project will apply for a second funding period, in which, amongst other things, the further development of domain-specific NLP methods shall be addressed.

NER – Underrated for Literary Studies in the Classics?

The methods of information retrieval (IR) and linguistics share a common fate in (traditional) literary studies in (German) Classics. Although they might be called the foundation ('ground truth') of any qualitative analysis – in the form of an interpretation of a passage or a text, they are often not explicitly mentioned or are even despised because of their data-driven approaches. Nevertheless, of course, they are used by philologists to analyse text patterns and extract information from texts about e.g. people, author, or historical context. That being said, the key challenge for establishing data-driven methods in the community of practice of Classical philologists is rather a question of awareness,[19] both of methodology and methodological competence. For this reason, *Daidalos* has introduced the so-called research tandems. Both partners know the domain, but they apply different methods for solving their research questions – *close reading* vs. *distant reading*.[20]

In one of these case studies, NER is part of a workflow to find answers to the underlying research question: How can 'gaps' in historiographical texts be found using digital methods?[21] For modelling reasons this question is reduced to a manually analysed example: Can the absence of the historical event 'Conference of Luca' in Cassius Dio's *Roman History* be detected with digital methods?[22] The

19 De Carvalho-Filho et al. (2020).

20 Schubert (2015).

21 This research question addresses the problem whether an omission is based on a source problem or is intentionally done as part of a conscious decision (literary function).

22 The Conference of Luca was held 56 BC by the triumvirs Caesar, Pompey, and Crassus. It is mentioned by Cicero, Plutarch, Velleius Paterculus, Suetonius, and Appian. In contrast, Cassius Dio does not refer to it.

test corpus consisted of nine references, both in Latin and Ancient Greek texts.[23] Thus, two language specific NER taggers were applied (see tab. 3).

	Latin	Ancient Greek
Model Name	la_core_web_lg	UGARIT/flair_grc_bert_ner
Publication	Burns, 2023	Yousef et al., 2023
NLP Software	spaCy	Flair NLP
Architecture	floret vectors Transition-based Parser	BERT (transformer) vectors long short-term memory network conditional random field
Training Data	Caesar, Ovid, Pliny (Elder & Younger)	Homer, Herodotus, Athenaeus
Tagset	persons, locations	persons, locations, peoples

Tab. 3: Reuse of resources in *Daidalos*: NER taggers for Latin and Ancient Greek.

As it is shown in fig. 1, it was possible to find the passage containing the conference of Luca through visualising the NE.[24] Nevertheless, there was also one false positive (Plut. Caes. 21, 4) that could be explained by the literary scholar, but proved what is generally accepted about NLP tasks: having a human in the loop is essential for evaluating the results. In relation to the research request, the results were good enough for a proof-of-concept state, esp. because the known gap (Cass. Dio 39, 24–36) stayed blank.

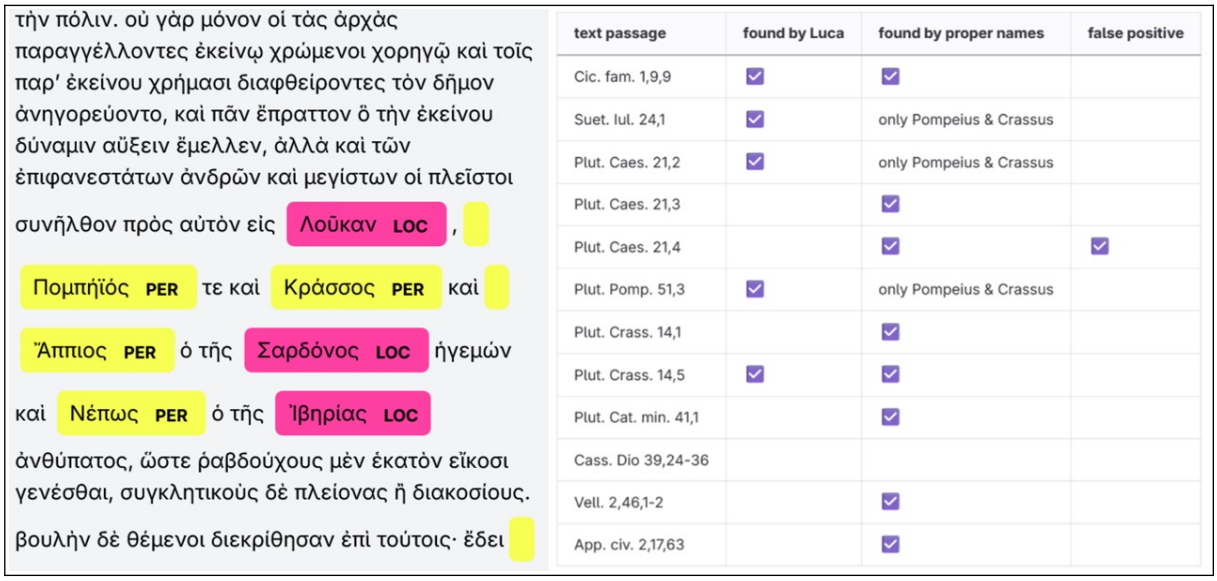


Fig. 1. Left: Example (Plut. Caes. 21) of NER visualisation with *Daidalos*; Right: Overall evaluation of NER results for the test corpus.

23 Cass. Dio 39,24–36; Cic. Fam. 1,9,8–9; Suet. Iul. 24,1; Plut. Caes. 21; Plut. Pomp. 51; Plut. Crass. 14–15; Plut. Cat. min. 41,1–2; App. Civ. 2,17; Vell. 2,46,1–2.

24 If someone already knows all NEs beforehand, he/she could use regular expressions as well. However, using the NER-tagger is a proof-of-concept for finding specific passages without knowing all NEs.

Based on these findings, the research tandem decided to continue further with a combination of NER and sentiment analysis to test a second hypothesis[25] that explains why Cassius Dio might not have mentioned the conference of Luca. The first findings are encouraging, namely that the polarity of the Cassius Dio passage is slightly more negative than any of the others.[26] Thus, the test corpus for this analysis is going to be expanded to the whole work of Cassius Dio, in order to evaluate whether the workflow 'NER and sentiment analysis' paves the way sufficiently to tackle the overall research question, i.e. finding omissions in historiographical Latin and Ancient Greek literature.

Obviously, the presented case study is just one possible application of NER in Classics. Inherently, NER techniques support complex information extraction and classification in corpora too big to read in the provided amount of time (cf. distant reading). This way, they enable literary scholars to visualise frequencies, sequences, recurrent patterns, and hot spots to gain systematic and replicable insights into specifically assembled corpora. Such being the case, some other potential literary use cases are outlined below:

Authorship Attribution

- Corpus: Ps.-Sallust, *Invectiva in M. Tullium Ciceronem*.
- Background: By now, most researchers agree that this speech is an example of a *declamatio* of the late Augustan era, but esp. in Germany some researchers still try to attribute it to Sallust.
- Research question: If the speech is situated in 54 BC, do the timeline and chronology of appearing persons match this date? What evidence is there for a *terminus post quem*?
- Requirements: Digitised Text and a knowledge base of events and persons up to 54 BC.

Relationship of Title Heroes and Protagonists

- Corpus: Sallust, *Bellum Catilinae* & *Bellum Iugurthinum*.
- Background: Both monographs are named after one person (Catilina & Iugurtha) exemplary for the danger they posed to Rome. However, there are other main characters (Caesar & Cato; Marius & Sulla) who drive on the events.
- Research question: How is the relationship between eponymous hero and protagonists weighted? How similar are the two monographs in this respect to each other?
- Requirements: In addition to NER, social network analysis is required.

Time as a Device for Framing

- Corpus: e.g. Cicero, *De Amicitia*, *Tusculanes*; Plato, *Protagoras*, *Politeia*.
- Background: In antiquity, the genre *dialogus* is common to 'teach' philosophy. Some are explicitly set in the past, others span multiple days in a row.

25 Cassius Dio attributes early on a negative relationship between Pompey and Caesar, which might be the reason for Luca's 'gap'. If so, this should be reflected by the (latent) emotions/moods in the text, i.e. more negative polarity in Cassius Dio than in the other passages of the test corpus.

26 The sentiment analysis for both languages is a challenge in it-self, esp. for comparable resources and models. Regardless, this NLP task is introduced in this paper only for the reason of demonstrating a pipeline in which NER is one of the first steps to discover new insights in the classical-philological literary studies.

- Research question: How is time used in the *dialogi*? Do references to time statements fit the fictitious frame of a dialogue? How do time statements support the narration? What can be inferred about the meaning of time in literature of the same era?

- Requirements: <DATE> annotation in a training corpus and a newly trained model.

Stereotypes of Organisations or People over Time

- Corpus: Historiographical texts from 100 BC to 300 AD.

- Background: Different political systems and becoming an empire may have had an influence on Latin literature and the ideas conveyed by the historiographical works.

- Research question: How do political organisations and their representatives evolve over time? Are the descriptions rather like stereotypes in the beginning? Do they evolve into stereotypes? Are certain peoples more prone to be stereotyped than others?

- Requirements: <ORG> and <PEOPLE> annotation in a training corpus and a newly trained model. In addition to NER, social network analysis and word embeddings are recommended.

These examples are only thought experiments, of course, and each idea would need more consideration before being applied in literary research. But they hint at the potential value of an IR method like NER in Classical literary studies, if both types of researchers work together on research questions.

NER as Part of Teaching Digital Literacies

As shown above, it is worth knowing how to use NER in the Classical Studies, but the majority of researchers are lacking the necessary skills. Above all, this poses a challenge for the future of Classics in itself, because the researchers are also teachers. So, if the researchers are too unskilled in digital research methods, they will not or cannot teach in a way that enables students to acquire a certain level of digital literacies. Consequently, future researchers and teachers will still have the same problem they face right now. That is why the *Daidalos* project has also developed a framework for describing domain-specific and use-case-sensitive levels of three digital literacies, i.e. digital literacy, data literacy, and AI literacy,[27] which are aggregated as digital literacies. This framework is used to define what is needed for accomplishing digital-based research. Thus, it helps in providing support to literary scholars, and in developing learning material like the curated Jupyter Notebooks.

For example, in the case study presented above, the literary scholar did not know anything about NER (which was expected), but lacked also ‘basic’ knowledge about file formats or even the term ‘annotation’ (which was unexpected, see tab. 4). For one thing, why should a literary scholar be expected to know about annotations? On the other hand, these expectations are reasonable from the view of those researchers who are accustomed to NLP and surrounded by digital devices and data-driven approaches. In fact, both ‘sides’ would be right in their own perspective, illustrating the aforementioned two types of researchers who have difficulties to understand each other. For that reason, although it is hard work to adapt the framework for each case study, it has proven very useful for communication and collaboration in the research tandems.

27 There is an abundance of literature on each of the mentioned literacies and some more that might also be part of the digital literacies framework, e.g. media literacy or information literacy. However, deploying NLP services boils down to the fact that the focus is set on AI and supportive competences related to data and general digital abilities.

Dimension	Get to know	Acquire	Deepen	Create
AI literacy	NER concepts, e.g. taggers, model cards, datasheets	Comparing NER taggers and results	Evaluating NER errors systematically	Refining NER taggers through feedback
Data literacy	Knowledge about annotations in general	Annotating a test corpus	Using a multi-layer annotation scheme	Creating annotation guidelines
Digital literacy	Difference DOCX & TXT	Converting DOCX to TXT	Identifying challenges for pre-processing in TXT	Modifying TXT depending on the use case

Tab 4: Domain-specific and use-case-sensitive framework to describe the needed level of digital literacies for working on an NER task. Not all characteristics need to be acquired for accomplishing a task, but the full overview breaks down the complex requirements into learnable modules.

Obviously, the framework might also be applied to teaching and could facilitate setting the goals of tasks, courses or curricula by structuring explicitly the needed skills. Even if it is mostly not necessary to acquire the full pattern of digital literacies for the selected use case, providing a variety of frameworks for different use cases in Latin and Ancient Greek philology might make it easier for traditionally trained scholars to get an idea of what level of digital literacies is required to apply digital research methods. Consequently, it would be possible to address this need and create learning units with authentic research questions, in order to improve digital literacies in Classics and therefore break the vicious circle of lacking digital competence.

Generative AI and the Future of NER and NLP in Classics

Finally, it is necessary to talk about the implications of the frantic development of generative AI related to NLP tasks and Classics as well. Some people might ask why they should bother to acquire digital literacies, if they just as well could use an AI chatbot, i.e. a Large Language Model (LLM)[28] with a graphical user interface. This opens up a discussion that not only concerns the quality of these models and further technology advancements, but also core values of research like autonomy, replicability, transparency, and openness: In a nutshell, what we can do and what we should do.

Getting back to NER, there is some evidence that LLMs, namely the GPT models, can be deployable for NER tasks (Wang et al. 2023). Although the "intrinsic gap between the two tasks of NER and LLMs"[29] and "the hallucination issue, where LLMs have a strong inclination to over-confidently label NULL inputs as entities"[30] still exist, the researchers showed that using in-context learning – primarily few-shot prompting – combined with a self-verification process of the LLMs is especially beneficial "in the low-resource scenario […] when the amount of training data is extremely scarce"[31]. Consequently, this paper indicates that, by prompting, it is possible to structure unstructured data and process this data in such a way and with such results that are comparable to what specialised taggers

28 As mentioned above, deep learning approaches are being tested and adapted right now for Latin and Ancient Greek. Thus, LLMs are already being applied. In contrast to this meaning, in the following context, the term LLM is related to foundation models that are domain-unspecific and accessed through a chatbot interface by prompting a question or task in the natural language of the user.

29 Wang et al. (2023), 1: "NER is a sequence labeling task in nature, where the model needs to assign an entity-type label to each token within a sentence, while LLMs are formalized under a text generation task."

30 Wang et al. (2023), 2.

31 Wang et al. (2023), 11.

would do. Nevertheless, it should be added that, in the light of the evolution of LLMs to Large Action Models (LAM), these findings might soon be outdated. With the arrival of agents[32], a task could be broken up into a lot of individual actions and binary decisions which follow an even better workflow than the prompt engineering so far. Accordingly, the need for training models or curating resources could be reduced, especially for low-resource languages like Latin and Ancient Greek, though, in fact, nobody can really make an accurate estimation for the future.

Despite this, a much more intricate problem is involved when talking about LLMs – in particular LAMs – or AI-driven technology in general. As it is known,[33] the costs for training a LLM, but also for fine-tuning, applying techniques like retrieval-augmented generation (RAG) and low-rank adaptation (LoRA), and for operating the LLMs are extremely high. Thus, the usage of LLMs creates new boundaries, because only global private players are able to finance the 'whole AI ecosystem' so far. This ever-growing dependency on a non-scientific, profit-oriented community comes with crucial drawbacks for an independent research community:

- reduced accessibility due to high prices, costly infrastructure and lack of data privacy,
- restricted replicability due to a lack of openness of models and training data,
- unclear biases due to training the models on the whole internet,
- limited or no transparency at all due to privately owned solutions and the DL approach.

Under these conditions, small domains like Classics have an even harder time working autonomously with digital-based research methods. This means that abstaining from 'buying' seemingly easy access to LLMs, and keeping instead to the old-fashioned NLP approaches might be the only salubrious solution yet. Consequently, researchers still need to learn about NER as an NLP task and become digitally literate.

32 Xi et al. (2023), 1: "AI agents are artificial entities that sense their environment, make decisions, and take actions."

33 Maslej et al. (2024).

Online Sources

https://pleiades.stoa.org/places (last access 11.07.2025).

https://www.lgpn.ox.ac.uk/ (last access 11.07.2025).

https://pir.bbaw.de/ (last access 11.07.2025).

https://trismegistos.org (last access 11.07.2025).

https://github.com/Ivona221/LOCALE (last access 11.07.2025).

http://cltk.org/ (last access 11.07.2025).

http://daidalos-projekt.de (last access 11.07.2025).

References

Bamman / Burns (2020): D. Bamman / P. J. Burns, Latin BERT: A Contextual Language Model for Classical Philology, online 2020, http://arxiv.org/abs/2009.10053 (last access 11.07.2025).

Beersmans et al. (2023): M. Beersmans / E. de Graaf / T. Van de Cruys / M. Fantoli, Training and Evaluation of Named Entity Recognition Models for Classical Latin. Proceedings of the Ancient Language Processing Workshop, online 2023, 1–12, https://aclanthology.org/2023.alp-1.1.pdf (last access 11.07.2025).

Beersmans et al. (2024): M. Beersmans / A. Keersmaekers / E. De Graaf / T. Van De Cruys / M. Depauw / M. Fantoli, "Gotta catch 'em all!": Retrieving people in Ancient Greek texts combining transformer models and domain knowledge. Proceedings of the 1st Workshop on Machine Learning for Ancient Languages, online 2024, 152–164, https://doi.org/10.18653/v1/2024.ml4al-1.16 (last access 11.07.2025).

Beyer (2026, forthcoming): A. Beyer, Textanalyse mit KI: Wie kommt sie in die Klassische Philologie?, in: S. Faller / W. Polleichtner (ed.), Digitalität im Unterricht der Alten Sprachen. (Digitale) Chancen für den Lateinunterricht, Baden-Baden 2026.

Beyer / Schulz (2023a): A. Beyer / K. Schulz, DAIdalos: Forschen und Lernen zugleich?, online 2023, 391–393, https://doi.org/10.18420/inf2023_42 (last access 11.07.2025).

Beyer / Schulz (2023b): A. Beyer / K. Schulz, Data Literacy für die Klassische Philologie: d^{AI}dalos – eine interaktive Infrastruktur als Lernangebot, online 2023, https://doi.org/10.5281/zenodo.8420565 (last access 11.07.2025).

Beyer / Schulz (2024): A. Beyer / K. Schulz, Daidalos: Wie viel Methodenkompetenz braucht ein User? Book of Abstracts – DHd2024, online 2024, 336–338, https://doi.org/10.5281/zenodo.10698299 (last access 11.07.2025).

Broux / Depauw (2015): Y. Broux / M. Depauw, Developing Onomastic Gazetteers and Prosopographies for the Ancient World Through Named Entity Recognition and Graph Visualization: Some Examples from Trismegistos People, in: L. M. Aiello / D. McFarland (eds.), Social Informatics, Heidelberg 2015, 304–313, https://doi.org/10.1007/978-3-319-15168-7_38 (last access 11.07.2025).

Burns (2019): P. J. Burns, Building a Text Analysis Pipeline for Classical Languages, in: M. Berti (ed.), Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution, Berlin / Boston 2019, 159–176, https://doi.org/10.1515/9783110599572-010 (last access 11.07.2025).

Burns (2023): P. J. Burns, LatinCy: Synthetic Trained Pipelines for Latin NLP, online 2023, https://doi.org/10.48550/arXiv.2305.04365 (last access 11.07.2025).

Chastang et al. (2021): P. Chastang / S. O. Torres Aguilar / X. Tannier, A Named Entity Recognition Model for Medieval Latin Charters, Digital Humanities Quarterly 15/4 (2021).

de Carvalho-Filho et al. (2020): M. A. de Carvalho-Filho / R. A. Tio / Y. Steinert, Twelve tips for implementing a community of practice for faculty development, Medical Teacher 42/2 (2020), 143–149, https://doi.org/10.1080/0142159X.2018.1552782 (last access 11.07.2025).

Ehrmann (2008): M. Ehrmann, Les Entités Nommées, de la linguistique au TAL: Statut théorique et méthodes de désambiguïsation. Paris 2008, https://hal.science/tel-01639190v1/document (last access 11.07.2025).

Ehrmann et al. (2021): M. Ehrmann / A. Hamdi / E. L. Pontes / M. Romanello / A. Doucet, Named Entity Recognition and Classification on Historical Documents: A Survey, ACM Computing Surveys 56/2 (2021), 1–47, https://doi.org/10.1145/3604931 (last access 11.07.2025).

Erdmann et al. (2016): A. Erdmann / C. Brown / B. Joseph / M. Janse / P. Ajaka / M. Elsner / M.-C. De Marneffe, Challenges and Solutions for Latin Named Entity Recognition. Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), online 2016, 85–93.

Feng et al. (2018): Feng X. / Feng X. / Qin B. / Feng Z. / Liu T., Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge Transfer. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, online 2018, 4071–4077. https://doi.org/10.24963/ijcai.2018/566 (last access 11.07.2025).

Maslej et al. (2024): N. Maslej / L. Fattorini / R. Perrault / V. Parli / A. Reuel / E. Brynjolfsson / J. Etchemendy / K. Ligett / T. Lyons / J. Manyika / J. C. Niebles / Y. Shoham / R. Wald / J. Clark, Artificial Intelligence Index Report 2024, online 2024, https://arxiv.org/abs/2405.19522v1 (last access 11.07.2025).

Palladino et al. (2020): C. Palladino / F. Karimi / B. Mathiak, NER on Ancient Greek with minimal annotation, online 2020, https://doi.org/10.17613/j7jt-b052 (last access 11.07.2025).

Palladino / Yousef (2024): C. Palladino / T. Yousef. Development of Robust NER Models and Named Entity Tagsets for Ancient Greek, LT4HALA, online 2024, https://aclanthology.org/2024.lt4hala-1.11.pdf (last access 11.07.2025).

Riemenschneider / Frank (2023), F. Riemenschneider / A. Frank, Exploring Large Language Models for Classical Philology, online 2023, https://doi.org/10.48550/arXiv.2305.13698 (last access 11.07.2025).

Sankarapu et al. (2024): V. K. Sankarapu / C. Chitroda / Y. Rathore / N. K. Singh / P. Seth, DLBacktrace: A Model Agnostic Explainability for any Deep Learning Models, online 2024, https://doi.org/10.48550/arXiv.2411.12643 (last access 11.07.2025).

Schubert (2015): C. Schubert, Close Reading und Distant Reading. Methoden der Altertumswissenschaften in der Gegenwart, Digital Classics Online 1,1 (2015), 1–6, https://doi.org/10.11588/dco.2015.1.20483 (last access 11.07.2025).

Schulz / Deichsler (2024): K. Schulz / F. Deichsler, SEFLAG: Systematic Evaluation Framework for NLP Models and Datasets in Latin and Ancient Greek, in: M. Hämäläinen / E. Öhman / S. Miyagawa / K. Alnajjar / Y. Bizzoni (eds.), Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, online 2024, 247–258, https://aclanthology.org/2024.nlp4dh-1.24 (last access 11.07.2025).

Wang et al. (2023): Wang S. / Sun X. / Li X. / Ouyang R. / Wu F. / Zhang T. / Li J. / Wang G., GPT-NER: Named Entity Recognition via Large Language Models, online 2023, https://doi.org/10.48550/arXiv.2304.10428 (last access 11.07.2025).

Xi Z. et al. (2023): Xi Z. / Chen W. / Guo X. / He W. / Ding Y. / Hong B. / Zhang M. / Wang J. / Jin S. / Zhou E. / Zheng R. / Fan X. / Wang X. / Xiong L. / Zhou Y. / Wang W. / Jiang C. / Zou Y. / Liu X. / Yin Z. / Dou S. / Weng R. / Cheng W. / Zhang Q. / Qin W. / Zheng Y. / Qiu X. / Huang X. / Gui, T., The Rise and Potential of Large Language Model Based Agents: A Survey, online 2023, https://doi.org/10.48550/arXiv.2309.07864 (last access 11.07.2025).

Yousef et al. (2023): T. Yousef / C. Palladino / S. Janicke, Transformer-Based Named Entity Recognition for Ancient Greek, Digital Humanities 2023: Book of Abstracts, online 2023, 420–422, https://zenodo.org/records/8107629 (last access 11.07.2025).

Figure References

Fig. 1. Left: Example (Plut. Caes. 21) of NER visualisation with *Daidalos*; Right: Overall evaluation of NER results for the test corpus.

Author Contact Information[34]

Dr. Andrea Beyer
Humboldt-Universität zu Berlin
Sprach- und Literaturwissenschaftliche Fakultät
Unter den Linden 6
10099 Berlin
E-mail: beyeranz@hu-berlin.de

34

DIGITAL CLASSICS ONLINE

Opera Graeca Adnotata: Building a 40M+ Token Multilayer Corpus for Ancient Greek

Giuseppe G. A. Celano

Abstract: In this article, the beta version 0.2.0 of *Opera Graeca Adnotata* (*OGA*), the largest open access multilayer corpus for Ancient Greek (AG), is presented. *OGA* consists of 1,999 literary works and 40M+ tokens sourced from the *canonical-greekLit*, *First1KGreek*, and *PatristicTextArchive* GitHub repositories, which together host AG texts ranging from approximately 900 BCE to 1400 CE. The texts have been enriched with nine annotation layers: (i) tokenization; (ii) sentence segmentation; (iii) lemmatization; (iv) morphology; (v) dependency structure; (vi) dependency function; (vii) IPA transcription; (viii) composition date; and (ix) CTS structure. The layers are described by highlighting the main technical and annotation-related issues encountered. The corpus is released in the standoff formats PAULA XML and its derivative LAULA XML and is queryable online through ANNIS.

1. Introduction[1]

Multilayer corpora contain a variety of annotations modelled as independent layers. Unlike corpora with inline annotations, multilayer corpora have the unique advantage of scalability, as a potentially infinite number of annotation layers can be added using a standoff approach, where layers are interconnected through references to base texts in a graph structure.[2] An example of an open access multilayer corpus for a modern language is the National Corpus of Polish,[3] which is encoded in a standoff format according to the P5 TEI formalism: Polish texts mostly sourced from newspapers and magazines are tokenized, sentence-segmented, and annotated for morphosyntax, named entities, and word sense disambiguation.

A number of historical language corpora have also been annotated with different layers of linguistic information, such as Coptic Scriptorium[4] and RIDGES Herbology,[5] which are both provided in standoff PAULA XML. RIDGES Herbology, for example, contains German herbal texts ranging from 1478 to 1870, annotated with three different transcription layers, namely a diplomatic one, which is the closest to the original text, and two normalization layers called 'clean' and 'norm', respectively: the former aims to address the issue of a few character variations of the diplomatic transcription, while the latter

1 Acknowledgements: This work has been supported by the German Research Foundation (DFG project number 408121292).

2 While Zeldes (2018) proposes a definition of 'multilayer' with reference to independent annotation types, I adopt a definition where independence simply refers to formally separate standoff annotation layers, regardless of how content-wise independent layers are.

3 Przepiórkowski et al. (2011).

4 Schroeder / Zeldes (2016).

5 Odebrecht et al. (2017).

offers a higher-level normalization according to the principles of modern German orthography. Besides morphosyntactic annotation, it is noteworthy that the standoff nature of the RIDGES corpus allows the addition of, for example, lexical layers, such as the one specifying a token's language (e.g., whether it is German or Latin) and the one containing a person's full name, both of which would be difficult to add to syntactic annotation inline.

In the present article, I describe the beta version 0.2.0 of *Opera Graeca Adnotata* (*OGA*),[6] a multilayer corpus for Ancient Greek (AG),[7] focusing on its design as well as the technical and annotation-related issues encountered while working with a large dataset consisting of 1,999 texts and 40,105,221 tokens, which is currently by far the largest open access annotated corpus for AG.[8]

The paper is organized as follows: in Section 2, related work is presented, while Section 3 introduces the standoff formalism of PAULA XML and its derivative LAULA XML. Section 4 and its subsections describe the original texts (Section 4.1) and the annotation layers: tokenization (Section 4.2); sentence segmentation (Section 4.3); morphosyntax and lemmatization (Section 4.4); IPA transcription (Section 4.5); composition date (Section 4.6); and Canonical Text Services (CTS) structure (Section 4.7). Section 5 concludes the article with a brief summary and final remarks.

2. Related Work

There are a few noteworthy corpora of literary AG works. *Thesaurus Linguae Graecae*[9] is the largest non-open access corpus for AG (110M+ words); its texts are accessible only via a query interface with limited functionality, primarily supporting word form and lemma search. The open access counterpart of *TLG* is represented by *Perseus Digital Library* (*PDL*)[10] and its derivative *Scaife Viewer* (*SV*),[11] whose collection of literary AG texts largely coincides with that of *OGA*: both websites aim to offer a reading environment for literary texts (with limited search capability). *PhiloLogic4*[12] and *Diogenes*[13]—which are based (also) on the texts of *PDL*/*SV*—are also open access resources designed as reading environments, with integration of morphological and lexical information. *eAQUA*[14] offers a number of tools to extract information from classical texts. A pioneering corpus for fragmentary authors is *Digital Fragmenta Historicorum Graecorum*:[15] it provides the texts of 636 fragmentary Greek historians, along with their translations and commentaries, all of which can be searched for word forms.

Among the ever-growing number of open access resources, treebanks, such as the *Ancient Greek Dependency Treebank* (*AGDT*),[16] are particularly worth mentioning, as they contain a variety of texts

6 A previous release, *OGA v0.1.0*, is available on Zenodo at https://doi.org/10.5281/zenodo.8158675 (last access 11.07.2025) and is described in the preprint Celano (2024a), on which the present paper is based.

7 Celano (2024b).

8 The data is available on Zenodo at https://doi.org/10.5281/zenodo.14206061 (last access 11.07.2025) and can be queried online at https://annis.varro.informatik.uni-leipzig.de (last access 11.07.2025). Note that this latter website hosts the latest version of the corpus, which is meant to change over time as new releases become available.

9 https://stephanus.tlg.uci.edu (last access 11.07.2025).

10 https://www.perseus.tufts.edu/hopper (last access 11.07.2025). The website is a legacy one, the work continuing on the *SV*.

11 https://scaife.perseus.org (last access 11.07.2025).

12 https://perseus.uchicago.edu (last access 11.07.2025).

13 https://d.iogen.es (last access 11.07.2025).

14 http://www.eaqua.net (last access 20.01.2026).

15 https://www.dfhg-project.org (last access 11.07.2025); Berti (2021).

16 See Celano (2019) for an overview of existing treebanks.

manually annotated for morphosyntax, which many other resources, including *OGA*, rely on. The *Diorisis corpus*[17] offers lemmatization and morphological analyses of 820 texts and 10M+ tokens.[18] More recently, the *GLAUx* project aims to provide a larger morphosyntactically and semantically annotated corpus, i.e., the *GLAUx* corpus:[19] the latest version[20] consists of 20M+ morphosyntactically annotated tokens.[21] The above-mentioned annotated corpora share the characteristic of being encoded in a project-specific format, which is meant to provide consumable, but hardly extensible, data.

3. The Formats: Paula XML and LAULA XML

PAULA XML (Potsdamer AUstauschformat Linguistischer Annotationen) is an established open access standoff format for linguistic annotation,[22] which was inspired by LAF (Linguistic Annotation Framework).[23]

In PAULA XML, a base text is directly or indirectly referenced by identifiers contained in annotation layers, each of which is stored in a separate file.[24] Altogether, the files form an acyclic graph. A base text embeds the transcription of an original text within a shallow XML structure, so that it can typically be referenced by at least one annotation layer, i.e., the tokenization layer, which identifies tokens by referencing character offsets. Each thus identified token is associated with an ID, which can in turn be referenced in other annotation layers.

17 https://doi.org/10.6084/m9.figshare.6187256.v1 (last access 11.07.2025). *Diorisis* does not seem to be a currently maintained resource.

18 Vatri / McGillivray (2018).

19 Keersmaekers (2021).

20 https://glaux.be/; https://github.com/alekkeersmaekers/glaux (last access 11.07.2025).

21 More recently, an attempt to annotate Greek Papyri is detailed in Keersmaekers / Van Hal (2024).

22 https://github.com/korpling/paula-xml (last access 11.07.2025).

23 https://www.iso.org/standard/37326.html (last access 11.07.2025).

24 Standoff annotation is typically performed by using separate files for each annotation layer, and indeed, this is the PAULA XML model. However, it is to be noted that standoff annotation is primarily defined by a referencing mechanism keeping markup data and text to markup separate, without this necessarily implying separation in different files. However, for scalability purposes, separate files are typically used.

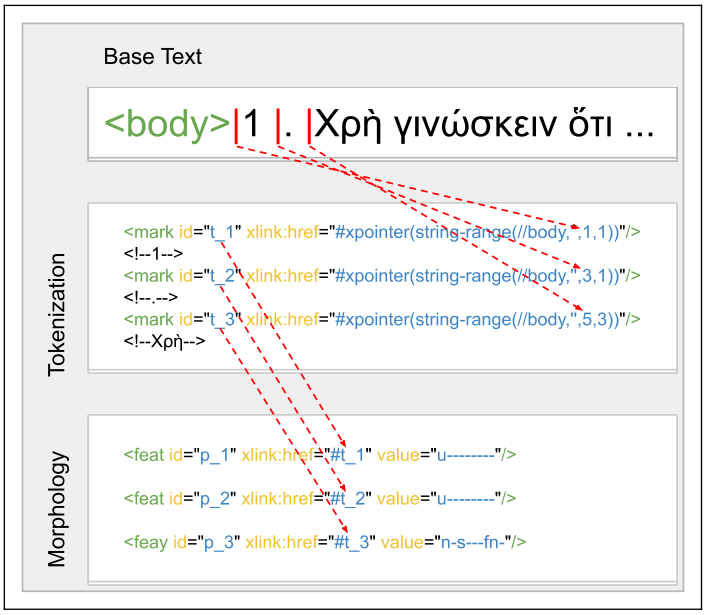


Fig. 1: Standoff annotation layers in PAULA XML.

As shown in fig. 1, the XPointer expressions within the tokenization layer reference the base text by specifying the start offset of a token – numbering starts with 1 and not 0 – and its length. The XPointer expressions are associated with IDs, which can then be used, for example, in the morphological layer to associate each of them with the corresponding morphological annotation contained in the `value` attribute.

Such a model has the advantage of offering an elegant solution to the issue of overlapping markup. For example, a layer for prosody annotation could reference tokens that differ from morphosyntactic tokens. Ensuring that different tokenization layers reference the same base text guarantees that their schemes can be compared.

Currently, *OGA* contains the following annotation layers: (i) tokenization; (ii) sentence segmentation;[25] (iii) lemmatization; (iv) morphology; (v) dependency structure; (vi) dependency function; (vii) IPA transcription; (viii) composition date; and (ix) CTS structure.

Notably, PAULA XML requires a base text to be inserted in the `<body>` element of an XML file, without any XML markup within it. PAULA XML is part of a set of technologies developed for annotation of multilayer corpora, which includes ANNIS,[26] a query engine, and its related technologies: Salt, a meta model for manipulating linguistic data, Pepper,[27] a converter between different annotation formats, and Hexatomic, an annotation editor.

One disadvantage of PAULA XML is that the size of a corpus tends to grow significantly as new texts and/or annotation layers are added. For example, the directory containing the PAULA XML files of *OGA 0.2.0* is as large as about 23GB (unzipped). For this reason, *OGA* files are processed using a lighter and more efficient XML structure, called LAULA XML (*Leipziger AUstauschformat Linguistischer Annotationen*), which retains the logic of PAULA XML, but its repeating element and attribute names are shortened to one character (e.g., `<mark>` becomes `<m>` and `<feature>` `<f>`); moreover, information that in PAULA XML can only be added inside XML comments is conveniently encoded, in LAULA XML, within XML attributes, whose contents can typically be processed by XPath parsers much more efficiently.

25 As explained below, this layer is available in LAULA XML, but not in PAULA XML.

26 Krause / Zeldes (2014); https://corpus-tools.org/annis (last access 11.07.2025).

27 https://corpus-tools.org/pepper (last access 11.07.2025).

More importantly, LAULA XML allows original TEI XML files to be directly referenced, without the need of modifying texts, a particularly convenient property for corpora such as *OGA*, whose main texts consist solely of TEI XML files, sometimes with heavy paratext markup. This means that, for example, if there is an interest in connecting tokens to the critical apparatus encoded in original files, LAULA XML—but not PAULA XML—supports this. A sentence segmentation layer is also provided in LAULA XML, but not in PAULA XML, in that the latter is used for conversion into relANNIS, the file format for ANNIS, which only allows token- and text-based queries.

4. The *OGA* Pipeline

Due to the high number of texts and annotation layers to process, *OGA* is created automatically with minimal human inspection: it is the output of several scripts, whose content is summarized in the following subsections. Because of its large size, the corpus, a few related resources, and its documentation are made available on Zenodo;[28] for the latest updates, the reader is referred to the associated GitHub repository.[29]

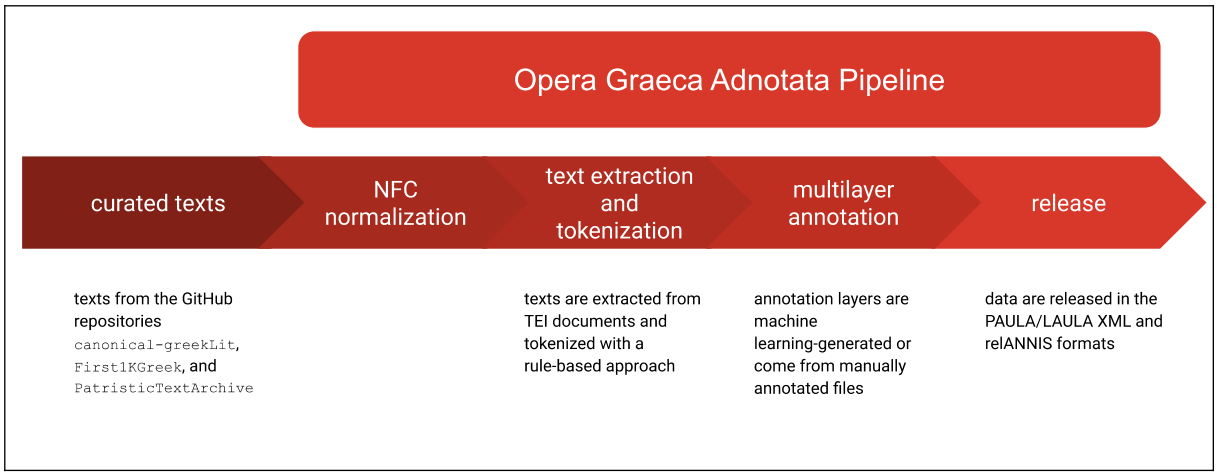


Fig. 2: Pipeline for the creation of *OGA*.

4.1. The Texts

The main texts of *OGA 0.2.0* are sourced from three independently managed GitHub repositories: (i) *canonical-greekLit*;[30] (ii) *First1KGreek*;[31] and (iii) *PatristicTextArchive*.[32] While the first repository contains Classical Greek literary texts, which mostly coincide with the ones available in PDL 4.0,[33] the second one aims to complement *canonical-greekLit* by adding one edition of any Greek literary work composed until about 250 CE. *PatristicTextArchive* is a more recent effort, which aims to create digital critical editions of patristic texts.

Since at least 2017, the texts in *canonical-greekLit* and *First1KGreek* have been edited actively for correction of OCR errors and, more in general, compliance with more recent standards. In particular, there has been an ongoing effort to transition older XML files into EpiDoc P5 TEI XML files and

28 https://doi.org/10.5281/zenodo.14206061 (last access 11.07.2025).

29 https://github.com/OperaGraecaAdnotata/OGA (last access 11.07.2025).

30 https://github.com/PerseusDL/canonical-greekLit (last access 11.07.2025).

31 https://github.com/OpenGreekAndLatin/First1KGreek (last access 11.07.2025).

32 https://github.com/PatristicTextArchive (last access 11.07.2025).

33 https://www.perseus.tufts.edu/hopper (last access 11.07.2025).

provide each text with CTS structure. New releases of both repositories are issued on a frequent basis. *PatristicTextArchive* encodes texts in TEI XML files with the specification of the same CTS structure found in *canonical-greekLit* and *First1KGreek*. Of the texts of these repositories, *OGA 0.2.0* contains 1,999, with a total of 40M+ tokens. *OGA* is conceived in such a way that, when new releases of the above-mentioned repositories become available, the corpus, along with its annotation layers, can be re-built from scratch, so as to incorporate newly added or modified texts.

It is to be noted that, to the best of my knowledge, an evaluation of the accuracy of the digitized texts (especially those in *canonical-greekLit* and *First1KGreek*) is still lacking.[34] While the digitized texts, in general, seem to represent the main texts of the original print editions accurately,[35] there still are various errors and inconsistencies that are going to affect tokenization and sentence segmentation (see section 4.2).

Although correction of the original texts is outside the scope of *OGA*, some automatic encoding normalization can be performed. More precisely, NFC normalization is applied to address the issue of characters such as 'epsilon with an acute accent', which can be encoded as the Unicode codepoint U+03AD or U+1F73, the former being in the "Greek and Coptic" chart, while the latter in the 'Greek Extended' one: only the codepoint U+03AD is, however, the NFC-normalized codepoint, and therefore NFC normalization ensures uniform encoding of this character.

4.2. Text Extraction and Tokenization

Since the original texts are encoded as TEI XML texts, preprocessing is needed to separate the text of a work, which annotation layers reference, from its paratext, which is not annotated.[36] This is a non-trivial task with heavily marked-up literary texts, because the distinction between (main) text and paratext is signalled in TEI documents via use of many different XML elements that serve different functions.

For example, the element `<note>`, as the name itself suggests, contains a note, and therefore its content can be safely identified as paratext. Similarly, the contents of elements such as `<app>` and `<bibl>` can be discarded, in that they unambiguously identify paratext related to critical apparatus and bibliography, respectively.

The content of other TEI elements is, however, part of a text: for example, `<foreign>` contains a foreign language term, while `<add>` signals a text addition by an editor, which should arguably be considered as part of the main text.

In a few cases, the semantics, and consequently the structure, of a TEI element are complex. For example, `<choice>` presents a number of alternative readings for a specific passage: it can contain `<sic>` to highlight a certain word form whose correction is given in its sibling element `<corr>`; or it can contain the sibling elements `<abbr>` and `<expan>` for an abbreviation and its expansion, respectively.

After a main text has been extracted, a rule-based tokenization is applied to it, as there is almost a one-to-one correspondence between graphic words and morphosyntactic tokens in AG. The most notable exception to this occurs in the case of a few conjunctions,[37] such as οὐδέ, and crasis, i.e., the phonological phenomenon whereby two words can be univerbated: for example, the word κἀκεῖνος consists of

34 An attempt to evaluate OCR for Ancient Greek is documented in Boschetti et al. (2009).

35 However, accuracy seems to greatly vary from text to text.

36 Currently, a way to keep a reference to paratext in PAULA XML would be to convert paratext TEI elements into annotation layers, as in the GUM corpus (see Zeldes [2017]). However, this solution turns out to be cumbersome, especially for heavily marked-up files. This issue does not arise in LAULA XML, however, in that the tokenization layer references an original file.

the words καὶ (‘and’) and ἐκεῖνος (‘that’): since they belong to different POS, they need to be separated in a morphosyntactic tokenization scheme.

Luckily, most cases of crasis can be unambiguously identified because of the punctuation mark ‘coronis’ (i.e., the Unicode codepoint COMBINING COMMA ABOVE, U+0313) placed on a non-initial vowel. On the basis of this formal criterion, the texts were searched for words with a coronis and a list of them, which is also made available in the *OGA* release, was compiled and used for tokenization. Cases where a coronis is more difficult to identify, as when a smooth breathing is on a word-initial vowel, were left untreated.

On a preliminary test based on 38,710 tokens from 2,234 sentences randomly selected from all texts in *OGA*, 281 erroneous tokens were found (i.e., error rate of about 0.0073): however, all tokenization errors except one[38] were due to OCR/encoding errors in the original texts: for example, paratext content is sometimes encoded as text or tokens contain wrong characters or missing accents.

Notably, XPointer expressions in a PAULA XML tokenization layer reference a base text that only consists of a long string contained in a `<body>` element (see fig. 1), in that no XML markup is allowed in it; on the other hand, since base texts in LAULA XML coincide with the original TEI XML texts, tokenization layers can reference tokens conveniently through XPath expressions following CTS structures.

4.3. Sentence Segmentation

Similarly to tokenization, sentence segmentation is achieved in *OGA* through a rule-based approach. In fact, sentence boundaries are regularly signalled in modern editions of AG texts by the period, semicolon, and middle dot punctuation marks.

The algorithm also addresses the issue of use of parentheses – mostly of editorial meaning – in conjunction with other sentence-final punctuation marks, in that their order is not standardized, but usually follows a modern language’s style rules. It has been found that 9 sentences out of 2,234 (i.e., error rate of about 0.0040) are wrongly segmented because of OCR errors in the original texts.

The sentence segmentation layer is made available only in LAULA XML; as appears in Section 3, PAULA XML is used as a serialization format for conversion into relANNIS, i.e., the file format for ANNIS, which displays and queries annotations without the help of sentence boundaries.

4.4. Morphosyntactic Annotation and Lemmatization

The morphosyntactic annotation in *OGA 0.2.0* is performed automatically using Trankit.[39] Trankit is a transformer-based parser, which, relying on the pretrained model XLM-RoBERTa, proved to deliver state-of-the art results for UD treebanks.[40] Moreover, it is very time-efficient because only the weights of a few layers are trained, while all the others remain fixed.

Trankit delivered the best results for AG (see tab. 1) when compared to three other parsers,[41] i.e., a baseline LSTM-based parser with randomly initialized character embeddings called Dithrax, and two parsers updating the token embeddings provided by GreBERTa and PhilBERTa,[42] respectively.

37 The list of them can be found at https://github.com/OperaGraecaAdnotata/OGA (last access 11.07.2025), subdirectory *tokenize*.

38 A RIGHT SINGLE QUOTATION MARK used to mean elision is separated from the preceding token.

39 Nguyen et al. (2021).

40 https://trankit.readthedocs.io/en/latest/performance.html (last access 11.07.2025).

41 See Celano (2025) for the detailed analysis of the comparison of the parsers as well as lemmatizers.

42 Riemenschneider / Frank (2023).

Lemmatization was performed using GreTa, an encoder-decoder transformer, which resulted in the best performing model (see Tab. 1), when compared to Dithrax and PhilTa.[43]

POS	XPOS	Feats	AllTags	UAS	LAS	Lemmata
96.41	91.90	94.77	91.56	82.60	77.10	91.41

Tab. 1: F1 Scores for the Trankit parser and the GreTa lemmatizer.

Trankit and GreTa were trained on (i) the AGDT v2.1,[44] (ii) the Gorman Trees,[45] and (iii) the Pedalion Trees.[46] All these treebanks adopt the same annotation scheme (i.e., the AGDT one) and altogether constitute by far the largest morphosyntactically annotated corpus for AG (1.2M+ tokens). Since the corpus comprised texts of different genre and age (from approximately 900 BCE to 400 CE), the Trankit and GreTa models[47] are expected to be able to generalize much better than other existing models, such as, for example, the ones trained on UD treebanks, whose sizes are much smaller – the token count for the UD Perseus Treebank plus the UD PROIEL treebank is about 416K tokens.

Although all the above-mentioned treebanks follow the same annotation scheme, their texts, which were annotated by many different annotators, needed to be made consistent internally, among themselves, and compliant with the *OGA* tokenization scheme. All treebank fields, such as word form, lemma, POS, etc., required a non-trivial normalization because of some clearly erroneous or null values. All apostrophe-looking characters were converted into MODIFIER LETTER APOSTROPHE (U+02BC). Some tokens, such as the coordinate conjunctions οὐδὲ and εἴτε, were tokenized. Sentences with syntactic cycles were corrected or deleted.[48]

The AGDT annotation scheme derives from that of the Prague Dependency Treebank[49] (Hajič et al., 2018) and consists of four annotation layers: (i) morphological layer; (ii) lemma layer; (iii) dependency structure layer; and (iv) dependency function layer. The morphological annotation is represented as a 9-character string, where the first character is the part of speech of a token and the remaining characters its morphological features. The lemma annotation refers to the dictionary entry for a token: notably, the AGDT annotation scheme follows the convention of representing lemmata as single word forms, even if, in traditional grammar, lemmata consist of more word forms, which describe a token more accurately: for example, a lemma for a noun in an AG dictionary consists not only of its nominative form, as in the AGDT, but also of its genitive, which conveys relevant information about its declension.

The syntactic annotation follows a dependency formalism, which consists in the identification of directed relations between a head token and its dependent tokens – a head token can have more than one dependent, but not vice versa. When considered together, all relations form a directed acyclic graph, more commonly described as an (upside-down) syntactic tree. Each syntactic dependency is also typed with a syntactic label expressing the function a dependent has with reference to its head (e.g., subject or attribute). Dependency grammar formalism is quite popular in computational linguistics because it represents a balanced trade-off between syntactic analysis precision and annotation feasibility. De-

43 Riemenschneider / Frank (2023).

44 https://github.com/PerseusDL/treebank_data (last access 11.07.2025).

45 https://github.com/vgorman1/Greek-Dependency-Trees (last access 11.07.2025); Gorman (2020).

46 https://github.com/perseids-publications/pedalion-trees (last access 11.07.2025).

47 https://git.informatik.uni-leipzig.de/celano, subdirectories `morphosyntactic_parser_for_OGA` and `lemmatizer_for_OGA` (last access 11.07.2025).

48 See Celano (2025) for more details.

49 Hajič et al. (2018).

scribing the many details of the syntactic annotation is outside the scope of the article: the reader is referred for them to Celano (2019) and the annotation guidelines.[50]

4.5. The IPA Transcription Layer

OGA contains an experimental IPA transcription layer (for tokens). IPA transcription allows querying the prosodic structure of AG, which could be enormously beneficial to studies on AG poetry or prose rhythm.

The annotation is the output of a ByT5 model trained on *Wiktionary* lemmata,[51] achieving an accuracy of 0.83 in producing correct IPA transcriptions. The training dataset contained both AG and Latin IPA transcriptions. *Wiktionary* provides IPA transcriptions corresponding to different historical periods: *OGA* currently provides the '5th century BCE Attic' pronunciation for AG and the 'Classical Latin' one for Latin. For example, the word ἁβρός corresponds to the IPA transcription /ha.brós/, while the word ἀασθείς to /a.aːs.tʰĕːs/.

It is to be noted that, while AG orthography easily allows conversion to IPA transcriptions of many graphemes (because of one-to-one correspondences), this does not hold true for a few more complex cases: for example, the length of the vowel α is not marked, and more complex rules are needed to treat the conversion of diphthongs such as ει. Moreover, identification of syllable division, which is marked by full stops in the IPA transcriptions above, is not a trivial task. For all these reasons, the task was approached using machine learning rather than a rule-based method.

4.6. The Composition Date Layer

Being able to identify when a text was composed is of crucial importance for query purposes. For this reason, texts in *OGA* were annotated for estimated composition dates. As is well known, the chronology of ancient works cannot always be precisely determined. Sometimes, a composition date is highly disputed or unknown, and this represents a modelling issue. To address this, composition dates were annotated by one student expert in AG literature following academic reference works or Wikipedia: since different modern authors can suggest different dates, the chosen dates have been documented with their sources,[52] so that querying of the corpus and future changes or corrections can be facilitated.[53]

4.7. The CTS Structure Layer

CTS structure generally refers to a hierarchical citation system used by the CTS protocol to retrieve passages from a literary work by means of URNs,[54] which include identifiers for an author, work, edition, and passage.

In reference to the annotation layer, however, the phrase 'CTS structure' is used with a narrower scope, indicating passage tags assigned to tokens. For example, since Herodotus' *Histories* are divided into books, chapters, and sections, a CTS structure tag provides each token of this work with the number of the book, chapter, and section it belongs to.

---

50 https://github.com/PerseusDL/treebank_data, subdirectories `v1/greek/docs` and `AGDT2/guidelines` (last access 11.07.2025).

51 https://git.informatik.uni-leipzig.de/celano, subdirectory `ipa_transcription_for_OGA` (last access 11.07.2025).

52 https://github.com/OperaGraecaAdnotata/OGA, subdirectory `work_chronology` (last access 11.07.2025).

53 The addition of composition dates represented a complex modelling problem also because ANNIS does not currently support data types such as numbers or dates; see more details at https://github.com/OperaGraecaAdnotata/OGA, subdirectory `query` (last access 11.07.2025).

54 Blackwell / Smith (2020).

Indeed, *OGA* base texts are based only on those TEI XML texts containing a `<refsDecl n="CTS">` element (within the `<encodingDesc>` element), in which an XPath expression is provided for the identification of work divisions. For example, in the file `tlg1600.tlg001.perseus-grc2.xml` (corresponding to *Flavii Philostrati Opera*, Vol. 2), the following XPath expression is given:

```
/tei:TEI/tei:text/tei:body/tei:div/tei:div[@n="$1"]/tei:div[@n="$2"]
```

The variables `$1` and `$2` stand for the numbers of, respectively, the first and second kinds of division (the `<div>` elements), which, in this case, correspond to "book" and "chapter"—the names of these divisions are specified within a `<refsDecl>` element.

The importance of the CTS citation layer for philological, historical, and linguistic studies cannot be overstated, considering how heavily they rely on the network of references made possible through such a passage numbering system.

5. Conclusion

In this paper, the architecture of the beta version 0.2.0 of the *OGA* corpus was presented. The base texts and their nine annotation layers have been described: (i) tokenization; (ii) sentence segmentation; (iii) lemmatization; (iv) morphology; (v) dependency structure; (vi) dependency function; (vii) IPA transcription; (viii) composition date; and (ix) CTS structure. The layers are serialized as standoff PAULA XML and standoff LAULA XML and can be queried online through ANNIS.[55]

OGA is generated automatically through a series of scripts, which can be re-executed to reproduce it or update it, if its base texts, which derive from independently managed GitHub repositories, change or new texts are added.

Base texts in *OGA* are extracted from original TEI XML files, in which text is encoded together with paratext, and tokenized into morphosyntactic tokens with a rule-based system. The corpus presents a morphosyntactic annotation and lemmatization based on models trained on treebank data: a significant challenge was ensuring that the tokenization schemes of treebank texts and *OGA* base texts matched as closely as possible. The corpus is enriched with an experimental IPA transcription layer based on a ByT5 model trained on *Wiktionary* data and with a composition date layer that associates each work with an estimated composition date based on manual annotation. Finally, since all original TEI XML files specify the internal structure of a work, *OGA* comprises an annotation layer that allows retrieval of passages following a canonical citation scheme (CTS structure layer).

Building a multilayer corpus is challenging because a number of issues arise that do not have definitive answers. For example, identification of what counts as text and what counts as paratext is not always clear-cut (for example, should the title of a work be considered as text?)[56], and annotation schemes, including the tokenization one, are questionable on many points. This holds particularly true when annotation is applied to texts belonging to very different ages and genres.

It is also difficult to guarantee consistency at any level: the quality of the original texts, some of which still contain many errors, significantly impacts the quality of annotation layers. Similarly, trying to train models on data that are as similar to the *OGA* ones as possible requires intense normalization work and extensive manual annotation effort. Evaluating annotation quality on such a large corpus as *OGA* is also arduous, and more work is needed in the future.

55 https://annis.varro.informatik.uni-leipzig.de (last access 11.07.2025).

56 The choice has an impact on how the final text is represented and can be annotated.

Finally, it is important to note that annotation modeling choices are also dependent on the technologies and standards available: *OGA* adopts PAULA XML, a de facto standard standoff format, which provides a number of advantages, including the possibility of making the data easily queryable through ANNIS. As noted above, however, a few issues with PAULA XML arise in terms of efficient parsing, which are bound to be exacerbated as the size of the corpus increases.

For all of these reasons, a multilayer corpus should be considered work in progress, continuously evolving with the addition of new texts, more accurate annotations, and the adoption of new technologies and standards.

List of Abbreviations

AG	Ancient Greek
AGDT	Ancient Greek Dependency Treebank
CTS	Canonical Text Services
OGA	Opera Graeca Adnotata
PDL	Perseus Digital Library
SV	Scaife Viewer
TLG	Thesaurus Linguae Graecae

Sources

Online Sources

https://corpus-tools.org/annis (last access 11.07.2025).

https://corpus-tools.org/pepper (last access 11.07.2025).

https://github.com/alekkeersmaekers/glaux (last access 11.07.2025).

https://github.com/korpling/paula-xml (last access 11.07.2025).

https://github.com/OpenGreekAndLatin/First1KGreek (last access 11.07.2025).

https://github.com/OperaGraecaAdnotata/OGA (last access 11.07.2025).

https://github.com/PatristicTextArchive (last access 11.07.2025).

https://github.com/vgorman1/Greek-Dependency-Trees (last access 11.07.2025).

https://github.com/perseids-publications/pedalion-trees (last access 11.07.2025).

https://github.com/PerseusDL/canonical-greekLit (last access 11.07.2025).

https://github.com/PerseusDL/treebank_data (last access 11.07.2025).

https://git.informatik.uni-leipzig.de/celano (last access 11.07.2025).

https://trankit.readthedocs.io/en/latest/performance.html (last access 11.07.2025).

https://www.dfhg-project.org (last access 11.07.2025).

https://www.iso.org/standard/37326.html (last access 11.07.2025).

Digital Corpora

ANNIS = https://annis.varro.informatik.uni-leipzig.de (last access 11.07.2025).

Diogenes = https://d.iogen.es (last access 11.07.2025).

Diorisis = https://doi.org/10.6084/m9.figshare.6187256.v1 (last access 11.07.2025).

Opera Graeca Adnotata v0.1.0 = https://doi.org/10.5281/zenodo.8158675 (last access 11.07.2025).

Opera Graeca Adnotata v0.2.0 = https://doi.org/10.5281/zenodo.14206061 (last access 11.07.2025).

GLAUx = https://glaux.be/ (last access 11.07.2025).

Perseus Digital Library = https://www.perseus.tufts.edu/hopper (last access 11.07.2025).

Philologic4 = https://perseus.uchicago.edu (last access 11.07.2025).

Scaife Viewer = https://scaife.perseus.org (last access 11.07.2025).

Thesaurus Linguae Graecae = https://stephanus.tlg.uci.edu (last access 11.07.2025).

References

Blackwell / Smith (2020). C. W. Blackwell / N. Smith, The CITE Architecture (CTS/CITE) for Analysis and Alignment. it-information Technology 62/2 (2020), 91–98, https://doi.org/10.1515/itit-2019-0044 (last access 11.07.2025).

Berti (2021): M. Berti, Digital Editions of Historical Fragmentary Texts, Heidelberg 2021.

Boschetti et al. (2009): F. Boschetti / M. Romanello / A. Babeu / D. Bamman / G. Crane, Improving OCR Accuracy for Classical Critical Editions, in: Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (2009), 156–167, https://dl.acm.org/doi/10.5555/1812799.1812822 (last access 11.07.2025).

Celano (2019): G. G. A. Celano, The Dependency Treebanks for Ancient Greek and Latin, in: Monica Berti (ed.), Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution, Berlin / Boston (2019), 279–298, https://doi.org/10.1515/9783110599572-016 (last access 11.07.2025).

Celano (2024a): G. G. A. Celano, Opera Graeca Adnotata: Building a 34M+ Token Multilayer Corpus for Ancient Greek, ArXiv (2024), https://arxiv.org/abs/2404.00739 (last access 11.07.2025).

Celano (2024b): G. G. A. Celano, Opera Graeca Adnotata 0.2.0, Zenodo (2024), https://doi.org/10.5281/zenodo.14206061 (last access 11.07.2025).

Celano (2025): G. G. A. Celano, A State-of-the-Art Morphosyntactic Parser and Lemmatizer for Ancient Greek, in: Proceedings of the First Workshop on Natural Language Processing and Language Models for Digital Humanities (2025), 48–65, https://aclanthology.org/2025.lm4dh-1.5/ (last access 30.03.2026).

Gorman (2020): V. B. Gorman, Dependency Treebanks of Ancient Greek Prose, Journal of Open Humanities Data 6/1 (2020), https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.13 (last access 11.07.2025).

Hajič et al. (2018): J. Hajič / E. Bejček / A. Bémová / E. Buráňová / E. Hajičová / J. Havelka / P. Homola / J. Kárník / V. Kettnerová / N. Klyueva / V. Kolářová / L. Kučová / M. Lopatková / M. Mikulová / J. Mírovský / A. Nedoluzhko / P. Pajas / J. Panevová / L. Poláková / M. Rysová / P. Sgall / J. Spoustová / P. Straňák / P. Synková / M. Ševčíková / J. Štěpánek / Z. Urešová / B. Vidová Hladká / D. Zeman / Š. Zikánová / Z. Žabokrtský, Prague Dependency Treebank 3/5 (2018), http://hdl.handle.net/11234/1-2621 (last access 11.07.2025).

Keersmaekers (2021): A. Keersmaekers, The GLAUx Corpus: Methodological Issues in Designing a Long-Term, Diverse, Multi-Layered Corpus of Ancient Greek, in: Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change (2021), 39–50, https://aclanthology.org/2021.lchange-1.6/ (last access 11.07.2025).

Keersmaekers / Van Hal (2024): A. Keersmaekers / T. Van Hal, Creating a Large-Scale Diachronic Corpus Resource: Automated Parsing in the Greek Papyri (and Beyond), Natural Language Engineering 30/5 (2024), 1035–1064, https://doi.org/10.1017/S1351324923000384 (last access 11.07.2025).

Krause / Zeldes (2014): T. Krause / A. Zeldes, ANNIS3: A New Architecture for Generic Corpus Query and Visualization, Digital Scholarship in the Humanities 31/1 (2014), 118–139, https://doi.org/10.1093/llc/fqu057 (last access 11.07.2025).

Nguyen et al. (2021): M. V. Nguyen / V. D. Lai / P. B. Veyseh / T. H. Nguyen, Trankit: A Light-Weight Transformer-Based Toolkit for Multilingual Natural Language Processing, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (2021), 80–90, https://aclanthology.org/2021.eacl-demos.10/ (last access 11.07.2025).

Odebrecht et al. (2017): C. Odebrecht / M. Belz / A. Zeldes / A. Lüdeling / T. Krause, RIDGES Herbology: Designing a Diachronic Multi-Layer Corpus, Language Resources and Evaluation 51 (2017), 695–725, https://link.springer.com/article/10.1007/s10579-016-9374-3 (last access 11.07.2025).

Przepiórkowski et al. (2011): A. Przepiórkowski / M. Bańko / R. L. Górski / B. Lewandowska-Tomaszczyk / M. Łaziński / P. Pęzik, National Corpus of Polish, in: Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań 2011, 259–263.

Riemenschneider / Frank (2023): F. Riemenschneider / A. Frank, Exploring Large Language Models for Classical Philology, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (2023), 15181–15199, https://aclanthology.org/2023.acl-long.846/ (last access 11.07.2025).

Schroeder / Zeldes (2016): C. T. Schroeder / A. Zeldes, Raiders of the Lost Corpus, Digital Humanities Quarterly 10/2 (2016), https://www.digitalhumanities.org/dhq/vol/10/2/000247/000247.html (last access 11.07.2025).

Vatri / McGillivray (2018): A. Vatri / B. McGillivray, The Diorisis Ancient Greek Corpus: Linguistics and Literature, Research Data Journal for the Humanities and Social Sciences 3/1 (2018), 55–65, https://doi.org/10.1163/24523666-01000013 (last access 11.07.2025).

Zeldes (2017): A. Zeldes, The GUM Corpus: Creating Multilayer Resources in the Classroom, Language Resources and Evaluation 51 (2017), 581–612, http://dx.doi.org/10.1007/s10579-016-9343-x (last access 11.07.2025).

Zeldes (2018): A. Zeldes, Multilayer Corpus Studies, New York 2018.

Figure References

Fig. 1: Standoff annotation layers in PAULA XML.

Fig. 2: Pipeline for the creation of *OGA*.

Author Contact Information[57]

Dr. Giuseppe G. A. Celano
Universität Leipzig
Institut für Informatik
Augustusplatz 10
04109 Leipzig
E-mail: celano@informatik.uni-leipzig.de

57 The rights pertaining to content, text, graphics, and images, unless otherwise noted, are reserved by the author. This contribution is licensed under CC BY-SA 4.0.

DIGITAL CLASSICS ONLINE

Automatic Annotation of *Nomina Sacra*

Carina Geldhauser

Abstract: *Nomina sacra* are a specific kind of named entities appearing in biblical manuscripts. Due to the large amount of biblical manuscripts, many questions about *nomina sacra* could not be answered to the present time. In order to use the methods of Digital Humanities for research questions on *nomina sacra*, they need to be consistently and accurately annotated. We report on our recent efforts on combining Handwritten Text Recognition (HTR) with annotation for biblical manuscripts written in Greek majuscule script. We reflect on the lessons learned from this work, especially on the technical aspects such as the available NER algorithms for classical languages, the performance of machine-learning based tools in comparison to rule-based annotation algorithms. We also discuss the pro's and con's of the approach we chose in our work.

Nomina sacra

A big part of the work in creating digital editions of ancient text is the standardised annotation of relevant features in the text. Which features are 'relevant' depends on the scope of the edition, the nature of the text and its linguistic features. Often, proper names are a relevant category, and this work deals with the annotation of a very specific class of proper names, the so-called *nomina sacra*.

A *nomen sacrum* is an abbreviation of a specific name or word carrying a meaning (e.g. The Spirit) that appears systematically in manuscripts of the Bible, most prominently in Greek manuscripts, some as old as *Papyrus Bodmer* II (known as P66). It is characterised by an overline bar spanning two or more letters from the original word, for example ΘΣ for Θεός.

The origin of *nomina sacra* as a scribal practice is not totally clarified as of today. Ludwig Traube, who coined the usage of *nomina sacra* as a technical term in his 1907 monograph,[1] conjectured the origin of *nomina sacra* in Hellenistic Judaism and their usage of the Tetragramm *JHWH*, but the current hypothesis is that *nomina sacra* are a characteristic Christian phenomenon.[2]

Nomina sacra as a phenomenon in Biblical manuscripts has been described by many scholars, and a number of theories on their origin and usage were developed. Within the manuscripts that were investigated by theologians, abbreviation by contraction, meaning that the first and last letter (at least) of each word are used, prevails. However, especially in earlier manuscripts,[3] we also find the alternate practice of abbreviation by suspension, meaning that the initial two letters of the word are used. We may, at the moment, only speculate on the prevalence of abbreviation by contraction – it certainly was of advantage as it indicated the case of the abbreviated noun.

1 Traube (1907).

2 Hurtado (2017), 127 even uses the variations in spelling (contraction, suspension, or mixed abbreviation) as evidence for this hypothesis.

3 For example, Jesus Christ is abbreviated as *IH XP* in the opening verses of Revelation in P18.

The current hypothesis is that *nomina sacra* were not merely used to save space, but as an act of reverence,[4] evidenced by the observation that *nomina sacra* were sometimes used to distinguish between 'mundane' and 'sacred' usages of the same word (e.g., spirit vs The Spirit), and employed even when other common abbreviations were not used.

The picture is yet not so simple as the observations mentioned above do not hold for all manuscripts, and there are different abbreviated forms used for the same word in different manuscripts. Moreover, there were different ways that scribes used to deal with the Hebrew *tetragrammaton JHWH*, spanning from unabbreviated forms in the Greek text to the peculiar double zeta with a horizontal line through the middle in the *Septuagint* manuscript *Papyrus Oxyrhynchus* 1007.[5]

Methods of Digital Humanities could help to shed further light into the usage of *nomina sacra*, to bolster or to reject the above-mentioned hypotheses, and to help getting new insight. For this, we need a reliable way to mark/annotate *nomina sacra* within a (digitalised or) transcribed text – which motivated our work.

The Annotation Problem for Greek Manuscripts

Naively, one might believe that all questions regarding *nomina sacra*, as a particular case of Named Entity Recognition (NER), might be solved within minutes. Indeed, for modern high-resource languages, named entity recognition and other Natural Language Processing (NLP) tools have been successfully used to annotate texts, allowing authors to explore huge datasets through statistical methods.

However, scholars in Classics and Biblical Theology have a harder life here, as Ancient Greek is, from the Machine Learning point of view, a low-resource language.[6] Not because there would not exist enough literary texts, prosaic and poetic, in Greek language – far from it! However, there exist few and mainly[7] quite small annotated text corpora that can be used as training sets for data-hungry ML models.

Furthermore, there might not even exist reliable digitalisations of the texts that one intends to annotate. Sometimes, but not always, a digitalisation can be generated quickly through readily available software such as *transcribus* and *escriptorium*. In general, we need to be cautious when dealing with manuscripts[8], as often, the automatic post-processing step removes diacritics or hyphens, meaning a significant piece of information is lost.

Another cause of trouble arises when using readily digitised texts from different sources in one dataset: first, different OCR software supply different output formats, and unification might be difficult or prone to create errors in many instances of the dataset. Also, copyright or licensing restrictions are a big obstacle in gathering available datasets for training of NLP tools.

Second, even if technical obstacles are overcome, the ML model might still pick up on specific, technical cues such as encodings and output file characteristics, instead of textual cues. Some even have problems with punctuation signs.[9] Furthermore, the heterogeneity in encoding causes current ML tools

4 See, e.g. Hurtado (2017), 100, 104–106.

5 Wilkinson (2015), 55.

6 See e.g. Kostkan et al. (2023), 128.

7 The recent *Opera Graeca Adnotata* by Celano (2024) could be a gamechanger here for certain applications (not *nomina sacra*).

8 See Geldhauser / Malyshev (2024).

9 Vatri / McGillivray (2020), 189 reported that GLEM was unable to identify a word followed by a punctuation mark as equal to the same form in the middle of a sentence.

to fail. It is known[10] that features characteristic of a philological, palaeographical, or diplomatic edition, such as special characters, textual gaps, abbreviations, headlines, and orthographic redundancy may "confuse" the algorithm or at least present additional degrees of freedom that create the need for a much larger training dataset in order to be classified as noise in the training data. Clearing these features from the dataset, or to diminish it by automatic and manual cleaning, may help to resolve the issue, but may also remove important cues like the overline bar characterising a nomen sacrum. Specific to Greek texts, let us mention the apostrophe issue that greCy developer J. Myerston remarked in his readme file:

> "Unfortunately, there is no consensus among the different internet projects that offer ancient Greek texts about how to represent the Ancient Greek apostrophe. Modern Greek simply uses the regular apostrophe, but ancient texts available in *Perseus* and *Perseus* under Philologic use various unicode characters for the apostrophe. Instead of the apostrophe, we find the Greek *koronis*, modifier letter apostrophe, and right single quotation mark. Provisionally, I have opted to use modifier letter apostrophe in the corpus with which I trained the models. This means, that if you want the greCy models to properly handle the apostrophe you have to make sure that the Ancient Greek texts that you are processing use the modifier letter apostrophe ʼ (U+02BC). Otherwise the models will fail to lemmatize and tag some words in your texts that ends with an 'apostrophe'".[11]

Named Entity Recognition for Classical Languages: Progress and Challenges

Named Entity Recognition is a task that grew out of the more general task of information extraction since the 1990s. Initially, handcrafted rule-based algorithms were used, but later machine-learning techniques became more and more popular. Nowadays Natural Language Processing (NLP) models such as NER are everywhere, and advertised with amazing performance. However, users of these technology observe that as soon as an off-the-shelf algorithm is applied to 'the real world', performance drops.[12] The reason for this problem is the training of NLP algorithms on a very specific, standard dataset, or basically a limited set of canonical varieties of it, used as benchmark corpora. So, essentially all NLP models are trained on English newswire. Real world data differs radically from the benchmarks, and although "any domain can be reasonably supported, porting a system to a new domain or textual genre remains a major challenge".[13] Constructing a robust model requires finding a technical and intellectual compromise between collecting sufficiently varied data and avoiding extreme training on some similar structures.[14]

10 See, e.g., Chastang et al. (2021), paragraph 73.

11 See https://github.com/jmyerston/greCy?tab=readme-ov-file (last access 28.07.2025).

12 Plank (2016a), 1; Nadeau / Sekine (2007).

13 Nadeau / Sekine (2007), 2. They report that some tests reveal up to 20% to 40% of precision drop when applying an algorithm to a different genre.

14 Plank (2016a), 1.

Challenges in NER for Classical Languages

In order to properly understand and appreciate the developments in NER for classical languages, its specific tools and their performances, we need to take a closer look at the challenges that NER faces on historical documents in general. The challenges can be roughly divided into four types:[15]

1. The historical variety space.
2. Noisy input.
3. Language dynamics.
4. Lack of resources.

The *variety space* is a term coined by Barbara Plank, which she defined as an "unknown high-dimensional space, whose dimensions contain fuzzy aspects such as language (or dialect), topic or genre, and social factors (age, gender, personality etc.) amongst others".[16] The usual *terminus technicus* is 'domain', but Plank remarks that this term is ill-defined in the literature and overloaded. Thinking of the huge variety of historical texts, from administrative documents, correspondences, archives of all sorts, literary works, articles, reports, memoirs etc, spanned over thousands of years and countless languages, it is easy to see that no NER tool can be equally capable of handling all of these.

NER is classically trained on standardised datasets containing English-language data.[17] Some research, always on modern texts, has been dedicated to test the ability of these systems to generalise to a different genre, unseen mentions, another domain or document type. All studies reveal a 'NER transfer gap' already for modern texts, and given the needs of humanities research, which is much broader than the typical contemporary NLP applications (that are often motivated by commercial interests, narrowing the scope of the tool), we need to be very careful when using an off-the-shelf tool for our work.

The term *noisy input* is another way of saying OCR errors. Like for all efforts to annotate a digitised text, the accuracy of the annotation highly depends on the accuracy of the OCR or HTR step that created the text basis. While human understanding is quite robust in the sense that it does not bother about tokenisation problems or character misinterpretations (e.g. the characters 'e' and 'c' in printed Latin texts might be mixed up, especially for printed input), the 'OCR noise propagation' may cause a drastic decline in F-score of up to 30%,[18] making annotations as random as tossing a coin.

The term 'language dynamics' summarises the variations in spelling, naming conventions, and in general the differences between modern and historical languages, which affect the performance of NLP tools. A machine-learning based algorithm trained on Ionic Greek forms[19] as found in Herodot. There are some works which study the difference in the structure of entity names, e.g. between historical and contemporary news,[20] but further research is needed: even relatively simple sounding issues such as correctly linking, e.g., 'Madame Pierre Curie' to 'Marie Curie' seem to be left unaddressed for now.

A big challenge for NER of historical texts in general is the heterogeneity of typologies: The mainstream typologies for modern documents have a few 'universal' classes, e.g. *Person*, *Organisation*,

15 We follow here the systematisation of Ehrmann et al. (2023), section 4.

16 Plank (2016a).

17 Standard datasets include CoNLL-2003 (English news texts from Reuters in the year 1996) BERT was originally trained on the *BookCorpus* and English *Wikipedia*: see Devlin et al. (2019).

18 E.g., van Strien et al. (2020), 195 reported a drop from 87% to 63% for person identities.

19 E.g. the lemmatizer *GLEM* was trained on Ionic prose texts by Herodotus, see also the discussion in Vatri / McGillivray (2020).

20 See Rosset (2012).

Location, which could, in principle, be re-used for historical documents, but in general the entity types might require adaptations to fit to the application at hand. An off-the-shelf NER tool is unlikely to capture all entities of interests in a historical document, or certain classes like *Person* need to be divided into *Historical Person*, *Literary Person* or even finer into *Greek/Latin (Half-)God*, *Mythical Creature* etc. – depending on the research question. Also inside a topology, spelling variations are a problem for NER, and so-called 'historical normalisation' might be necessary.

Finally, a big challenge for NER for historical languages is the lack of financial resources, but also digitally usable lexica and annotated corpora. Most NER methods use supervised learning, i.e. they depend on already annotated corpora (= labelled data) to train their models. When annotated corpora are scattered over time and domains, the models can neither be improved nor their generalisation capacity increased. That 'state-of-the-art'-results on specific tasks can be achieved for classical languages if that dataset is large enough is shown by LatinBERT,[21] which was trained on the *Perseus Digital Library*, Latin *Wikipedia* and Latin texts from the Internet Archive.

We should, however, be aware of the "news bias"[22] in NER: Most annotated corpora for (standard) NER consist of news texts. As historical newspaper collections were massively digitised during the last years, it was logical to also base the annotated corpora for historical NER on those historical newspaper collections. These corpora also benefit from available word embedding technologies that are able to flag potential OCR misspellings and therefore allow for post-OCR correction. But then historical NER will run eventually in the same "news bias" as modern NER.

New Developments

In recent years there have been several approaches to improve NER for classical languages. Let us recall that common difficulties for historical NER are spelling variations, including punctuation, capitalisation and person name abbreviation, unknown names, and in general the complexity of entities. To give some very simple examples from a Latin language project:[23] Gaius Iulius Caesar should be marked as one entity, not as a "partial match". The algorithm may learn well compound entities of clear forms, e.g. name + de + toponym, such as *Bertrannus de Verziaco*, but already *Gariardus de loco Antimiano* may be difficult to resolve as one person name.

Chastang et al. (2021), who worked on a model for the automatic recognition of named entities in medieval Latin charters, describe very well the need for careful pre-processing to detect nested, overlapping or ill-formed annotations. Overlapping annotations stem from names that serve different functions: a saint's name can denominate the historical person, an abbey, a feudal territory, a festivity date, etc. As a standard machine learning classifier is not designed to attribute more than one class to each instance, the confusion between the multiple usage of a name (i.e., the *overlapping entities*) must be solved by creating designated classes for each entity. However, the more entities created, the more choice and potential ambiguity is created.

Moreover, Latin is a very versatile language, which makes NER a difficult task *in se*:

> The overgeneralization of very common particles (such as *de* in compound entities), as well as of location trigger words (such as *terra, serum, pars, domus, manus, apud*) and also of personal co-occurrent words (such as *episcopus, beatus, dominus, sacerdos, miles*, etc.) can lead to false positives when the model finds an entity different than that expected. (…) Latin phrase order is irregular, and exceptions in medieval variants are almost infinite; consequently, training taking

21 See Bamman / Burns (2020) and https://github.com/dbamman/latin-bert (last access 28.07.2025). LatinBERT achieved "state-of-the-art" performance on POS tagging.

22 Used as *terminus technicus* in Plank (2016a).

23 Chastang et al. (2021).

> into account grammatical rules, co-occurrences, and context can generate many false positives (…) The difficulty in recognizing entities (...) lies not so much in their quantity as in their extensive consequences. The percentage of complex entities does not exceed 11% of the total in our corpora, but the statistical impact on results due to bad recognition of such entities is more elevated [with growing complexity].[24]

With this in mind, it is understandable that the transformer-based NER for ancient Greek by Yousef et al. (2023b) performed poorly on multi-token entities: the available training data was composed of single-token entities.

The work of Berti (2019) started from large annotated datasets of specific sources and used semantic annotation platforms and Machine Learning. Vatri /McGillivray (2020) compared several lemmatizers and concluded that those based on large lexica are still producing better results than the ML-based lemmatizers.

The interesting work of Yousef et al. (2023a) suggests to use cross-lingual annotation projection to transfer NER annotations, done on translations, to Greek and Latin. The idea here is not to machine-translate detected name entities and to look them up in the "more difficult" language, but to employ automatic word alignment to find the equivalents of the detected entity in the parallel sentence and then to project the annotation.[25] The precondition for the success of such an approach is the alignment of the text corpora in both languages on the sentence or paragraph level. This makes the Bible as an obvious demonstration example. The perceived accuracy of the annotations was 86% for English-Ancient Greek.[26] Misclassifications appeared as a consequence of the translation, which adopted ad different type of entity, and multi-token entities such as 'Jesus Christ' or 'Pontius Pilate' were frequently misaligned or only partially aligned. Nevertheless, their result is a promising sign, if one happens to have a fine-tuned model for the languages and sentence-level alignment at hand.

Which Tool to Use?

Considering the above-mentioned challenges, what shall we do when we have a NER task to do? The answer is, as always 'it depends' – in particular, it depends what we want to annotate, and which factors are important to us.

First of all, what is our goal with the annotation? Which question do we wish to answer? Do we have a huge unexplored, but fully digital text corpus at hand, and we wish to get a rough overview? Are we aiming for a fine-grained annotation to answer delicate research questions? Is the goal a digitally enhanced edition, which is easily searchable?

Second, what is our raw data? In which documents and for which usecases do we intend to use our desired annotation? Of which century or which type of language are we concerned? Does there already exist a high-quality OCR of the desired texts? If not, which century and which writing style do the manuscripts-to-be-annotated have? Do we already have a robust OCR model that we can use?

Third, we need to decide which role does accuracy play for us: Do we need a highly accurate annotation within a digital edition, with (ideally) no false-positives or false-negatives? Or do we merely wish to get an overview, a visualization or another type of digital enhancement of the text?

To give a comprehensive decision 'algorithm' in answer to the above-mentioned questions is beyond the scope of this work. A lesson-learned from our work is the following rule of thumb: The more accuracy plays a role, the more specific a tool has to be, and the more likely we will end up with pro-

24 Chastang et al. (2021), paragraphs 72, 83, 84.

25 This approach was suggested in the computational linguistics community by Ni et al. (2018).

26 We use 'perceived accuracy' here as there was no standardised scoring employed, but qualitative inspection of random sample verses by humans. It is not known if partial matches were perceived as accurate or as not accurate.

gramming a rule-based algorithm ourselves, potentially including a large lexicon as feed-in data to it if necessary. But if our aim is to quickly get an overview or explore a huge corpus, the more likely it is that we are OK with a couple of false positives, as they do not skew a statistical analysis of a text too much.

Of course, with limited time, technical expertise and resources, we may often be inclined to use off-the-shelf NER tools for our data at hand. This is perfectly reasonable, but the question is whether we can reasonably expect accuracies[27] when our dataset is very different from the standard CoNLL-03 English corpus. Despite the advertisement, all-purpose general NER may have significant drop in accuracy in our application. Already an accuracy of 80% means that every fifth annotation is wrong. As we may, after hasty employment of an off-the-shelf tool, not expect much more than 70%, rather less, we need to ask ourselves if an error rate of 3 matches out of 10 is still giving us a meaningful output of our research question.

Recognition and Annotation of *Nomina Sacra*

Depending on the factors above, we may decide which tool to turn to. There are already good tools like *spaCy*, *greCy* or *odyCy*, that do give annotations of reasonable quality on certain Greek texts. But we need to keep in mind that Ancient Greek is a highly fragmented language with plenty of variants, both regional and temporal. Currently, each available model is limited to the particular dialect or morphology of their training dataset and generalization is rather poor.[28]

To our best knowledge, the current available NER tools do not recognize or evaluate *nomina sacra*. We guess that probably no available model has ever seen the original abbreviated form, as available datasets are made with a different scope than ours: We are interested (also) in the temporal development of scribal practice, hence, the comparison of manuscripts from different centuries is an important tool for our research question. The majority of scholars interested in texts are, however, dealing with questions on the content of the text, and therefore are not concerned with potential scribal mistakes or incomplete textual transmission, they merely need 'the text', by which they mean 'the original / true text'. Which makes total sense as there are not too many cases of complex transmission histories that are well-studied.

In case of biblical texts, 'the text of the Bible' is usually defined as Nestle-Aland's latest edition, which is also available in digital form. Nestle-Aland's edition is the output of decade-long research on the largest possible amount of available manuscripts[29] (papyri, majuscules, minuscules, lectionares, talismans), it is not 1:1 aligned to the text of any particular manuscript, but a carefully curated edition that aims to get as close as possible to the 'Ausgangstext', the best approximation to the original Bible text. While being an incredible piece of scholarly work, it is, unfortunately, not helpful for us.

The ideal (non-existing) machine-learning algorithm to recognize *nomina sacra* would scan the manuscript image for overline bars, check whether the letters under the overline bar are connected to a 'sacred word', in the sense that the letters form a reasonable abbreviation of a dictionary word that is a named entity in the Bible, and then would finally expand the abbreviation into that word.[30]

To train a model for such a goal, we would need a large set of manuscript-specific transcriptions of high-quality manuscript scans that contain *nomina sacra*. That led us to use available transcribed co-

27 It is hard to give reliable accuracy numbers, as the versions of commercial tools put out in the web change fast, and their underlying datasets are not always revealed. Just to give an idea, BiLSTM achieved an F-score of 90,1% in Huang et al. (2015) and, when improved with subword representations, of 91,2%, see Ma / Hovy (2016).

28 See Kostkan et al. (2023).

29 The editing institution, the Institute For New Testament Textual Research, has collected approximately 5800 manuscripts up to the present: see https://www.uni-muenster.de/INTF/ (last access 28.07.2025) for more information.

30 More technical details can be found in Geldhauser / Malyshev (2024).

dices such as *Codex Sinaiticus* as a ground truth: we started fine-tuning an *escriptorium* model with 50 pages of artificially created ground truth from *Codex Sinaiticus*, for which we provided a ground-truth annotation, including an annotation of *nomina sacra*.

We used this 'data augmentation' approach as it was the fastest way to ensure we had a lot of examples of *nomina sacra* correctly annotated. Despite 50 pages should have been sufficient to adapt an existing model, according to the *escriptorium* user community's 'rule of thumb', the model failed to recognize *nomina sacra* – indeed, it seems one needs much more data to preserve the overline bar as a cue.

Therefore, it became clear that our aim was far more difficult to achieve than we thought: after recognising the overline bar as a cue, an imaginary future algorithm also has to understand, given the heterogeneity of the abbreviations used, that two or more different sets of characters may correspond to the same expanded word form.

Moreover, to successfully annotate *nomina sacra* of transcriptions of biblical texts originating in a different century, a ML model has to be general enough to deal with a variety of different writing styles, page formats etc.

Hence, we subsequently turned to a sequential method, disentangling recognition and annotation. For annotation, we decided for a rule-based approach.[31] Here, we define rule-based approach in the pure sense of the word: the user creates a list of relevant abbreviations for annotation, using a tool of their choice, and the algorithm contains a function that finds and expands abbreviated *nomina sacra* from the transcribed text. Here, no 'learning' of the algorithm is required, it is a simple search script.

This 'disentangling'-approach is possible as Bruce Metzger compiled what seems to be a complete list of words treated as *nomina sacra* from Greek papyri: The Greek counterparts of God, Lord, Jesus, Christ, Son, Spirit, David, Cross, Father, Israel, Saviour, Man, Jerusalem, and Heaven[32] all occur as *nomina sacra* in Greek language biblical manuscripts of the 3rd century and often earlier, and, according to Comfort and Barrett, also Mother is consistently used as nomen sacrum from the 4th century onwards.[33] These 15 *nomina sacra*, together with their relevant forms for genitive and other cases, and both in contraction and suspension abbreviation, are used as rulebook for annotation.

The advantage of our method is its simplicity, fastness and precision. Of course, our method is not universal, but tailored to the problem: the working of our algorithm employs 'expert knowledge' in the sense that we use the special characteristics of *nomina sacra*, which were discovered and listed by humans. In case the OCR were flawless, we would get a perfect accuracy thanks to the rule-based approach that we take in the annotation step.

A tailored approach is by design not thought to generalise well to other situations, hence, we do not claim that our approach will be suitable for a vast amount of other purposes. However, our concept will be helpful also to scholars that look for a fine-grained annotation of named entities in other settings: Indeed, our approach can be used whenever a complete list of entities of a given class is readily available, independently of the century, the genre, the type of text etc. For example, reference works such as the *Genealogisches Handbuch des Adels* provide a list of noble people in the former *Sacrum Imperium Romanum*, and new insights on power structures, alliances and networks of the nobility of a certain time could be gained when annotating all appearances of noble people listed in this reference work.

---

31 Historically, the terminus rule-based approach was attributed to methods that exploit grammar and regularities in the data. In the 1990s, these were state-of-the-art algorithms that rely on linguistic pre-processing, such as morpho-syntactic tagging, tokenisation and sentence splitting, and that often require external resources such as trigger words in gazetteers.

32 Metzger (1981), 36.

33 Comfort / Barrett (2001), 34.

Summary and Conclusion

NER techniques are an excellent way to provide an overview of a corpus, and can provide added value to many datasets in general and digital editions in particular.

We reviewed a couple of existing methods and concluded that there is no all-purpose tool that always gives good results. A big obstacle for the application of the most widely known and used NER tools to texts and research questions of scholars in Classics is their unsuitable training databases, normally modern newspaper collections. Moreover, the vast variety of texts and the variability of the Latin and Greek language over the course of the centuries implies the need for a multitude of tools.

In our specific application, we were interested in the annotation of a specific named entity that is not among the standard list, and for which there is not yet a suitably large training dataset. Hence, there was no potential basis of success for a machine-learning based algorithm. Nevertheless, we could get quite promising results with a tailored rule-based approach.

In terms of high-level recommendations for users, one takeaway from our project is that NLP tools, at their current state-of-the-art, are not the right tool for a highest-accuracy project. This is currently an issue ‘by design’: An LLM tool was built to produce an output that is ‘natural‘, by whatever standard, it uses probabilities and predictions on words based on their underlying dataset. In this sense, it is designed to output a ‘good’ text, but it is not designed to perform a very precisely defined ‘algorithmic’ task with 100% accuracy. When deciding which tool to (further) develop, digital humanists should not let themselves be blinded by the current AI hype, but decide strategically which kind of working mechanism is suitable for their research question and the specific task to be performed algorithmically.

References

Agerri et al. (2018): R. Agerri / Y. Chung / I. Aldabe / N. Aranberri / G. Labaka / G. Rigau, Building named entity recognition taggers via parallel corpora, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki 2018.

Bamman / Burns (2009): D. Bamman / P. J. Burns, LatinBERT: A contextual language model for classical philology, in: arXiv:2009.10053.

Berti (2019): M. Berti, Named entity annotation for ancient Greek with INCEpTION, in: CLARIN Annual Conference Proceedings, Leipzig 2009, 1–4.

Celano (2024): G. G. A. Celano, Opera Graeca Adnotata: Building a 34M+ Token Multilayer Corpus for Ancient Greek, in: arXiv:2404.00739.

Chastang et al. (2021): P. Chastang / S. Torres Aguilar / X. Tannier, A Named Entity Recognition Model for Medieval Latin Charters, DHQ 15/4 (2021), https://www.digitalhumanities.org/dhq/vol/15/4/000574/000574.html (last access 28.07.2025).

Comfort / Barrett (2001): P. W. Comfort / D. Barrett, Text of the Earliest New Testament Greek Manuscripts (2nd ed.), Wheaton (IL) 2001, 34–35.

Devlin et al. (2019): J. Devlin / M.-W. Chang / K. Lee / K. Toutanova, BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis (MN) 2019, 4171–4186.

Ehrmann et al. (2023): M. Ehrmann / A. Hamdi / E. L. Pontes / M. Romanello / A. Doucet, Named entity recognition and classification in historical documents: A survey, in: ACM Computing Surveys 56/2 (2023), 1–47.

Geldhauser / Malyshev (2024): C. Geldhauser / K.A. Malyshev, Semi-automatic annotation of Greek majuscule manuscripts: Steps towards integrated transcription and annotation, in: Annals of Computer Science and Information Systems 41 (2024), 37–44.

Huang et al. (2015): Huang Z. / Xu W. / Yu K., Bidirectional LSTM-CRF Models for Sequence Tagging, in: arXiv:1508.01991.

Hurtado (2017): L. W. Hurtado, Texts and Artefacts: Selected Essays on Textual Criticism and Early Christian Manuscripts, London 2017.

Kostkan et al. (2023): J. Kostkan / M. Kardos / J. P. B. Mortensen / K. L. Nielbo, OdyCy – A general-purpose NLP pipeline for Ancient Greek, in: Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Dubrovnik 2023, 128–134.

Ma / Hovy (2016): Ma X. / E. Hovy, End-to-End Sequence Labeling via Bi-Directional LSTM-CNNs-CRF, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin 2016, 1064–1074.

Metzger (1981): B. M. Metzger, Manuscripts of the Greek Bible: An Introduction to Palaeography. London 1981.

Nadeau / Sekine (2007): D. Nadeau / S. Sekine, A survey of named entity recognition and classification, in: Lingvisticae Investigationes 30 (2007), 3–26.

Ni et al. (2017): Ni J. / D. Georgiana / F. Radu, Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver 2017, 1470–1480.

Plank (2016a): B. Plank, What to do about non-standard (or non-canonical) language in NLP, in: arXiv:1608.07836.

Plank (2016b): B. Plank, What to Do about Non-Standard (or Non-Canonical) Language in NLP, in: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), Bochum 2016, 13–20.

Rosset (2012): S. Rosset / C. Grouin / K. Fort / O. Galibert / J. Kahn / P. Zweigenbaum, Structured Named Entities in Two Distinct Press Corpora: Contemporary Broadcast News and Old Newspapers, in: 6th Linguistics Annotation Workshop (The LAW VI), Jeju 2012, 40–48.

van Strien et al. (2020): D. van Strien / K. Beelen / M. Coll Ardanuy / K. Hosseini / B. McGillivray / G. Colavizza, Assessing the Impact of OCR Quality on Downstream NLP Tasks, in: Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Valletta 2020, 484–496.

Traube (1907): L. Traube, Nomina sacra. Versuch einer Geschichte der christlichen Kürzung, München 1907.

Vatri / McGillivray (2020): A. Vatri / B. McGillivray, Lemmatization for ancient Greek: An experimental assessment of the state of the art, in: Journal of Greek linguistics 20/2 (2020), 179–196.

Wilkinson (2015): R. J. Wilkinson, Tetragrammaton: Western Christians and the Hebrew Name of God: From the Beginnings to the Seventeenth Century, Leiden 2015.

Yousef et al. (2023a): T. Yousef / C. Palladino / G. Heyer / S. Jänicke, Named entity annotation projection applied to classical languages, in: Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 175–182.

Yousef et al. (2023b): T. Yousef / C. Palladino / S. Jänicke, Transformer-based named entity recognition for ancient Greek, in: Digital Humanities 2023, University of Graz 2023, 1–3.

Author Contact Information[34]

Dr. Carina Geldhauser
ETH Zürich
Rämistr. 101
8092 Zürich
Schweiz
E-mail: carina.geldhauser@math.ethz.ch

34

DIGITAL CLASSICS ONLINE

Greek and Latin Proper Names in Georgian Scholarship: Epigraphic, Lexicographic and Encyclopedic Traditions, Their Standardisation and Digitisation

Irine Darchia

Abstract: Greek and Latin proper names are embedded across many Georgian scholarly traditions, including epigraphic corpora, lexicographic works, and encyclopedic projects. Their rendering into Georgian has varied over time, shaped by Byzantine, Russian, and European influences, and by differing translational practices. As a result, multiple versions coexist, often creating inconsistencies in scholarship and pedagogy. This article examines three major resources: the *Encyclopedia Caucasus Antiquus*, the corpora of Greek inscriptions discovered in Georgia, and the *Orthographic Dictionary of Greek and Roman Proper Names*. It traces the historical stages of translating Greek and Latin names into Georgian and discusses current attempts of Standardisation. Particular attention is given to ongoing digitisation initiatives, including adopting the Cadmus platform for the *Digital Caucasus Antiquus*. The article argues that digitisation and standardisation of proper names is not simply a technical matter but a cultural and linguistic imperative. For languages with limited global digital presence such as Georgian, the creation of structured digital resources is essential for safeguarding scholarly traditions and ensuring visibility within international digital humanities.

Introduction

Studying proper names has always stood at the crossroads of language, culture, and identity. In the case of Georgia, a country situated at the meeting point of East and West, Greek names, and to a lesser extent Latin ones, occupy a special place in intellectual history. They appear in inscriptions scattered across the territory of ancient Colchis and Iberia, in lexicographic works that reflect centuries of linguistic adaptation, and in encyclopedic projects that attempt to systematise the ancient Caucasian world. These names serve as markers of cultural contact, transmission, and translation.

Yet their rendering into Georgian has been far from uniform. Across centuries, Georgian scholars and translators have employed different models, sometimes privileging vocative forms rooted in oral transmission, at other times adopting nominative endings aligned with Byzantine or European practice, and later borrowing root forms introduced through Russian mediation. The result has been a layered and sometimes inconsistent tradition, visible in classical translations, scholarly works, and reference texts.

In the present digital age, these inconsistencies take on new significance. The issue is not merely philological but also cultural. For Georgia to maintain its scholarly visibility, it must establish a strong digital presence. This involves not only digitising primary sources and secondary scholarship but also standardising the representation of proper names across resources. Proper names are central to historical, geographical, and literary texts, and their accurate and consistent rendering determines the usability of digital corpora for both national and international audiences.

This article focuses on three major resources that together illustrate both the richness of Georgian classical scholarship and the challenges it faces:

1. The *Encyclopedia Caucasus Antiquus* – the most comprehensive encyclopedic project on the ancient Caucasus, which contains a vast corpus of Greek and Latin names.

2. The various epigraphic corpora, including the *Corpus of Greek Inscriptions of Georgia* and other digitisation projects.

3. The *Orthographic Dictionary of Greek and Roman Proper Names*, which provides standardised forms and addresses long-standing translational inconsistencies.

The paper also considers the broader historical traditions of translating Greek and Roman names into Georgian and explores the digitisation initiatives that seek to integrate these resources into accessible, interoperable, and sustainable digital platforms. The central argument advanced here is that digitisation must be understood as a cultural strategy: a means of safeguarding Georgia's intellectual heritage, ensuring linguistic visibility, and positioning Georgian scholarship within the global digital humanities.[1]

Although this article primarily focuses on three major scholarly resources rather than on individual names themselves, it is important to underline that Greek and Latin proper names form a substantial part of the material preserved in Georgian classical scholarship. Personal names (e.g. Alexander, Plato, Homer), ethnonyms (e.g. Iberians, Colchians, Scythians), and geographical names (e.g. Phasis, Pontus) appear in inscriptions, historical sources, and lexicographic works. Their transmission into Georgian reflects complex processes of linguistic adaptation, cultural mediation, and scholarly interpretation. The coexistence of multiple forms of the same name illustrates the broader challenges of transliteration, translation and normalisation across centuries.

Encyclopedia Caucasus Antiquus

The *Encyclopedia Caucasus Antiquus* (ECA) is the first and only comprehensive encyclopedic study of the ancient history and culture of the Caucasus, spanning the second millennium BC to the 5th–6th centuries AD. Conceived and led by the late Academician Rismag Gordeziani at the Institute of Classical, Byzantine, and Modern Greek Studies at Ivane Javakhishvili Tbilisi State University, the project engaged nearly fifty scholars over two decades. It was supported by multiple funding schemes of the Ministry of Education and Science of Georgia, the Shota Rustaveli National Science Foundation, the Research Centre of Kartvelology of the Patriarchate of Georgia, and the University itself.

The Encyclopedia is divided into five volumes:

Volume I (Sources): Includes 97 Greek and 67 Latin authors, presented in bilingual Greek–Georgian and Latin-Georgian translations. It also incorporates Hittite, Assyrian, Urartian, Persian, and biblical texts (in Georgian translation only).[2]

Volumes II–IV (Articles): Comprising approximately 2,500 entries, these volumes cover general topics on the Caucasus as well as specific countries, ethnic groups, geographical locations, and historical or mythological figures.[3]

1 ChatGPT (OpenAI) was used for the translation of selected Georgian secondary sources into English, and Grammarly for proofreading language and style. All content was verified by the author.

2 Gordeziani et al. (2022).

3 Gordeziani et al. (2014); Gordeziani et al. (2016); Gordeziani et al. (2018); Gordeziani et al. (2020); Gordeziani et al. (2021).

Volume V (Maps and Illustrations): It contains 26 original maps produced by the project's cartographer in close collaboration with the historians and philologists engaged in the research. These include: a. Maps focused on the Caucasus region itself; b. Maps illustrating the Caucasus in relation to the wider ancient world; c. Synthesising maps that position the Caucasus within the context of neighbouring regions, such as Achaemenid Persia and its satrapies, Greek colonisation, the Hellenistic world, the Arsacid kingdom, the Sassanid empire, and beyond. This volume of the Encyclopedia provides a comprehensive visual exploration of the historical and geographical dimensions of the Caucasus.[4]

The *Encyclopedia Caucasus Antiquus* – as expected – provides rich and well-systematized material on Greek and Latin proper names, including geographical/toponyms, historical, and mythological ones. One vivid example is a lemma presented in Georgian with its English translation, which illustrates the richness of the material:

> "პლატონი (Πλάτων) – ძვ.წ. V-IV სს. ათენელი ფილოსოფოსი. მის დიალოგებში (მაგ., «ფედონი», «კანონები», «ევთიდემოსი») გაკვრით არის ნახსენები: ფასისი, პონტოსი, სკვითები, იბერები, სავრომატიდები, კოლხი მედეა".

> "Plato (Πλάτων) – 5th–4th centuries BC, Athenian philosopher. In his dialogues (e.g., *Phaedo*, *Laws*, *Euthydemos*), Phasis, Pontos, Scythians, Iberians, Savromats, and Medea are briefly mentioned".

This short encyclopedic entry contains eight toponyms, ethnonyms, and personal names.

The *Index Nominum et Locorum* provided in the *Encyclopedia* spans 30 pages. It includes several types of personal names, presented with their alternative forms in both the original language and the Georgian translation (e.g., აიეტი – Αἰήτης / Αἰᾱ́τᾱς – Aeeta/Aeetes). Historical figures, literary characters, and mythological personages are all represented, though their exact number remains undetermined and is the subject of separate study.

After extensive discussions and consultations with leading experts in digital humanities, various options were explored and analysed. For example, one of the possibilities considered was the Software MediaWiki platform (which underlies both *Wikipedia* and the *Digital Classicist Wiki*). It is particularly well-suited for collaborative editing and can be configured so that only official authors have editing rights. Various plugins are also available to add functionality and customise the display. However, as one digital humanities expert noted, it "would not have the streamlined feel of a printed Encyclopedia."

Other options included traditional and prestigious publishers such as Brill, Oxford, and Cambridge, yet these proved financially unfeasible. Ultimately, we identified the *Digital Encyclopedia of Atticism* as the closest model, both in its concept and in the structure and visual appearance, to the *Encyclopedia Caucasus Antiquus*.[5]

The *Digital Encyclopedia of Atticism* is one of the research outputs of *PURA*, a five-year ERC Consolidator Project (grant agreement no. 865817), which commenced in January 2021 at Ca' Foscari University of Venice. *PURA* investigates the theories of linguistic purism developed in ancient Greek culture and their reception in later periods. The primary focus of the analysis is the Atticist lexica – ancient 'dictionaries' that collect linguistic features to be cultivated or avoided in correct Greek. All sections have been created with the Cadmus program, developed by Daniele Fusi in collaboration with the Venice Centre for Digital and Public Humanities. Cadmus is an open-source, lightweight frame-

4 Gordeziani et al. (2023).

5 Tribulato (2022): https://atticism.eu/ (last access 10.09.2025).

work designed for building web-based, layered, modular, and user-friendly content creation systems for highly structured data in the field of Digital Humanities.[6]

After long and thorough discussions, the Encyclopedia team decided to adopt the Cadmus program, which will be specially adjusted to meet the project's needs.[7] The *Digital Encyclopedia Caucasus Antiquus* – following the structure of the printed version – will consist of the following lemma categories:

- *Sources:* Primary sources in Greek and Latin with Georgian translations. (At a later stage, translations into English, Greek, and German will most likely be provided as well.)
- *Encyclopaedic articles:* Initially in Georgian. (At the next stage, translations into English will be added, along with new, original articles contributed by Georgian and international scholars).
- *Maps:* A total of 26.
- *Additional information:* Abbreviated cited scientific literature; authors of scientific articles; abbreviations of ancient authors and works, etc.

The structure of each article will be as follows: Term/Lemma in Georgian; Equivalent in the ancient languages (Greek/Latin); Scientific elaboration/article itself (potentially sub-structured); Bibliography; Initials of the article authors.

A key added value of the *Digital Caucasus Antiquus* will be its advanced *search functionality.* In the *Sources* section, users can filter materials by categories: Ancient Oriental sources, Hittite texts, Akkadian texts, Urartian texts, Persian cuneiform inscriptions, Ancient Greek sources, Latin sources, and the Old Testament. In the *Articles* section, users can filter terms by categories such as biographies of ancient authors, geographical locations/toponyms, ethnonyms, hydronyms, oronyms, mythological figures, historical figures, and literary personages.

Thus, once the Cadmus program is adapted and the existing material (five volumes of the printed Encyclopedia) is uploaded, the *Digital Caucasus Antiquus* will become a valuable open-source resource, providing thousands of Greek and Latin proper names and documenting their usage in different historical and literary contexts related to the Ancient Caucasus.

The *Digital Caucasus Antiquus* will replicate the printed structure while adding search functionalities, cross-references, and multilingual translations (Georgian, English, Greek, German). It will also integrate with digitised inscriptions, provide automatic translation options through AI tools, and allow continuous updating. This will transform the Encyclopedia from a monumental printed achievement into a living, open-access resource for the international scholarly community.

---

6 Apart from the *Digital Encyclopedia of Atticism*, the following projects also make use of the Cadmus program: ERC Consolidator Grant *PURA* (*Purism in Antiquity: Theories of Language in Greek Atticist Lexica and their Legacy*); ERC *PAGES* (*Priscian's Ars Grammatica in European Scriptoria*); PRIN *TAL* (*The Transmission of Ancient Linguistics: Texts and Contexts of the Roman Grammatical Studies*); Sapienza Università di Roma/von Humboldt Stiftung *ThDS* (*Thesaurus Dubii Sermonis: Digital Critical Collection of Ancient Latin Linguistics – 1st century BC–8th century CE*); PRIN Petrarch's *ITINERA* (*Italian Trecento Intellectual Network and European Renaissance Advent*); PRIN *Re.Novella* (*The Genre of the Novella in the Italian Renaissance: Repertoire, Database and Historiographical Framework*); PRIN *MQDQ* (*Musisque Deoque*); Horizon *GISARC* (*Greek In Sicily After the Roman Conquest*); Horizon *MapAeg* (Cristoforo Buondelmonti's *Liber Insularum*); etc.

7 The TSU Institute of Classical, Byzantine, and Modern Greek Studies is highly grateful to Elena Spangenberg Yanes for facilitating communication and coordinating the potential collaboration with the University of Sapienza of Rome, and personally with Daniele Fusi, the developer of Cadmus. We greatly appreciate his kind acceptance of our proposal to use the Cadmus program for the *Digital Encyclopaedia Caucasus Antiquus* and to adapt the software to its specific needs.

Epigraphic Corpora

Greek inscriptions discovered in Georgia represent another vital source for studying proper names and cultural contacts. They testify to the historical interactions between Georgia (ancient Colchis and Iberia) and the Greek world.

The foundational work is the *Corpus of Greek Inscriptions of Georgia*, compiled by Tinatin Kaukhchishvili and later edited by Levan Gordeziani.[8] The first volumes appeared in 1999–2000, followed by a revised single-volume edition in 2004 and the third in 2009.[9] The volume consists of photos of Greek inscriptions, their transliteration and Georgian translation, commentaries, an index (with more than 1200 entries), and a comprehensive summary in German.

The *Corpus of Greek Inscriptions of Georgia* contains over 100 inscriptions, dating from the 6th–5th centuries BC. to the present day, found on various materials such as stone, objects, and frescoes. In terms of content, these include epitaphs, building inscriptions, dedications, and religious texts. The *Corpus* is a significant source for researching the history of ancient Colchis and Iberia (the historical names of Western and Eastern Georgia). The exact number of Greek proper names it contains is still unknown, but it may serve as an essential resource given the specific nature of the epigraphic material.

A writing set from Mtskheta, discovered in 2001 during excavations at Svetitskhoveli Cathedral (Stonetomb No. 14) and dated to the 3rd–4th centuries, is a notable example of using proper names.[10] The set is decorated with relief representations of the nine Muses in silver-gilt, miniature silver figures of Homer, Demosthenes, and Menander, and an openwork gold plaque bearing the Greek inscription: Βασιλέως Ουστάμου του και Ευγενίου. This writing set – featuring images of the Muses, Greek writers, and Greek personal names – symbolically reflects the historical role of the Greek language and culture in Georgia as vehicles for the dissemination of writing and knowledge.

From 2015 to 2017, the TSU Institute of Classical, Byzantine, and Modern Greek Studies carried out a research project led by Levan Gordeziani and titled *Online Catalogue of Greek Inscriptions of Georgia*, which the Shota Rustaveli National Science Foundation of Georgia funded.[11] Within the framework of this project, 150 inscriptions were digitised in EpiDoc – the best available toolset and community of practice for encoding digital editions of ancient texts (including inscriptions, papyri, seals, coins, and related objects) in TEI XML, the de facto standard for digital literary and historical editions.[12] However, these inscriptions are not yet available online due to technical and financial issues, which are expected to be resolved soon. Most probably, the 150 digitised Greek inscriptions will be made accessible online in 2026, offering valuable insights not only into various aspects of Georgian history and its historical and cultural connections with Greece, but also serving as a source for identifying Greek proper names that were common outside Greece, in ancient Colchis and Iberia. It should be mentioned that once the *Digital Caucasus Antiquus* is finalised, these 150 digitised inscriptions will be linked to it.[13]

---

8 Darchia (2007); Wyles et al. (2016), 25.

9 Kaukhchishvili (1999); Kaukhchishvili (2000); Kaukhchishvili (2000a); Kaukhchishvili (2009).

10 Apakidze et al. (2004), 104–123.

11 It should be underlined that digital humanities tools were introduced into the field of classics in Georgia through collaboration with the Humboldt Chair of Digital Humanities at the University of Leipzig, as well as through two seasonal schools in digital humanities organised by Irine Darchia in Tbilisi in 2013–2014 with funding from the Ministry of Education and Science of Georgia and the Shota Rustaveli National Science Foundation. Special thanks are due to Gregory Crane, Monica Berti, Gabriel Bodard, Simona Stoyanova, Dimitar Iliev, and Gian Paolo Renello for their valuable help and support in introducing digital classics in Georgia.

12 https://ics.sas.ac.uk/ics-digital/epidoc (last access 10.09.2025).

13 It is worth mentioning that 28 Greek inscriptions discovered in Georgia and containing proper names are available through the Packard Humanities Institute's project: http://epigraphy.packhum.org/inscriptions/ (last access 10.09.2025).

Orthographic Dictionary of Greek and Roman Proper Names

The Orthographic Dictionary of Greek and Roman Proper Names, compiled by Nana Tonia in 2023, is the most extensive single collection of Greek and Latin names in Georgian.[14] This revised and updated edition builds on earlier dictionaries from 1980 and the 1970s volumes of the Georgian Soviet Encyclopedia.[15] It contains over 4,000 entries and was prepared in collaboration with the Department of State Language of Georgia.

The dictionary provides lemmas of Greek and Latin proper names with standardised Georgian transliterations. It establishes rules for translating names, thereby addressing long-standing inconsistencies. One striking example concerns Homer: his name appears in Georgian sources in numerous forms – ჰომეროსი (Homerosi), ჰომერე (Homere), ომიროს (Omiros), ომირი (Omiri), უმიროს (Umiros), უმირი (Umiri). The dictionary sets forth a consistent approach, helping to unify scholarship and pedagogy.

Thus, the dictionary functions both as a practical tool for scholars and as a symbolic step towards linguistic normalisation. Its digitisation will be essential, not only for cross-linking with encyclopedic and epigraphic corpora but also for ensuring consistency in digital editions of classical texts.

Historical Traditions of Translating Greek and Latin Proper Names into Georgian

The rendering of Greek and Latin proper names into Georgian has followed two broad historical stages.

The *first stage* (from the earliest centuries to the 18th century) unfolded under the influence of Byzantium, fostering both written and direct cultural exchange. When translating early Christian writings, Georgian translators inevitably faced the challenge of rendering proper names, which were often transmitted orally without linguistic standardisation. From the 11th century onward, the translation and commentary of philosophical and theological works intensified, especially at the Gelati school. The complexity of Greek terminology and the need to refine Georgian theological language encouraged more systematic approaches. By the Middle Ages, two distinct tendencies had emerged: (a) orally transmitted names became established in Georgian with vocative endings (თევდორე [tʰevdore], არისტოფანე [aristopʰane]); (b) literary scholarship favored more academic forms (ჰომეროსი [h̲omerosi], პინდაროსი [pindarosi], არტემისი [artemisi]).

The *second stage* (from the 18th century onward) coincided with increasing Georgian-Russian cultural interaction. Numerous classical works entered Georgian through Russian, leaving a strong imprint on terminology. Russian practice encouraged the use of root forms and introduced gendered endings, such as აფროდიტა [apʰrodita], which Georgian adopted, although older Georgian forms like აფროდიტე [apʰrodite] also persisted, resulting in frequent parallel usages.

As a result of these two stages and the diverse approaches they introduced, different viewpoints have emerged among Georgian scholars. Some researchers argue that the European tradition should be followed, rendering names in the nominative form. Others generally support this approach but consider names ending in *-ας* or *-ης* as exceptions, since their Georgian forms can suggest a genitive ending (e.g., ალკიბიადესი [alkibiadesi]). They therefore propose that only such names should be adapted,

Another 7 Greek inscriptions can be accessed online via the *Epigraphic Corpus of Georgia* project run by Ilia State University: http://v.epigraphy.iliauni.edu.ge/en-US (last access 10.09.2025).

14 Tonia (2023).

15 Gigineishvili (1985).

while others should remain non-inflected (for example, it should be rendered as ალკიბიადე [alkibiade] rather than ალკიბიადესი [alkibiadesi]).

Some scholars still follow the Russian-influenced approach, favouring the transfer of names in their root forms (e.g., არტემიდა [artemida], ვენერა [Venera]). Another significant tradition, rooted in centuries of Greco-Georgian contact, favours rendering names with vocative endings (ალექსანდრე [aleḳsandre], სოკრატე [sokrate], არისტოტელე [aristotele]), reflecting their oral transmission. Geographical names, however, remain especially problematic, and complete standardisation is impossible – just as in other languages. For example, Russian uses both Ахилл and Ахиллес; English and German use Pindar and Pindaros; and Modern Greek uses both Πλάτων and Πλάτωνας.

In general, the most recent Georgian publications reveal a clear tendency to render all foreign terms in the nominative case, which significantly simplifies the approach: there is no longer a need to deliberate whether one should write ჰომეროსი [ḫomerosi] (the academic form), ჰომერი [ḫomeri] (formed from the Greek root with the Georgian nominative ending), ჰომერე [ḫomere] (the Greek vocative form, as in ალექსანდრე [aleḳsandre]), or the various forms attested in older Georgian translations – ომიროს [omiros], ომირი [omiri], უმიროს [umiros], უმირი [umiri], and others.

After careful consideration of these inconsistencies, preference has been given to the Byzantine model of rendering Greek and Roman names, as it essentially corresponds to modern European practice. Accordingly, when proper names are rendered into Georgian, priority should be given to the nominative form, while retaining certain traditional and well-established forms that derive from the vocative.[16]

The most recent publications show a clear trend towards nominative forms, which aligns with modern European practice and simplifies usage. Yet the persistence of older variants highlights the cultural depth and complexity of Georgian engagement with antiquity.

The question of standardising the Georgian forms of Greek and Latin proper names has therefore become increasingly important in modern scholarship. Without a consistent approach, the same historical figure may appear under several different spellings in academic publications, translations, and digital databases. Such variation complicates indexing, digital searchability, and cross-referencing between scholarly resources. The recent *Orthographic Dictionary of Greek and Roman Proper Names* represents a major step toward resolving these issues by proposing unified forms based primarily on the nominative case while respecting well-established Georgian traditions. Standardisation is therefore not merely a matter of orthography but also a prerequisite for integrating Georgian classical studies into international digital infrastructures.

Digitisation Challenges and Implications

The digitisation of classical materials in Georgia is not simply a matter of technical preservation. It is a cultural and linguistic necessity in a world where English and other global languages dominate digital and AI-driven environments, smaller languages such as Georgian risk marginalisation unless their resources are made digitally accessible.

The *Digital Caucasus Antiquus* will provide structured, searchable, multilingual access to sources, articles, and maps. Integration with digitised inscriptions will create a powerful resource for studying Greek and Latin names in Caucasian contexts. The digitisation of the *Orthographic Dictionary* will serve as a foundation for consistency across all platforms.

Digitisation ensures visibility and accessibility. It allows Georgian scholarship to participate fully in international digital humanities projects and enables collaboration with global scholars. By embedding

16 Tonia (2023), 16–22.

translational variants and historical traditions into digital platforms, Georgian scholarship can showcase its unique cultural heritage while aligning with international standards.

Digitisation also addresses the survival of Georgian as a scholarly language. Beyond Georgia's borders, the language is spoken within historical diaspora communities in Turkey, Azerbaijan, and Iran. Ensuring its digital presence strengthens its resilience in an AI-driven era. Moreover, proper names – being at once linguistic, cultural, and historical signifiers – are an ideal focal point for this effort.

The issue of proper names therefore intersects with broader cultural and political questions. The way in which names are transmitted, translated, or standardised reflects historical patterns of cultural orientation and scholarly influence. In the Georgian context, the transmission of Greek and Latin names has historically been shaped by Byzantine, Russian, and Western European traditions. The current digitisation initiatives therefore represent not only a technical development but also a symbolic reaffirmation of Georgia's place within the broader intellectual heritage of classical and European scholarship.

Conclusion

The representation of Greek and Latin proper names in Georgian scholarship reveals the richness of Georgia's intellectual traditions and the challenges of inconsistency. The *Encyclopedia Caucasus Antiquus*, the epigraphic corpora, and the *Orthographic Dictionary* form a powerful resource triad. Each highlights different aspects of the problem: the encyclopedia provides systematised data, the inscriptions preserve authentic usage, and the dictionary establishes rules for standardisation.

Digitisation is the key to uniting these resources and securing their future. By adopting open-source platforms such as Cadmus, Georgian scholarship is taking decisive steps to integrate its classical heritage into global digital humanities. The *Digital Caucasus Antiquus*, linked with inscriptions and supported by the dictionary, will offer scholars worldwide a valuable resource for studying proper names and cultural interactions in the ancient Caucasus.

Ultimately, the study of Greek and Latin proper names in Georgian scholarship illustrates the intersection of linguistic tradition, cultural identity, and modern technological development. The digitisation and standardisation of these names are not purely technical processes but part of a broader effort to preserve and promote Georgian scholarly heritage within the global digital environment. By making its classical resources accessible and interoperable, Georgian scholarship can contribute more visibly to international research while maintaining its linguistic and cultural specificity. In this sense, the question of names – *nomina omnia* – truly reflects the deeper structures of intellectual tradition and cultural orientation.

In a broader perspective, digitisation is more than a technical project; it is a cultural strategy. It ensures the survival and visibility of Georgian in an era dominated by global languages. It safeguards a tradition intensely local and integrally connected to the broader world of classical studies. It also affirms Georgia's place in the ongoing dialogue between past and present, East and West, and tradition and innovation.

Sources

Online Sources

PURA. Purism In Antiquity: Theories Of Language in Greek Atticist Lexica and their Legacy: https://atticism.eu (last access 10.09.2025).

EpiDoc: https://ics.sas.ac.uk/ics-digital/epidoc (last access 10.09.2025).

Digital Corpora

PHI Searchable Greek Inscriptions: https://epigraphy.packhum.org (last access 10.09.2025).

The Epigraphic Corpus of Georgia 2.0: http://v.epigraphy.iliauni.edu.ge/en-US (last access 10.09.2025).

References

Apakidze et al. (2004): A. Apakidze / G. Kipiani / V. Nikolaishvili, A Rich Burial from Mtskheta (Caucasian Iberia), in: Ancient West & East 3/1 (2004), 104–123, https://doi.org/10.1163/9789047405139_009 (last access 10.09.2025).

Darchia (2007): Irine Darchia, Greek Palaeography and Epigraphy in Georgian Science, in: Neli Makharadze / Tina Dolidze (eds.), Byzantine Studies in Georgia, Tbilisi 2007.

Gigineishvili (1985): K. Gigineishvili, Orthographic Dictionary of Greek and Roman Proper Names, Tbilisi 1985.

Gordeziani et al. (2022): R. Gordeziani / M. Danelia / G. Ugulava (eds.), Encyclopedia Caucasus Antiquus, I, Sources, 2nd revised edition, Tbilisi 2022.

Gordeziani et al. (2014): R. Gordeziani / M. Danelia (eds.), Encyclopedia Caucasus Antiquus, II, 1, Tbilisi 2014.

Gordeziani et al. (2016): R. Gordeziani / M. Danelia (eds.), Encyclopedia Caucasus Antiquus, II, 2, Tbilisi 2016.

Gordeziani et al. (2018): R. Gordeziani / M. Danelia / G. Ugulava (eds.), Encyclopedia Caucasus Antiquus, III, Tbilisi 2018.

Gordeziani et al. (2020): R. Gordeziani / M. Danelia / G. Ugulava (eds.), Encyclopedia Caucasus Antiquus, IV, 1, Tbilisi 2020.

Gordeziani et al. (2021): R. Gordeziani / M. Danelia / G. Ugulava (eds.), Encyclopedia Caucasus Antiquus, IV, 2, Tbilisi 2021.

Gordeziani et al. (2023): R. Gordeziani / M. Danelia / E. Kvirkvelia / G. Ugulava (eds.), Encyclopedia Caucasus Antiquus. Maps and Illustrations, Tbilisi 2023.

Kaukhchishvili (1999): T. Kaukhchishvili, Corpus of Greek Inscriptions of Georgia, I, Western Georgia, Tbilisi 1999.

Kaukhchishvili (2000): T. Kaukhchishvili, Corpus of Greek Inscriptions Georgia, II, Eastern Georgia, Tbilisi 2000.

Kaukhchishvili (2000a): T. Kaukhchishvili, Corpus of Greek Inscriptions Georgia, Index, Tbilisi 2000.

Kaukhchishvili (2009): T. Kaukhchishvili, Corpus of Greek Inscriptions of Georgia, 3rd revised edition, Tbilisi 2009.

Tonia (2023): Nana Tonia, Orthographic Dictionary of Greek and Roman Proper Names, Tbilisi 2023.

Tribulato (2022): O. Tribulato (ed.), Digital Encyclopedia of Atticism. With the assistance of E. N. Merisio, Venice 2022.

Wyles et al. (2016): R. Wyles / E. Hall (eds.), Women Classical Scholars: Unsealing the Fountain from the Renaissance to Jacqueline de Romilly. Classical Presences, Oxford / New York 2016.

Author Contact Information[17]

Associate Professor, Dr. Irine Darchia
Institute of Classical, Byzantine and Modern Greek Studies
Ivane Javakhishvili Tbilisi State University
I. Chavchavadze Str., 1
0179 Tbilisi Georgia
E-mail: irine.darchia@tsu.ge

17

DIGITAL CLASSICS ONLINE

Hypotheseis, a Database of Named Entities Surrounding Greek Rhetorical Exercises

Camillo Carlo Pellizzari di San Girolamo

Abstract: The rhetorical exercises written in Greek from the Hellenistic to the Byzantine age constitute a wide corpus of texts that in the last two centuries have been the subject of few studies in comparison with many other areas of ancient Greek literature. As a result, there is no comprehensive list of Greek *progymnasmata* and declamations nor an overall study of the topics they cover and of the Named Entities they mention (mythological characters, historical persons, places, events, etc.) has already been published. The relational database *Hypotheseis*[1], founded in March 2024, is a *Wikibase* instance that aims to fill this gap, providing an easy way to describe the Greek rhetorical exercises and the Named Entities to which they are connected as structured data, available in CC0 license; a SPARQL endpoint allows making sophisticated analyses on the collected data. This paper presents the data model of the database, the corpus of texts it intends to catalogue, the work done so far and what could be done in the future.

Progymnasmata and Declamations: an Overview

Definitions and Classification

Progymnasmata and declamations are two types of rhetorical exercises used in Greek-language rhetorical teaching and as self-standing literary works from the first centuries AD to the Palaeologan age (15th century).

A *progymnasma*[2] (προγύμνασμα, 'preliminary exercise') is an exercise that prepares to the declamation. Four ancient manuals about *progymnasmata* survive (Theon, Pseudo-Hermogenes, Aphthonius, Nicholas of Myra);[3] the last three agree in defining 14 types of *progymnasmata*, arranged in a standard order of increasing difficulty: fable, narration, anecdote (*chreia*),[4] maxim, refutation, confirmation, common-place, praise, invective, comparison, *ethopoeia*,[5] description, thesis, introduction of a law.[6]

1 https://hypotheseis.wikibase.cloud/ (last access 11.07.2025).

2 Cf. Berardi (2017); for an extended bibliography, Chiron (2017).

3 Editions of reference: for Theon, Patillon / Bolognesi (1997) with French translation; for Pseudo-Hermogenes and Aphthonius, Patillon (2008) with French translation; for Nicholas of Myra, Felten (1913). All the manuals are translated in English by Kennedy (2003), the first three are translated in Spanish by Reche Martínez (1991).

4 Cf. Hock / O'Neil (1986); Hock / O'Neil (2002); Hock (2012).

5 Cf. Amato / Schamp (2005).

6 The SPARQL query on *Hypotheseis* https://tinyurl.com/2xjd9wg3 (last access 11.07.2025) shows the 14 types with Aphthonius' definitions.

A declamation[7] (μελέτη, 'exercise') is a speech that is imagined to be pronounced by a fictional speaker in a given context. It is the last and most important part of the rhetorical education. Declamations are usually classified by scholars according to their topic: mythological declamations, historical declamations, and *plasmata* (i.e. declamations with stock characters set in an undetermined classicizing city).[8]

Lists

The main lists of surviving *progymnasmata* and declamations by epoch are the following:

- Hunger (1978), 92–120 for the *progymnasmata* and the declamations from the Late Antiquity to the Byzantine age;
- Russell (1983), 3–6 for the surviving declamations up to the Late Antiquity;
- Hock / O'Neil (2002) for all the known anecdotes (*chreiai*), from the Imperial age to the Byzantine age;
- Amato / Ventrella (2005) for all the known *ethopoeiae*, both in Greek and Latin, from the Hellenistic age to the Byzantine age;
- Guast (2023), 2–9 for the declamations of the first three centuries AD.

There are also lists of *progymnasmata* and declamations organized thematically:

- Jacobs (1899) for mythological themes in *progymnasmata*;
- Kohl (1915) for historical themes in Greek and Latin declamations;
- Ureña Bracero (1999), 324–329 – with additions in Ureña Bracero (2005), 95 n. 4 – for Homeric themes in *ethopoeiae*; see also Fernández Delgado (2025), 112–116;
- Gibson (2004) for historical themes in the manuals about *progymnasmata*.

Classification

Progymnasmata and declamations constitute a very diverse corpus, which can be sorted according to various parameters:

- form: whilst extant *progymnasmata* and declamations are mostly in prose, a certain number of *ethopoeiae* in verses survive, both as independent short works and as parts of longer poems, as well as other types of *progymnasmata*, like praises;[9]
- literariness: extant *progymnasmata* and declamations are mostly literary works; nonetheless, there are also some *progymnasmata* composed as school exercises, mainly thanks to papyrus discoveries in Egypt;[10]

7 Cf. Russell (1983), Guast (2023).

8 Sophistopolis is the well-known nickname attributed to it by Russell (1983), 21–39.

9 Cf. Cribiore (2001), 225–230; Agosti (2005).

10 Cf. Cribiore (1996), 259–262 (n° 344–357); Cribiore (2001), 221–230; Agosti (2005), 55; Amato / Ventrella (2005), 223–225.

- tradition: most extant *progymnasmata* and declamations survived in medieval manuscripts, but a relevant number of *progymnasmata* have been discovered in papyri (and tablets);[11]
- completeness: apart from *progymnasmata* and declamations which are entirely or fragmentarily extant, hundreds of themes of *progymnasmata* and declamations are mentioned in other works (rhetorical manuals, commentaries to rhetorical manuals, biographies of rhetoricians etc.).[12]

Other criteria of classification include genre, author, epoch, character(s), and theme(s).

The Database *Hypotheseis*

Reason and Aims

The idea of starting a database of *progymnasmata* and declamations originated from two main needs:

- firstly, a comprehensive overview of Greek rhetorical exercises could allow to understand more deeply these literary genres, their role in rhetorical teaching, their relationship with other genres and their chronological evolution; however, the existing lists are partial, each covering only a part of *progymnasmata* and declamations (defined by genre and/or epoch), and a systematic extraction of themes mentioned in other works is still missing;
- secondly, existing lists of Greek rhetorical exercises are textual publications, so they sort the entries according to one criterion; consequently, sorting the entries according to a different criterion requires a relevant amount of work, and analysing them according to multiple criteria combined together is basically impossible without a tabular reshaping of the data.

Thus, the main aims of the database are, in the long-term, providing a comprehensive database of *progymnasmata* and declamations, both extant and mentioned in other works, and making it possible to analyse them according to the multiple classification criteria previously outlined.

Hypotheseis also aims to make data about *progymnasmata* and declamations FAIR,[13] specifically in the following ways:

- providing both *progymnasmata* and declamations and the Named Entities to which they are connected (authors, characters, places, etc.) with PIDs;
- making the data retrievable in multiple ways (mainly through a web interface and a SPARQL endpoint);
- interlinking all the entities, wherever possible, with *Wikidata*, in order to make it easier to compare the two databases and to make federated SPARQL queries on the two databases;
- releasing the data in CC0 license,[14] in order to allow their widest possible reuse.

11 Cf. Morgan (1998), 198–226; Cribiore (2001), 225–230; Hock / O'Neil (2002), 5–48, 56–66, 94–97.

12 Cf. Jacobs (1899) and Kohl (1915), which extract themes not only from extant texts but also from these sources.

13 Cf. https://www.go-fair.org/fair-principles/ (last access 11.07.2025).

14 https://creativecommons.org/public-domain/cc0/ (last access 11.07.2025).

Software Choice: *Wikibase*

The software chosen to create the database *Hypotheseis* is *Wikibase*,[15] a set of *MediaWiki* extensions allowing collecting structured data. *Hypotheseis* can thus be defined as a *Wikibase* instance, like *Wikidata*. The main reasons of this choice are the following:

- *MediaWiki* and its extensions, including *Wikibase*, are open source software, and they are regularly maintained[16] by a wide community of volunteers, by WMDE, which developed *Wikibase* (and *Wikidata*) from their launch[17] and has currently a dedicated development plan for *Wikibase* (and *Wikidata*),[18] and by the WMF;

- a *Wikibase* instance can have a SPARQL endpoint[19] allowing to query data, besides offering also other ways to retrieve them (the web interface, the special page EntityData,[20] the *MediaWiki* Action API,[21] and the REST API[22]); it is easy to make federated queries with *Wikidata*,[23] provided that the entities described in *Hypotheseis* declare their equivalents in *Wikidata*,[24] and also with other *Wikibase* (especially *Wikibase Cloud*) instances;

- the user interface is intuitive, which makes potentially easy to involve in the process of data curation scholars with little previous experience in digital humanities and databases; additionally, two user-friendly and open source tools are available to perform massive imports of data, i.e. *QuickStatements*[25] and the software *OpenRefine*[26], and it is also possible for more advanced users to program bots in Python using the *Pywikibot* library;[27]

- all edits are tracked in the page histories, so it is easy to reconstruct both the evolution of pages and the contributions of each editor, potentially also creating statistics about them; moreover, mistakes can easily be undone and no damage is permanent (at least considering damages made through the user interface), since also page deletions can be quickly reverted;

- data can be entered in hundreds of languages,[28] thus constructing a multilingual database which could potentially be a good hub for collaboration between scholars from different linguistic backgrounds;

---

15 Cf. https://wikiba.se/ (last access 11.07.2025) and https://www.mediawiki.org/wiki/Wikibase (last access 11.07.2025).

16 Cf. https://www.mediawiki.org/wiki/Version_lifecycle (last access 11.07.2025).

17 Cf. https://meta.wikimedia.org/wiki/Wikidata/Technical_proposal (last access 11.07.2025).

18 https://www.wikidata.org/wiki/Wikidata:Development_plan (last access 11.07.2025).

19 https://hypotheseis.wikibase.cloud/query/ (last access 11.07.2025).

20 https://hypotheseis.wikibase.cloud/wiki/Special:EntityData (last access 11.07.2025).

21 https://hypotheseis.wikibase.cloud/w/api.php (last access 11.07.2025).

22 Cf. https://doc.wikimedia.org/Wikibase/master/php/repo_rest-api_README.html (last access 11.07.2025).

23 https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/Federated_queries (last access 11.07.2025).

24 The equivalence with *Wikidata* is stated through the properties "Wikidata property" (P1) for properties and "Wikidata item" (P2) for items.

25 https://hypotheseis.wikibase.cloud/tools/quickstatements/ (last access 11.07.2025).

26 https://openrefine.org/ (last access 11.07.2025).

27 Cf. https://www.mediawiki.org/wiki/Manual:Pywikibot (last access 11.07.2025).

28 Specifically, in all the languages used in Wikimedia wikis (https://meta.wikimedia.org/wiki/Special:SiteMatrix [last access 11.07.2025]) and, more generally, in all languages which have a Wikimedia language code (https://www.wikidata.org/wiki/Help:Wikimedia_language_codes/lists/all [last access 11.07.2025]).

- the properties used to construct statements are not predefined, but created by users according to their needs, so the data model is highly flexible (cf. below *Data model*);

- all properties can have more than one value on the same item, also when it would be logically or empirically wrong (e.g. two birth dates), because of the high flexibility of the software. Consequently, it is possible to store redundant and obsolete values, besides the best ones, whenever needed (e.g. when important sources have made influential mistakes or when, despite their obsolescence, it is anyway useful to store them for historical purposes). In order to mark the quality of concurring values, it is possible to use ranks,[29] specifically marking the best value(s) as preferred and wrong value(s) as deprecated, whilst the standard rank is normal; SPARQL queries allow selecting only 'truthy' statements (i.e., for each property, all the preferred-ranked values, or, if absent, all the normal-ranked values), or all statements, or only the statements with a specific rank;

- statements can be provided with qualifiers[30] and references:[31] qualifiers are used to provide context on statements, enabling the expression of nuances and details; references provide sources for the data, and the possibility of associating them with each statement, rather than with the entire entity, makes them much easier to verify.

Presently, the two ways available to create a *Wikibase* instance are *Wikibase Suite* and *Wikibase Cloud*:

- *Wikibase Suite*[32] consists of the possibility of downloading *Wikibase* and managing it independently to create a self-hosted *Wikibase* instance; it allows high margins of personalisation (e.g. through the installation of extensions[33]), but it also implies that the creator of the instance is fully in charge of its maintenance, both technically (i.e. software updates) and financially (i.e. server cost);

- *Wikibase Cloud*[34] is a platform hosted by WMDE since 2022,[35] where each user can freely create up to six cloud-hosted *Wikibase* instances; thus, WMDE manages the technical and financial costs of maintaining these instances; the main drawback, in comparison with *Wikibase Suite*, is the limited degree of possible personalisation for each instance.

According to data extracted from *Wikibase World*,[36] a *Wikibase* instance designed to be a census on a voluntary basis of *Wikibase* instances, as of February 16th 2025 the number of existing *Wikibase* instances with properties was 777, most of them (711) hosted in *Wikibase Cloud*.[37]

The growing interest for *Wikibase* in the field of Digital Humanities is proven also by the establishment in September 2024 of the DHwiki Working Group inside DARIAH-EU, which "is a space for

29 Cf. https://www.wikidata.org/wiki/Help:Ranking (last access 11.07.2025).

30 Cf. https://www.wikidata.org/wiki/Help:Qualifiers (last access 11.07.2025).

31 Cf. https://www.wikidata.org/wiki/Help:Sources (last access 11.07.2025).

32 https://www.mediawiki.org/wiki/Wikibase/Docker (last access 11.07.2025).

33 For the list of extensions used in *Hypotheseis* (and in all *Wikibase Cloud* instances), cf. https://hypotheseis.wikibase.cloud/wiki/Special:Version (last access 11.07.2025).

34 https://www.wikibase.cloud/ (last access 11.07.2025).

35 For historical remarks, cf. https://addshore.com/2024/09/2-years-of-wikibase-cloud-by-wmde/ (last access 11.07.2025).

36 https://wikibase.world/ (last access 11.07.2025).

37 Cf. https://addshore.com/2025/02/visualizing-wikibase-ecosystem-using-wikibase-world/ (last access 11.07.2025).

discussion and dissemination of use case experiences and best practices around *Wikibase* and *Wikidata* in a DH context, and for contributions to further developing the *Wikibase* software and related tools".[38]

Hypotheseis was created by the author of this paper on March 27th 2024 using *Wikibase Cloud*[39], since it did not rely on any financing, and the limited personalisation options were considered a minor issue. Its data are released in CC0 license.[40]

Data Entry and Statistics

As of March 13th 2025, the database *Hypotheseis* has been edited only by the author of this paper, with the username Epìdosis;[41] the total number of registered users is five.[42] According to the automatic statistics,[43] 11440 edits have been made and the number of pages is 1753, including 48 properties[44] and 1636 items.[45] The total number of triples is 63832 and they can all be downloaded from the SPARQL endpoint.[46] The most significant steps in data entry, and in general in the evolution of the database, are recorded in Italian in the page "Cronistoria".[47]

Most data have been added using the aforementioned tool *QuickStatements*, which allows making massive additions of data previously collected in a CSV file. Specifically, *QuickStatements* has been used to import massively most of the *hypotheseis*, providing them with basic data; the "tagging" with "character(s)" (P25), "object(s)/concept(s)" (P35), "action(s)/state(s)" (P26), "place" (P27), and "time" (P28) has been done manually in the great majority of cases. Labels have been added systematically in Italian first; since April 2024, they have also been added in English. Most entities (excluding *hypotheseis*, works, and expressions) also have labels in Ancient Greek.

The order in which statements appear in the items is not casual, nor it is based on the order of addition, but it relies on the order of properties established through the page "MediaWiki:Wikibase-SortedProperties", through which it can also be quickly modified (the properties not listed in that page appear after the ones listed in the page, in order of addition).[48] Conversely, qualifiers and references presently appear, for each statement, in the order in which they have been added.[49]

For 523 items the equivalent *Wikidata* item is specified through "Wikidata item" (P2);[50] they represent the majority of items, excluding *hypotheseis* (because they would not fall into *Wikidata*'s notability

38 Cf. https://www.dariah.eu/activities/working-groups/dhwiki/ (last access 11.07.2025) and https://dhwiki.wikibase.cloud/ (last access 11.07.2025).

39 https://hypotheseis.wikibase.cloud/ (last access 11.07.2025).

40 https://hypotheseis.wikibase.cloud/wiki/Project:About (last access 11.07.2025).

41 Cf. userpage: https://hypotheseis.wikibase.cloud/wiki/User:Epìdosis (last access 11.07.2025).

42 Cf. https://hypotheseis.wikibase.cloud/wiki/Special:ListUsers (last access 11.07.2025). New accounts can be requested using https://hypotheseis.wikibase.cloud/wiki/Special:RequestAccount (last access 11.07.2025).

43 https://hypotheseis.wikibase.cloud/wiki/Special:Statistics (last access 11.07.2025).

44 https://hypotheseis.wikibase.cloud/wiki/Special:ListProperties (last access 11.07.2025).

45 https://hypotheseis.wikibase.cloud/wiki/Special:AllPages?namespace=120 (last access 11.07.2025).

46 Cf. SPARQL query: https://tinyurl.com/2xq49u2l (last access 11.07.2025).

47 https://hypotheseis.wikibase.cloud/wiki/Cronistoria (last access 11.07.2025).

48 https://hypotheseis.wikibase.cloud/wiki/MediaWiki:Wikibase-SortedProperties (last access 11.07.2025).

49 Cf. request to order also qualifiers and references in the same way as main statements: https://phabricator.wikimedia.org/T169960 (last access 11.07.2025).

50 Cf. SPARQL query: https://tinyurl.com/24m56wup (last access 11.07.2025).

policy[51]) and expressions (because they are not currently recognised in *Wikidata*'s data model for books, which collapses the two concepts of "expression" and "manifestation" into "edition"[52]).

The works already consulted to extract *hypotheseis* are listed in the page "Opere".[53] As of March 2025, all the extant *progymnasmata* and declamations (including four still unpublished[54]) known to the author of this paper have been added to *Hypotheseis*, with two exceptions: *progymnasmata* in verses and *progymnasmata* discovered in papyri.

Data Model

The definition of the data model was the initial focus of the editing activity in *Hypotheseis*[55] and is presently defined in Italian in the page "Modello dei dati".[56]

The basic unit of the data model is the ὑπόθεσις (*hypothesis*) of a *progymnasma* or a declamation, i.e. the topic of the rhetorical exercise; it is used as a title in most extant *progymnasmata* and declamations,[57] although it is sometimes absent.[58] The choice of using the *hypothesis* as basic unit, instead of the entire *progymnasma* or declamation, is motivated by the attempt to model similarly extant *progymnasmata* and declamations and the ones whose themes are only mentioned in other works, for which therefore only the *hypothesis* is known.[59]

An item about a *hypothesis* has mainly the following statements:

- "instance of" (P4) with one of the four recursive subclasses of "hypothesis" (Q2) as value;[60]
- "text of the hypothesis" (P6) with the ancient Greek text of the *hypothesis* as value; the edition used to extract the text is specified through the qualifier "transcribed from the expression" (P7); it is possible to add multiple values, extracting the *hypothesis* from different editions and marking the value extracted from the reference edition with preferred rank;[61]
- "taken from" (P10) with the work from which the *hypothesis* is taken as value; the qualifiers "citation (hypothesis only)" (P11) and "citation (entire *progymnasma*/declamation)" (P36),

51 https://www.wikidata.org/wiki/Wikidata:Notability (last access 11.07.2025).

52 https://www.wikidata.org/wiki/Wikidata:WikiProject_Books (last access 11.07.2025).

53 https://hypotheseis.wikibase.cloud/wiki/Opere (last access 11.07.2025).

54 Two *progymnasmata* of Konstantinos Akropolites (https://hypotheseis.wikibase.cloud/entity/Q792 [last access 11.07.2025] = https://pinakes.irht.cnrs.fr/notices/oeuvre/16480/ [last access 11.07.2025] and https://hypotheseis.wikibase.cloud/entity/Q793 [last access 11.07.2025] = https://pinakes.irht.cnrs.fr/notices/oeuvre/20274/ [last access 11.07.2025]) and two declamations of Thomas Magistros (https://hypotheseis.wikibase.cloud/entity/Q1461 [last access 11.07.2025] = https://pinakes.irht.cnrs.fr/notices/oeuvre/75/ [last access 11.07.2025] and https://hypotheseis.wikibase.cloud/entity/Q1462 [last access 11.07.2025] = https://pinakes.irht.cnrs.fr/notices/oeuvre/7664/ [last access 11.07.2025]). I am currently working on the critical edition and translation of the two declamations of Thomas Magistros.

55 Cf. first edit: https://hypotheseis.wikibase.cloud/w/index.php?title=Pagina_principale&oldid=1 (last access 11.07.2025).

56 https://hypotheseis.wikibase.cloud/wiki/Modello_dei_dati (last access 11.07.2025).

57 E.g. Lib. Eth. 1 (https://hypotheseis.wikibase.cloud/entity/Q453 [last access 11.07.2025]).

58 E.g. G.Kyp. Prog. 1 (https://hypotheseis.wikibase.cloud/entity/Q729 [last access 11.07.2025]).

59 E.g. Aphth. Prog. p. 35 Rabe = 11.2 Patillon (https://hypotheseis.wikibase.cloud/entity/Q1213 [last access 11.07.2025]).

60 Cf. SPARQL query presenting a scheme of the types of *hypothesis*: https://tinyurl.com/29onqdo5 (last access 11.07.2025).

61 E.g. Aphth. Fab. 8 Hausrath – Hunger = Ps.Nicol. Fab. 1 Walz (https://hypotheseis.wikibase.cloud/entity/Q133 [last access 11.07.2025]).

the second one used only for extant *progymnasmata* and declamations, specify as a string of text the passage of the work where the *hypothesis* is found;

- "literary genre of the hypothesis" (P13) with the literary genre of the *hypothesis* as value; "literary subgenre of the hypothesis" (P47) can also be added with one or more values;[62]

- "character(s)" (P25), "object(s)/concept(s)" (P35), "action(s)/state(s)" (P26), "place" (P27), and "time" (P28) with one or more items as values, to tag the *hypothesis* by topic; as specified in references through the property "determination method" (P34), most of these tags have been assigned on the basis of the *hypothesis* (Q83, "inferred from the hypothesis"); the qualifier "role" (P24) allows giving further details about the role of the value of the statement, e.g. "fictional speaker" (Q101) for characters and "setting place" (Q1344) for places;[63]

- "same hypothesis as" (P19), "comparable with" (P20) and "opposite of" (P30) with one or more *hypotheseis* as values, to connect *hypotheseis* that are thematically related.

The Italian and English label of the *hypotheseis* contain the abbreviated bibliographical reference to the *hypothesis*; the Italian aliases also include other possible bibliographical references to the *hypothesis* (especially in cases of reattributed texts) and a brief summary of the *hypothesis* in Italian.[64]

Works have "instance of" (P4) "work" (Q3), are connected to their author through "author" (P21) and always have one or more abbreviations associated through "abbreviation" (P31). The abbreviations are usually derived from a reference work, specified through the qualifier "reference work" (P17); however, the abbreviation is created *ex novo* if absent in the reference works considered (LSJ and LBG, consulted in their TLG digitisations),[65] and it is readapted if stylistically incoherent with the other abbreviations in the use of dots.[66]

Editions have "instance of" (P4) "expression" (Q4), are connected to the corresponding work through "expression of" (P23) and can have a "transcription URL" (P9: used to link to a transcription of the edition in the TLG) and/or a "digitization URL" (P32: used to link to a downloadable PDF file of the edition in Internet Archive or Google Books).[67] The choice of modelling editions using the term "expression" is based on its definition as given in the conceptual model IFLA LRM (s.v. Expression, LRM-E3):[68]

> "An *expression* is the specific intellectual or artistic form that a work takes each time it is 'realized'. *Expression* encompasses, for example, the specific words, sentences, paragraphs, etc. that result from the realization of a *work* in the form of a text, or the particular sounds, phrasing, etc. resulting from the realization of a musical work. The boundaries of the entity *expression* are defined, however, so as to exclude incidental aspects of physical form, such as typeface and page layout for a text, unless, due to the nature of the work, these are integral to the intellectual or artistic realization of the work as such."

62 E.g. Lib. Decl. 26 (https://hypotheseis.wikibase.cloud/entity/Q1393 [last access 11.07.2025]).

63 E.g. Lib. Decl. 18 (https://hypotheseis.wikibase.cloud/entity/Q1385 [last access 11.07.2025]).

64 E.g. Sev. Eth. 10 Amato = Lib. Eth. 26 (https://hypotheseis.wikibase.cloud/entity/Q250 [last access 11.07.2025]).

65 E.g. G.Kyp. Decl.av. (https://hypotheseis.wikibase.cloud/entity/Q1253 [last access 11.07.2025]).

66 E.g. Pach. Decl. (https://hypotheseis.wikibase.cloud/entity/Q1250 [last access 11.07.2025]).

67 E.g. the text of Pach. Decl. established by Boissonade (https://hypotheseis.wikibase.cloud/entity/Q1280 [last access 11.07.2025]).

68 IFLA LRM 2017, 23.

Editions of rhetorical works, in *Hypotheseis*, are considered as realizations of the respective works in a concatenation of "specific words, sentences, paragraphs" as defined by the critical editor(s), but excluding the "incidental aspects of physical form", since it does not differentiate between expressions as materialized in their printed manifestation and in their different digital manifestations.

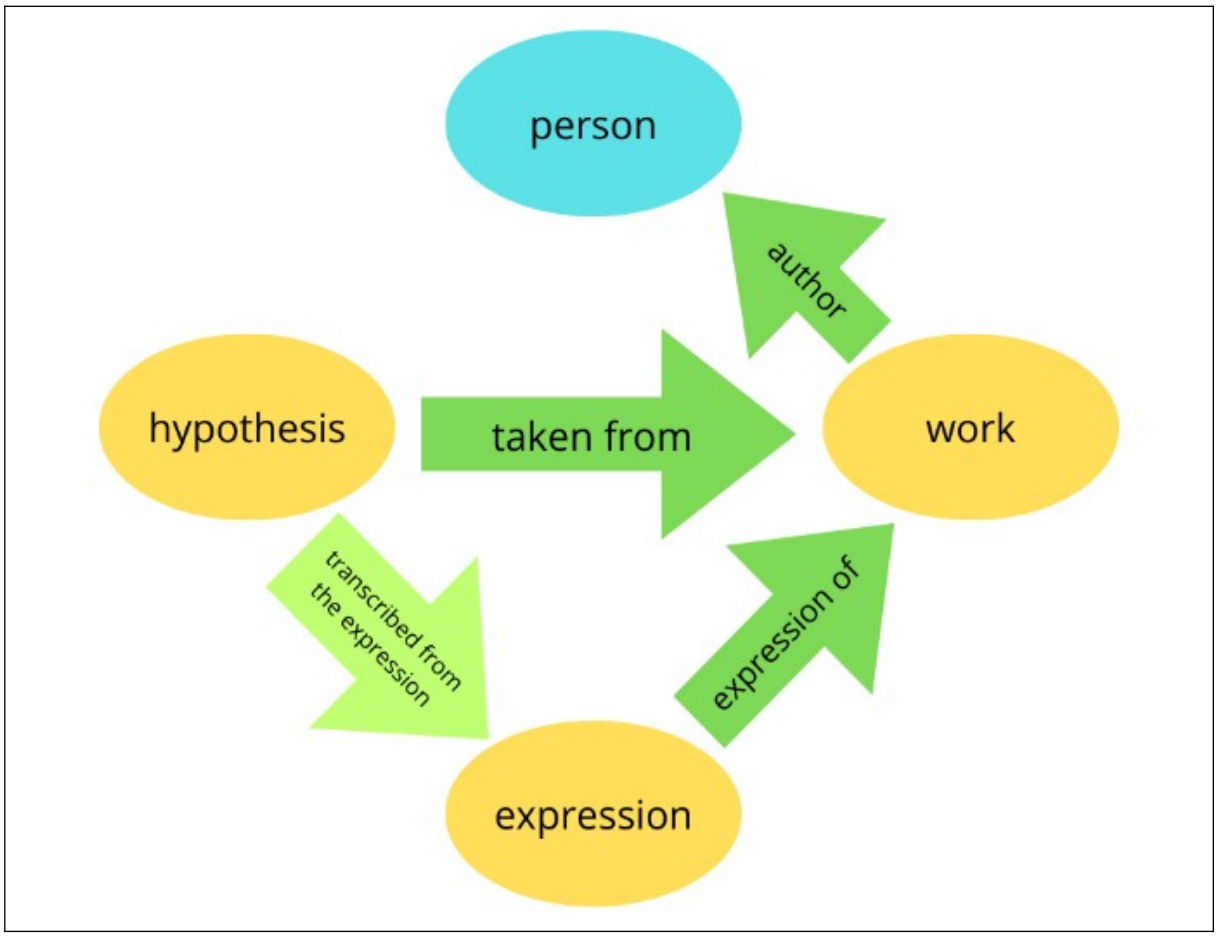


Fig. 1: Schema of the relations between *hypothesis*, work, expression and author.

For all other types of entities, statements are mostly limited to "instance of" (P4) and/or "subclass of" (P5), which are used as in *Wikidata*:[69]

- "instance of" means that the subject of the item is an individual appertaining to the class defined by the object; e.g. "war hero" (Q998) is an instance of "simple human type" (Q59) and "son-war hero" (Q1506) is an instance of "composite human type" (Q591);

- "subclass of" means that the subject of the item is a class appertaining to the class defined by the object; e.g. "son-war hero" (Q1506) is a subclass of "son" (Q1330) and of "war hero" (Q998).

The use of "subclass of" allows the inference of which items appertain, directly or recursively, to a certain class, e.g. all types of birds;[70] it is possible to use these inferences to group *hypotheseis* according to certain criteria, e.g. all *hypotheseis* mentioning birds.[71]

On authors, the use of "epoch of the author" (P48) allows a classification by epoch, and consequently also to filter works and *hypotheseis* by epoch; for the 34 authors added to the database so far, the four

69 Cf. https://www.wikidata.org/wiki/Help:Basic_membership_properties (last access 11.07.2025).

70 Cf. SPARQL query: https://tinyurl.com/243kfqj5 (last access 11.07.2025).

71 Cf. SPARQL query: https://tinyurl.com/2xmdgtue (last access 11.07.2025).

epochs currently in use are “imperial age” (Q1676), “late antiquity” (Q1677), “middle-Byzantine age” (Q1678), and “late-Byzantine age” (Q1679).[72]

Data Analysis and Visualisation

As previously said, the main – although not sole – way to query the structured data in a *Wikibase* instance is its SPARQL endpoint. In order to use the SPARQL endpoint of a *Wikibase Cloud* instance, it is necessary to explicitly define the prefixes used at the start of each query, or to use the full URIs, which is very inconvenient for readability.[73] A collection of precompiled SPARQL queries on *Hypotheseis*, and a list of all usable prefixes, is available in the page “Query” with Italian titles.[74] Results mentioned in this paragraph have been obtained running queries on March 14th 2025.

The main aim of queries on *Hypotheseis* is making analyses of the structured data. A first possible type of analysis is creating lists of entities:

- using one criterion, e.g. all extant *progymnasmata* (451 results)[75] and all extant declamations (120 results);[76]
- using two criteria, e.g. all extant *progymnasmata* and declamations mentioning Achilles (24 results)[77] or war heroes (10 results);[78]
- using more criteria, e.g. all extant late antique *progymnasmata* and declamations regarding the Trojan war (51 results)[79] or all the extant Byzantine *progymnasmata* and declamations with Biblical characters (33 results).[80]

A second possible type of analysis is making statistics on entities; these statistics can be visualised both as tables and as graphs inside the SPARQL endpoint, or data can be exported and used to create different visualisations with external tools. Some meaningful statistics can already be drawn from the entered data about extant *progymnasmata* and declamations (for the completeness of these data, cf. above *Data entry and statistics*), dividing them by author, by epoch and by genre:

- by author: considering *progymnasmata*,[81] the biggest corpus is Libanius (144), followed by Pseudo-Nicholas (111), Nikephoros Basilakes (56), and Aphthonius (54), with a total of 18 authors plus anonymous texts;[82] considering declamations,[83] the biggest corpus is again

72 Cf. SPARQL query: https://tinyurl.com/26ncftsb (last access 11.07.2025).

73 Cf. request to allow defining a complete list of prefixes for SPARQL queries from the dashboard of each *Wikibase Cloud* instance: https://phabricator.wikimedia.org/T335448 (last access 11.07.2025).

74 https://hypotheseis.wikibase.cloud/wiki/Query (last access 11.07.2025).

75 SPARQL query: https://tinyurl.com/2xs5h4gm (last access 11.07.2025).

76 SPARQL query: https://tinyurl.com/238bo7fo (last access 11.07.2025).

77 SPARQL query: https://tinyurl.com/2dafhs45 (last access 11.07.2025).

78 SPARQL query: https://tinyurl.com/24p4t9sp (last access 11.07.2025).

79 SPARQL query: https://tinyurl.com/2bs6bvhk (last access 11.07.2025).

80 SPARQL query: https://tinyurl.com/2dq5qe8w (last access 11.07.2025).

81 Cf. SPARQL query: https://tinyurl.com/237fl5ho (last access 11.07.2025).

82 It should be considered that 18 *progymnasmata* attributed to both Libanius and Pseudo-Nicolas are counted for both authors, as well as 12 *progymnasmata* attributed to both Aphthonius and Pseudo-Nicolas, and 2 *progymnasmata* attributed to both Severus of Alexandria and Libanius; cf. SPARQL query for a list of the 32 double attributions (https://tinyurl.com/24k5hfm9, [last access 11.07.2025]).

83 Cf. SPARQL query: https://tinyurl.com/25gqshu2 (last access 11.07.2025).

Libanius (51), followed by Georgios Pachymeres (13), Choricius and Aelius Aristides (12), with a total of 16 authors; these data can also be visualised as bubble charts;

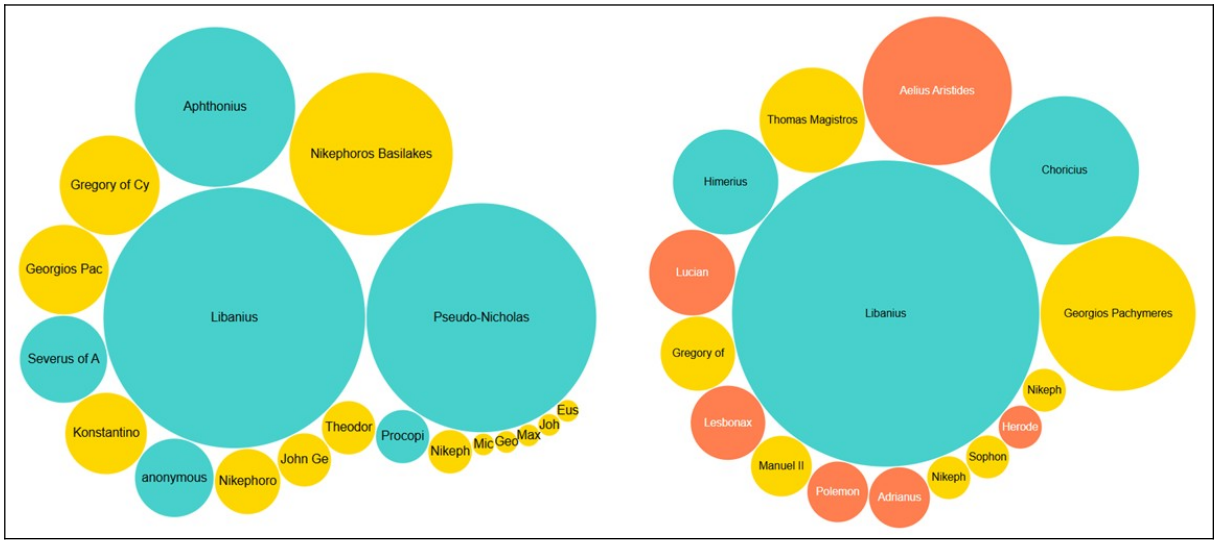


Fig. 2: Bubble charts representing the authors of extant *progymnasmata* (on the left) and of extant declamations (on the right) by number of texts and coloured by epoch.

- by epoch: for both extant *progymnasmata*[84] and extant declamations[85] there is a significant prevalence of late antique texts (313 out of 451, i.e. 69.4%, for *progymnasmata*, and 69 out of 120, i.e. 57.5% for declamations), followed by late-Byzantine texts (73, i.e. 16.2% for *progymnasmata*, and 26, i.e. 21.7%, for declamations); the share of middle-Byzantine texts is relevant for *progymnasmata* (65, i.e. 14.4%) but nearly non-existent for declamations (1, i.e. 0.8%); imperial age declamations complete the overview (24, i.e. 20%);

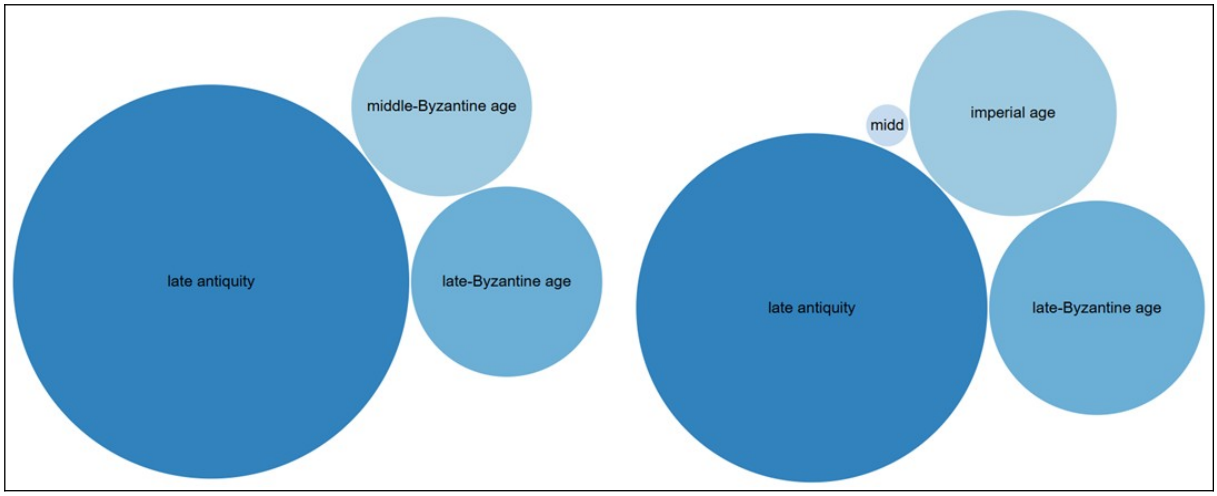


Fig. 3: Bubble charts representing the epochs of extant *progymnasmata* (on the left) and of extant declamations (on the right) by number of texts; the colours are the default ones.

- by genre: for *progymnasmata*,[86] the most common genres for extant texts are by far *ethopoeia* (94 out of 451, i.e. 20.8%), narration (86, i.e. 19.1%), and fable (81, i.e. 18.0%), whilst the least common are thesis (7, i.e. 1.6%) and introduction of a law (4, i.e. 0.9%); for de-

84 Cf. SPARQL query: https://tinyurl.com/28j3tus7 (last access 11.07.2025).

85 Cf. SPARQL query: https://tinyurl.com/26lho4aa (last access 11.07.2025).

86 Cf. SPARQL query: https://tinyurl.com/299svjxs (last access 11.07.2025).

clamations,[87] historical (57, i.e. 47.5%) and stock characters (52, i.e. 43.3%) declamations are far more common than mythological ones (11, i.e. 9.2%).

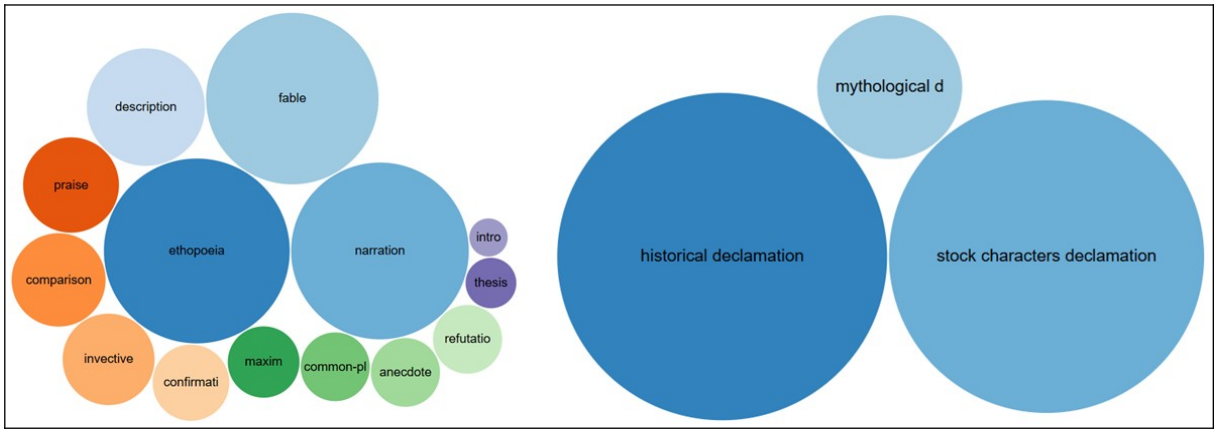


Fig. 4: Bubble charts representing the genres of extant *progymnasmata* (on the left) and of extant declamations (on the right) by number of texts; the colours are the default ones.

An important caveat to the data previously shown regards the attribution of these texts: each *hypothesis* is presently considered according to its belonging to a collection (e.g. all the 27 texts in the corpus of Libanius' *ethopoeiae* are counted as by Libanius), but many judgements of inauthenticity have been expressed by modern scholars for single texts in these collections, and taking these judgements into account could significantly modify these numbers. This is one of the possible future improvements to the database (see below *Possible future developments*).

Other interesting statistics can be drawn about the characters of the extant *progymnasmata* and declamations, e.g. exploring the most common characters in different contexts:

- in extant *progymnasmata* and declamations regarding the Trojan war,[88] Achilles is by far the most frequent character (23 appearances), followed distantly by Odysseus (9), Ajax son of Telamon (8) and Hector (5);

87 Cf. SPARQL query: https://tinyurl.com/279ptags (last access 11.07.2025).

88 Cf. SPARQL query: https://tinyurl.com/26tu2o7q (last access 11.07.2025).

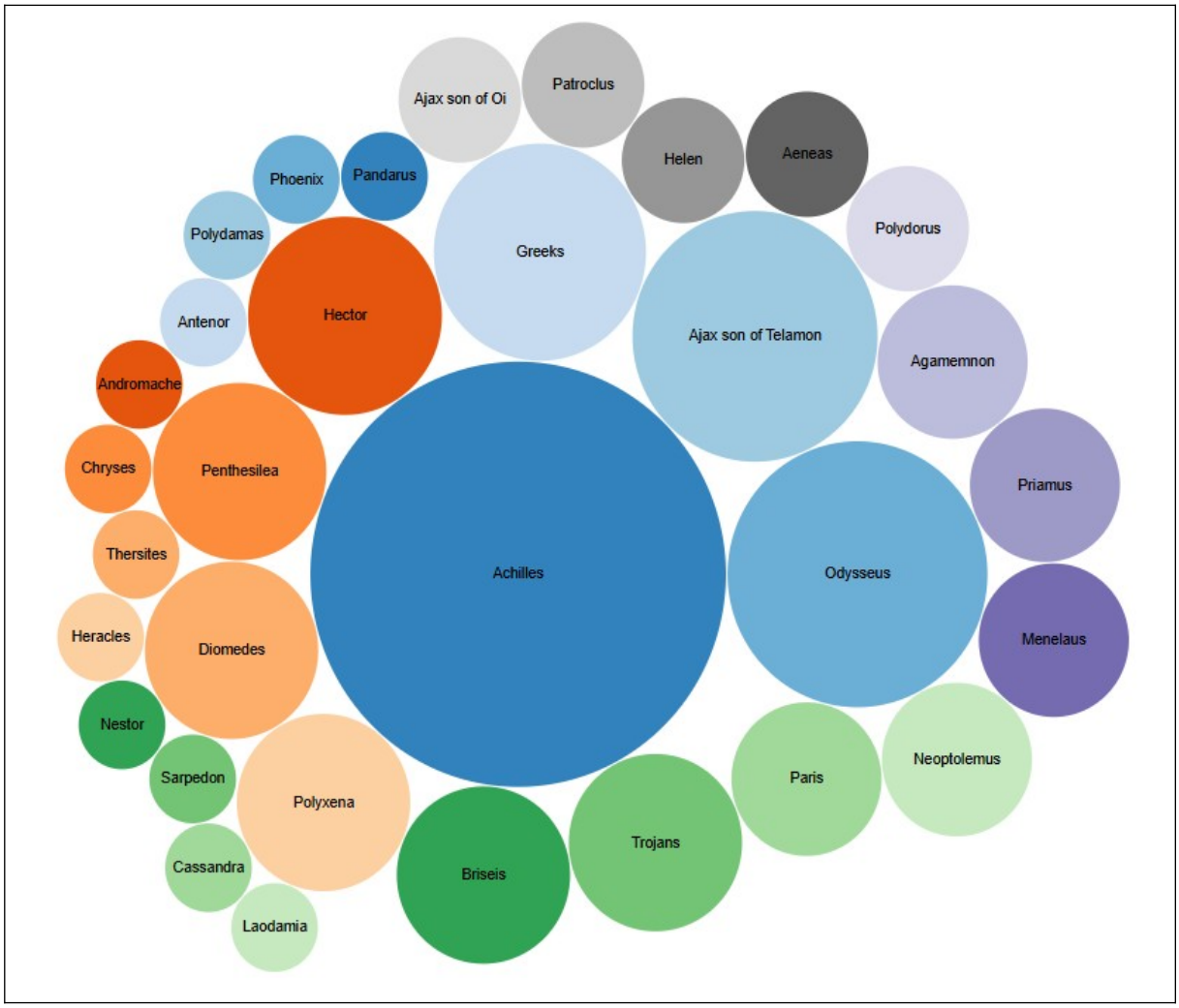


Fig. 5: Bubble chart representing the most frequent characters of extant *progymnasmata* and declamations regarding the Trojan war by number of appearances; the colours are the default ones.

- in extant *ethopoeiae* and declamations,[89] the most common named fictional speakers are Demosthenes (9 appearances), Achilles (8), Heracles (5), and Odysseus, Menelaus and Ajax son of Telamon (4).

89 Cf. SPARQL query: https://tinyurl.com/2cesrkj4 (last access 11.07.2025).

I conclude with an interesting visualisation of the most common values of P26 ("action(s)/state(s)") in all the extant *progymnasmata* and declamations:[90]

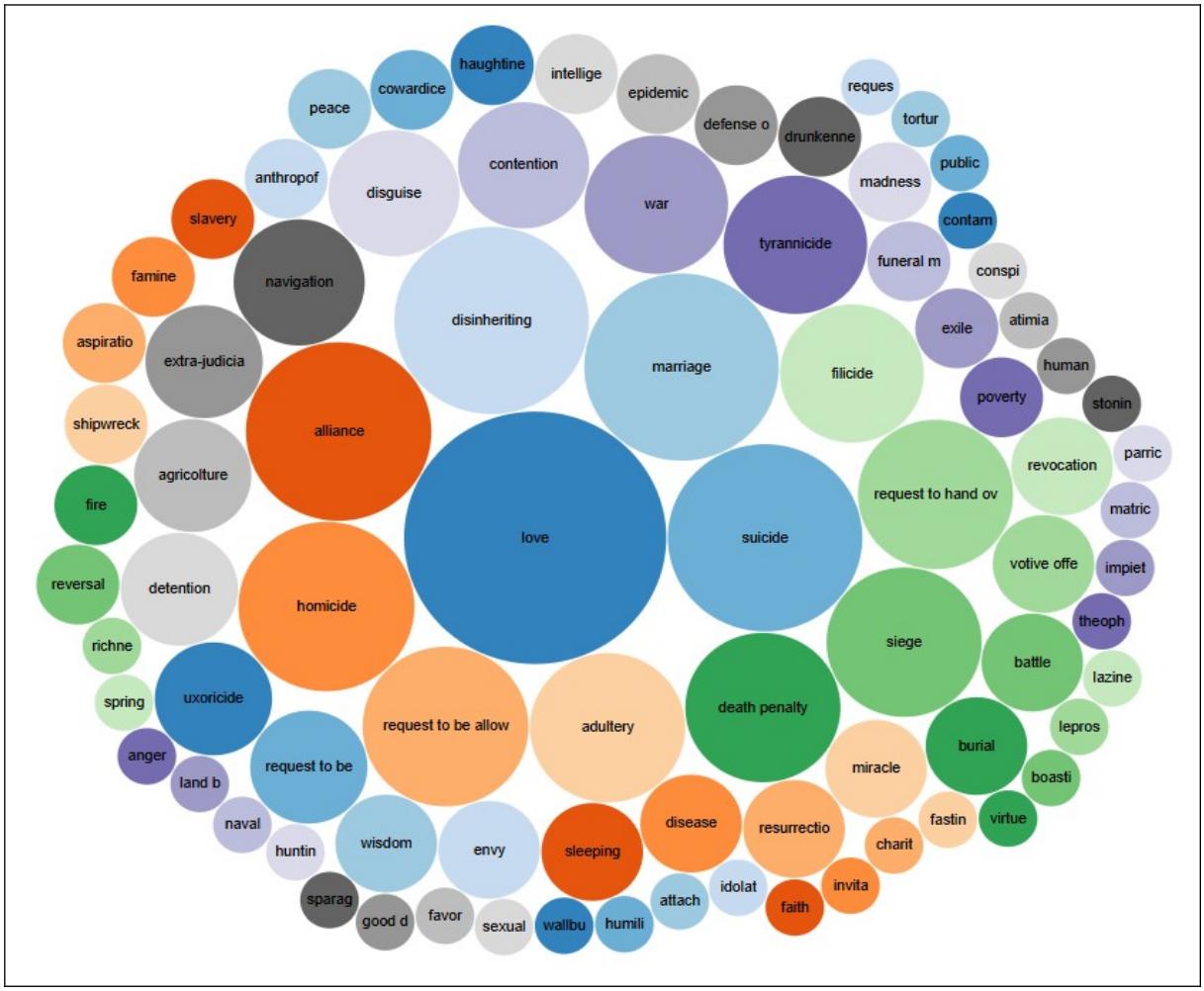


Fig. 6: Bubble chart representing the most frequent actions/conditions in the extant *progymnasmata* and declamations by number of appearances; the colours are the default ones.

The above data, despite the described limitations, already show that SPARQL queries on *Hypotheseis* could be an effective mean to statistically analyse this corpus of texts through the combination of multiple criteria and to visualise data about it.

Finally, SPARQL queries can also be used to find inconsistencies, incoherences, and mistakes in the data, so as to be an effective tool for the routine maintenance of the data quality in the database.

Possible Future Developments

The main path for expanding *Hypotheseis* is surely continuing the additions of *hypotheseis*, starting from the already mentioned groups of *progymnasmata* that are still missing in the database: *progymnasmata* in verses and *progymnasmata* discovered in papyri. Moreover, a systematic research of themes of *progymnasmata* and declamations mentioned in other works is yet to begin, but is one of the long-term aims of the project.

Other possible additions to the database include translating labels in other modern languages besides Italian and English and providing structured data about the editions – as of now, for expressions, only the name of the critical editor(s) is specified – and the translations of the works. The data about the translations could also be added directly to *Wikidata* and retrieved through federated SPARQL queries.

90 Cf. SPARQL query: https://tinyurl.com/2xjgb3hq (last access 11.07.2025).

It would also be possible to improve the data model of *Hypotheseis* to store data about the judgements of authenticity and inauthenticity that have been pronounced about specific *progymnasmata* and declamations, since the authorship of many of them is debated by modern scholars, and manuscripts themselves often show different attributions (see above *Data analysis and visualisation*).

The scope of *Hypotheseis* could also be expanded to include rhetorical exercises in languages other than Greek, especially Latin declamations, which would give interesting possibilities in terms of analyses of common themes. The addition of Armenian *progymnasmata* started in May 2025. The possibility of including also Greek literary works whose literary genre has some resemblances to *progymnasmata* and declamations, e.g. literary letters (Aelian, Alciphron, Philostratus, Aristaenetus, Theophylact Simocatta), could also be considered.

In order to ensure the accuracy and the coherence of the data of the database, it would also be useful to involve more editors, which could check already entered data and discuss more deeply the data model presently used. In fact, *Wikibase* is designed to create collaboratively edited databases, so its current use by one single editor unleashes only a fraction of its potential. When other editors will join the database, it will be a priority to translate documentation pages into English, since most of them are presently available only in Italian;[91] it is already possible to show a textual page in Italian or in English according to the interface language chosen by the user.[92]

As of now, *Wikibase* is not designed to host long texts as structured data, for various reasons. Firstly, the maximum length of labels, descriptions, and aliases, as well as of string and monolingual text values in statements, is 2500 characters. Secondly, inside these fields no markup is visually rendered (e.g. the string text is visualised in this exact way)[93] and it is impossible to store multiline texts.[94] However, it would be possible to store texts in *MediaWiki* textual pages and link them from items;[95] these options could be explored e.g. to store complete translations of rhetorical exercises made by the editors of the database. However, translations of *hypotheseis* only could well be stored directly as values of monolingual text statements, as they would remain within the 2500-character limit (the longest *hypothesis* presently stored in *Hypotheseis* is 1309 characters long,[96] and it is unlikely that longer ones will ever be added).

91 Cf. https://hypotheseis.wikibase.cloud/wiki/Category:Documentazione (last access 11.07.2025) and https://hypotheseis.wikibase.cloud/wiki/Category:Documentation (last access 11.07.2025).

92 Compare the first parts of https://hypotheseis.wikibase.cloud/wiki/Pagina_principale?uselang=it (last access 11.07.2025) and https://hypotheseis.wikibase.cloud/wiki/Pagina_principale?uselang=en (last access 11.07.2025).

93 Cf. request to create "a datatype capable of interpreting XML markup" in *Wikibase*: https://phabricator.wikimedia.org/T372326 (last access 11.07.2025).

94 Cf. request to create a datatype for multiline text in *Wikibase*: https://phabricator.wikimedia.org/T323705 (last access 11.07.2025).

95 To make this linking easier, it would be useful to have the possibility to create sitelinks from items to textual pages of the same *Wikibase* instance; cf. request to add this feature: https://phabricator.wikimedia.org/T330672 (last access 11.07.2025).

96 It is the *hypothesis* of Manuel Decl.: https://hypotheseis.wikibase.cloud/entity/Q1460 (last access 11.07.2025).

Conclusions

Greek *progymnasmata* and declamations are a wide but understudied corpus of texts, and structured data is required in order to apply a wide range of statistical analyses, which would have been significantly difficult or impossible with only the existing textual studies.

The software *Wikibase* has proven to be a good choice for creating databases in the field of Digital Humanities for a wide range of reasons: among the most relevant ones, being open-source and community-managed, giving the opportunity to create a customisable data model, enabling data entry through user-friendly interfaces (both manually and massively) and providing various ways to retrieve data (most notably a SPARQL endpoint). The platform *Wikibase Cloud* allows freely creating up to six cloud-hosted *Wikibase* instances, thus permitting scholars to focus exclusively on data curation, without worrying about costs and technical issues. This platform, in fact, has already been used to create many *Wikibase* instances in the field of classical studies,[97] and is likely to host others in the future.

The database *Hypotheseis*, started on March 27th 2024, aims to collect structured data about Greek *progymnasmata* and declamations, with the long-term target to describe also the themes mentioned in other works (especially rhetorical manuals) and not only extant texts. A data model has been designed to structure these data and has already been experimented by collecting data about most of the extant prose *progymnasmata* and declamations; SPARQL queries show that this data model is effective in allowing multiple kinds of statistical analyses about these texts.

The database could be developed in many possible ways, including cataloguing modern editions and translations of these texts and the scholarly debate about their authenticity, although the main purpose remains the extraction of themes mentioned in rhetorical manuals. However, in order to accomplish these aims, it will be fundamental to involve more editors in data curation, since as of now the database has been curated solely by the undersigned.

97 In 2024, apart from *Hypotheseis*, *DataLib* about Libanius' *Letters* (https://datalib.wikibase.cloud/ [last access 11.07.2025]) and *Greek Metrical Inscriptions* (https://greek-metrical-inscriptions.wikibase.cloud/ [last access 11.07.2025]) have also been created.

List of Abbreviations

API	Application Programming Interface
FAIR	Findable Accessible Interoperable Reusable
CSV	Comma-Separated Values
Fig.	Figure
LBG	Lexikon zur byzantinischen Gräzität
LSJ	Liddell-Scott-Jones
PID	Permanent Identifier
SPARQL	SPARQL Protocol and RDF Query Language
URI	Uniform Resource Identifier
WMDE	Wikimedia Deutschland
WMF	Wikimedia Foundation

Sources

Online Sources

https://hypotheseis.wikibase.cloud/ (last access 11.07.2025).

https://www.wikidata.org/ (last access 11.07.2025).

Digital Corpora

TLG database

Text Editions

Felten (1913): I. Felten (ed.), Nicolai progymnasmata, Leipzig 1913.

Kennedy (2003): G. Kennedy, Progymnasmata: Greek Textbooks of Prose Composition and Rhetoric, Leiden / Boston 2003.

Patillon / Bolognesi (1997): M. Patillon / G. Bolognesi (eds. trans.), Aelius Théon, progymnasmata, Paris 1997.

Patillon (2008): M. Patillon (ed. trans.), Corpus rhetoricum: Anonyme, Préambule à la rhétorique; Aphthonios, progymnasmata; Pseudo-Hermogène, progymnasmata, Paris 2008.

Reche Martínez (1991): M. D. Reche Martínez, Teón. Hermógenes. Aftonio. Ejercicios de retorica, Madrid 1991.

References

Agosti (2005): L'etopea nella poesia greca tardoantica, in: Amato / Schamp (2005), 34–60.

Amato et al. (2015): E. Amato / F. Citti / B. Huelsenbeck (eds.), Law and Ethics in Greek and Roman Declamation, Berlin 2015, https://doi.org/10.1515/9783110401882 (last access 27.12.2025).

Amato / Schamp (2005): E. Amato / J. Schamp (eds.), Ethopoiia: la représentation de caractères entre fiction scolaire et réalité vivante à l'époque impériale et tardive, Salerno 2005.

Amato / Ventrella (2005): E. Amato / G. Ventrella, L'éthopée dans la pratique scolaire et littéraire, in: Amato / Schamp 2005, 213–231.

Berardi (2017): F. Berardi, La retorica degli esercizi preparatori: glossario ragionato dei Progymnásmata, Hildesheim 2017.

Chiron (2017): P. Chiron, Les progymnasmata de l'Antiquité gréco-latine, Lustrum 59 (2017), 7–129.

Cribiore (1996): R. Cribiore, Writing, Teachers, and Students in Graeco-Roman Egypt, Atlanta 1996.

Cribiore (2001): R. Cribiore, Gymnastics of the Mind: Greek Education in Hellenistic and Roman Egypt, Princeton 2001.

Fernández Delgado (2025): J. A. Fernández Delgado, Homero en la enseñanza práctica de la retórica griega, Anuario de Estudios Filológicos 48 (2025), 109–130, https://doi.org/10.17398/2660-7301.48.109 (last access 27.12.2025).

Gibson (2004): C. A. Gibson, Learning Greek History in the Ancient Classroom: The Evidence of the Treatises on progymnasmata, Classical Philology 99/2 (2004), 103–129, https://doi.org/10.1086/423858 (last access 27.12.2025).

Guast (2023): W. Guast, Greek Declamation and the Roman Empire, Cambridge 2023, https://doi.org/10.1017/9781009297158 (last access 27.12.2025).

Hock / O'Neil (1986): R. F. Hock / E. N. O'Neil, The Chreia in Ancient Rhetoric 1: The progymnasmata, Atlanta, 1986.

Hock / O'Neil (2002): R. F. Hock / E. N. O'Neil, The Chreia and Ancient Rhetoric: Classroom Exercises, Atlanta 2002.

Hock (2012): R. F. Hock, The Chreia and Ancient Rhetoric: Commentaries on Aphthonius's progymnasmata, Atlanta 2012.

Hunger (1978): H. Hunger, Die hochsprachliche profane Literatur der Byzantiner, Munich 1978.

IFLA LRM (2017): P. Riva / P. Le Bœuf / M. Žumer, Consolidation Editorial Group of the IFLA FRBR Review Group, IFLA Library Reference Model. A Conceptual Model for Bibliographic Information, Den Haag 2017.

Jacobs (1899): J. Jacobs, De progymnasmaticorum studiis mythographicis, Marpurgi Cattorum 1899.

Kohl (1915): R. Kohl, De scholasticarum declamationum argumentis ex historia petitis, Paderborn 1915.

Morgan (1998): T. Morgan, Literate Education in the Hellenistic and Roman Worlds, Cambridge 1998.

Russell (1983): D. A. Russell, Greek Declamation, Cambridge 1983, https://doi.org/10.1017/CBO9780511897887 (last access 27.12.2025).

Ureña Bracero (1999): J. Ureña Bracero, Homero en la formación retórico-escolar griega: etopeyas con tema del ciclo troyano, Revista de Lingüística y Filología Clásica 67/2 (1999), 315–338, https://doi.org/10.3989/emerita.1999.v67.i2.178 (last access 27.12.2025).

Ureña Bracero (2005): J. Ureña Bracero, El uso de fuentes literarias, recursos retóricos y técnicas de composición en etopeyas sobre un mismo tema, in: Amato / Schamp (2005), 93–111.

Figure References

Fig. 1 Camillo Carlo Pellizzari di San Girolamo.

Fig. 2 On the left: from the SPARQL query https://tinyurl.com/262we83r (last access 11.07.2025); on the right: from the SPARQL query https://tinyurl.com/22p6y32l (last access 11.07.2025).

Fig. 3 On the left: from the SPARQL query https://tinyurl.com/259x5c6n (last access 11.07.2025); on the right: from the SPARQL query https://tinyurl.com/25neodcy (last access 11.07.2025).

Fig. 4 On the left: from the SPARQL query https://tinyurl.com/27f732ec (last access 11.07.2025); on the right: from the SPARQL query https://tinyurl.com/27886ong (last access 11.07.2025).

Fig. 5 From the SPARQL query https://tinyurl.com/2b3976m2 (last access 11.07.2025).

Fig. 6 From the SPARQL query https://tinyurl.com/2db9k66r (last access 11.07.2025).

Author Contact Information[98]

Camillo Carlo Pellizzari di San Girolamo
Scuola Normale Superiore
Piazza dei Cavalieri 7
56126 Pisa
E-mail: camillo.pellizzaridisangirolamo@sns.it

98

DIGITAL CLASSICS ONLINE

Annotating Named Entities in the Trilingual Inscription at Ka'ba-ye Zartošt (ŠKZ)

Farnoosh Shamsian, Monica Berti

Abstract: This study examines proper names in the trilingual inscription of Shapur I at Ka'ba-ye Zartošt (ŠKZ) located in Naqsh-e Rustam, Fars province, Iran. We introduce a corpus of Greek, Middle Persian, and Parthian versions of the inscription aligned at both sentence and word levels, using the Ugarit translation alignment tool. Through manual extraction and categorization, nearly 400 Named Entities (i.e., proper names) were identified and classified as persons (PER), locations (LOC), or location derivatives (LOCderiv). The paper addresses methodological challenges encountered during the alignment of the text, as well as the extraction and classification of Named Entities, including ambiguities in determining proper names, variations in how some names have been recorded across different versions, and complexities in maintaining consistency in categorizing names across various languages. Additionally, we highlight the value of the aligned corpus as a lexicographical resource beyond Named Entity annotation. All datasets, including the aligned versions of the text and the extracted Named Entities, are openly accessible via GitHub and Zenodo to provide a foundation for further historical and computational research. Lastly, we explore the possibility of adding further annotation layers and linking the corpus to other datasets.

Introduction

The inscription of Shapur I at the Ka'ba-ye Zartošt, also referred to as ŠKZ or *Res Gestae Divi Saporis*, is a trilingual inscription carved on Ka'ba-ye Zartošt, an ancient building located at Naqsh-e Rustam, near Persepolis, in today's Fars province, Iran. The inscription is in three versions: Middle Persian, Parthian, and Greek. Though there are some differences, the content of the three versions is comparable. The Parthian version, in 30 lines, and the Middle Persian version, in 35 lines, are both carved in scripts explicitly used for royal inscriptions. The Greek version consists of 70 lines and was written in a late imperial script.[1] The Greek and Parthian versions are better preserved than the Middle Persian version, which is partially damaged.

The inscription was created by Shapur I (240–270 AD), the second Sasanian king of Persia in the 3rd century A.D. In the inscriptions, Shapur is identified as "King of kings of the Iranians and non-Iranians". In addition to listing the extensive territories he ruled, it outlines the details of his military triumphs, including his defeat and killing of Gordian III, his negotiations with Philip the Arab, the capture of Emperor Valerian, and his conquest of thirty-six Roman cities. The text then recounts the Zoroastrian sacred fires and religious sacrifices that Shapur supplied, and documents administrative structure and noble courtiers from the reigns of Papag, Ardashir, and Shapur I.[2]

1 Huyse (1999), 9–10.

2 Daryaee (2018).

The ŠKZ inscription contains a rich repository of nearly 400 proper names of both individuals and geographical locations. This paper introduces a parallel corpus of word-level alignment of the ŠKZ inscription, accompanied by a dataset of manually extracted Named Entities. Moreover, we will discuss the workflow for preparing the corpus and report on the challenges of classifying and annotating Named Entities within the inscription.

Building upon established philological scholarship, the contribution of this paper lies not in new textual interpretations but rather in the digital humanities approach to this well-studied inscription. All data, including the parallel corpus and the Named Entity dataset, are openly accessible on both GitHub[3] and Zenodo.[4]

Previous Projects and Studies

Erich F. Schmidt uncovered the inscription at Ka'ba-ye Zartošt in 1939.[5] In *Acta Iranica* 18, Michael Back provides a thorough explanation of the inscriptions' Middle Persian phonology and orthography, together with a discussion of the historical contexts, an etymological index, and the text of the inscription. Within the text, he offers an aligned version of ŠKZ in a tabular format, along with other inscriptions.[6] The Greek text is included only in its diplomatic version. The Middle Persian and Parthian versions are given solely in transliteration with no accompanying transcription. In addition to the three columns for the three versions of the inscription, two columns are also designated for the interpretative rendering of the Middle Persian and Parthian versions. The columns, however, do not provide exact word-level alignment. The word order of all versions has been kept unchanged, and in many cases where the word order differs between the versions, there is no semantic relevance between the words of one row. To give a clear example, in line 51 of the inscription (35/29/69 according to Back), the Greek phrase "εἰς βοήθιαν τῶν θεῶν" is in parallel with the Parthian "pty y'ztn 'dywrpy". The Greek word "βοήθιαν" is in the same row as "y'ztn" while "τῶν θεῶν" is parallel to "'dywrpy".[7] However, if the alignment were based on semantic relevance, the opposite would be true, which would require changing the word order in one of the versions.

The most notable study of this inscription is the work of Philip Huyse in two volumes as part of the *Corpus Inscriptionum Iranicarum*.[8] Published in 1999, the first volume provides a critical edition of the ŠKZ inscription, which includes all three versions aligned at the line level.[9] The Parthian and Middle Persian versions are given in both transliteration and transcription. Additionally, Huyse provides a separate German translation for each version of every line, offering the reader a reliable framework for comparing variants between versions of the inscription. The second volume offers a comprehensive commentary, detailed analysis of the phonology and morphology of the Greek version, an index, and images of the inscription. The extensive information on proper names and Greek recordings of them has been essential for this study.

The parallel corpus presented in this paper is compiled from several sources. The Parthian version is taken from Jake Nabel's digital resource,[10] which is based on the edition of Huyse. The Greek text is taken from the digital epigraphy collection of the Packard Humanities Institute, which uses the edition

3 https://github.com/farnoosh-shamsian/SKZ (last access 26.06.2025).

4 https://zenodo.org/records/15050878 (last access 26.06.2025).

5 Schmidt (1970), 39.

6 Back (1978), 284–371.

7 Back (1978), 369.

8 https://www.iranicaonline.org/articles/corpus-inscriptionum-iranicarum-c (last access 26.06.2025).

9 Huyse (1999).

10 http://parthiansources.com (last access 26.06.2025).

by Canali De Rossi. We added the Middle Persian version ourselves based on the edition of Huyse. Since the original line numbers in different versions of the inscription do not align, we have used paragraph numbers from Huyse's edition as the foundation for aligning the three versions of the inscription and for all references throughout this paper.

As for attempts at Named Entity annotation of the ŠKZ inscription, the results are inconsistent. The project ToposText provides automated annotation of the Greek version; however, the results are not entirely accurate when it comes to name disambiguation. For instance, distinguishing between different individuals named "Peroz" in the text presents a challenge, given that this name appears multiple times referring to various persons and also has the meaning "victor".[11] For instance, the word "Peroz" in "Π]η[ρ]ωσ-σαβουρ" in line 4 of the Greek text is mistakenly linked to the Wikidata entry Q310233 for "Peroz", the 18th Sasanian king.[12] However, in the context of the inscription, "Pērōz-Šābuhr" is a place name and also bears the meaning "Shapur the victor".[13] The inscription is also registered in the Trismegistos database with the identifier text/818077.[14]

Data Preparation and Translation Alignment

The alignment process began with the segmentation of all three versions – Parthian, Middle Persian, and Greek – into 51 parallel sentences, following Huyse's edition. To facilitate accurate tokenization, editorial signs in the text, such as square brackets indicating damaged sections in the inscription, were removed to prepare the text for word-level alignment. These signs, however, are present in the parallel corpus, which is aligned at the sentence level. The tokenization and word-level alignment were then performed using the Ugarit translation alignment tool developed at Leipzig University by Tariq Yousef.[15] The alignment of the versions of the inscription in Ugarit was the foundation for subsequent Named Entity extraction (fig. 1).

Fig. 1: Trilingual alignment of the first lines of ŠKZ in Ugarit.

Given the limited size of the corpus, it was not feasible to develop detailed guidelines for aligning the texts. Therefore, we adopted an adaptable yet consistent approach to alignment that adhered to the same principles outlined in the alignment guidelines for Greek-Persian alignment projects, with some

11 https://topostext.org/work/561 (last access 26.06.2025).

12 https://www.wikidata.org/wiki/Q310233 (last access 26.06.2025).

13 Huyse (1999), 24.

14 https://www.trismegistos.org/text/818077 (last access 26.06.2025).

15 Yousef et al. (2022).

flexibility to accommodate the specific characteristics of the languages involved in this corpus.[16] The Greek version of the inscription was used as the first text in the alignment process; however, this was purely for practical purposes and does not reflect a hierarchical prioritization among the three versions. Moreover, we emphasize that paired words or phrases in the alignments represent practical correspondences across the three languages and are not meant to be taken as equivalents. Here we distinguish between equivalence and functional translation equivalence, recognizing that the latter is not a precise semantic equivalent but rather a practical matching of corresponding textual elements.[17] For example, in tab. 1, we see that a Greek word in the genitive case, "θεῶν", has been aligned with the Parthian and Middle Persian "až yazdān" and "az yazdān". While this alignment reflects a translation choice in this context, it should not be taken as establishing "az" as a full equivalent of the Greek genitive. The Greek genitive can express a wide range of relations beyond origin, and "az" itself has multiple meanings, many of which do not correspond to the broader semantic scope of the genitive case.

Although the alignments did not follow a strict set of guidelines, our goal was to maintain consistency in alignment decisions, ensuring that words and sentences with comparable syntax are aligned similarly. One significant challenge in maintaining alignment consistency involves handling grammatical features that exist in only one of the languages of the inscription. For instance, in the phrase 'ἐκ γένους θεῶν" from line 1, we needed to determine how to align each Greek word with the Middle Persian phrase "kē čihr az yazdān" and the Parthian "kē čihr až yazdān". When adpositions or particles exhibited a similar grammatical relationship to that expressed by Greek cases, we aligned Greek words with their functional counterparts in Parthian or Middle Persian. This decision was then applied consistently throughout the entire corpus to ensure coherence across all three versions.

Greek	Parthian	Middle Persian	Type of pairing
ἐκ γένους	čihr	čihr	N-1-1
θεῶν	až yazdān	az yazdān	1-N-N
Ø	kē	kē	Ø-1-1

Tab. 1: Alignment pairs of a phrase in Line 1 of ŠKZ across three versions.

In cases where a word or a phrase is missing an equivalent in one version, it is aligned in the other two versions. In one example at line 10, the word "Wardyāz" is missing in the Greek but is aligned in the Middle Persian versions of the text:

> Greek :"ὅσα ἐπ' αὐτὴν ἔθνη καὶ περίχωροι ἦσαν, πάντα ἐκαύσαμεν καὶ ἠρημώσαμεν καὶ ἐκρατήσαμεν"
>
> Parthian: "čē abar Asūriyā šahr parβēr būd, hamag ādurwaxt, awērān ud wardyāz kerd"
>
> Middle Persian: "čē abar Asūriyā šahr parwār būd hamag ādursōxt ud awērān ud wardyāz kerd"

The alignment of function words – particularly particles and prepositions – presented occasional challenges requiring case-by-case decisions based on editorial commentaries. For example, at line 16 of the inscription, we encounter a situation where the adposition "pad" in both Parthian and Middle Persian is aligned with a Greek genitive case:

16 Shamsian (2023).

17 Munday (2016), 81.

Greek: “καὶ τῆς Καππαδοκίας Σάταλα πόλιν σὺν τῇ περιχώρῳ”

Parthian: “pad Kap<p>ōdakiyā: Sātal šahrestān aδ parβēr hamgōs”

Middle Persian: “pad Kap<p>ōdakiyā Sātal šahrestān az parwār hammis”

Greek	Parthian	Middle Persian	Type of pairing
καὶ	Ø	Ø	1-Ø-Ø
τῆς Καππαδοκίας	pad Kap<p>ōdakiyā	pad Kap<p>ōdakiyā	N-N-N
Σάταλα	Sātal	Sātal	1-1-1
πόλιν	šahrestān	šahrestān	1-1-1
σὺν	aδ	az	1-1-1
τῇ περιχώρῳ	parβēr hamgōs	parwār hammis	N-N-N

Tab. 2: Alignment pairs of a phrase in Line 1 of ŠKZ across three versions.

This decision is based on the German translation of Huyse in which the Middle Persian and Parthian are translated as “in Kappadokien” and the Greek as “von (=in) Kappadokien”.[18]

Following the alignment process, we extracted all alignment pairs from Ugarit in CSV format. Named Entities were then manually identified and annotated throughout the corpus. The Named Entities are categorized with three tags used in computational linguistics: PER for personal names, LOC, and LOCderiv for places and derivatives.

Although we did not align words without equivalents in two of the three languages and they are therefore absent from the alignment pairs dataset, we ensured to extract all Named Entities. For instance, “Ray” in line 50, which appears only in the Middle Persian text, is included in the named entity dataset; however, it does not appear in the alignment pairs.

Manual Extraction of Named Entities

Named Entity Recognition (NER) is a task of information extraction that involves finding mentions of Named Entities in a text and classifying their types corresponding to proper names and quantities of interest, such as people, places, organizations, time expressions, monetary amounts, and percentages. NER is a relatively mature technology in Natural Language Processing (NLP), whose goal is to extract semantic content from texts by acquiring structured data from unstructured information. NER is also showing a great interest from scholars working on historical languages, although in these cases, this technique presents significant challenges, if we consider the complexities of past languages and the fact that we don’t have new data from native speakers.[19]

The primary reasons for extracting Named Entities from historical sources are centered on their significance for textual analyses involving research in onomastics, prosopography, and historical geography. Moreover, digital resources today require more data in machine-readable formats. In this regard, Named Entities are significant because they can be extracted from our corpora and function as

18 Huyse (1999), 32.

19 Berti (2021), 398 with further bibliographic references.

anchors in the text for further linguistic analyses.[20] We therefore decided to extract Named Entities from the inscription of Shapur I, which is rich in proper names across the three linguistic versions of the text, to create new digital data and use it to present methodological challenges for future data extraction and annotation of proper names in historical inscriptions.

The process for extracting Named Entities from the inscription began with retrieving the alignment pairs of word-level alignments for all paragraphs in Ugarit. Then, the Named Entities within the text were identified and manually extracted. This process, while seemingly straightforward, required careful consideration of what constitutes a proper name.

To maintain consistency with our source editions, we have adopted the convention of treating capitalized words as proper names, as per the edition of Huyse. We also emphasize several complexities in the identification of Named Entities across the three versions of the inscription. Ambiguities can occur in several contexts: within a single language version, certain words fall somewhere in between proper names and other lexical categories. Furthermore, some Named Entities show significant variation in their representation across different versions of the inscription or are documented in multiple ways throughout the text. We also note that the capitalization of certain words varies between different scholarly editions. One instance is the word "Μασδαασνης" in the first line of the inscription, which is capitalized in PHI,[21] but not in the Greek version available in the edition of Huyse.[22]

To prepare the Named Entity dataset, we excluded articles and other enclitics from the alignment pairs, since they do not constitute part of the names themselves. It is also worth noting that we have not lemmatized the Named Entities in our dataset. This decision was largely driven by the ambiguity in determining the lemma for some names in the Greek version. Without sufficient comparative data on the morphological forms, lemmatization would have required an additional layer of interpretation. All Greek names in the dataset are preserved in the original morphological form within the text. References to the line numbers are provided for each occurrence, enabling a clearer understanding of the morphology within the syntactic and semantic context of the Greek sentence.

The categorization of Named Entities across the three languages occasionally presents classification challenges due to linguistic differences. A notable example occurs in line 38, where a geographical reference is categorized as a location (LOC) in both Parthian and Middle Persian, while the Greek version employs a plural adjective derived from a place name, which should be annotated as an instance of LOCderiv. In such cases, both categories are used as illustrated in tab. 3.

Greek	Parthian	Middle Persian	Line Number	Wikidata	Categorization
Μησανηνῶν	Mēšān	Mēšān	Line 38	Q3843570	LOC, LOCderiv

Tab. 3: Different recordings of the toponym "Meshan" in line 38 of ŠKZ.

The ŠKZ inscription presents unique challenges for Named Entity annotation beyond the typical disambiguation issues. While it is easier to classify the named entities in the Middle Persian and Parthian versions, the Greek text contains numerous ambiguous instances that fall into a gray area of classification. There are considerable examples of Persian words and terms transcribed into Greek, particularly in royal titles and official designations. In these cases, the Greek text preserves the Middle Persian word through transcription rather than translation. One instance is the word "Πιτιξιγαν" in line 49 of the inscription. While the Middle Persian and Parthian versions of this word are not capitalized in Huyse's edition and are written as "bidaxšgān", the Greek version is capitalized. "Bidaxš" is an Ira-

20 Berti (2019).

21 https://epigraphy.packhum.org/text/314697 (last access 26.06.2025).

22 Huyse (1999), 22.

nian title of high position that has also been attested in different languages.[23] The following is how the phrase that contains the word is translated to German by Huyse:[24]

Greek: Ἀρταξάρου Πιτιξιγαν", Translation: "Ardašīr, (den Sohn) des Bidaxš ('Vizekönigs')"

Parthian: "Ardašīr bidaxšgān", Translation: "Ardašīr, den Sohn des Vizekönigs"

Middle Persian: "Ardašīr ī bidaxšgān", Translation: "Ardašīr, den Sohn des Vizekönigs"

In instances similar to "Πιτιξιγαν," where one version is assumed to be a named entity due to capitalization, we have included it in our Named Entity dataset. When none of the versions include a capitalized word, similar to the case of "βιδιξ", it only appears in the alignment pair dataset.

Greek	Parthian	Middle Persian	Line Number
βιδιξ	bidaxš	bidaxš	Line 42
πιτιάξου	bidaxš	bidaxš	Line 45
πιτυάξου	bidaxš	bidaxš	Line 47
Πιτιξιγαν	bidaxšgān	bidaxšgān	Line 49

Tab. 4: Different recordings of the word "bidaxš" and its related terms in the Greek version of ŠKZ.

It is important to note that the alignment data can be employed for purposes beyond extracting Named Entities. The trilingual alignment itself serves as a lexicographical resource, providing valuable information for understanding certain words. One example that demonstrates the application of the alignments, as well as the complexity of the Greek version, is the word "dastgerd", which appears in the Greek text both as a transcription and as a translation and has been the subject of extensive scholarly debate.[25] Instances of this word are found in the alignment data but not in the Named Entity dataset, in line with our general approach of excluding uncapitalized terms from Named Entities. The word "dastgerd" has been defined as "estate",[26] "landed estate",[27] and "mansion"[28] in different lexicons. Recent studies by Jam[29] (in Persian) and Panaino[30] offer diachronic investigations of this term through a comparative analysis of attestations in the ŠKZ inscription alongside parallel occurrences in various texts in Arabic, Syriac, Armenian, and multiple Iranian languages. While the focus of this paper is on Named Entities, such philological studies underscore the importance of parallel corpora for terminological research.

23 Sundermann (1989).

24 Huyse (1999), 61.

25 Skalmowski (1993); Gignoux (1994); Dhabhar (1930).

26 Mackenzie (1971), 25.

27 Nyberg (1974), 59.

28 Durkin-Meisterernst (2004), 142.

29 Jam (2019).

30 Panaino (2022).

Greek	Parthian	Middle Persian	Line Number
κτίσματα	dastgerd	dastgerd	Line 30
δαστικερτας	dastgerd	dastgerd	Line 32
δαστικιρτ	dastgerd	dastgerd	Line 44
κτίσμα	dastgerd	Ø	Line 51
δαστικιρτην	dastgerd	Ø	Line 51

Tab. 5: Different recordings of the word "dastgerd" in the Greek version of ŠKZ.

Conclusions and Future Work

This paper introduces a word-aligned parallel corpus and a manually annotated Named Entity dataset for the trilingual inscription of Shapur I at Ka'ba-ye Zartošt (ŠKZ). The dataset, comprising nearly 400 Named Entities within the inscription, contains information on proper names and geographical locations attested across the three versions of ŠKZ in Greek, Middle Persian, and Parthian. We have described the methods and tools used in preparing the corpus and alignments, as well as the identification, extraction, and categorization of Named Entities in the texts. Moreover, we have discussed various challenges inherent in extracting Named Entities from a multilingual corpus, including cases in which a proper name is recorded differently in the same language or across languages, as well as cases where classification decisions are complicated.

The corpus, alignment data, and Named Entity dataset are openly accessible through GitHub and Zenodo. By making our dataset openly available, we hope to provide a foundation for further historical and computational research on this inscription. Future enhancements to this work include linking each Named Entity to corresponding Wikidata entries, adding diplomatic versions, transliterations, and original scripts for the Iranian languages, and incorporating line references based on Back's edition of the inscription, which would facilitate further comparison and documentation. Furthermore, adding contextual information extracted from Huyse's commentary and other sources would facilitate linking to external resources. Other steps include visualization of the geographical data and publishing the dataset alongside high-resolution photographs on digital platforms.

To conclude, we present all the Named Entities extracted from the inscription in its three languages: Middle Persian, Parthian, and Greek. The table is sorted by the Middle Persian version of each name as the primary column, followed by the Parthian and then the Greek version. Where multiple variations of a name appear in the Greek text, these are consolidated into a single cell rather than given separate rows; all attested Greek variants are included in sequence. The table thus serves as a concise reference for examining the correspondence of personal and place names between Middle Persian, Parthian, and Greek, including the internal variation found within the Greek text. More extensive versions of the datasets are available on Zenodo and GitHub.[31]

31 Data and paper have been prepared and written by Farnoosh Shamsian. Monica Berti has contributed to the selection of data, annotation of Named Entities, and discussion of related paragraphs of the paper.

Middle Persian	Parthian	Greek	Line Number
Abaršahr	Abaršahr	ἀνωτάτω ἔθνη	Line 3
Abrēnag	Abrēnag	Ἀβρηναχ	Line 41
Abursām Šabuhr	Abursām Šābuhr	Ἀβουρσαμ-σαβωρ	Line 48
Abursān	Abursām	Ἀβουρσαμ	Line 42
Adāniyā	Adāniyā	Ἄδανα	Line 25
Ādur-Anāhīd	Ādur-Anāhīd	Ἀδουρ-αναιδ	Line 33, 36
Ādurbāyagān	Ādurbādegān	Ἀδουρβαδηνήν	Line 2
Afrikē	Afrikiyā	Λυσιτανίας	Line 19
Alānān	Alānān	Ἀλανῶν	Line 2
Alexsandariyā	Alexsandariyā	Ἀλεξάνδριαν, Ἀλεξάνδριαν	Line 14, 24
Ānāt	Ānăt	Ἀναθαν	Line 11
Anazarbos	Anazarbos	Ἀγρίππαν	Line 26
Andēgān	Andēgān	Ἰνδηγαν, Ἀνδηγαν	Line 42, 46
Andiyōk	Andiyŏk	Ἀντιόχιαν	Line 13, 27
Anērān	Anērān	Ἀναριανῶν	Line 1, 30
Anīmūrīn	Anīmūrīn	Ἀνεμοῦριν	Line 27
Anōšag	Anōšag	Ἀνωσακ	Line 37
Apōmiyā	Apōmiyā	Ἀπαμίαν	Line 12
Ar<r>ān	Ardān	Ἀλβανίαν	Line 2
Arabiyā	Arabiyā	Ἀραβίας	Line 21
Arbāyestān	Arbāyestān	Ἀραβίαν	Line 2
Ardašīr	Ardašīr	Ἀρταξάρου, Ἀρταξάρου, Ἀρταξάρου, Ἀρταξιρ, Ἀρταξάρου, Ἀρταξάρου, Ἀρταξιρ, Ἀρταξάρου, Ἀρταξιρ, Ἀρταξιρ, Ἀρταξάρου, Ἀρταξάρου, Ἀρταξάρου	Line 1, 36, 41, 41, 41, 41, 42, 44, 45, 46, 48, 49, 50

Ardašīr-Farr	Ardašīr-Farr	Ἀρταξιρουφρ	Line 42
Ardašīr-Šnōm	Ardašīr-Šnōm	Ἀρταξαρισνουμ	Line 46
Ardawān	Ardaβān	Ἀρταβάνου, Ἰρδουαν	Line 47, 50
Aristōn	Aristōn	Ἀριστίαν	Line 15
Armenāz	Armenāž	Λαρμέναζα	Line 13
Armin	Armin	Ἀρμενίαν, Ἀρμενίαν	Line 2, 9
Arminān	Arminīn	Ἀρμενίας, Ἀρμενίας, Ἀρμενίων	Line 33, 37, 38
Aršād	Aršād	Ἀστατ	Line 50
Artangiliyā	Artangiliyā	Ἀρτανγιλλα	Line 17
Āsiyā	Āsāyā	Ἀσίας	Line 20
Aspōrag	Aspōrag	Ἀσπωρικ	Line 40
Aspōragān	Aspōragān	-, Ἀσπωριγαν	Line 11, 40
Astriyā	Astriyā<->	Ἀμαστρίας	Line 19
Asūrestān	Asūrestān	Ἀσσυρίαν, Ἀσσυρίας, Ἀσσυρίαν, Ἀσσυρίᾳ	Line 2, 6, 6, 30
Asūriyā	Asūriyā	Συρίας	Line 10, -
Aygiyā	Āygā	Αἰγέαν	Line 24
Balāsagān	Balāsagān	— — —γηνήν	Line 2
Bandagān	Bandagān	Βανδιγαν	Line 50
Barēsagān	Barēsagān	Βερησιγαν	Line 43
Barragān	Barragān	Βαρριγαν	Line 48
Batnān	Batnān	Βάτναν	Line 16
Baydād	Baγdād	Βαδου	Line 50
Bebālis	Bēbāliš	Βαρβαρισσῷ, Βαρβαλισσὸν	Line 9, 11
bidaxšgān	bidaxšgān	Πιτιξιγαν	Line 49

Bīrt	Bīrt	-, Βίρθαν	Line 11, 29
Bīrt Arūbān	Bīrt-Arūbān	Βίρθαν Ἀσπωράκου	Line 11
Bitūniyā	Bitūniyā	Βιθυνίας	Line 20
Čāčestān	Čāčestān	Τσατσηνῆς	Line 3
Čašmag	Čašmag	Τιεσμακ, Τιασμικ	Line 37, 46
Čihrag	Čihrag	Τζερικ	Line 43
Dākiyā	Dākiyā	Δακείας	Line 19
Dēhēn	Dēhēn	Δηην	Line 42
Dēnag	Dēnag	Δηνακης, Δηνικ, Δηνακης, Δηνακη<ς>	Line 36, 41, 42, 44
Dīkōr	Dīkōr	Διχωρ	Line 15
Dirām	Dirān	Δηραν	Line 43
dizbedgān	dizbedgān	Δησβηδιγαν	Line 49
Dōlīx	Dōlōx	Δολίχην	Line 15
Domān	Domān	Δόμαν	Line 16
Dumbāwan	Dumbāwand	Τουμβαούντων	Line 47
Dūrā	Dūrā	Δοῦραν	Line 15
Ēpīfaniyā	Ēpīfaniyā	Ἐπιφάνιαν	Line 27
Ērān	Ērān	Ἀριανῶν	Line 1
Ērānšahr	Ērānšahr	τὸ τῶν Ἀριανῶν ἔθνος, Ἀριανῶν ἔθνους, Ἀριανῶν	Line 6, 1, 30
Ēwagān	Ēwagān	Ἀβγαν	Line 47
Ēwaxš	Abdaxš	Ἀβ<δ>αγας	Line 49
Farrag	Farrag	Φαρρεκ	Line 40
Farragān	Farragān	Φαρρικαν, Οὐιφεριγαν, Παρικαν	Line 40, 43, 45
Filip<p>os	Filip<p>os	Φίλιππον, Φίλιππος	Line 7, 8
Flāwiyās	Flāwiyās	Φλαυιάδα	Line 26

Fōnikiyā	Fonikāyā	Φοινείκης	Line 21
Frāt	Frāt	Φρέατα	Line 17
Frīg	Frīg	Φρείκου	Line 46
Frūgē	Frūgāyā	Φρυγίας	Line 20
Galātiyā	Galātīniyā	Γαλατίας	Line 20
Garmān	Garmāniyā	Γερμανῶν	Line 6
Garmāniyā	Garmāniyā	Γερμανῶν, Γερμανίας	Line 19, 21
Garmanos	Garmaniyos	Γερμανείκιαν	Line 15
Gay	Gaβ	Γη	Line 47
Gēlān	Gēlān	Γεληνῶν	Line 36
Gindaros	Gindaros	Γίνδαρον	Line 13
Gō<y>mān	Gō<y>mān	Γωμαν	Line 46
Gōg	Gōg	Γωωκ	Line 42
Gōrdanyos	Gōrdanyos	Γορδιανὸς	Line 6, 7
Gōt	Gōt	Γούθθων	Line 6
Gulag	Wardag	Οὐαρδικ	Line 50
Gundifarr	Gundifarr	Γυνδιφερ	Line 47
Gurgān	Wurgān	Γουργαν	Line 3
Halab	Halab	Βέρροιαν	Line 12
Hamadān	Hamadān	Ἀμιδαν	Line 48
Hamāt	Hamāt	Χαμαθ	Line 14
Harēw	Harēw	Ἡην	Line 3
Harrān	Harrān	Κάρρας, Καρρῶν	Line 18, 22
Hind	Hind	Ἰνδίας	Line 34
Hindestān	Hindestān	Ἰνδίαν	Line 3
Homfrād	Hōmfraδād	Χουμαφρατ	Line 43
Hōragān	Hōragān	Ὡριγαν	Line 40

Hormezd	Hormezd	Ὡρμισδ, Ὡρμίζ<δ>ου	Line 38, 49
Hormezdag	Hormezdag	Ὡρμισδακ	Line 38
Hrōm	Frōm	Ῥωμαίων ἀρχῆς	Line 6
Hrōmāyin	Frōmāyīn	Ῥωμαῖοι, Ῥωμαίων	Line 7, 7, 9, 9, 10, 24, 30
Hudug	Hudug	Χουδικ	Line 43
Husraw-ādur-Anāhīd	Husraw-Ādur-Anāhīd	Χοστρω-αδουραναιδ	Line 33
Husraw-Narseh	Husraw-Narseh	Χοστρω-ναρση	Line 34
Husraw-Ohrmezd-Ar-dašīr	Husraw-Ohrmezd-Ar-dašīr	Χοστρω-ορμισδαρταξειρ	Line 33
Husraw-Šābuhr	Husraw-Šābuhr	Χοστρω-σαβουρ	Line 33, 34
Īkōniyā	Īkōniyā	Ἰκόνιν	Line 29
Isawriyā	Isawriyā	Συρίας	Line 20
Ispāniyā	Ispaniyā	Ἱσπανίας	Line 19
Jahēn	J̌ahēn	Διεην	Line 43
J̌ōymard	J̌ōymard	Διωμερδου	Line 48
Kadugān	Kadugān	Κιδουκαν	Line 45
Kaf	Kaf	Καπ	Line 2
Kap<p>ōdakiyā	Kap<p>ōdakiyā	Καππαδοκίας	Line 16, 20, 23
Kārin	Kārin	Καρην	Line 42, 46
Kastābalāy	Kastābalā	Καστάβαλα	Line 26
Katabalā	Katabalā	Κατάβολον	Line 24
Katis<s>os	Katis<s>os	κατ’ Ἰσὸν	Line 24
Kerdīr	Kerdīr	Καρτειρ, Κιρδειρ	Line 49, 50
Kerdsraw	Kerdsraw	Κιρδιστρω	Line 47
Kermān	Kermān	Κερμαν, Κιρμανζηνῆς	Line 41, 44
Kēsar	Kēsar	Καῖσαρ, Καῖσαρ, Καῖσαρ, Καῖσαρ, Καῖσαρ, Καίσαρα, Καίσαρος	Line 6, 7, 8, 9, 18, 22, 22

Kēsariyā	Kēsariyā	Μηιακαριρη	Line 28
Kīlikiyā	Kīlikiyā	Κιλικίας	Line 20, 23
Kīlindiros	Kīlindiros	Κελένδεριν	Line 27
Kinasrā	Kinašrā	Χαλκίδα	Line 12
Kir<r>os	Kir<r>os	Κύρρον	Line 13
Kirkīsiyā	Kirkisiyā	Κορκουσίωνα	Line 15
Kirmān	Kermān	Κερμανζηνήν	Line 3
Kōmānāy	Kōmānāyā	Κόμανα	Line 28
Kōrikos	Kōrikos	Κώρυκον	Line 25
Kūbistariyā	Kūbistariyā	Κύβιστρα	Line 28
Kušānšahr	Kušānšahr	Κουσηνῶν ἔθνη	Line 3
Lārandiyā	Lārandiyā	Λάρανδα	Line 29
Likōniyā	Likōniyā	Λυκαονίας	Line 20
Lūkiyā	Lūkiyā	Λυκίας	Line 20
Māh	Māδ	Μαδηνήν	Line 3
Mak<u>rān	Mak<u>rān	Μακραν	Line 3
Māl<l>os	Māl<l>os	Μαλλὸν	Line 25
Māmasastiyā	Māmāstiyā	Μομψουεστίαν	Line 24
Manbūg	Manbūg	Ἱεράπολιν	Line 11
Mard	Mard	Μαρδ	Line 42
Mardēn<a>gān	Mardēn<a>gān	Μερδιγαν	Line 40
Marw	Maryγ	Μαρου	Line 3, 41
Mayānrōdān	Maδyānrōdān	Μεσοποταμίας	Line 21
Mazūnšahr	Mazūnšahr	Μιζουν ἔθνος	Line 3
Mēšān	Mēšān	Μησανηνήν, Μησανηνῶν, Μησανηνῶν, Μησανηνῶν, Μησων	Line 2, 34, 36, 38, 44

Mihrag	Mihrag	Μεερ<ι>κ	Line 43
Mihrān	Mihrān	Μεεραν	Line 50
Mihrōzān<a>gān	Mihraβōzān<a>gān	Μεερωζινηγαν	Line 40
Mihrxwāst	Mihrxwāšt	Μερχουάστου, Μεερχουαστ	Line 43, 49
Mišīk	Mišīk	Μησιχίῃ, Μισιχην	Line 6, 8
Miyanpolos	Mīyanpolos	Μυῶν	Line 27
Mōrān	Mōrān	Μαυριτανίας	Line 21
Mōsiyā	Mōsiyā	Μυσίας	Line 19
Mōstinopolos	Mōstinopolos	Δομετίου	Line 28
Murrōd	Murrōd	Μυρρωδ	Line 37
Nādug	Nāδug	Ναδωκ	Line 49
Narseh	Narseh	Ναρσαίου	Line 34, 37, 45, 46, 48, 49
Narsehgān	Narsehgān	Ναρσηγαν	Line 48
Nāsbed<a>gān	Nāšbed<a>gān	Νασπαδιγαν	Line 50
Nerōniyās	Nerōniyās	Νερωνιάδα	Line 26
Nēw-Šābuhr	Nēw-Šābuhr	Νι-σαβωρ	Line 46
Nīkopolos	Nīkopolos	Νεικόπολιν, Νεικόπολιν	Line 14, 26
Nīrīz	Nīrīz	Νηρηζ	Line 50
Nodšīragān	Nōdšīragān	Ἀδιαβηνήν, Ἀδιαβηνῆς	Line 2, 44
Nōrikos	Nīrakos	Νωρικοῦ	Line 19
Ōdābaxt	Ōdābaxt	Ὀδαβαχθ	Line 38
Ohrmezd-Ardašīr	Ohrmezd-Ardašīr	Ὡρμισδαρταξιρ, Ὡρμισδαρταξιρ, Ὡρμισδ-αρταξάρου	Line 4, 33, 37
Ohrmezd<d>uxtag	Ohrmezdduxtag	Ὡρμισδ-δουκτακ	Line 38
Orsigān	Orsigān	Ἀρνηγαν	Line 40
Pā<k>čihr	Pāčihr	Παζηρου, Παζηρ	Line 43, 46

Pābag	Pābag	Παπάκου, Παπάκου, Παπάκου, Παβάκου, Παβακ, Παβάκου, Παβάκου, Παπάκου, Παπάκου, Παπακ	Line 1, 36, 40, 41, 42, 42, 45, 47, 48, 49
Pābagān	Pābagān	Παπακαν, Παπακαν, Παβάκου	Line 36, 42, 44
Pābīg	Pābič	Παβις	Line 47
Pahlaw	Parθaw	Παρθίαν, Παρθίᾳ	Line 2, 30
Pamfiliyā	Pamfilāyā	Καμπανίας	Line 20
Pannaniyā	Pannaniyā	Παννονίας	Line 19
Pār<a>dān	Pār<a>dān	Παραδηνήν	Line 3
Parišxwār	Parišxwār	Πρεσσουαρ	Line 2
Pārs	Pārs	Περσίδα, Περσίδα, Περσίδει	Line 2, 22, 30
Pāsfal	Pāsfard	Πασφερδ	Line 49
Pāsfalgān	Pāsfardgān	Πασφερδιγαν	Line 49
Paškabūr	Paškabūr	Πασκιβουρων	Line 3
Pērōz	Pērōz	Πηρώζου, Πηρωζ, Πηρωζ, Πηρώζου	Line 37, 38, 42, 45
Pērōz-Šābuhr	Pērōz-Šābuhr	Πηρωσ-σαβουρ, Πηρωσαβωρ	Line 4, 8, 47
Pērōzgan	Pērōzgan	Πηρωζιγαν	Line 45
Puhrag	Puhrag	Πωρικ	Line 40
Rākūndiyā	Rākūndiyā	Ῥακουνδιαν	Line 29
Rastag	Rastag	Ῥαστακ	Line 49
Rastagān	Rastagān	Ῥαστιγαν	Line 48
Raxš	Raxš	Ῥοξ	Line 42
Ray			Line 50
Refaniyos	Refaniyos	Ῥεφανέαν	Line 12
Rēšiyā	Rēšiyā	Ῥετίας	Line 19

Rind	Rind	Ῥινδ	Line 48
Rōdag	Rōdag	Ῥωδακης	Line 41
Rōdduxt	Rōdduxt	Ῥωδ-δουκτκ<α>κ	Line 37
Rōdōs	Rōdōs	Λυδίας	Line 21
Šābuhr	Šābuhr	Σαπώρης, Σαβουρ, Σαπώρου, Σαπώρου, Σαβουρ, Σαπώρου, Σαπωρ, Σαπωρ, Σαβωρ, Σαπώρου	Line 1, 34, 36, 37, 38, 40, 44, 45, 48, 49
Šābuhr šāhān šāh	amā xwadāyīf	δεσποτείαν ἡμῶν	Line 44
Šābuhr-Šnōm	Šābuhr-Šnōm	Σαπωρ-σνουμ	Line 47
Šābuhrduxtag	Šābuhrduxtag	Σαβουρ-δουκτακ	Line 37, 38
Sadāluf	Sadāluf	Σαταροπτ	Line 41
Sagān	Sagān	Σεγιστηνῶν, Σιγαν	Line 37, 37, 38, 41
Sagbus	Sagbus	Σαγβους	Line 43
Sagestān	Sagestān	Σεγιστανήν, Σεγιστηνῆς	Line 3, 34
Šāhmust	Šāmust	Σαιμούστου	Line 46
Šahrkerd	Šahrkerd	Σαρακάρτων	Line 47
Šamīšāt	Šamšāt	Σαμόσατα	Line 24
Šanbidgān	Šanbidgān	Σονβεδηγαν	Line 47
Sāsān	Sāsān	Σασάνου, Σασάνου, Σασαν, Σασάνου, Σασαν, Σασάνου	Line 36, 40, 42, 45, 50, 50, 50
Sāsāngān	Sāsāngān	Σασανγαν	Line 50
Sātal	Sātal	Σάταλα	Line 16
Sebastiyā	Sebastiyā	Σηβαστὴν, Σηβάστιαν	Line 25, 29
Selīnūs	Selīnūs	Σελινοῦν	Line 27
Selūkān	Selūkān	Σλωκαν	Line 48
Selūkiyā	Selūkiyā	Σελεύκιαν	Line 13, 14, 27
Sīgān	Sīgān	Μαχελονίαν	Line 2

Sinzar	Sīzar	Σίνζαρα	Line 14
Sridōy	Sridōy	Στρηδω	Line 46
Staxryād	Staxryād	Σταριαδ	Line 38
Šūd	Šūd	Σουιδ	Line 17
Sugd	Suγd	Σωδικηνῆς	Line 3
Sūrā	Sūrā	Σοῦραν	Line 11
Sūrēn	Sūrēn	Σουρην	Line 42, 46
Sūriyā	Sūriyā	Συρίας	Line 20, 23
Sūš	Sūš	Σουισαν	Line 17
Tahm-Šābuhr	Tahm-Šābuhr	Ταμ-σαβουρ	Line 46
Tarsos	Tarsos	Ταρσὸν	Line 25
Tīrmihr	Tīrmihr	Τιρμερ	Line 47
Tīyanā	Tūyanā	Τύανα	Line 28
Tiyānag	Tiyānag	Τιανικ	Line 48
Tōsar<a>gān	Tōsar<a>gān	Τουσσεριγαν	Line 43
Trākiyā	Trākiyā	Θρᾳκίας	Line 20
Tūrān	Tuγrān	Τουρηνήν	Line 3
Tūrestān	Turγestān	Τουρηνῆς	Line 34
Umā	Urnā	Οὐριμα	Line 12
Urhā	Urhā	Ἔδεσσα<ν>, Ἐδέσσων	Line 18, 22
Wala<x>š	Wala<x>š	Οὐαλάσσου, Οὐαλάσου	Line 44, 48
Waliyāranos	Wālaraniyos	Οὐαλεριανος, Οὐαλεριανοῦ, Οὐαλεριανὸν	Line 18, 22, 22
Wārāz	Wārāz	Γοράζου, Γουραζ	Line 42, 45
Warāzduxt	Warāzduxt	Γοραζ-δουκτ	Line 37
Wardān	Wardān	Οὐαρδαν, Οὐαρδάνου	Line 43, 50
Wardbed	Wardbed	Γουλβαδ	Line 48

Wardbed<a>gān	Wardbed<a>gān	Γουλιβηγαν	Line 50
Warhrān	Warhrān	Γουαραθραν, Γουαραθρανου, Γουαραθραν	Line 32, 36, 38
Warhrānbād	Warhrānbād	Γοαρθανιπατ	Line 40
Wārzan	Wārzan	Γουαρζιν	Line 47
Weh-Andiyōk-Šābuhr	Weh-Andiyŏk-Šābuhr	Γουε-αντιοχ-σαβωρ	Line 46
Weh-Ardašīr	Weh-Ardašīr	Γυε-αρταξάρων	Line 49
Wēzān<a>gān	Wēzān<a>gān	Γουεζηνιγαν	Line 40
Wifr	Wifr	Οὐιφερου	Line 43
Wifr<a>gān	Wifr<a>gān	Γυιφεριγαν	Line 48
Winnār	Winnār	Γυινναρ	Line 50
Wīrōy	Wīrōδ	Οὐορωδ	Line 50
Wiruzān	Wiržān	Ἰβηρίαν, Ἰβηρίας	Line 2, 44
Wisfarr<a>gān	Wisfarr<a>gān	Γουασπεριγαν, Οὐισπερηγαν	Line 43, 47
Wohnām	Wohnām	Γοαννάμου, Γοανναμ	Line 46, 47
Xānar	Xănar	Χαναρ	Line 16
Xūzestān	Xūzestān	Οὐζηνήν, Οὐζηνῇ	Line 2, 30
Xwar<r>ānzēm	Xwar<r>ănzēm	Χορνανζημ	Line 36, 37
Yahūdiyā	Yūdāyā	Ἰουδαίας	Line 21
Yazadbed	Yaz<a>dbed	Ἰησδιβαδ	Line 48
Zabr<a>gān	Zabr<a>gān	Ζαβρικαν	Line 43
Zādspraxm<a>gān	Šābuhr<a>gān	Σαβουργαν	Line 45
Zefīrōn	Zefīrōn	Ζεφύριν	Line 25
Zig	Zīg	Ζιγ, Ζηκ, Ζικ	Line 40, 43, 47
Zōmā	Zōmā	Ζεῦγμα	Line 12
Zurwāndād	Zurwāndād	Ζαρουανδατ	Line 50
		Ἀσίας	Line 21

≺H>amāzāsp	<H>amāzāsp	Ἀμαζασπου	Line 44
	Dunbāwand	Δουμβαουνδ	Line 42
	Gēlmān	Γηλιμαν	Line 42
	Razmayōd	Ῥισμαωδ	Line 47
	Mānzag	Μανζικ	Line 50
	Mānzag	Μανζικ	Line 50

Tab. 6: List of Named Entities in the Middle Persian, Parthian, and Greek versions of ŠKZ.

Sources

Online sources and digital corpora

Datasets on GitHub = https://github.com/farnoosh-shamsian/SKZ (last access 26.06.2025).

Datasets on Zenodo = https://zenodo.org/records/15050878 (last access 26.06.2025).

Encyclopaedia Iranica = https://www.iranicaonline.org (last access 26.06.2025).

Parthian Sources = http://parthiansources.com/ (last access 26.06.2025).

PHI Epigraphy Project = https://epigraphy.packhum.org (last access 26.06.2025).

Topos Text = https://topostext.org/ (last access 26.06.2025).

Trismegistos = https://www.trismegistos.org/ (last access 26.06.2025).

Ugarit Alignment Tool = http://ugarit.ialigner.com/ (last access 26.06.2025).

Wikidata = https://www.wikidata.org (last access 26.06.2025).

Text editions

Back (1978): M. Back (ed.), Die sassanidischen Staatsinschriften, Acta Iranica 18, Teheran / Leiden 1978.

Canali De Rossi (2004): F. Canali De Rossi (ed.), Iscrizioni dello estremo oriente greco. Un repertorio. Inschriften griechischer Städte aus Kleinasien 65, Bonn 2004.

Huyse (1999): P. Huyse (ed.), Die dreisprachige Inschrift Šābuhrs I. an der Kaʿba-i Zardušt (ŠKZ), Corpus Inscriptionum Iranicarum III/I/I, London 1999.

References

Berti (2019): M. Berti, Named Entity Annotation for Ancient Greek with INCEpTION, in: K. Simov / M. Eskevich (eds.), Proceedings of CLARIN Annual Conference 2019, Leipzig 2019, 1–4

Berti (2021): M. Berti, Digital Editions of Historical Fragmentary Texts, Digital Classics Books 5, Heidelberg 2021, https://doi.org/10.11588/propylaeum.898 (last access 26.06.2025).

Daryaee (2018): T. Daryaee, Res Gestae Divi Saporis, in: The Oxford Dictionary of Late Antiquity, Oxford 2018.

Dhabhar (1930): B. N. Dhabhar, Pahlavi dstkrt' or YDHkrt' – Dastkart, in: Modi Memorial Volume, Bombay 1930.

Durkin-Meisterernst (2004): D. Durkin-Meisterernst, Dictionary of Manichaean Middle Persian and Parthian, Dictionary of Manichaean Texts, Vol. III, Texts from Central Asia and China, Part 1, Turnhout 2004.

Gignoux (1972): P. Gignoux (ed.), Glossaire des Inscriptions Pehlevies et Parthes, Corpus Inscriptionum Iranicarum Supplementary Series I, London 1972.

Gignoux (1994): P. Gignoux, Dastgerd, in: Encyclopaedia Iranica VII, London 1994.

Jam (2019): P. Jam, Dastgerd and Daskara, Farhang Nevisi 15 (2019).

MacKenzie (1971): D. N. MacKenzie, A Concise Pahlavi Dictionary, London 1971.

Merkelbach / Stauber (2005): R. Merkelbach / J. Stauber (eds.), Jenseits des Euphrat: griechische Inschriften. Ein epigraphisches Lesebuch, München / Leipzig 2005.

Munday (2016): J. Munday, Introducing Translation Studies: Theories and Applications, Fourth Edition, New York 2016.

Nyberg (1964–1974): H. S. Nyberg, A Manual of Pahlavi, Vol. I / II, Wiesbaden 1964 / 1974.

Panaino (2022): A. Panaino, Between Semantics and Pragmatics: Origins and Developments in the Meaning of dastgerd. A New Approach to the Problem, Sasanian Studies: Late Antique Iranian World / Sasanidische Studien: Spätantike Iranische Welt 1 (2022), 215–242.

Rubin (2002): Z. Rubin, Res Gestae Divi Saporis, Greek and Middle Iranian in a Document of Sasanian Anti-Roman Propaganda, in: J. N. Adams / M. Janse / S. Swain (eds.), Bilingualism in Ancient Society, Oxford 2002.

Schmidt (1970): E. F. Schmidt, Persepolis III: Royal Tombs and Other Monuments, Chicago 1970, https://isac.uchicago.edu/research/publications/oip/persepolis-iii-royal-tombs-and-other-monuments (last access 26.06.2025).

Shamsian (2023): F. Shamsian, Alignment Guidelines for Classical Greek-Persian, online 2023, https://doi.org/10.5281/zenodo.8039931 (last access 26.06.2025).

Skalmowski (1993): W. Skalmowski / A. Van Tongerloo (eds.), Medioiranica: Proceedings of the International Colloquium Organized by the Katholieke Universiteit Leuven, Leuven 1993.

Sundermann (1989): W. Sundermann, BIDAXŠ, in: Encyclopaedia Iranica IV/3, 242–244, https://www.iranicaonline.org/articles/bidaxs-title-of-iranian-origin (last access 26.06.2025).

Yousef et al. (2022): T. Yousef / C. Palladino / F. Shamsian / M. Foradi, Translation Alignment with Ugarit, Information 13/2 (2022), 65, https://doi.org/10.3390/info13020065 (last access 26.06.2025).

Figure References

Fig. 1: Trilingual alignment of the first lines of ŠKZ in Ugarit.

Author Contact Information[32]

Dr. Farnoosh Shamsian
Universität Leipzig
Institut für Angewandte Linguistik und Translatologie
Geisteswissenschaftliches Zentrum
Beethovenstr. 15
D-04107 Leipzig
E-mail: farnoosh.shamsian@uni-leipzig.de

PD Dr. Monica Berti
Universität Leipzig
Lehrstuhl für Alte Geschichte
Beethovenstraße 15
04107 Leipzig
E-mail: monica.berti@uni-leipzig.de

32

DIGITAL CLASSICS ONLINE

LGPN-Ling for the Preservation of Greek Personal Names in a Digital Environment

Matilde Garré

Abstract: The aim of this paper is to illustrate the preservation of Greek personal names in a digital environment, going beyond the concept of 'name entity' and taking into account the morphological, morphosyntactic and lexical entities involved in the analysis of these personal names. To achieve this, I will focus on the output of the *LGPN-Ling* project, directed by Sophie Minon: the *LGPN-Ling* search site and the printed volume of *Lexonyme*. I will first examine the structure of the website and its conception, taking into account the uncertainties that can arise in the analysis of personal names and that can be difficult to convey in a clear and concise manner. I will then use representative examples to illustrate how the structure of the *LGPN-Ling* site allows Greek personal names to be preserved in a digital environment not only as *nomina omina*, but also as even smaller entities such as morphemes and lexemes. Indeed, each entry in the database is connected and linked to the others in all aspects of its analysis. This includes gender, geolinguistics, bases, suffixes, morphosyntax, semantic and even lexical correspondences in Greek. Names are thus linked on the basis of one or more of these features, creating a complex network of interrelationships between the registered forms.

Introduction

From *LGPN* to *LGPN-Ling*, from a name entity to an even smaller linguistic – lexical, morphological, morphosyntactic, and semantic – entity: this is what has been achieved through the research and the elaboration of the *LGPN-Ling* website, under the direction of Sophie Minon. The history of the project, its objectives and its design will be considered in the first instance. The discussion will then shift to the *LGPN-Ling* website, examining how its architecture ensures the preservation of ancient Greek personal names in a digital environment. Afterwards, we will examine the concept of *nomen omen*, and discuss how the morphological, syntactic, semantic, and lexical analyses conducted in this project allow us to move past it and focus on even smaller entities.

1. Conceiving Tools for a Linguistic Analysis of Greek Personal Names

How was *LGPN-Ling* conceived and developed? This is the question that we will tackle in this paragraph. To do so, we will first focus on the history of the project and on its goals. We will then examine the issues that had to be taken into account in the creation of the database itself, due to the specific linguistic features of Greek personal names.

1.1 From *LGPN* to *LGPN-Ling*

The adventure of the *LGPN-Ling* project started in 2013, when the principal investigator, Sophie Minon, presented to Robert Parker – director of the *LGPN* series in Oxford, at the time – the idea of a linguistic analysis of the names included in the volumes of the *Lexicon of Greek Personal Names* (*LGPN*).[1] The volumes of the Oxford's *Lexicon of Greek Personal Names* thus constitute the corpus on which *LGPN-Ling* is built. Indeed, the primary objective of this project was to leverage the extensive and comprehensive onomastic data collected in the various volumes of the *Lexicon of Greek Personal Names*.[2] This data would then be used to create a new and updated version of *Die historischen Personennamen des Griechischen bis zur Kaiserzeit*, which Friedrich Bechtel first published in 1917,[3] as reminded by Minon in 2020:

> "Le projet *LGPN-Ling* […] a donc pour objectif de doter le *Lexicon of Greek Personal Names* d'Oxford d'une contrepartie qui ne peut être, bien sûr, exclusivement linguistique, dans le cas de noms de personnes, mais doit tenir compte aussi de l'ancrage historique, géographique et socio-culturel de chaque individu porteur d'un nom: c'est cette complémentarité des approches, que nous nous sommes efforcés de rendre par cet acronyme duel."[4]

From its inception, this initiative had a dual objective: to establish an online platform for the *LGPN-Ling* search engine and to publish three printed volumes of the *Lexonyme*, with the first volume appearing in 2023.[5] On the one hand, the website allows access to the detailed (and up-to-date) analysis of more than 36,000 Greek personal names thanks to the search options proposed. The *Lexonyme*, on the other hand, is a *dictionnaire raisonné*,[6] organised by etymological *lemmata*. While these printed volumes aim to produce an updated version of Bechtel's book, the *Lexonyme* is characterised by a completely new structure, which is the result of more than a century of research in Ancient Greek personal onomastics.[7] The *LGPN-Ling* search site and the printed volumes of the *Lexonyme* were thus conceived as two complementary tools, each one of them allowing for a different perspective on the linguistic analysis of Greek personal names.

---

1 So far, nine volumes have been published, according to a geographical distribution: 1 (Aegean Islands, Cyprus and Cyrenaica); 2 (Attica); 3a (Peloponnese, Western Greece, Sicily, Magna Graecia); 3b (Megarid, Oropos, Boeotia, Lokris, Phokis, Doris, Acarnania, Aetolia and Thessaly); 4 (Macedonia, Thrace, Northern Shores of the Black Sea); 5a (Coastal Asia Minor: Pontos to Ionia); 5b (Coastal Asia Minor: Caria to Cilicia); 5c (Inland Asia Minor). See also: https://www.lgpn.ox.ac.uk/volumes (last access 20.03.2025).

2 The *LGPN* series has been published since 1987. As a consequence, for certain regions, a great number of epigraphic discoveries and publications was made in the last decades Therefore, one of the biggest achievements of the *LGPN-Ling* team-work has been to update the onomastic data, thanks to the systematic verification of the indexes of the *Supplementum Epigraphicum Graecum* and the *Bulletin épigraphique*, as well as the newest *corpora*. This incredible amount of work is still ongoing thanks to the dedication and unwavering efforts of Gérard Genevrois. See also below §2.1. The achievements accomplished owe a lot to the entire *LGPN-Ling* team including, in addition to Sophie Minon and Gérard Genevrois, Jean-Claude Chuat, Dan Dana, Enrique Nieto Izquierdo, Florian Réveilhac, as well as the developer Magdalena Turska. The site (https://lgpn-ling.huma-num.fr/about.html [last access 31.01.2026]) is most easily accessible using Firefox and Google Chrome browsers.

3 Bechtel (1917).

4 Minon (2020), 256.

5 Minon (2020), 257–258; Minon (2023), XV–XVIII.

6 Minon (2023), XVII.

7 De Lamberterie (2023).

In 2015, the connection with the Oxford's *LGPN* lead to the elaboration of the *LGPN-Ling* website, developed by Magdalena Turska from *Exist Solutions*, under the direction of Sophie Minon, in TEI-xml. Opened to the public in 2023, the elaboration of its final structure has required the development of several preliminary and intermediary versions, aiming at understanding how to present all the linguistic features of an ancient Greek personal name.

1.2 Conveying the Analysis of a Personal Name into a Strict Structure

When analysing Greek personal names, it is not uncommon to encounter certain degrees of uncertainty, which can be challenging to convey in a concise and clear manner. When scientifically we are not even sure about the analysis of a name, how can we present it clearly to make it acceptable to a *machine*? This question was one of the most demanding issues that had to be tackled during the conception phase of the *LGPN-Ling* website. Therefore, the primary challenge in developing a database and search engine was not just designing and structuring it, but rather the ability to dissect and categorise each component of a personal name, ensuring that they aligned with the database's architecture. Displaying and organising the database in a way that allows for the inclusion of all morphological and semantic information in a specific 'column' represents a significant challenge in itself. This is particularly evident in cases such as Ἀριστῶναξ, for example, where a contraction has occurred between the final vowel of the first base (Ἀριστο°) and the initial vowel of the second base (°ϝανακτ-), following the reduction of intervocalic *w*. These details concerning its analysis must in fact be presented in the clearest and most concise way possible. But in addition to these specific and punctual issues, concerning the phonological analysis of particular names, other – more general – questions had to be addressed.

The first issue concerns the classification of the different kinds of personal names depending on their composition process. In 1917, Bechtel[8] decided to organise his book into two main sections: the first one concerns compounds and abbreviated compounds, while the second one is devoted to *die übrigen Namen* (meaning: the remaining names!). This second section includes the *simplicia*: names whose base – typically a name or an adjective – originate directly from the lexicon. This implies that the lexical component of these names is not usually attested as a compound. Nevertheless, as reminded by Minon,[9] the label of these different parts of his book clearly implies a hierarchy between the compounded forms, inherited from Indo-European, on one hand, and the personal names derived directly from the Greek lexicon, on the other. Furthermore, as Dobias-Lalou and Dubois stated clearly in their introduction to Olivier Masson's *Onomastica Graeca Selecta*,[10] in some cases, the distinction between abbreviated compounds and *simplicia* is not as clearcut as it seems. For example, an adjective like ἀγαθός (good, noble) could potentially have been used to form a personal name, such as Ἀγάθων. Therefore, a name such as Ἀγάθων does not have to be necessarily interpreted as a shortened form of a compound, such as Ἀγαθο-κλῆς. Both analyses could, in fact, be true at the same time for the same individual, or one option could be valid for one Ἀγάθων, while the other for another. This example raises another question, which can be challenging to summarise concisely: the morphological and semantic analysis of a name may differ based on the person who bears it (§3.1).

The terminology of the analysis of the construction of personal names can also be problematic. This is evident from the wide, often imprecise, use of the term 'hypocoristic', as noted by Dubois in 2017.[11] That is why one of the biggest efforts undertaken by the *LGPN-Ling* team was to determine a classification and a terminology that could be used to describe Greek personal names solely depending on their

8 Bechtel (1917).

9 Minon (2020), 263–264.

10 Dobias-Lalou / Dubois (1990), XI–XIV.

11 Dubois (2017), 6–11.

composition process.[12] To gain a deeper understanding of the morphological formation of Greek personal names, this study proved vital. It served as the foundation for creating an exhaustive and accurate structure on the *LGPN-Ling* platform. As a result, the database structure was designed to accommodate up to two bases, or a prefix and a base (§2.1). However, it is possible that in certain cases, one of the bases may contain a compounded form. This could be interpreted as proof that we are dealing with a three-base compound.[13] Therefore, we had to create a solution to accommodate the three-base compounds in the database without adding an additional column, which could have been labelled 'base 3'. Doing so we would have conveyed the misleading idea that the three bases function independently, when in reality, they do not. For example, the name Φιλ-έφηβος[14] consists of a compound base, ἔφηβος, including the prefix ἐπί and the root ἥβη. Consequently, it was determined that an equal sign should be introduced between the two members of the compounded base. As a result, the Base 2 appears in *LGPN-Ling* as επ(ι)=(h)εβ(α/ο) (in youth), a compound (XY) originating from the syntagm ἐφ'ἥβης (during adolescence). This allowed a more precise correlation to be established between the morphological structure of the bases and their morphosyntactic analysis, since ἔφηβος is here the whole head (X) of the compound (§2.1).

The second issue concerns the derivational process and the segmentation of the suffixes attested by the personal name. Indeed, it is of critical importance that those who analyse Greek personal names be able to identify the endings of names and the suffixes or suffixes used in their formation. To this day, aside from *La formation des noms en grec ancien* by Pierre Chantraine (1933),[15] no exhaustive analysis of Greek suffixes has been undertaken. Such a study should include not only literature, but also epigraphy and papyrology. Additionally, it should compare the lexical and onomastic uses of the suffixes. Consequently, one of the initial scientific activities undertaken as part of the *LGPN-Ling* project was a conference held in Lyon in 2015 on the suffixation of Greek personal names, later published in 2017 as *La suffixation des anthroponymes grecs antiques*.[16] The papers included in this book provide a preliminary overview of the issue, including different suffixes and contact areas, as well as features concerning specific dialects. This research also enabled Minon to develop a preliminary catalogue of the different types of these suffixes, and, most notably, the so-called *chaînes suffixales*,[17] as discussed in the aforementioned book. To design the database structure, it was crucial to understand the maximum number of suffixes that could be used in a chain and the sequence in which they could be used. Through this research, it was established that in Greek personal names there is the potential for up to four different suffixes to be employed.[18] In certain instances, the same suffix may be found in multiple positions. The derivational suffix -ων, -ωνος, for example, could have been used in the fourth, third, second, and first position. The one furthest to the right, i.e. at the end of the name is in fact the position of the inflectional suffix, the only suffix that is necessary. The standard order is, from left to right, 4, 3, 2, 1; but the numbering is hierarchised. This research was therefore crucial for understanding how many columns were to be created in the database. Four columns were thus designed, corresponding to the different suffixes. The case of Φιλ-ων-ιχ-ίδ-ης,[19] is particularly noteworthy, as all four columns are used. Clearly, the creation and organisation of the different suffixes within the four columns required some level of abstraction and the making of difficult choices and concessions (§3.2). For instance, as

12 Minon (2023), XIV–XVI.

13 Minon (2020), 283.

14 https://lgpn-ling.huma-num.fr/Philephēbos (last access 23.11.2024).

15 Chantraine (1933).

16 Alonso Déniz et al. (2017).

17 Minon (2017).

18 In the *LGPN-Ling* website these suffixes have been given a number (1, 2, 3 or 4) depending on their position.

19 Minon (2017), 704; Minon (2020), 262; https://lgpn-ling.huma-num.fr/Philonichidēs (last access 21.05.2024).

illustrated by Φιλ-ων-ιχ-ίδ-ης, the complex derivational suffix -ίδᾱς has been presented here as -(ί)δ- and -ᾱς, despite the ongoing debate surrounding the origin and evolution of this suffix within the field of Greek derivational suffixation.[20]

The first steps undertaken by the *LGPN-Ling* team thus aimed at gaining more insight on the detailed analysis of a Greek personal name. This in-depth research was crucial in ensuring that every linguistic detail was accurately represented in the database. The team aimed to answer a multitude of questions, such as how many 'columns' were required, and what kind of data these columns should contain. How will the content of these columns interact with the other columns? It was by answering these questions and addressing issues such as the construction and suffixation of Greek personal names that the website's current structure was progressively developed and refined.

2. Preserving Greek Personal Names in a Digital Environment

In this paragraph, we will address the question on how the structure of *LGPN-Ling* allows to preserve Greek personal names in a digital environment. To do so, we will first focus on the structure of *LGPN-Ling*. We will then illustrate the internal and external connection and interconnections that have been implemented.

2.1 The Structure of *LGPN-Ling*

The *LGPN-Ling* website is organised in two main sections: 'Browse' and 'Catalogues', both accessible from the homepage.[21] While both sections provide access to the study of Greek personal names, they cater to distinct purposes.

The 'Browse' section was created with two primary objectives in mind: (1) offering a visually engaging representation of name analyses, and (2) facilitating full-text and faceted searches for these names. We will focus here on the analysis of personal names (1), as it can be accessed in the browse section. The current format of the analysis of all names included in the *LGPN-Ling* search page is organised into several columns. The interface's columns were designed and organised to facilitate user research and to enable the most accurate possible scientific analysis of personal names. This presentation of personal name analysis aims to address the challenges and uncertainties identified in Greek name analysis (§1.2). The existing columns are the following, presented from left to right (cf. fig. 1):[22]

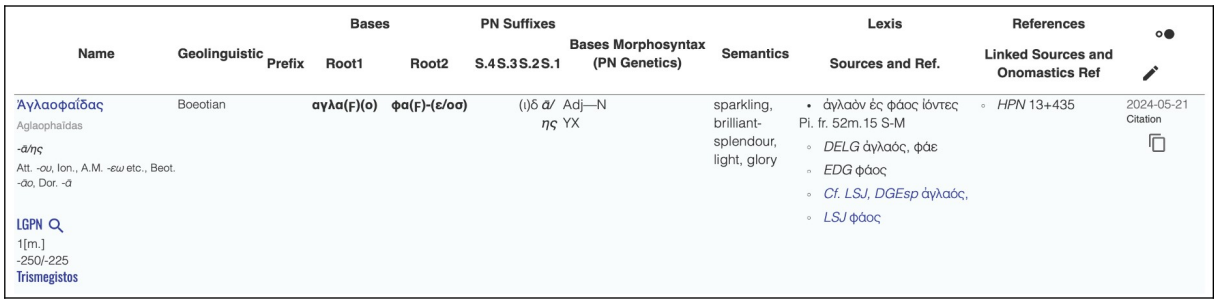
1. The 'Name' column, which includes: the name in the Greek alphabet (e.g. Ἀγλαοφαΐδας);[23] its transcription in Latin characters (e.g. *Aglaophaïdas*); the inflectional suffix and its genitive, as well as the dialectal variants of the genitive; the number of individuals bearing the entry name, according to *LGPN*; the time range (from the oldest to the most recent attestation of the entry name); the connections to *LGPN* and *Trismegistos* (that will be discussed in further detail below §2.2).

20 Garré (2022), 132–156.

21 https://lgpn-ling.huma-num.fr/about.html (last access 20.03.2025).

22 See also: Minon (2020), 281–292; the entry 'How to Read the Analysis of a Name' on the *LGPN-Ling* website (https://lgpn-ling.huma-num.fr/analysis.html [last access 15.03.2025]).

23 https://lgpn-ling.huma-num.fr/Aglaophaïdas (last access 21.05.2024).



Name	Geolinguistic	Bases			PN Suffixes		Lexis		References	
		Prefix	Root1	Root2	S.4 S.3 S.2 S.1	Bases Morphosyntax (PN Genetics)	Semantics	Sources and Ref.	Linked Sources and Onomastics Ref	
Ἀγλαοφαΐδας Aglaophaïdas *-ά/ης* Att. *-ου*, Ion., A.M. *-εω* etc., Beot. *-ᾱο*, Dor. *-ᾱ* LGPN 1[m.] -250/-225 Trismegistos	Boeotian		αγλα(ϝ)(ο)	φα(ϝ)-(ε/οσ)	(ι)δ ᾰ/ης	Adj—N YX	sparkling, brilliant-splendour, light, glory	• ἀγλαὸν ἐς φάος ἰόντες Pl. fr. 52m.15 S-M ◦ *DELG* ἀγλαός, φάε ◦ *EDG* φάος ◦ *Cf. LSJ, DGEsp* ἀγλαός, ◦ *LSJ* φάος	◦ *HPN* 13+435	2024-05-21 Citation

Fig. 1: The Presentation of the Name Ἀγλαοφαΐδας.

2. The ‘Geolinguistic’ column includes the areal, dialectal or interlinguistic characterisation of the name. More than one label can be presented here in chronological order, the first label corresponding to the oldest occurrence of a name.[24]

3. The ‘Bases’ column, which, as explained above (§1.2), is organised into three sub-columns, ‘Prefix’, ‘Root1’, and ‘Root2’ (cf. fig. 1).

4. The ‘Personal Names Suffixes’ column, structured in four additional sub-columns, as clarified above (§1.2), which allow for the presentation of up to four derivational and inflectional suffixes. It should be noted that, when the inflectional suffix of the name is the same as the inflectional suffix of the base in the Greek lexicon, all the ‘Personal Names Suffixes’ columns will appear empty.[25] As a consequence, for a name such as Θεο-γένης[26], a compounded name of a base 1 Θεο°, from θεός (god), and a base 2 °γενεσ-, from γένος (kin, birth), the columns concerning the suffixes will thus be empty. The inflectional suffix (-εσ-) is, in fact, already presented as a part of the base 2 °γενεσ-.

5. The ‘Bases Morpho-syntax’ includes several details concerning the morphosyntactic analysis of personal names. The visual presentation on the *LGPN-Ling* site is organised on several lines. The first one clarifies the morphological value of the bases, as for the following examples: Δαμο-κυδ-ίδας,[27] Noun – Noun; Διο-κύδης,[28] Proper Name – Noun, and Ἐχε-κύδης,[29] Verb – Noun. But we can also find adjectives, adverbs, etc. The second line clarifies the syntactical order in which the bases have to be read. One of the two bases is the head of the compound (X) and the core of the syntagm, while the other (Y) is the tail of the compound, syntactically dependent on the head. For example, the names Δᾱμο-κυδ-ίδᾱς and Διο-κύδης mentioned above are accompanied by YX: this means that the names must be read from right to left. Conversely, Ἐχε-κύδης is analysed as XY, meaning that the head of the compound is on the left, in this case, the verb ’Εχε-. However, in certain instances, although this presentation may provide an account of the syntactical order of the bases in synchrony, it may be insufficient to explain the diachronic development of the compound. Occasionally, personal names display an unexpected morphosyntactic order of the bases, such as Ἀνδρ-άγαθος (cf. fig. 2),[30] a compound of the name ἀνήρ (male, warrior), and the adjective ἀγαθός (good, brave). In this case, although the adjective is usually expected as base 1,

24 Minon (2020), 282.

25 Minon (2020), 284–285.

26 https://lgpn-ling.huma-num.fr/Theogenēs (last access 21.05.2024).

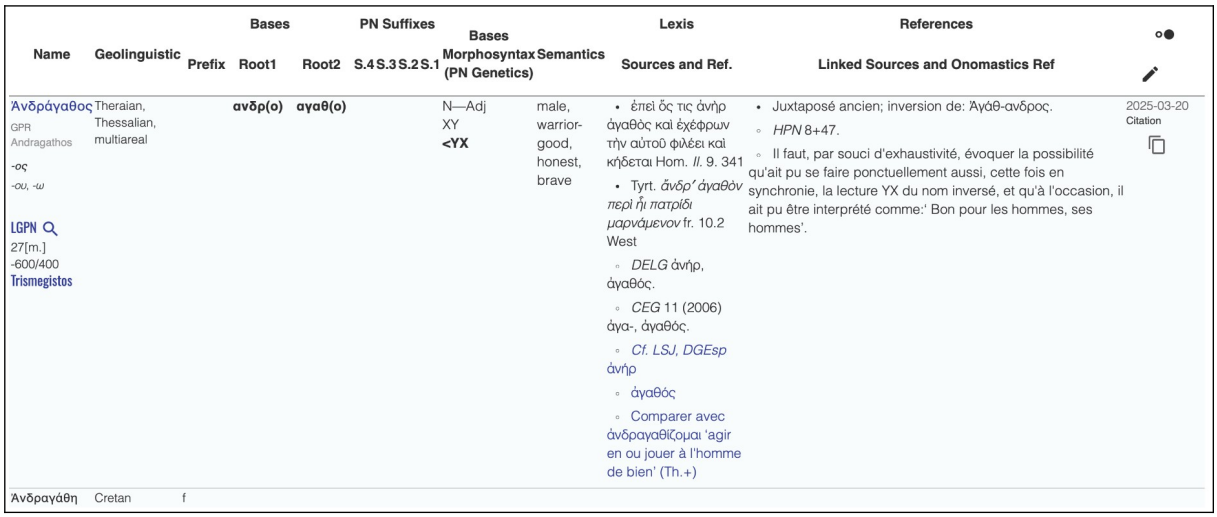
27 https://lgpn-ling.huma-num.fr/Damokydidas (last access 21.05.2024).

28 https://lgpn-ling.huma-num.fr/Diokydēs (last access 21.05.2024).

29 https://lgpn-ling.huma-num.fr/Echekydēs (last access 21.05.2024).

30 https://lgpn-ling.huma-num.fr/Andragathos (last access 21.05.2024).

it occupies, in fact, here the second position. This could be explained as a result of the inversion of a name such as Ἀγάθ-ανδρος. The inversion of an older compound Ἀγάθ-ανδρος could have been supported by the existence of the syntagm ἀνὴρ ἀγαθός. Additionally, as it is clarified by the columns 'Lexis' and 'References', illustrated below, the syntagm ἀνὴρ ἀγαθός also contributed to the creation of the verb ἀνδραγαθίζομαι (play the honest man). A close examination of the content of the 'Morpho-syntax' column reveals that the aforementioned compound should, in fact, be read from left to right, XY. The reason is that, diachronically, an adjective cannot serve as the head of a compound syntactically. Additionally, the third line describes the process of forming this compound by inverting the order of the roots of the older and more typical Ἀγάθ-ανδρος.[31] Nevertheless, it cannot be excluded that for certain individuals, in synchrony, their name should actually be read from left to right (good for the/his men). In other instances, this column allows to examine as well the question of abbreviated compounds, both limited to a single base and truncated: for an abbreviated compound Ἄλεξις, we will only find X followed by 'Abbr'; while in the case of truncated compounds, such Ἀνδροκλος, we have YXAbbr. In addition to these features, numerous others have been incorporated into the 'Morpho-syntax' column. The objective is to provide a comprehensive account of the intricate process of Greek personal name creation, which necessitates accessibility in order to facilitate a full understanding of the name as a whole.



Name	Geolinguistic	Bases Prefix	Root1	Root2	PN Suffixes S.4	S.3	S.2	S.1	Bases Morphosyntax (PN Genetics)	Semantics	Lexis Sources and Ref.	References Linked Sources and Onomastics Ref
Ἀνδράγαθος GPR Andragathos -ος -ου, -ω LGPN 27[m.] -600/400 Trismegistos	Theraian, Thessalian, multiareal		ανδρ(ο)	αγαθ(ο)					N—Adj XY <YX	male, warrior-good, honest, brave	• ἐπεὶ ὅς τις ἀνὴρ ἀγαθὸς καὶ ἐχέφρων τὴν αὐτοῦ φιλέει καὶ κήδεται Hom. *Il.* 9. 341 • Tyrt. *ἄνδρ' ἀγαθὸν περὶ ἧι πατρίδι μαρνάμενον* fr. 10.2 West ◦ *DELG* ἀνήρ, ἀγαθός. ◦ *CEG* 11 (2006) ἀγα-, ἀγαθός. ◦ *Cf. LSJ, DGEsp* ἀνήρ ◦ ἀγαθός ◦ Comparer avec ἀνδραγαθίζομαι 'agir en ou jouer à l'homme de bien' (Th.+)	• Juxtaposé ancien; inversion de: Ἀγάθ-ανδρος. ◦ *HPN* 8+47. ◦ Il faut, par souci d'exhaustivité, évoquer la possibilité qu'ait pu se faire ponctuellement aussi, cette fois en synchronie, la lecture YX du nom inversé, et qu'à l'occasion, il ait pu être interprété comme:' Bon pour les hommes, ses hommes'.	2025-03-20 Citation
Ἀνδραγάθη	Cretan	f										

Fig. 2: The Analysis of the Name Ἀνδράγαθος.

6. The 'Semantics' column includes the semantic value (in French or in English, depending on the language chosen for the interface) for every base. Usually, more than one semantic value (or possible translation) is given for every base. As for Θεο-γένης mentioned above, the base 2 γένος will be accompanied by the following semantic values: kin, birth, origin and born.

7. The 'Lexis' column includes the lexical references and sources necessary to the understanding of the name. Another challenge encountered during the *LGPN-Ling* project was, in fact, the issue of semantic analysis of Greek personal names and their translation. The process of rendering a proper name with a translation is a complex issue, particularly given the potential for semantic interpretation to differ between individuals sharing the same name, but even for the abstract personal name (§3.1). As Minon elucidates in the introduction to the first volume of the *Lexonyme*,[32] this is a particularly challenging aspect of the project. As a

31 On this topic, see also: Minon (2020), 285–288; Minon (2023), XXXVI–XXXVIII.

32 Minon (2023), XXXIII–XL.

consequence, it was chosen to give a translation of the bases in the column 'Sematics' mentioned above, without, at first, suggesting a real translation of the compound itself. The 'Lexis' column provides in fact insights into the semantic and, on occasion, phraseological analysis of compounds (cf. fig. 3). For instance, for Διο-κύδης,[33] a compound of the Proper Name Διο° (Zeus), and °κυδεσ- (divine power, glory, fame), it has been decided to quote the Homeric recurrent syntagm Ζεῦ κύδιστε (e.g. *Iliad* 2, 412), which illustrates how Zeus and Fame have be associated in ancient, especially literary, sources.

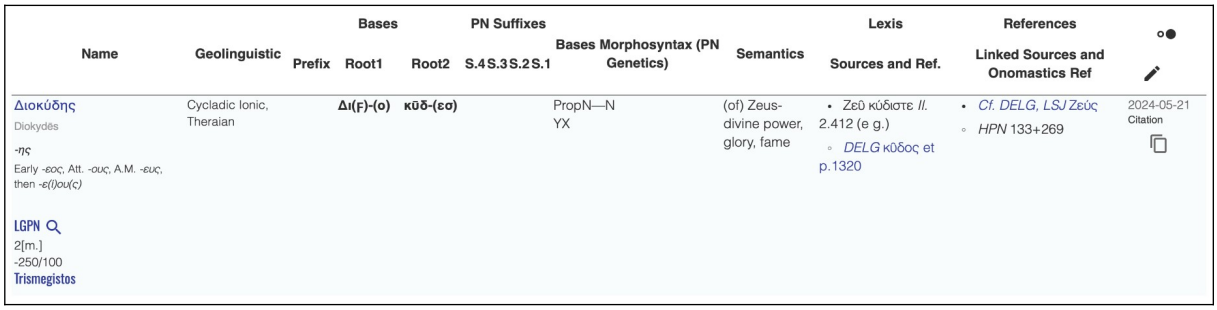


Fig. 3: The Semantic Analysis of the Name Διοκύδης.

8. The 'References' column contains several references pertaining to the onomastic analysis and interpretation of the name, as well as some comments concerning its formation. This column illustrates also one of the most significant challenges of the *LGPN-Ling* project, already mentioned above: the inclusion in the *LGPN-Ling* onomastic corpus of the Greek personal names published after the completion of the *LGPN* volumes. It was thus decided to clarify in this column the references to the names that do not appear on the *LGPN* database. For instance, under the name Ὑπάτης, the Boeotian variant from Tanagra Οὐπάττει – presented in the *editio altera* of the Boeotian corpus concerning Tanagra and its surroundings,[34] published in 2025 – has recently been added. This is clarified in the column 'References' by the following note: "Add *IG* VII^2 3, 626 (Οὐπάττει, Tanagra, 3a)".

9. The final column enables the user to copy and paste the link to the analysis of the name, as well as to modify the entry (this option being reserved to the members of the *LGPN-Ling* team).

By the end of 2017, an earlier version of the website had undergone a major overhaul, involving the merging of multiple variants of a name into one entry. As part of this restructuring, the team faced the challenge of deciding how to arrange these names. This led to an overview of the criteria that should guide the ordering process. Ultimately, it was decided to present the variants in chronological order, with the oldest attested variant first and the most recent last.[35] Therefore, the entry name is not necessarily the most common nor the best attested variant of the name. On the contrary, it can happen that the oldest attestation of a name is, in fact, *hapax*. One example of this phenomenon is the entry for the name Fεργ-αένετος (cf. fig. 4),[36] a compound of the noun (F)εργο° (work, worker), and the verbal adjective °αινετο- (praiseworthy, meaningful). In this case, the Boeotian occurrence, being the oldest attested, was chosen as the entry name, while the more 'standard' variant, without any specific dialectal phonological features – Ἐργ-αίνετος (found in Euboea) only comes as the second variant of the name, under the entry name. Among the variants, one can find forms of the same personal name as attested in different dialects – and often characterised by phonetic and phonological differences (such as the

33 https://lgpn-ling.huma-num.fr/Diokydēs (last access 21.05.2024).

34 Hallof et al. (2025).

35 Minon (2023), XXI.

36 https://lgpn-ling.huma-num.fr/Wergaenetos (last access 21.05.2024).

couple Boeotian Fεργ-αένετος and Ἐργ-αίνετος). Additionally, the masculine/feminine variants of the name are represented, as evidenced by the third variant of the entry name Fεργ-αένετος, which is the Boeotian female name Fεργ-αινέτᾱ. This functionality of the search site is especially significant when considering dialectal variants. Indeed, they can sometimes be challenging to identify and could go unnoticed through simple research on *LGPN Database Search*. If the Oxford's *LGPN* wishes, in fact, to implement their search interface, taking into account the dialects, currently dialectal variants are recorded on *LGPN Database Search* as separate entries.[37] The reorganisation of *LGPN-Ling* – undertaken in 2017 – was thus a crucial step in the development of its current version. It enabled the reduction of the number of analysed names, which, as shown above, can contain a multitude of elements. Additionally, it simplifies the identification of all the dialectal variants, even those that are less obvious.

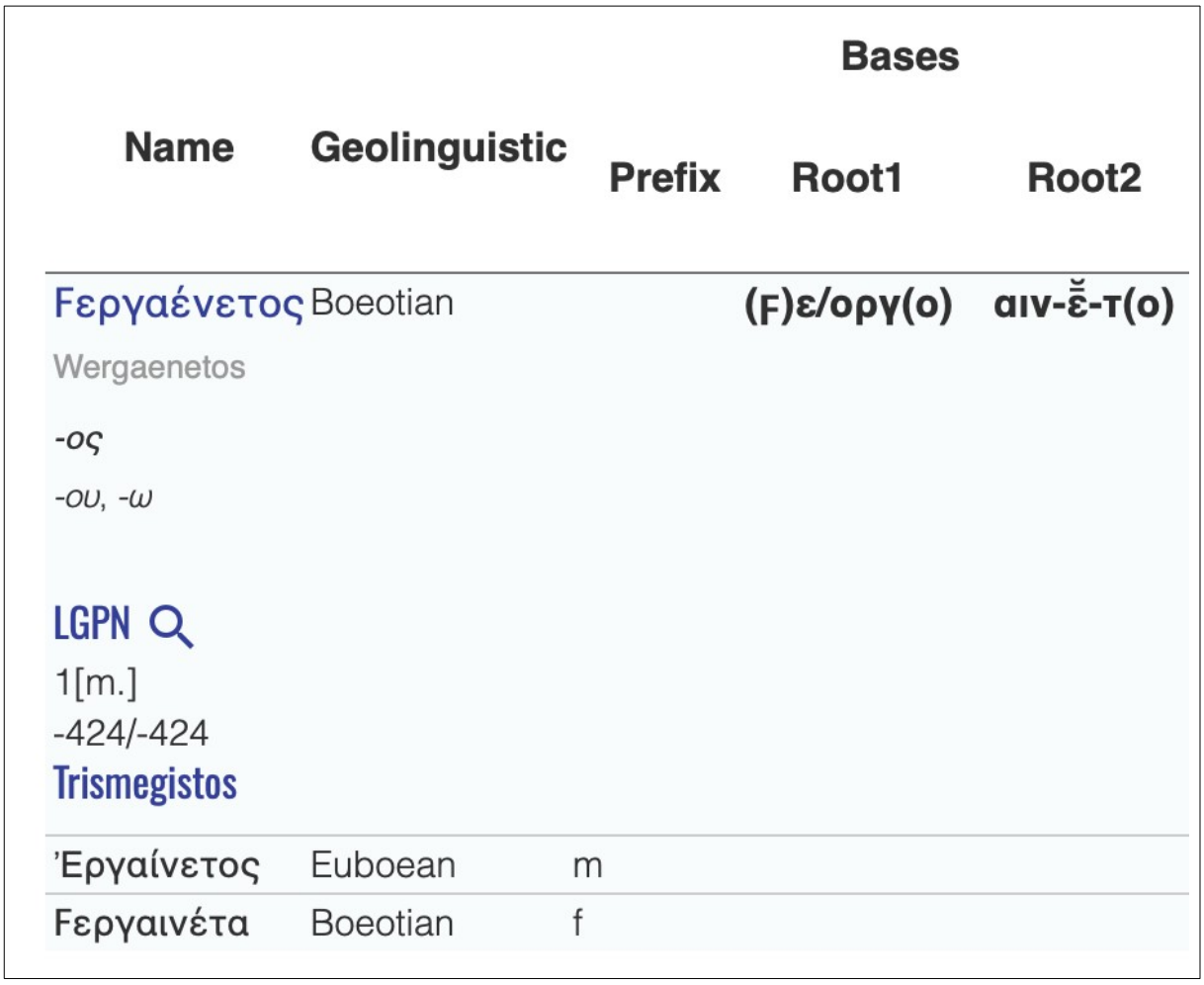


Fig. 4: The Name Fεργαένετος and its Variants.

Another important feature of the *LGPN-Ling* search site is its 'Catalogue' section. This section is useful not only for studying Greek personal names, but also for investigating the Greek lexicon. The bases identified in the name analysis (column 3, 'Bases') and their semantic values have been organised into two catalogues: one of Greek bases, and another listing all the meanings.

The 'Bases' catalogue[38] is structured alphabetically by the Greek letter that each base begins with. Every entry is accompanied by the translation of the base, and thus gives access to all the personal names – analysed in the database – featuring a specific base. For example, the entry αυγσι (cf. fig. 5) provides direct access to all the names that contain the base αὐξι°, which is derived from the verb αὔξω (to increase). These include the compounds Αὐξί-βιος and Αὐξί-θεμις, as well as the abbreviated

37 https://www.lgpn.ox.ac.uk/search (last access 12.03.2025).

38 https://lgpn-ling.huma-num.fr/lexemes.html (last access 15.03.2025).

compound Αὐξίᾱς. The content of this catalogue is closely connected to the printed volume, *Lexonyme*, discussed above (§1.1), and the two resources can be used together. The entries in the *Lexonyme* volumes follow the principles of an etymological dictionary, including all the different forms derived from the main entry. For example, names whose base is αὐξι° will appear under the heading αὔξω.[39]

Fig. 5: The Entry αυγσ(ι) in the 'Bases' Catalogue.

In contrast to an etymological dictionary, this compilation only includes the derivated forms that are attested as bases in the formation of personal names. The systematic presentation of the personal names, which is duly elucidated in the introduction to the book,[40] facilitates the understanding of the derivation process. Conversely, although *LGPN-Ling* incorporates the same data, its presentation does not permit the examination of the etymological derivation of the names or the connections between the different bases derived from the same root. The 'Bases' section does not organise the entries in a hierarchical manner, but instead lists them in alphabetical order. As a result, it is challenging to discern the etymological and derivational relationships among these disparate roots using the *LGPN-Ling* database alone. Nevertheless, the comprehensive and thorough examination of each name offered by *LGPN-Ling* is not incorporated into the printed volume. These two instruments, which were developed as part of the *LGPN-Ling* project, can be used independently. However, they can also be combined to provide a more complete understanding of the linguistic, etymological and semantic analysis of Greek personal names.

The 'Meanings' catalogue[41] contains an extensive alphabetical list of the meanings of the bases, arranged in alphabetical order in English. For example, under the entry 'glory' (cf. fig. 6), one can notice that seven bases are associated with this meaning: δοκ-σ(α/ο), from δόξα (good opinion, fame, glory); ευ=δοκ-σ-ι(ᾱ), from εὐδοξία (good fame, glory); κλε(ϝ)-(ε)(σ), from κλέος (fame, glory); κῡδ-(εσ) and κῡδ-(ι/ο), from κῦδος (divine power, glory, fame); στεφ-αν(ο), from στέφανος (crown, glory); φα(ϝ)-(ε/οσ), from φάος (splendour, light, glory). This catalogue offers an instrument that can be used in a wide range of studies on Greek onomastics and lexical matters.

Fig. 6: The Entry 'glory' in the 'Meanings' Catalogue.

This catalogue can facilitate the study of personal names associated with a particular semantic field, such as 'glory' – as in the example – but also in relation to animals or plants, for instance. In the past, such research often required the analysis of the entire corpus of Greek personal names in a specific region, as in Vottéro's study of Boeotian names and their connection to the natural and social environment of Boeotia.[42] The catalogue provides a convenient means of examining these questions, offering direct access to the onomastic corpus via the *LGPN-Ling* website. An instrument such as this catalogue can help us explore these kinds of questions, examining how lexemes that are absent from the ancient Greek lexicon are used in anthroponymic onomastics. The significance of this approach only increases

39 Minon et al. (2023), 189.

40 Minon (2023), XXI–XXVI.

41 https://lgpn-ling.huma-num.fr/senses.html?category=d&search (last access 12.03.2025).

42 Vottéro (1993).

when considering that onomastics – particularly anthroponymy – often retains bases that are no longer used in the lexicon.

2.2 Connecting and Interconnecting

The *LGPN-Ling* website features a complex network of internal and external connections and interconnections. Each element of the analysis of a name (such as geolinguistic, bases, suffixes, morpho-syntax, lexical references, etc., mentioned above §2.1) is linked to the others and to other names that share at least one of the features of another name. The intricate web of connections between the individuals analysed on the *LGPN-Ling* website is readily apparent to a website user who navigates the 'Browse' interface's left-hand 'Faceted Research' option. One can easily see how each name is linked to multiple features and other names, creating a complex network of relationships.

Moreover, since its initial conception phases, *LGPN-Ling* has been designed as connected to the Oxford's *LGPN* interface. In January 2023, when *LGPN-Ling* was open to the public, two separate *LGPN* search interfaces existed: *LGPN Database Search* and *LGPN Name Search*. At the time, a connection from *LGPN-Ling* to *LGPN Database Search* and an interconnection with *LGPN Name Search* were created, not only for the entry name of *LGPN-Ling*, but also for the variants. The ultimate objective was to enable users to access all *LGPN* resources, regardless of their initial search. However, due to the development of a new user interface for the Oxford's *LGPN*, which is currently in beta testing,[43] these links and relationships are temporarily unavailable. Nevertheless, the interoperability of the two systems remains a central goal of the project. It aims to provide users with direct access to both the full onomastic data and their linguistic analysis. Indeed, these interconnections between the two interfaces will allow linking the general linguistic analysis of a name (*LGPN-Ling*) to its particular manifestations borne by unique individuals (*LGPN Database Search*). Furthermore, an interconnection also exists between *LGPN-Ling* and *Trismegistos People*. To facilitate a comprehensive study of Greek personal names and to lay the groundwork for interoperability among databases specialising in the same field, a network of interconnected databases has thus been developed.

3. *Nomina omina* and Beyond

In this final section of our paper, we will explore the notion of *nomen omen* in connection with *LGPN-Ling*. We will do so by examining two perspectives. Initially, we will highlight the distinction between analysing an anthroponym and an idionym. We will then go beyond the concept of 'name entity' to concentrate on the morphological entities that arise in the linguistic analysis illustrated on the *LGPN-Ling* search site. These two examples will allow us to examine how the *LGPN-Ling* team's research and the project's outcomes can be applied and reused in other scientific studies on Greek personal names.

3.1 The Individuals behind the Names

If both the *LGPN-Ling* website and the *Lexonyme* volumes allow the linguistic analysis of a name as an abstract entity, neither provides systematically[44] a detailed analysis of the idionym, meaning the name as carried by a specific individual. As reminded by Minon,[45] this is partly due to the lack of information about most of the individuals. To truly comprehend why someone chooses a particular name, one must delve into their personal history. Who were their relatives? Where did they grow up?

43 https://search.lgpn.ox.ac.uk/index.html (last access 14.03.2025).

44 This functionality is being implemented, especially for *hapax*.

45 Minon (2023), XVII.

What was their profession? Despite this, many crucial aspects of an individual's past, such as the historical, cultural, societal, and spiritual environments that shaped them, may be lost to time. However, these elements are essential for understanding the naming process. They can also affect the linguistic, especially semantic, interpretation of a name, particularly if multiple interpretations are possible.

A linguistic analysis of a name can thus yield more than one plausible interpretation. Occasionally, it proves impossible to choose between two competing hypotheses. One interpretation may, in fact, apply to one person, while another is suitable for another person with the same name. Additionally, the linguistic analysis of a name can be problematic in itself and yield several acceptable interpretations. In order to take into account the possibility of multiple hypotheses, the *LGPN-Ling* interface allows the presentation of more than one linguistic analysis for the same name. In such cases, the different hypotheses are thus hierarchically presented, from the more likely to the less likely.[46] However, it is important to note that, from the start of the project, the in-depth study of Greek personal names has enabled the *LGPN-Ling* team to significantly reduce the number of alternative interpretations on the website.

However, understanding the reasons behind the attribution of a name to an individual does not always require multiple linguistic hypotheses. For example, in the case of theophoric personal names, the reference to a deity can, in fact, be made for different reasons. The personal names, including an element Δη/ᾱλιο° are well attested in the Greek world, due to the connection to the cult of Apollo Delian. In the area of Tanagra (Boeotia), we encounter several personal names featuring the element Δᾱλι(ο)° (non ionic-attic form of Δηλι(ο)°), such as masculine Δᾱλιό-δωρος[47] and feminine Δᾱλι-κκώ. In this case, the existence in the area of Tanagra of a sanctuary consecrated to Apollo Delian[48] could imply that these names were chosen, having in mind the cult of Apollo located in the vicinity of their city, and not the great sanctuary situated on the island of Delos. Therefore, if we do not have to propose two distinct hypotheses on *LGPN-Ling*, where the base 1 consistently represents the same linguistic form (Δᾱ/ηλιο°), one can argue that a Tanagrean Δᾱλιό-δωρος would likely have connected the epithet part of his name to the local Apollon Delian. As a consequence, the deeper understanding of an individual and the reasons behind the naming process do not always imply the necessity of suggesting a different linguistic interpretation of a name. The external references connected to a base – as in this instance – can, in fact, variate without having any impact on the linguistic comprehension of a personal name.

Occasionally, a pattern in the naming process might be discernible, but the motivations behind the particular choice might remain a mystery. This is the case of the slave names in Σω-, whose linguistic analysis is not problematic in itself as much as their distribution. Indeed, several scholars have already mentioned names, such as Σωσίᾱς, as being particularly common among the slaves, despite not being exclusive to them.[49] Masson, following Lambertz,[50] included them in the category of the 'auspicious names' ('Wunschnamen', following Lambertz, 'noms de souhait', according to Masson). The same conclusions were more recently shared by Kanavou, [51] who remarks that these names must have been perceived as common names for the slaves. Moreover, the recent publication of the corpus of the Delphian manumission records allowed Mulliez[52] to identify a similar pattern among the slaves freed

46 The printed volumes, Minon et al. (2023), also take into account this possibility. As a consequence, in such instances, the name has been presented under different entries.

47 The references for these names can be found in the volume 3b of the *Lexicon of Greek Personal Names*, as well as on the *LGPN* search site (https://search.lgpn.ox.ac.uk/index.html [last access 29.08.2025]).

48 Schachter (1981), 45–47.

49 Masson (1973), 15.

50 Lambertz (1907).

51 Kanavou (2011), 202.

52 Mulliez (2023), 301–303.

in Delphi: 15% of the slaves carried a name in either Σω° (< Σαϝο°, σῶς [intact, safe, healthy]) or Σωσι° (< Σαϝωσι°, σῴζω [to keep safe, to save]). However, if these names were often used for slaves due to their positive and auspicious semantic, why were they so common compared to other 'auspicious names'? As for many other names, the deep reasons behind the high frequency of names in Σω° and Σωσι° for slaves still elude us.[53] It is nevertheless clear that – without being exclusivity used for slaves – these names could be somehow perceived as 'slave names'.

3.2 Morphological Entities

If by *nomen omen* one refers to the name conceived as a unique entity, it is clear that the structure and the analysis illustrated on the *LGPN-Ling* website include a further segmentation of this name entity in even smaller units. Among these 'smaller entities', two allow the comprehension of a personal name through its morphological features: the bases (or compounds) and the suffixes. The global understanding of these elements has greatly increased, not only due to the research conducted for this project, but also thanks to the search engine itself.

Indeed, although creating the database required some compromises (§1.2), using this 'strict' structure often facilitated the identification of facts that could only be observed with such a database at hand. The case of the -ώνδᾱς/-ώνδης suffix provides a clear example of this phenomenon. The origin of the derivational suffix -ώνδᾱς, mostly attested in central Greece, has not been unanimously explained by scholars.[54] One of the key issues is the relationship between names ending in -ώνδᾱς and those ending in -ωνίδᾱς, meaning: do the names in -ώνδᾱς correspond in fact to those in -ωνίδᾱς? Since the organisation of the database demanded to fit the derivational suffixes within separate columns, it was possible to notice how several names in -ώνδᾱς had to be analysed exactly as those in -ωνίδᾱς. Therefore, it was decided to combine them under one entry. Furthermore, the 'Geolinguistics' column allows for observation of a geographical and dialectal distribution. For example, in places where -ώνδᾱς is not found, -ωνίδᾱς is used (Boeotian Δαιτώνδᾱς next to an Attic Δαιτωνίδης). This suggests that we are probably dealing with two variants of a single chain of suffixes.

If these morphological features can be completed – in a TEI-xml context – through the markup process, they can also live on their own. Indeed, the suffixes are one of the elements presented in the faceted search on the 'Browse' section of the website. Currently, the research of the suffixes is available only for the entry name. It is conceivable to enhance the functionality of this feature, particularly in relation to gender-specific variants. These may have distinct suffixes, such as the feminine variant of Ἄγν-ων, which is listed as Ἀγν-ώ.[55] In such cases, refined searching capabilities can help locate the root term Ἄγνων, while filtering out names characterised by a suffix-ων. On the other hand, a faceted search for the suffix -ώ(ι) will not allow users to find the feminine variant Ἀγνώ. The development of such a functionality could ultimately lead to creating an additional catalogue. Indeed, since the markup of the bases and the meanings contributed to the creation of the corresponding catalogues (§2.1), a similar tool would be also feasible for suffixes. Therefore, all the morphological (and lexical) entities characterising Greek personal names could exist on their own within the *LGPN-Ling* interface, allowing for further studies. Just as the current catalogues have been instrumental in analysing the use of

53 One could wonder if behind the choice of these slave names in Σω° and Σωσι° there would not be a word-play with σῶμα, due to the assonance between the initials in σω-. The word σῶμα developed progressively form at least the 5th c. BCE onwards, the meaning of 'slave' (cf. Chantraine [2009], *s.v.* σῶμα; Kretschmer [1928], 80–81). This interpretation remains nevertheless hypothetical at this stage.

54 On this suffix, see Garré (2022), 148–156.

55 https://lgpn-ling.huma-num.fr/Hagnōn (last access 21.05.2024). The organisation of male/female variants bearing respectively -ων, -ωνος and -ώ(ι) suffixes is due to the association of the two suffixations from the late Hellenistic period onwards.

bases, they could also help shed light on the use of suffixes and the derivational process of Greek personal names.

Conclusion

Due to the very specific nature of this linguistic material, the development of the *LGPN-Ling* search site had to take into account not only the synchronic analysis of the Greek personal names, but also their diachronic development. The *LGPN-Ling* team's committed work, under the direction and supervision of Sophie Minon, has achieved the goal of preserving not only Greek personal names, but also their complete linguistic analysis in a digital environment.

This research can count on several achievements, including the creation of a database that provides the clearest possible presentation of the morphological and semantic analysis of Greek personal names. It also allows for the possibility of multiple analyses that are likely to be true at the same time. However, since, in a database, the results must be presented on a single and unified level, it becomes more challenging to consider the analysis of a personal name simultaneously in both diachronic and synchronic contexts. This is why, as we have shown, the columns of 'Morpho-syntax' and 'References' permit the presentation, in abbreviated format, of the diachronic development of many of these names.

Moreover, each entry in the database has been connected and linked with the others in every aspect of its analysis. This includes gender, geolinguistics, bases, suffixes, morphosyntax, semantic and even lexical correspondences in Greek. As a result, the names are linked to all others in accordance with one or more of these features, thereby creating a complex network of interconnections between the registered forms. This becomes evident when utilising the research interface, which is structured on two distinct yet complementary levels: the full text and the faceted search. This design has resulted in the development of two separate catalogues, facilitating new perspectives on the analysis and study of Greek personal names. By implementing annotation standards for suffixes, including the variants, there is potential to create a specialised directory.

Furthermore, the connections and interconnections with other *LGPN* sites, which will be gradually improved over time, allow for a linguistic study of Greek personal names, as well as a direct access to the data itself. Moreover, the structure of the entries on *LGPN-Ling* is of great importance in order to identify all the dialectal variants of an anthroponym. Some of these variants may contain dialectal features, making it difficult for a user of *LGPN Database Search* to identify them. The hope is that the three *LGPN* sites, all developed by the same team, and aiming at being interconnected, will help preserve the sites and their onomastic data. It is thus thanks to this connection with *LGPN Database Search* that it will be possible, more comfortably, to reach the idionym – the *nomen omen* – and evaluating how (or which) linguistic analysis is applicable to a specific individual.

In conclusion, the design of the *LGPN-Ling* website, including its search tool and database, demonstrates an impressive commitment to preserving anthroponyms in their entirety. This is achieved by capturing every aspect that defines them, but from any element characterising them and, all the way down to the smallest entities, such as morphemes and bases.

Sources

Online sources

LGPN (Beta version): https://search.lgpn.ox.ac.uk/index.html (last access 20.03.2025).

LGPN-Ling: https://lgpn-ling.huma-num.fr/about.html (last access 31.01.2026).

References

Alsonso Déniz et al. (2017): A. Alonso Déniz / L. Dubois / C. Le Feuvre / S. Minon (eds.), La suffixation des anthroponymes grecs antiques, SAGA: actes du colloque international de Lyon, 17–19 septembre 2015, Genève 2017.

Bechtel (1917): F. Bechtel, Die historischen Personennamen des Griechischen bis zur Kaiserzeit, Halle 1917.

Chantraine (1933): P. Chantraine, La formation des noms en grec ancien, Paris 1933.

Chantraine et al. (2009): P. Chantraine / J. Taillardat / A. Blanc / O. Masson / J.-L. Perpillou / C. de Lamberterie, Dictionnaire étymologique de la langue grecque: histoire des mots, Paris 2009[2].

Dobias-Lalou / Dubois (1990): C. Dobias-Lalou / L. Dubois, Introduction, in: Masson (1990), I–XVI.

Dubois (2017): L. Dubois, La notion de dérivation «hypocoristique», in: Alsonso Déniz et al. (2017), 6–11.

Garré (2022): M. Garré, Morphologie des inscriptions béotiennes (fin VIII[e]-II[e] s. av. J.-C.), PhD dissertation defended at the École Pratique des Hautes Études (Paris), 10.12.2022.

Hallof et al. (2025): K. Hallof / Y. Kalliontzis / A. Charami, Inscriptiones Graecae Megaridis, Oropiae, Boeotiae. Editio altera, pars III. Tanagra et ager Tanagraeus, Berlin / Boston 2025.

Kanavou (2011): N. Kanavou, Aristophanes' Comedy of Names, Berlin / New York 2011.

Kretschmer (1928): E. Kretschmer, Beiträge zur Wortgeographie der altgriechischen Dialekte, Glotta (1928), 67–100.

de Lamberterie (2023): C. de Lamberterie, Hommage de M. Charles de Lamberterie, Comptes Rendus de l'Académie des Inscriptions et Belles-Lettres (2023), 1084–1086.

Lambertz (1907): M. Lambertz, Die griechischen Sklavennamen, Wien 1907.

Masson (1990): O. Masson, Onomastica Graeca Selecta 1, C. Dobias-Lalou / L. Dubois (eds.), Nanterre 1990.

Masson (1973): O. Masson, Les noms des esclaves dans la Grèce antique, Actes du Groupe de Recherches sur l'Esclavage depuis l'Antiquité 2/1 (1973), 9–23 (= Masson [1990], 143–161).

Minon (2017): S. Minon, Préfixes, suffixes et chaînes suffixales identifiés dans les anthroponymes, in: Alonso Déniz et al. (2017), 687–704.

Minon (2020): S. Minon, Le projet LGPN-Ling: analyse étymologique et sémantique des anthroponymes grecs antiques, Bulletin de la Société de Linguistique de Paris 115/1 (2020), 253–297.

Minon (2023): S. Minon, Introduction, in: Minon et al. (2023), I–XL.

Minon et al. (2023): S. Minon (dir.) / G. Genevrois / E. Nieto Izquierdo / F. Réveilhac / J.-C. Chuat, Lexonyme. Dictionnaire étymologique et sémantique des anthroponymes grecs antiques, Volume 1, A–E, Génève 2023.

Mulliez (2023): D. Mulliez, En marge du corpus des actes d'affranchissement delphiques. Quel nom pour un esclave ?, Journal des Savants (2023), 281–349.

Schachter (1981): A. Schachter, Cults of Boiotia 1: Acheloos to Hera, Bulletin Supplement 38/1 (1981).

Vottéro (1993): G. Vottéro, Milieu naturel, littérature et anthroponymie en Béotie à l'époque dialectale (VII^{e}-II^{e} s. av. J.-C.), in: E. Crespo / J. L. García Ramón / A. Striano, Dialectologica Graeca. Actas del II Coloquio Internacional de Dialectología Griega (Miraflores de la Sierra [Madrid], 19-21 de junio de 1991), Madrid 1993, 339–381.

Figure References

Fig. 1: The Presentation of the Name Ἀγλαοφαΐδας.

Fig. 2: The Analysis of the Name Ἀνδράγαθος.

Fig. 3: The Semantic Analysis of the Name Διοκύδης.

Fig. 4: The Name Ϝεργαένετος and its Variants.

Fig. 5: The Entry αυγσ(ι) in the 'Bases' Catalogue.

Fig. 6: The Entry 'glory' in the 'Meanings' Catalogue.

Author Contact Information[56]

Dr. Matilde Garré
École française d'Athènes
Didotou 6
106 80 Athina
Greece
Tel: 0033(0)769318087
E-mail: matilde.garre@gmail.com

56

DIGITAL CLASSICS ONLINE

Altinum: a *Wikidata* Project for Digital Epigraphy and Prosopography

Anna Clara Maniero Azzolini

Abstract: *Altinum* is the first Latin epigraphy project to be hosted on *Wikidata*, the collaborative database maintained by the Wikimedia Foundation. Launched in 2024, the project involves importing information about epigraphic artefacts from the Roman period originating in Altinum, a municipality in eastern Veneto, being also the first attempt of a digital prosopographic corpus of the municipality. The data have been sourced from *EDR*, *EDCS*, analogue catalogues, and unpublished theses. Once the data are integrated into *Wikidata*, users can formulate queries to generate graphs and tables, obtain statistics, and reconstruct family trees. As *Altinum* demonstrates, this approach not only expands the epigraphic *corpus* and prosopographical data but also makes it increasingly accessible in a collaboratively editable, multilingual and interdisciplinary database with the possibility of highly customisable queries.

Introduction

This paper presents the first publication on the *Altinum* project, which was conceived in 2024 as part of my Master's thesis at Ca' Foscari University.[1] The title was quite telling: *Altinum: a Wikidata Project for Digital Epigraphy*[2], focusing on a digital epigraphic study dedicated to Altinum, the municipal entity that originally gave rise to the Venetian settlement. Thanks to Professor Lorenzo Calvelli (Università Ca' Foscari) and Camillo Carlo Pellizzari di San Girolamo (Scuola Normale Superiore), the project developed to encompass a digital component, an epigraphic section, and a prosopographical survey. The aim was to catalogue inscriptions already published in existing epigraphic databases, transcode them, and channel the extracted data into a new database that has, for over a decade, demonstrated its potential across numerous fields of knowledge: *Wikidata*.

The rationale behind choosing a new database may not be immediately apparent. *Wikidata* serves as a data repository capable of providing customisable representations when queried by users. Moreover, it is an ecosystem that fosters collaborative, interdisciplinary, and multilingual engagement. Its structure enables a community to act, cooperate, correct, improve, query, compare, create, and delete. The advancement of humanities research, particularly the increasing interconnection between multiple fields of knowledge, necessitates the use of tools that can handle such a vast amount of information. Unlike traditional epigraphic databases, which provide only predefined queries and support a limited number of languages chosen by administrators, *Wikidata* overcomes these obstacles. The possibility of a project integrating multiple related fields of study is a goal we have pursued since the rise of the Digital Humanities[3]: one need only consider the challenges of conducting palaeographic or geological re-

1 Freely available and readable http://hdl.handle.net/10579/27792 (last access 11.07.2025).

2 https://www.wikidata.org/wiki/User:Anna_Clara_Maniero_Azzolini/Altinum (last access 11.07.2025).

search using existing epigraphic databases. The scholarly goal of *Altinum* is to assess the operability of *Wikidata* in the field of epigraphy, a domain where its application is still in its early stages.

A first part of this contribution will focus on a more detailed analysis of the internal structure of *Wikidata*, illustrating its technical components, advantages and even difficulties. The 'behind the scenes' of the *Altinum* project will then be explained in detail, outlining the obstacles and methodology adopted for the import of epigraphic and prosopographical data. Examples of RDF formulation and SPARQL queries will accompany the descriptions, as well as the prosopographical data modelling adopted for the project.

Architecture of *Wikidata*

Premise

Wikidata is a collaborative, multilingual knowledge base designed for both humans and machines.[4] It organises data to make it easily searchable and analysable. A logical statement such as 'Altinum is located in Veneto' contains data and relationships between them: Altinum (subject) is located in (predicate) Veneto (object). Data are then composed within the knowledge base to form an indissolubly logical connection (in *Wikidata*, a statement such as 'Altinum is located in stone' would not establish a logical subject-predicate-object relationship and would therefore be flagged as a constraint violation). Structured data, therefore, can include any kind of information, provided it follows a logical structure. The knowledge base can be queried to extract specific information and establish connections.

Another key feature of *Wikidata* is its multilingual nature. Unlike other databases, this resource is highly accessible, currently supporting 621 recognised languages.[5] This is unsurprising, as *Wikidata* is designed as a collaborative knowledge base that grants users full control over resources and data management. This multilingual approach is a necessary factor in fulfilling one of its primary objectives: accessibility. *Wikidata*'s collaborative model means that decisions and changes are driven by community contributions. This system ensures continuous data verification and, crucially, easy modification. If a user were to enter the incorrect statement 'Caesar was a king' into *Wikidata*, another user could quickly correct it to *'Caesar was a dictator'*. A more restrictive control system would slow down edits, increasing the risk of errors spreading.

Wikidata's utility extends beyond Wikimedia projects, as many external websites and databases draw upon its data. An institution might initially decide to import data from an external database into *Wikidata* and then extract it in an analytical format in order to use it in a new portal or database. However, few institutional projects focus exclusively on contributing data to *Wikidata* as an end in itself. Consider, for example, the epigraphic project IDEA, which compiles and analyses inscriptions from the archaeological site of Dura-Europos in Syria. Its creators agreed to use *Wikidata* as a tool for developing a new, independent website[6]: they extracted data from digital platforms, primarily within the academic domain, and imported them into *Wikidata* in a structured format according to a specific data model, before exporting them in an analytical format to IDEA.[7] Similarly, *Altinum* extracts data from

3 On the use of *Wikidata* in the Digital Humanities Zhao (2023) and, in the field of epigraphy, Orlandi (2021); Lorito (2018); Heřmánková et al. (2022).

4 See in particular the analysis by Kaffee et al. (2017). On *Wikidata* as a tool for research Mora-Cantallops et al. (2019).

5 https://www.wikidata.org/w/api.php?action=query&meta=wbcontentlanguages&wbclcontext=term&format=json&formatversion=2 (last access 11.07.2025).

6 https://duraeuroposarchive.org/ (last access 11.07.2025).

7 Thornton et al. (2024).

the epigraphic databases *EDR* and *EDCS*, as well as from analogue sources[8], to import them into *Wikidata* with the aim of obtaining structured data, structured queries, and ultimately contributing to a broader epigraphic project within this environment. Special mention must be given to Pietro Ortimini's project, *Greek Metrical Inscriptions (GMI)*[9], which catalogues Greek metrical inscriptions using the Wikibase software that underpins *Wikidata*.

From a licensing perspective, *Wikidata* is a free and open knowledge base, and its data are licensed under Creative Commons CC0.[10] This means that the data can be fully and freely reused. The only condition to massively import data into *Wikidata* is that the original source of the data must comply with the same licensing rule or grant *Wikidata* the legal right to operate under these terms.[11] This was precisely the case for the *Altinum* project: due to the incompatibility between *EDR*'s licence CC BY-NC-SA 4.0 and *Wikidata*'s CC0, it was necessary to obtain written permission from *EDR*'s management to export and import data. The data were then deposited in Zenodo under Creative Commons CC0[12] by *EDR* administration, which was cited as the source in the project to ensure both the legal and scholarly validity of the imported data within the user community.

Data Linking: Resource Description Framework

Extracting information is the primary reason why most users register data within a database. But how can this data be effectively retrieved? Let us imagine that we have entered details into *Wikidata* regarding archaeological artefacts from a museum. Our database would include the identifying name of each artefact, the date of discovery, the name of the archaeologist who supervised the excavations, and other relevant information, all interlinked. Depending on their needs, a user may wish to extract all artefacts discovered in a given year, those excavated by a specific archaeologist, or those belonging to the same museum.

The type of web that *Wikidata* belongs to is known as the Semantic Web, an extension of the broader web. In this environment, information within a text is encoded using ontologies and precise rules provided to the database[13], ensuring that all data are machine-readable. The semantic web can leverage the Linked Open Data system (LOD)[14], enabling data sharing and reuse. In 2006, Berners-Lee outlined the principles of Linked Data, emphasizing the use of URIs (Uniform Resource Identifiers) to uniquely identify resources. These URIs should be resolvable via HTTP, allowing standardized access to data. Additionally, data should be published in a structured, machine-readable method such as RDF (Resource Description Framework), and datasets should be interlinked to enable the construction of a global network of interconnected data, facilitating the expansion of available information. This framework is viable only if all nodes within the network meet the following conditions [15]:

8 Pivetta (1997/1998).

9 https://greek-metrical-inscriptions.wikibase.cloud/wiki/Main_Page (last access 11.07.2025).

10 https://www.wikidata.org/wiki/wikidata:Copyright (last access 11.07.2025).

11 https://www.wikidata.org/wiki/wikidata:Data_donation (last access 11.07.2025).

12 https://zenodo.org/records/11530904 (last access 11.07.2025).

13 Möller / Heath et al. (2007). In Digital Humanities Hyvönen (2020).

14 Hyvönen (2020), 3 tab. 1; Erxleben et al. (2014). In epigraphy: Tupman (2021).

15 Middle (2024).

- They are available on the web in any format and under an open licence.
- They are structured and machine-readable.
- They use a non-proprietary format.
- They employ RDF and SPARQL.
- They properly identify entities.

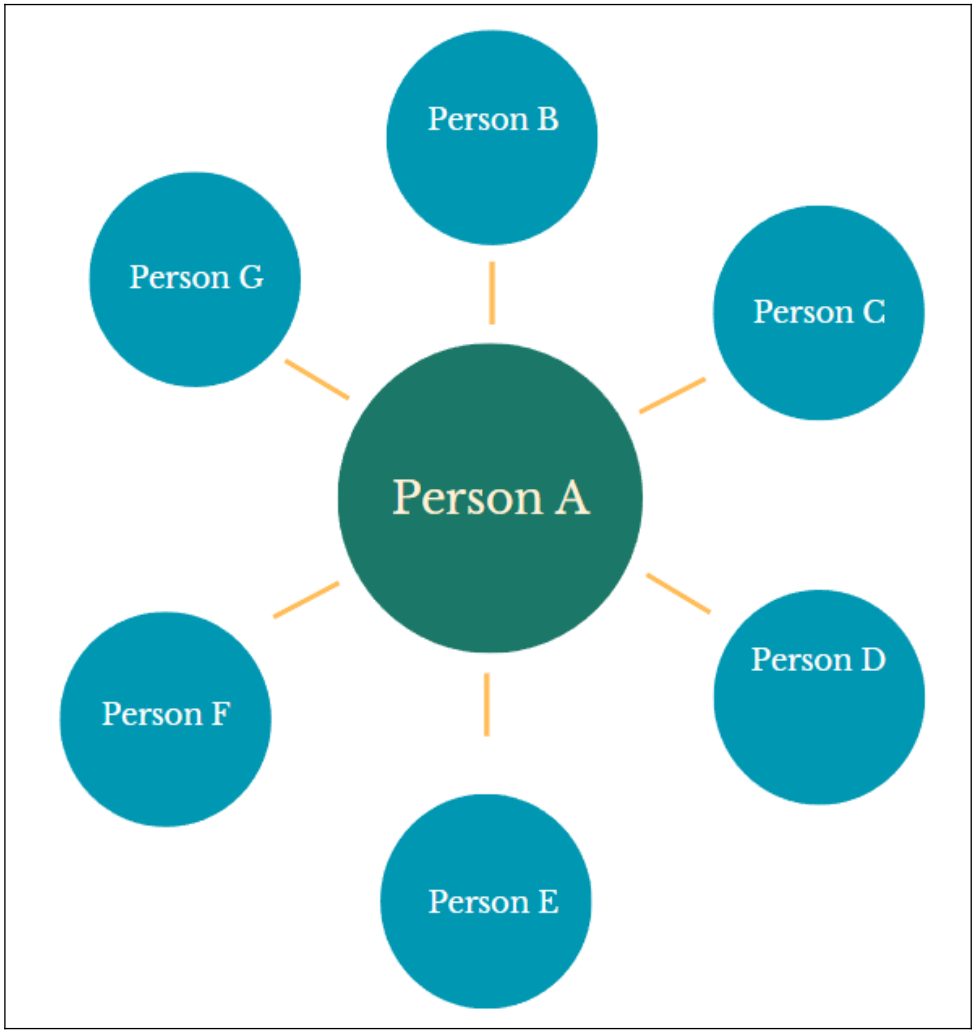


Fig. 1: Example of friendship graph.

To better understand how data are interrelated according to the RDF standard, a friendship relationship graph is often used as an illustrative example.

Each entity is connected by a link that identifies its relationship to another: Person A is a friend of Person B. However, if we were to structure such information in the network, we would have to resort to a machine-readable and therefore always valid scheme. In order to do so, a structure with logical meaning is required, i.e. one consisting of a subject, a predicate and an object. 'Person A is a friend of Person B' will thus be the minimum proposition or statement that cannot be further broken down without losing its logical and declarative functionality. Person A will be the subject, Person B the object while 'is friend of' will act as a bridge between the two, as a link in the direction Person A → Person B, a predicate making explicit the relationship between the former and the latter. Without a single element of what is called a 'triple' in computer language, the machine would not be able to interpret any information (and neither would a human). The minimum fundamental structure is represented below:

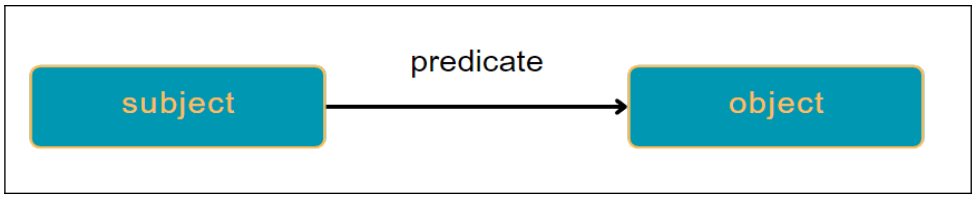


Fig. 2: Logical structure of an RDF triple.

At this point, it will be easily understood that the word 'person' alone does not constitute a machine-readable element of the declaration but is merely human-readable. This word must be matched by a readable code, i.e. a URI (a document, a postal code, a Uniform Resource Name or URN such as a particular ID or namespace defined by a programmer, or an ISBN: in short, anything that uniquely identifies a node and is machine readable[16]).

Virilis is a significant person for Publicia Amabilis:

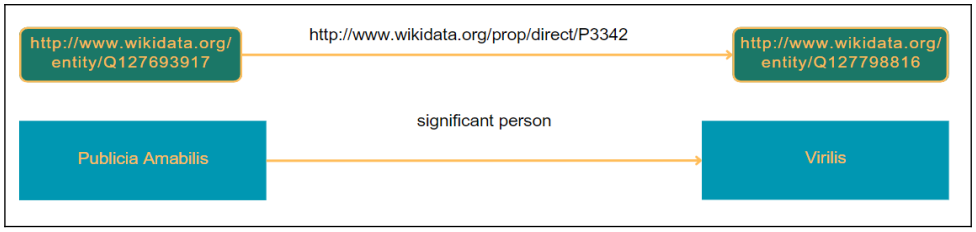


Fig. 3: Example of an RDF triple.

Overview of the Data Model (Items, Properties, Statements)

Having so far briefly described the RDF structure, it will be appropriate to present how it works within *Wikidata*, and thus how data are hierarchized not from the point of view of interconnections but from the ontological point of view. Let us imagine that we wish to enter prosopographical information on individuals who lived in a given *municipium* of the Roman era: this will include data such as name, era, friends and family, occupations and so on. Such data will need to be entered into *Wikidata* in a specific data model, collaboratively developed by the *Wikidata* community itself. First, an item describing the character must be created (a special page 'create a new item).[17] The item defining the individual will then be created along with its name (label) in whatever language or graphic system we want, and a summary description also in whatever language or alphabet we want. The item will be identified by an alphanumeric code consisting of the letter Q and an increasing number: this code is both a machine-readable and unambiguous identifier that distinguishes possible homonymy between several items. It is also possible to include alongside the item description its aliases, with which we imagine another user could search for its name.

From the creation of *Wikidata* on 29th October 2012 to 4th February 2013, an item only possessed the characteristics described so far, with additional links to any Wikipedia articles on the topic (sitelinks). Just four months after its creation, in 2013, the situation in the database changed considerably. New components were added: statements, i.e. information structured in triples concerning a given item ('Publicia Amabilis lived in Altinum'). Since then, it has been possible to add qualifiers and references to the statements (there is for example the possibility of entering a property as 'source of the statement', P248).[18]

16 Note that in an RDF triple, the subject and predicate must always be URIs, while the object may or may not be.

17 https://www.wikidata.org/wiki/Special:NewItem (last access 11.07.2025).

18 References also follow the logic of the subject-predicate-object triple or, better, item-property-value, but also property-property-value or lexeme-property-value. In the case of references, the subject is the reference itself, predicates are the various properties (one or more) used within it, and objects are the values of these properties.

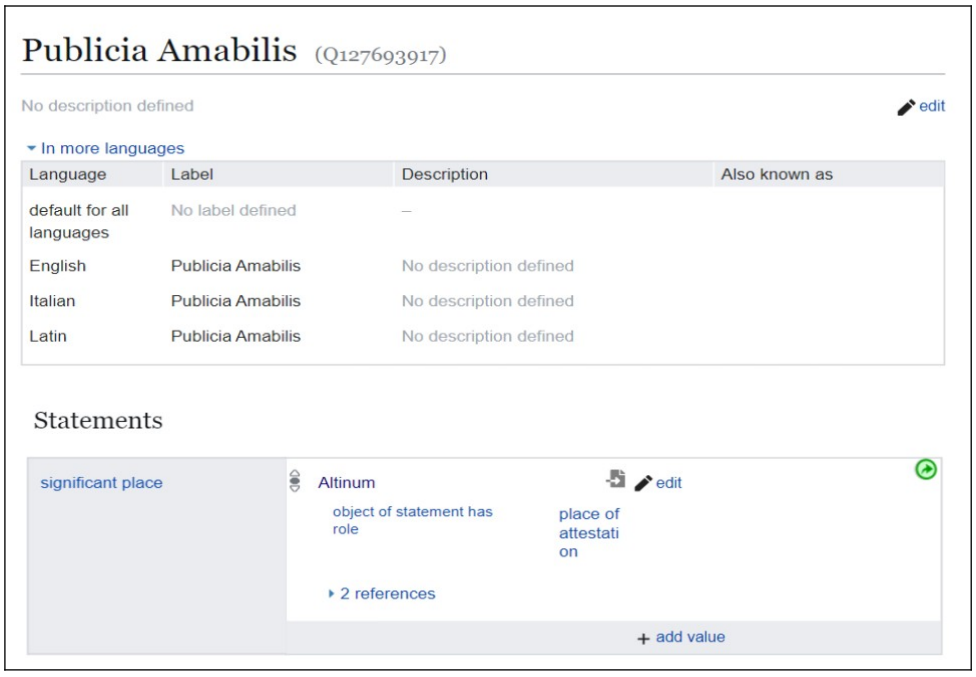


Fig. 4: Example of the layout of the statement 'Publicia Amabilis lived in Altinum'.

Tab. 1 provides a concise list of some of the properties in *Wikidata* that are particularly pertinent from a prosopographical perspective.[19] As can be seen, the property identifier consists of an alphanumeric code preceded by the letter P instead of Q. Whereas Q identifies an item, P identifies the property, i.e. the predicative function of our statement. However, this is not the only difference between the two. A further distinction emerges in the processes of creation and utilisation: items can be created arbitrarily by any user, whereas properties require proposal and can only be created by designated creators and administrators.[20] For example, to declare that an individual 'is a friend of' another in *Wikidata*, one uses the generic property 'significant person' (P3342) with the qualifier 'object of statement has role' (P3831) or 'subject has role' (P2868), specifying 'friend' (Q17297777) as a value. The *Wikidata* community usually prefers generic properties, avoiding 'friend of' in favour of the more versatile 'significant person'. An organized system facilitates queries whilst inconsistencies in the data modelling make them more difficult to write.[21]

***Wikidata* property**	***Wikidata* English label**
P31	instance of
P21	sex or gender
P569	date of birth
P3342	significant person
P106	occupation

19 Sandboxe items are useful for gaining experience in structuring data. There are some, e.g. https://www.wikidata.org/wiki/Q4115189 (last access 11.07.2025).

20 The page for requesting the creation of a property is: https://www.wikidata.org/wiki/wikidata:Property_proposal (last access 11.07.2025).

21 If the property 'friend of' existed, one user could use it, while another could enter a friend as 'significant person'. Whoever queries the database would have to consider both properties, risking overlooking one of them and altering the result. For further discussion, see https://www.wikidata.org/w/index.php?title=wikidata:Events/Data_Quality_Days_2022/Modeling_data&oldid=2018180212 (last access 11.07.2025).

P3831	object of statement has role
P248	stated in
P25	mother

Tab. 1: Selection of properties.

The Role of SPARQL in Data Querying: the *Wikidata* Query Service (WDQS)

Following the entry of data, the objective of *Altinum* is to query the database in order to conduct statistical and prosopographical investigations and to obtain visualisations of the data using to the wide variety of graphs available. The *Wikidata* Query Service[22] reaches its full potential when handling large-scale investigations. With two caveats: the variable of incompleteness of the result (e.g. if not all the data have been entered into the database) and the time factor (i.e. the more complex the bindings in the syntax of our statement are, the longer the query will take to extract the data[23]). Although a query builder is available[24], it is more efficient to interact directly with the machine when dealing with large datasets. It is important to note that, in order to achieve this, a fundamental understanding of the RDF standard alone is insufficient. It will be necessary to have a language through which to tell the database when we are talking about an item, a property, a reference and so on. The language that will be needed to extract data by querying the WDQS is SPARQL[25].

The following example illustrates the essential components of the SPARQL language through a prosopographical query. The query investigates the records of 'persons who lived in Altinum and held the office of military tribunes'. First of all, a SPARQL query necessitates fundamental components; in this instance, the most common SELECT and WHERE will be utilised. Specifically, SELECT denotes the variables to be returned, in this case 'person', with each variable preceded by a question mark.[26] The initial line of the query reads as follows[27]:

```
SELECT ?person
```

The WHERE section tells the database the meaning of the variables in the SELECT section and the conditions to be fulfilled in order to be extracted as results. In our case, therefore, the variable ?person must fulfil the two conditions 'held the office of military tribune' and 'lived in Altinum'. Curly brackets open and close the section.

```
SELECT ?person
WHERE
{
}
```

22 https://query.wikidata.org/ (last access 11.07.2025).

23 Since its inception, WDQS has always had a timeout limit of 60 seconds: thus, if the query does not succeed in less than one minute, it fails.

24 Which, as the page itself mentions, is 'ideal for users with little or no experience in SPARQL', https://query.wikidata.org/querybuilder/?uselang=it (last access 11.07.2025).

25 For an official guideline see https://www.w3.org/TR/sparql11-query/ (last access 11.07.2025). It was first introduced in 2008 by the World Wide Web Consortium (W3C).

26 Or $. Variables are named arbitrarily by the creator of the query, on the sole condition that they are preceded by ? or $ and do not contain the same symbols inside.

27 The subject is unified by convention in the singular.

Within the curly brackets, the two statements (the conditions) are formed into subject-predicate-object triples. In order to make the whole sentence machine-readable, it is necessary not only to use the identifiers of the properties and items, but also to respect an order within the sentence and to prefix the identifiers to indicate their role. The prefixes 'wd:' and 'wdt:' mark the object and the predicate respectively as known components of the triple[28] (in our case the subject ?person is a variable); if the object is also unknown, it must also be entered as a variable. Each triple is followed by a full stop (or by a semicolon if the query contains multiple triples with the same subject).[29] An example of a query with an unknown subject is[30]:

```
SELECT ?person
WHERE
{
?person wdt:P39 wd:Q849288; # Person holding military tribune office
wdt:P7153 wd:Q441542. # Person with significant place Altinum
}
```

While an example of a query with an unknown object is[31]:

```
SELECT ?position
WHERE {
wd:Q127694103 wdt:P39 ?position. # Position held by Manius Titius
}
```

By then including the ?personLabel in the SELECT section, the query service will return not only the identifier but also a human readable label. The user will also be able to indicate the preferred language of the result: with [AUTO_LANGUAGE] we will search all results in the interface language of WDQS. If English is our interface language, WDQS will extract the label 'en' if it exists, or otherwise extract the item's QID. In place of [AUTO_LANGUAGE], the language of preference may be entered between the inverted commas.

```
SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE]". }. #
Get labels
```

And thus, the final query:[32]

```
SELECT ?person ?personLabel
WHERE
{
?person wdt:P39 wd:Q849288; # Person holding military tribune office
wdt:P7153 wd:Q441542. # Person with significant place Altinum
```

28 wdt: links the item to the value (or values) that the property takes, considering only the value with the best rank and excluding the deprecated ones. https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format (last access 11.07.2025).

29 Accompanying the query is an explanation preceded by #, which is ignored by the query.

30 URL to the query https://w.wiki/DRFp (last access 11.07.2025); URL to the result https://w.wiki/DRFq (last access 11.07.2025).

31 URL to the query https://w.wiki/DPxq (last access 11.07.2025); URL to the result https://w.wiki/DPxt (last access 11.07.2025).

32 URL to the query: https://w.wiki/DRFy (last access 11.07.2025); URL to the result: https://w.wiki/DRFz (last access 11.07.2025).

```
SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE]". } #
Get labels
}
```

Or by changing language to Latin:[33]

```
SELECT ?person ?personLabel
WHERE
{
?person wdt:P39 wd:Q849288; # Person holding military tribune office
wdt:P7153 wd:Q441542. # Person with significant place Altinum
SERVICE wikibase:label { bd:serviceParam wikibase:language "la". }
}
```

Admittedly, this is a complicated language for information that we would be able to obtain in a much simpler way from the *EDR* database from which the information is retrieved. However, such a query is constructed according to the minimal basic structure needed to extract a much larger amount of data. An example is 'find inscriptions from Altinum that mention individuals belonging to the *gens* Iulia who are freedmen'[34]:

```
SELECT DISTINCT ?inscription ?inscriptionLabel ?person ?personLabel WHERE {
 ?inscription wdt:P31/wdt:P279* wd:Q1640824;
      wdt:P1071 wd:Q441542;
      wdt:P6568 ?person.
 ?person wdt:P7153 wd:Q441542;
       wdt:P5025 wd:Q127693269;
       wdt:P3716 wd:Q841571;
       wdt:P1343 ?inscription.
            SERVICE wikibase:label { bd:serviceParam wikibase:language
"[AUTO_LANGUAGE],mul,en". }
}
```

Another example is 'find significant persons related to Caetronia Maxima'[35]:

```
SELECT ?person ?personLabel WHERE {
 ?person wdt:P3342 wd:Q127693919 .
SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

The results can be visualised in graphs, grids, tables, maps, as well as in animated charts where the characteristics of each entity can be explored by moving the cursor over them:

33 URL to the query: https://w.wiki/DRF$ (last access 11.07.2025); URL to the result https://w.wiki/DRG5 (last access 11.07.2025).

34 URL to the query https://w.wiki/DPv9 (last access 11.07.2025); URL to the result https://w.wiki/DPvB (last access 11.07.2025).

35 URL to the query https://w.wiki/DPvC (last access 11.07.2025); URL to the result https://w.wiki/DPvF (last access 11.07.2025).

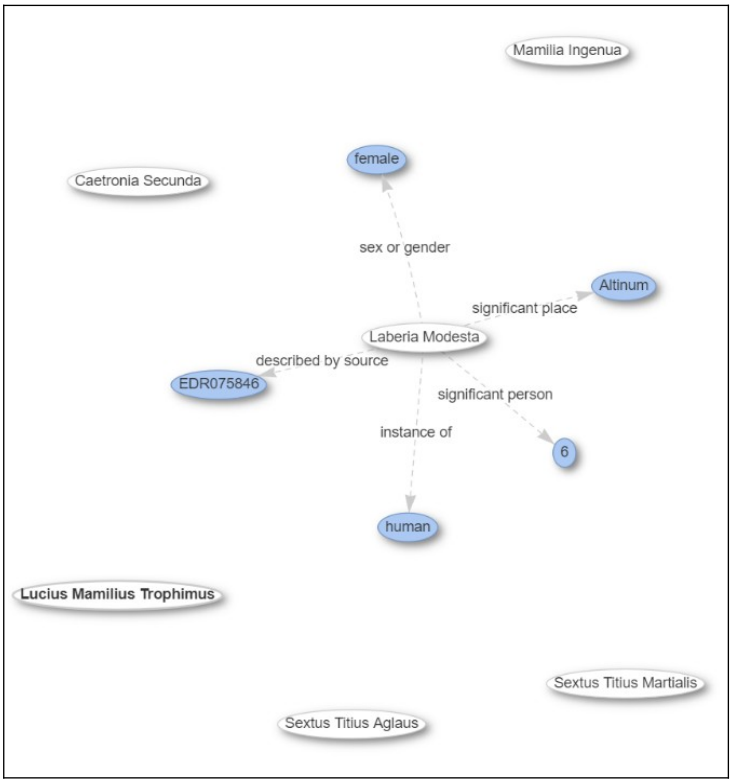


Fig. 5: Example of an animated graph (result to Caetronia Maxima query).

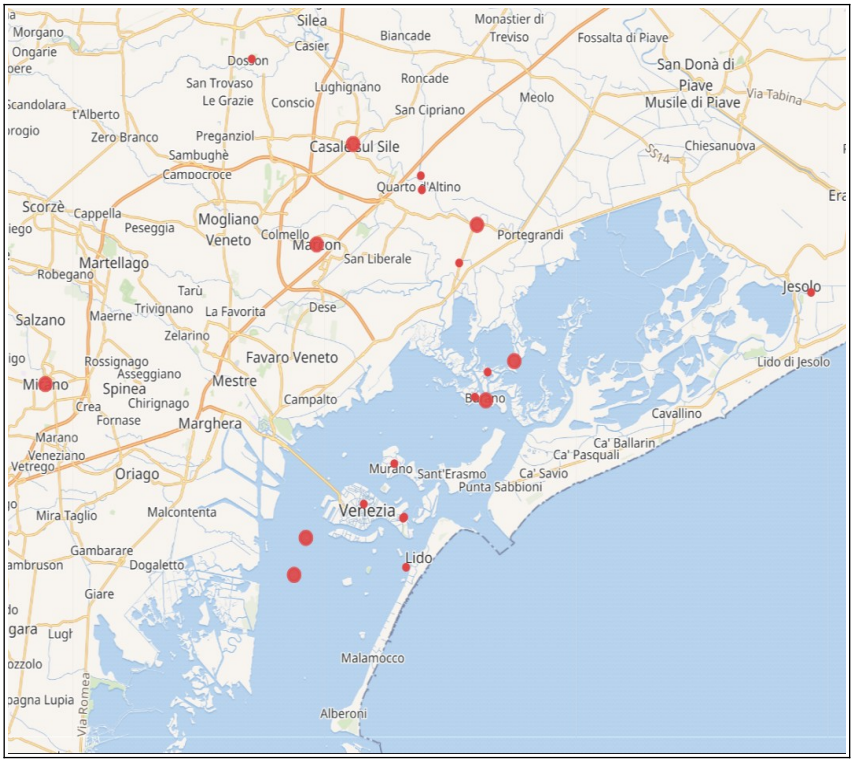


Fig. 6: Map of the sites of discovery of Altinum inscriptions. URL to the query https://w.wiki/DRHL (last access 11.07.2025); URL to the result: https://w.wiki/DRHN (last access 11.07.2025).

Behind the Scenes of the *Altinum* Project

At this stage, it is useful to outline the technical procedures that enabled the integration of epigraphic and prosopographic data from the inscriptions. As mentioned above, the *Altinum* project retrieved data from other databases, from unpublished analogue catalogues and from my personal examination of inscriptions to extract prosopographic information. With the help of tools for mass editing, the massive data import was successful: out of a total of 644 inscriptions, 665 individuals and 218 *gentes* imported,[36] almost 30,000 changes were made in *Wikidata*, a mass of data that would have made the manual entry process longer than necessary. The description of the import process is intended to demonstrate the replicability of the process in other contexts and projects, even from other databases. As said, *Wikidata* guarantees good reusability thanks to its CC0 licence, which allows free use of the data without restrictions: however, this licence is not always compatible with databases with more restrictive licences, requiring specific solutions for data integration.[37]

The *Altinum* project seeks to apply the FAIR principles (Findable, Accessible, Interoperable, Reusable) to the modelling of epigraphic data, taking advantage of the flexibility and accessibility of *Wikidata*. However, the lack of a unified controlled vocabulary and the absence of established epigraphic standards require further alignment to improve the interoperability of the dataset with other digital resources.

Entering Epigraphic Data

The first step was to extract the data into a CSV format. Data from *EDR* and *EDCS* had to be made compatible with the *Wikidata* ontology. To do this, another CSV was prepared containing the data to be transcoded. For each *EDR* field, a property in *Wikidata* was identified, or its creation was proposed to the community. An example of this is the property 'writing technique' (P12876), which corresponds to the *EDR* field *scriptura*: the properties 'writing system' (P282) or 'writing style' (P9302) were not at all relevant to epigraphy. The creation of the items in *Wikidata* required an effort of reflection from an ontological point of view, although it did not require prior discussion with the community.

Once the items and properties had been determined, it was possible to create statements, which govern the logical discourse in *Wikidata*, consisting of a property and a value (a QID or a text string if the property allows it or a URL etc. according to the property's datatype). Items about concepts in *Wikidata* should possibly be based on controlled vocabularies, in order to assure a consistent understanding of each term. The reference vocabulary for the project was therefore that of EAGLE (Europeana Network for Ancient Greek and Latin Epigraphy)[38] which, although extensive and epigraphically quite satisfactory, does not always seem to correspond the unambiguous understanding of its terms by the wider community (especially those not versed in epigraphy) nor does it follow either the ISO standard for multilingual thesauri (ISO 25964-2:2013) or the IFLA guidelines.[39]

Another problem was the translation of the *EAGLE* terms, which coincided with the *EDR* vocabulary, into *Wikidata* entities: fortunately, many had already been created and existed in *Wikidata*, and only in

36 As of 13.03.2025. The individuals catalogued in the project are actually more than 850, but those not mentioned in inscriptions published in *EDR* or *EDCS* were not included in *Wikidata* yet. URLs to the queries: (*gentes*) https://w.wiki/DRwc (last access 11.07.2025), (inscriptions) https://w.wiki/DRwb (last access 11.07.2025), (persons) https://w.wiki/DRwg (last access 11.07.2025). URLs to the results of queries: (*gentes*) https://w.wiki/DRwj (last access 11.07.2025), (inscriptions) https://w.wiki/DRwk (last access 11.07.2025), (persons) https://w.wiki/DRwi (last access 13.03.2025).

37 If a project requires a different licence, it will still be possible to create other *Wikibase* instances, which allow one to choose a licence compatible with any requirement.

38 https://www.eagle-network.eu/resources/vocabularies/ (last access 11.07.2025).

39 https://www.iso.org/standard/53658.html (last access 11.07.2025); https://www.ifla.org/wp-content/uploads/2019/05/assets/hq/publications/professional-report/115.pdf (last access 11.07.2025).

a few cases was it necessary to intervene in order to organise the information more effectively.[40] For the definitions of the terms I created ex novo I often made use of the descriptions offered by the online *EDR* Handbook[41], where for each section and lemma the summary and sufficient characteristics are presented.

Among the technical challenges of the transcoding process, the chronological issue deserves special mention. The chronological system in *Wikidata* works in an obsolete way in the epigraphic context: The system rightly provides the use of the qualifiers P1319 ("earliest date") and P1326 ("latest date") and aligns with the standard adopted by *EDR*. In order to include an interval in *Wikidata* it is necessary to specify an intermediate date, using P571 ("inception"), which is not at all consistent with the epigraphic perspective, but rather with the mechanical reasoning of the computer, which, given two extremes, chooses an intermediate date. A clearer example follows.[42]

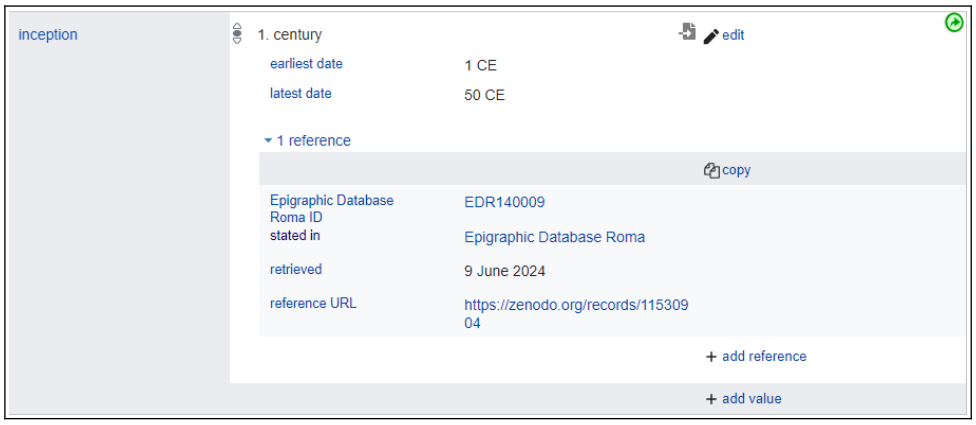


Fig. 7: Display of property P571 'inception' of *EDR*140009 (Q126687969).

As is frequently the case in epigraphy, it is indeed possible to refer to a specific time period in *Wikidata* without resorting to the use of extremes ('earliest date' and 'latest date'). However, it was necessary to adopt this solution in order to align as closely as possible with the data model of *EDR*, which relies on this fragmentation (without resorting to intermediate dates). Let us imagine that we want to search for all inscriptions datable to the 3rd century AD. We would have to resort to writing a date (not immediately accessible and understandable) according to the format of *Wikidata*: +200-01-01T00:00:00Z/9[43] as the earliest date and +300-01-01T00:00:00Z/9 as the latest date, where + stands for AD, T00:00:00Z indicates the time, which by convention is fixed at midnight UTC, and /9 indicates the precision of the dating: 6 - millennium, 7 - century, 8 - decade, 9 - year, 10 - month, 11 - day (and the expected, but not technically usable yet, 12 - hour, 13 - minute, 14 - second)[44]. The indication of precision is considered necessary for *Wikidata* coding and more faithful to the *EDR* standard, but not more searchable, given the complexity of the solutions adopted.

Concerning *EDCS*, the transcoding of its data was faster than that of *EDR* data. Primarily because *EDR* provides more diverse and larger quantities of structured data. Furthermore, unfortunately *EDCS* does not specify its license, which officially forbade me from importing its data into *Wikidata*. However, after creating a CSV file with all the entries not present in *EDR* (which were not downloadable from the site), it became evident that only a portion of them met the project's quality standards. This portion was small enough to avoid any concerns of copyright infringement. In fact, for the only in-

40 It should be noted that in some cases the items and properties had already been created by https://www.wikidata.org/wiki/wikidata:WikiProject_Epigraphy (last access 11.07.2025).

41 http://www.edr-edr.it/en/Guida_coll_en.php?lang=en (last access 11.07.2025).

42 https://www.wikidata.org/wiki/Q126687969#P571 (last access 11.07.2025).

43 Whereas in *EDR*, third century AD means the century starting from 201 AD.

44 https://www.wikidata.org/wiki/Help:Dates (last access 11.07.2025).

scriptions not found in *EDR* I used the *EDCS* ID[45], added as a label to the newly created entries, along with minimal data such as the dating and the limited terminology used to describe the inscription type. Other data were deduced a posteriori, such as certainly alphabet and language. In any case, it was stipulated that the source of each statement should always be cited and the link to the *EDCS* page included. The transcoding of *EDCS* data partly relies on the properties and items already established for *EDR*. It was decided not to import the bibliography either from *EDCS* or *EDR*: this is an operation that requires the input of thousands of items into *Wikidata* (as one can imagine, in order to cite an author and an article, it is first necessary to create all the metadata within the database), a time-consuming task, not necessary at the moment.

Once the fields and values of *EDR* had been transcoded into the corresponding properties and values of *Wikidata*, the next step was to import the data. For the massive operations, I used the tool designed by Magnus Manske QuickStatements (QS).[46] Examples of inscription pages follow.

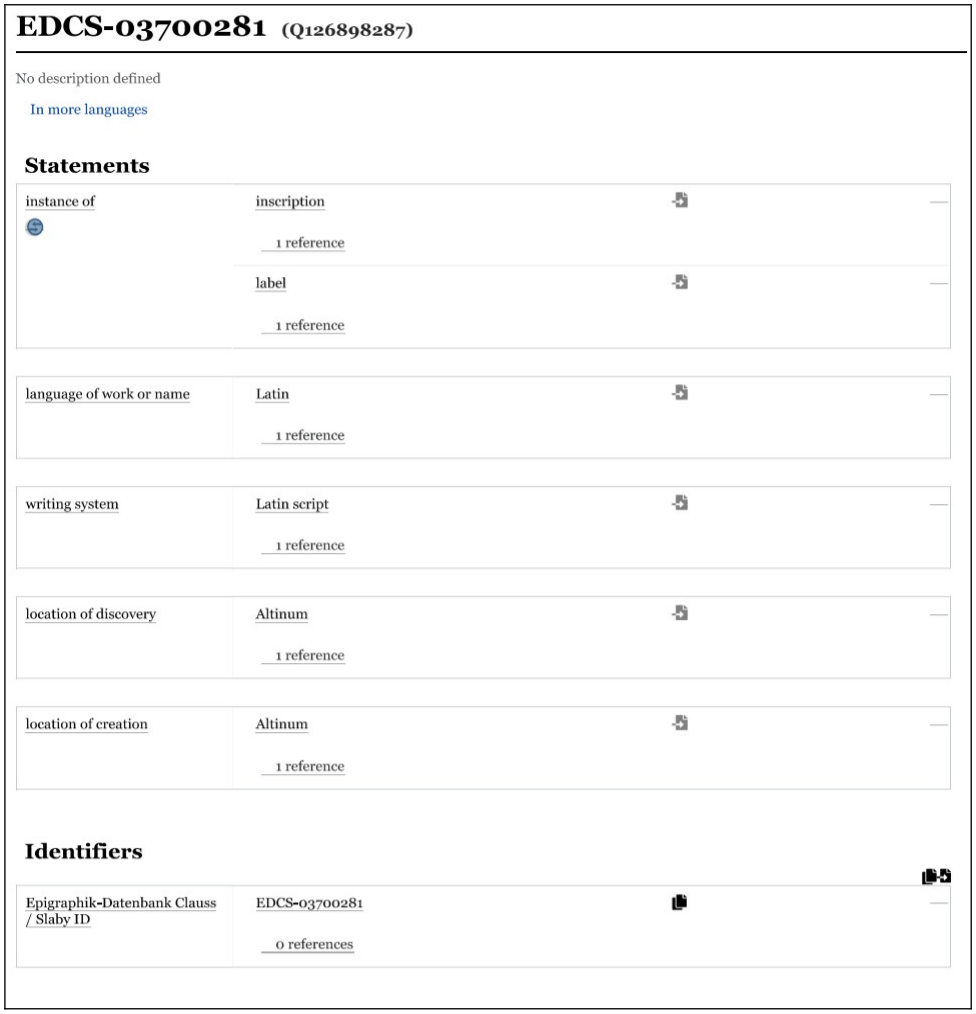
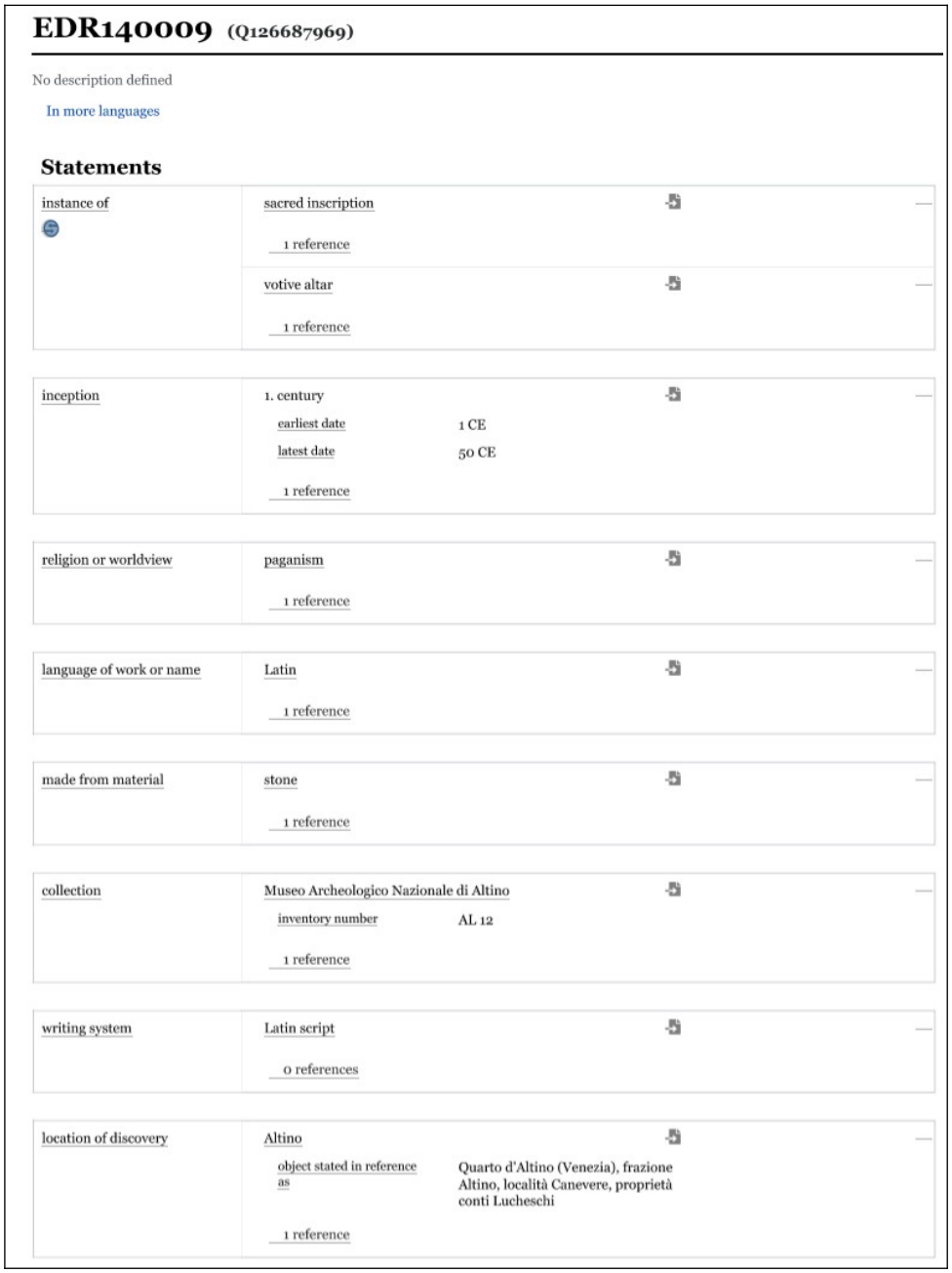


Fig. 8: Example of an *EDCS* inscription in *Wikidata*.

45 I obtained the correspondence of *EDR* and *EDCS* thanks to IDR, that associates *EDR*, *Trismegistos*, *EDCS*, *CIL* (etc.) IDs by entering a single identifier. https://id-resolver.inscriptiones.org/ (last access 11.07.2025).

46 https://quickstatements.toolforge.org/ (last access 11.07.2025). For massive import, the OpenRefine software (https://openrefine.org/ [last access 11.07.2025]) can be used and also bots can be programmed in Python using the Pywikibot library (documentation on https://doc.wikimedia.org/pywikibot/master/ [last access 11.07.2025]).



EDR140009 (Q126687969)

No description defined

In more languages

Statements

instance of	sacred inscription 1 reference	
	votive altar 1 reference	
inception	1. century earliest date: 1 CE latest date: 50 CE 1 reference	
religion or worldview	paganism 1 reference	
language of work or name	Latin 1 reference	
made from material	stone 1 reference	
collection	Museo Archeologico Nazionale di Altino inventory number: AL 12 1 reference	
writing system	Latin script 0 references	
location of discovery	Altino object stated in reference as: Quarto d'Altino (Venezia), frazione Altino, località Canevere, proprietà conti Lucheschi 1 reference	

Fig. 9a: Example of an *EDR* inscription in *Wikidata*, part 1.

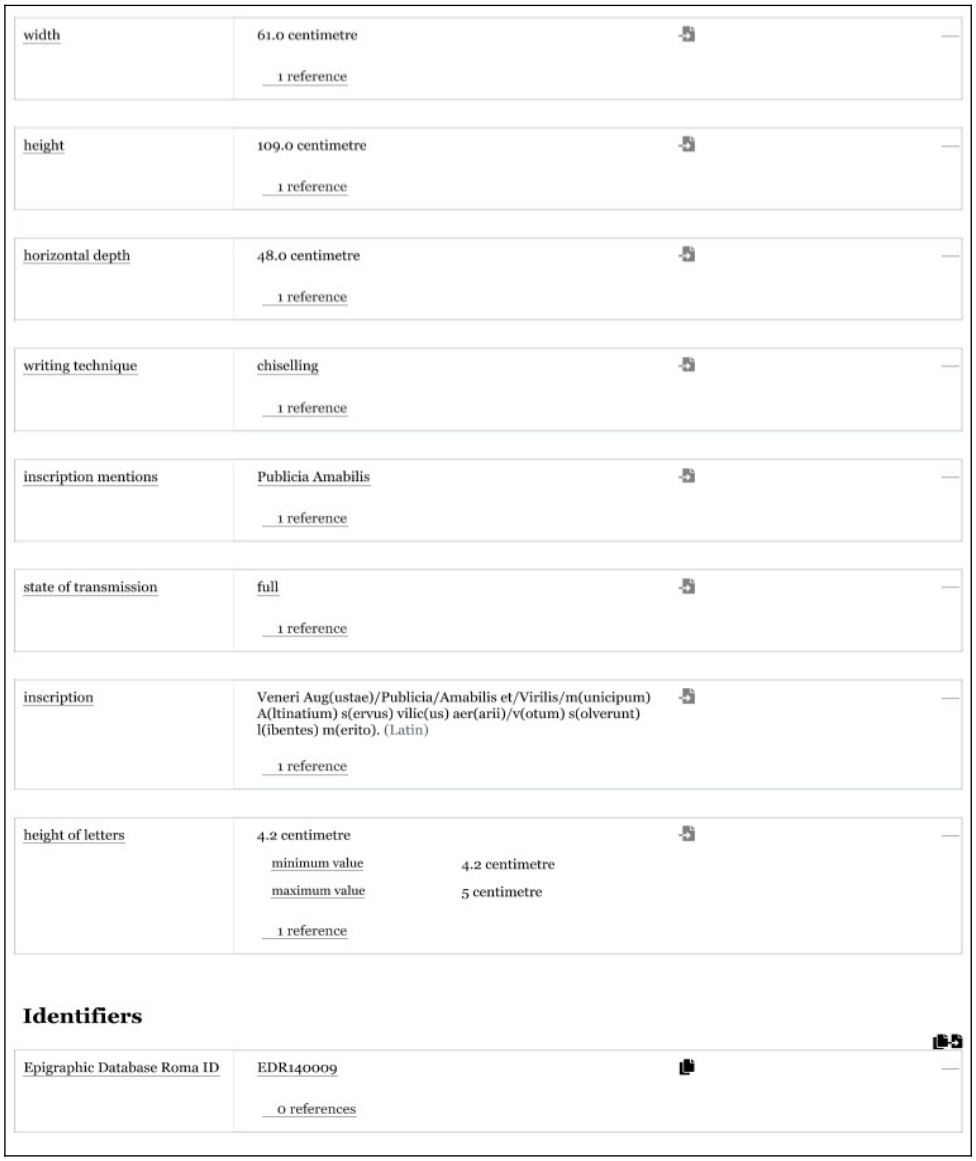


Fig. 9b: Example of an *EDR* inscription in *Wikidata*, part 2.

Entering Prosopographical Data

Once the insertion of the inscriptions was completed, the following step was then the final part of my work, namely the insertion and query of prosopographic data. I relied on the unpublished thesis of Barbara Pivetta[47], who catalogued the *gentes* and individuals mentioned in the inscriptions published up to that time (1998), as well as some that were yet to be published. Pivetta had catalogued 642 individuals, giving the bibliography of the inscriptions in which they were mentioned and, most importantly, any links with other individuals and the nature of their relationship. Each individual was catalogued by an identifier chosen arbitrarily by Pivetta, then retained as the value of P958 in references ("section, verse, paragraph or clause").

After incorporating the information from Pivetta's work, an additional 200 individuals could be added to the 642 catalogued here.[48] These were identified through onomastic data (father, master, and patron) and a closer analysis of the inscriptions – which sometimes revealed individuals not recorded by Piv-

47 Pivetta (1997/1998).

48 As said, only a few could be entered into *Wikidata* as many were recorded in inscriptions not found in *EDR* or *EDCS*.

etta, particularly servants. The first step was once again to create a CSV file structured according to the data required for import into *Wikidata*.[49] It was essential to encode key attributes that define an individual, such as name and gender: for instance, 'gender' corresponded to property P21, while the name was recorded both as a label (consistently in the three project languages – Italian, English, and Latin) and under property P1559, 'name in mother tongue'. Again, the project made use of *Wikidata*'s tools for massive import (QuickStatements), data extraction (*Wikidata* Query Service) and visualisation (such as EntiTree[50], an external tool for the construction of family trees and social networks). Currently, the focus of the *Altinum* project is on the survey of the female social network and marital relationships witnessed in Altinum inscriptions.

Below is a list of data transcoding based on the model adopted for the project.[51] While some descriptions were created ad hoc, others were used as they had already been established and discussed by the community.

Individuals Mentioned in the Inscriptions

Field name in Zenodo	***Wikidata* property**
ID *EDR*	To be used to create link with inscription • from person to entry: described by source (P1343): work where this item is described • from inscription to person: inscription mentions (P6568): item about a person or an object mentioned in the inscription's text. Use on Wikimedia Commons on media files
Name	Latin label + name in native language (P1559): name of a person in their native language
Gender	sex or gender (P21): sex or gender identity of human or animal. For human: male, female, non-binary, intersex, transgender female, transgender male, agender, etc. For animal: male organism, female organism. Groups of same gender use subclass of (P279)
Status	social classification (P3716): social class as recognized in traditional or state law
Age at the time of the event	age of subject at event (P3629): the age of the subject according to the cited source at the time of an event. Used as a qualifier of significant event property
Gens	gens (P5025): a clan or group of families from Ancient Rome who shared the same nomen
Provenance	place of birth (P19): most specific known birth location of a person, animal or fictional character
Significative Place	significant place (P7153): significant or notable places associated with the subject
Time	time period (P2348): time period (historic period or era, sports season, theatre season, legislative period etc.) in which the subject occurred or with which it is

49 Since *Wikidata* requires sources to be accessible and publicly available, the data I reused from Pivetta, together with the data I added, were published at https://zenodo.org/records/13773103 (last access 11.07.2025).

50 https://www.entitree.com/ (last access 11.07.2025).

51 https://zenodo.org/doi/10.5281/zenodo.12751850 (last access 11.07.2025).

	associated
Father	father (P22): male parent of the subject. For stepfather, use "stepparent" (P3448)
Mother	mother (P25): female parent of the subject. For stepmother, use "stepparent" (P3448)
Son/Daughter	child (P40): subject has object as child. Do not use for stepchildren – use "relative" (P1038), qualified with "type of kinship" (P1039)
Sibling	sibling (P3373): the subject and the object have at least one common parent (brother, sister, etc. including half-siblings); use "relative" (P1038) for siblings-in-law (brother-in-law, sister-in-law, etc.) and step-siblings (step-brothers, step-sisters, etc.)
Spouse	spouse (P26): the subject has the object as their spouse (husband, wife, partner, etc.). Use "unmarried partner" (P451) for non-married companions
Type of spouse	significant event (P793): significant or notable events associated with the subject
Other relative	relative (P1038): family member (qualify with "kinship to subject", P1039; for direct family member please use specific property)
Type of relative	kinship to subject (P1039): qualifier of "relative" (P1038) to indicate less usual family relationships (ancestor, son-in-law, adoptions, etc); indicate how the qualificator item is related to the main item (qualifier of relative (P1038): family member (qualify with "kinship to subject", P1039; for direct family member please use specific property))
Significant person	significant person (P3342): person linked to the item in any possible way
Object of statement has role	object of statement has role (P3831): (qualifier) role or generic identity of the predicate value/argument of a statement ("object") in the context of that statement; for the role of the item the statement is on ("subject"), use P2868 (qualifier of significant person (P3342): person linked to the item in any possible way)
Occupation	occupation (P106): occupation of a person; see also "field of work" (Property:P101), "position held" (Property:P39)
Member of	member of (P463): organization, club or musical group to which the subject belongs. Do not use for membership in ethnic or social groups, nor for holding a political position, such as a member of parliament (use P39 for that)
Position held	position held (P39): subject currently or formerly holds the object position or public office

Significant Place

Value name in Zenodo	*Wikidata* item
Altinum (default value)	Altinum (Q441542): *ancient city in Veneto and archaeological site in the Italian municipality of Quarto d'Altino (VE)*
Aquileia	Aquileia (Q2859274): *ancient Roman city now Aquileia, Province of Udine, Friuli–Venezia Giulia, Italy*
Roma	Rome (Q220): *capital and largest city of Italy*
Sardegna	geography of Sardinia (Q3760293): no description
Forum Cornelii	Forum Cornelii (Q3748870): *ancient Roman city (modern Imola)*
Luni	Luna (Q579763): *frazione of Italy*
Opitergium	Opitergium (Q130297514): *Ancient Roman settlement in the Venetia region*

Genders

Value name in Zenodo	*Wikidata* item
Male	male (Q6581097): *to be used in "sex or gender" (P21) to indicate that the human subject is a male or "semantic gender" (P10339) to indicate that a word refers to a male person*
Female	female (Q6581072): *to be used in "sex or gender" (P21) to indicate that the human subject is a female or "semantic gender" (P10339) to indicate that a word refers to a female person*
Not determinable	Unknown value

Status and Relationships

Reference property	*Wikidata* item
social classification (P3716)	• freedman: freedman (Q841571): *person who has been released from enslavement* • slave: slave (Q12773225): *person in a state of slavery* • ingenuus: ingenui (Q11926664): *legal term of ancient Rome indicating a person who was born free*
gens (P5025)	Abeia, Abidia, Abiria, Accia, Acellia, Acilia, Acutia, Aelia, Aemilia, Aeolia, Aequania, Aetriaca, Afinia, Ancharia, Annia, Antonia, Apertia, Aponia, Appuleia, Apronia, Aquilia, Aquilina, Aratria, Arcia, Aria, Arnia, Arruntia, Asconia, Asinia, Atia, Atilia, Attia, Auceia, Aulia, Aurelia, Avillia, Axia, Baebia, Baetia, Barbia, Braeta, Braetia, Caecilia, Caelia, Caesia, Caetronia, Calaecinia, Calventia, Cannusia, Cardia, Carminia, Cassia, Cassidia, Catius, Caulia, Caupia, Centia, Cervonia, Cethega, Ciceria, Cincia, Cleppia, Clodia, Cocceia, Coelia, Combulia, Cornelia, Cossutia, Crassicia, Cusonia, Didia, Domitia, Duronia, Egnatia, Elonia, Ennia, Epidia, Etuvia, Fabia, Fabricia, Faleria, Fannia, Faustina, Favonia,

	Firmia, Flavia, Folia, Fulvia, Furia, Gallia, Gavia, Grattia, Helvia, Helvidia, Herennia, Hostilia, Iulia, Iunia, Laberia, Laelia, Lartia, Lartidia, Latuonia, Licinia, Livia, Lollia, Lucana, Lucretia, Macioria, Maecenas, Maecia, Magia, Maicia, Mamilia, Manilia, Manlia, Mannia, Maria, Messia, Mestria, Mettia, Minucia, Mulvia, Munatia, Muria, Murria, Murtia, Mutia, Muttiena, Naevia, Nigidia, Nonia, Notellia, Novia, Numeria, Octavia, Ogia, Olia, Oppia, Ostilia, Ostoria, Paconia, Paescia, Paetinia, Papiria, Passena, Percennia, Peticia, Petronia, Pinnia, Pisidia, Plautia, Plotia, Poblicia, Pollia, Pompusia, Pontia, Popilia, Porcia, Postumia, Potia, Pupia, Putinia, Quinctia, Quinctilia, Remmia, Ruferia vel Rufertia, Sabina, Saenia, Safinia, Salvena, Satria, Saufeia, Seia, Sempronia, Senatia, Sescinia, Sevia, Sextia, Sextilia, Sicinia, Sintia, Sippia, Sosia, Statia, Tablinia, Tarutia, Tatia, Tattia, Tecina, Terentia, Tettenia, Titia, Titiena, Titurnia, Tommonia, Trebia, Trosia, Tufidia, Tullia, Turellia, Ulpia, Upsidia, Urtia, Vaccia, Valeria, Valgia, Varia, Veidia, Veronia, Vettia, Veturia, Viceria, Vilonia, Volumnia, Volusia
significant event (P793)	• contubernium • matrimonium • concubinatus • conubium
relative (P1038) with qualifier kinship to subject (P1039)	• father-in-law father-in-law (Q13204680): *male parent-in-law* • mother-in-law: mother-in-law (Q723868): *female parent-in-law* • amita: paternal aunt (Q5992509): *male parent's sister* • ancestor: ancestor (Q402152): *person from whom another person is descended* • daughter-in-law: child-in-law (Q2096646): *child's spouse* • grandfather: grandfather (Q9238344): *male grandparent* • maternal uncle: maternal uncle (Q4120409): *mother's brother* • paternal uncle: paternal uncle (Q12158205): *father's brother* • father-in-law: co-father-in-law (Q1498282): *father-in-law of one's child* • great-grandchild: great-grandchild (Q26237579): *grandchild's child* • grandson of grandfather: son's son (Q23684609): *male child of son* • uncle's nephew: niece or nephew (Q76477): *child of a sibling or half-sibling*
significant person (P3342)	• friend: friend (Q17297777): *companion or acquaintance*

with qualifier object of statement has role (P3831); use for servants owned by (P127) and for masters owner of (P1830)	*whom one regards with affection, affinity, or loyalty* • servant: item to be encoded with property owned by (P127): *owner of the subject* • freedman freedman (Q841571): *person who has been released from enslavement* • owner: to be encoded with the property owner of (P1830): *entities owned by the subject* • patron: patron (Q127800348): *individual who held a legal bond with his own freedman, now a member of his own gens* • delicatus: delicatus (Q130297560): no description or Q130297570: no description • collibertus: collibertus (Q127952883): *in ancient Rome, slave manumitted, along with others, by the same master*

Occupation/Other

Reference property	Name of item in inscriptions + identifier in *Wikidata*
position held (P39)	• sevir: seviratus (Q3958547): *magistracy of ancient Rome* • quattuorvir: Quadrumvir (Q23830356): *member of a college composed of four individuals holding institutional positions within a Roman municipality* • quattuorvir iure dicundo: quattuorvir iure dicundo (Q127638194): *Roman-era municipal magistrate in charge of the administration of justice* • quattuorvir aedilicia potestate: quattuorvir aedilicia potestate (Q127637725): *municipal magistrate from the Roman period with the role of overseer of the relevant municipality* • tribune: tribune (Q190401): *elected Roman officials* • decurion: decurion (Q1163056): *leader of ten legionaries* • augustalis: Augustalis (Q127690291): *Roman-era priest devoted to the cult of the imperial family* • veteran: veteran (Q4010462): in ancient Rome, soldier at the end of his service • praefectus fabrum: praefectus fabrum (Q3909815): *Roman military position* • priest: priest (Q42603): *person who consecrates his life to some divinity and whose main functions are to direct religious rites and offer sacrifices to the divinity (for a minister use Q1423891)*
member of (P463)	• Collegium funeraticium: Funerary institution (Q127633859): association raising funds for a collective burial • Collegium of lanarii purgatores: Q127701633: no description

	• Collegium of centonarii: Q127704335: no description If not just a member but a patron of the college then code with subject has role (P2868) and patronus (Q14052743): patron in ancient Rome
occupation (P106)	• physician: physician (Q39631): *professional who practices medicine* • pantomime: roman and greek pantomime (Q31183419): *a person who was involved in pantomime in ancient Greece and ancient Rome* • public freedman: public freedman (Q127701365): *former Roman slave belonging to and emancipated from a community* • mensor: agrimensor (Q396762): *surveyor in the Roman Empire* • sailor: sailor (Q45199): *person who navigates water-borne vessels or assists in doing so* • dispensator: dispensator (Q97190455): *butler or manager of payments, usually of a servile nature, in a private house or imperial office* • curator: legal guardian (Q157509): *person who has the legal authority to care for the personal and property interests of another person or comunity* • vilicus aerarii: vilicus aerarii (Q127797349): *in charge of the treasury of the municipia of the Roman Empire* • soldier: Roman legionary (Q17346959): *professional soldier of the Roman army* • veterinary: veterinarian (Q202883): *professional who treats disease, disorder, and injury in animals* • evocatus: Evocatus (Q568404): *a class of voluntarily reenlisted soldier in the Ancient Roman army* • procurator: Q3922473: no description
part of (P361)	imperial freedman: familia Caesaris (Q106602279): *set of slaves and freedmen who were in the service of the Roman emperor or under his patronage*
social classification (P3716)	• eques: equites (Q122166): *the lower of the two aristocratic classes of ancient Rome* • public slave: vilicus aerarii (Q127797349): *in charge of the treasury of the municipia of the Roman Empire*

Conclusions

This research methodology integrated traditional epigraphic sources with innovative digital tools, overcoming the limitations of conventional methods while encountering significant challenges. These included the compatibility of database licences with *Wikidata*, which required careful management work, highlighting the importance of ethical and legal use of the data, as well as the structuring and transcoding of the data: the creation of new entities in *Wikidata* required in-depth analysis on a controlled vocabulary to ensure terminological consistency. The EAGLE project provided a landmark, but its application in *Wikidata* required adaptations to meet academic and non-academic requirements.

The use of SPARQL showed the query potential of *Wikidata* for the study of the Altinum *corpus*, but also the difficulties related to the complexity of the language, which requires training or external support. This underlines the need to find more effective solutions for academic research through the WDQS.

Despite some challenges and the need for (albeit basic) technical training, *Wikidata* has proven to be an effective tool for epigraphic and prosopographical research, enabling flexible and multilingual data visualization while accommodating various research needs.

The next step will be collaborative: in fact, *GMI*, *Altinum*, *IDEA* and other projects shall work together to create a common data model that can lead to the expansion of the *Wikidata*:WikiProject Epigraphy[52] in alignment with FAIR Epigraphy principles.[53] This objective will result in the creation of a proper epigraphic dataset in *Wikidata*, expanded and categorised in the many disciplinary applications that epigraphy (not only Greek and Latin) entails.

52 https://www.wikidata.org/wiki/wikidata:WikiProject_Epigraphy (last access 11.07.2025).

53 Heřmánková et al. (2022); Cenati et al. (2021).

List of Abbreviations

CSV	Comma Separated Values
FAIR	Findable, Accessible, Interoperable, Reusable
Fig.	Figure
HTTP	Hypertext Transfer Protocol
IFLA	International Federation of Library Associations and Institutions
IDR	Inscriptiones Identifier Resolver
ISBN	International Standard Book Number
ISO	International Organization for Standardization
LOD	Linked Open Data
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language
Tab.	Table
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
QS	QuickStatements
WDQS	Wikidata Query Service

Sources

Online sources

https://dspace.unive.it/ (last access 11.07.2025).

https://duraeuroposarchive.org/ (last access 11.07.2025).

https://histropedia.com/ (last access 11.07.2025).

https://quickstatements.toolforge.org/ (last access 11.07.2025).

https://wikiba.se/ (last access 11.07.2025).

https://www.eagle-network.eu/ (last access 11.07.2025).

https://www.entitree.com/ (last access 11.07.2025).

https://www.ifla.org/ (last access 11.07.2025).

https://www.mediawiki.org/ (last access 11.07.2025).

https://www.w3.org (last access 11.07.2025).

https://zenodo.org/ (last access 11.07.2025).

https://query.wikidata.org (last access 11.07.2025).

https://www.wikidata.org (last access 11.07.2025).

https://doc.wikimedia.org/ (last access 11.07.2025).

https://openrefine.org/ (last access 11.07.2025).

https://id-resolver.inscriptiones.org/ (last access 11.07.2025).

Digital Corpora

EAGLE	Europeana Network for Ancient Greek and Latin Epigraphy
EDCS	Epigraphik-Datenbank Clauss / Slaby
EDR	Epigraphic Database Roma
GMI	Greek Metrical Inscriptions
IDEA	International (Digital) Dura-Europos Archive
TM	Trismegistos

References

Cenati et al. (2021): C. Cenati / G. Bodard, / H. Cayless / A. Cooley / T. Elliott / S. Evangelisti / A. Felicetti / P. Granados / F. Grieshaber / E. Gruber / A. Hershkowitz / T. Hill / H. Kiiskinen / T. Kollatz / A. Levivier / P. Liuzzo / F. Luciani / A. Mannocci / E. Mataix / F. Murano / O. Murphy / E. Mylonas / J. Prag / V. Razanajao / S. Stoyanova / G. Tsolakis / C. Tupman / I. Vagionakis / V. Vitale / F.. Weise, Modeling Epigraphy with an Ontology, online 2021, https://doi.org/10.5281/zenodo.4639507 (last access 11.07.2025).

Erxleben et. al. (2014): F. Erxleben / M. Günther / M. Krötzsch / J. Mendez / D. Vrandečić, «Introducing Wikidata to the linked data web», in: Mika et al. (2014), The Semantic Web. Lecture Notes in Computer Science – ISWC 2014 (LNCS 8796), Berlin 2014, 50–65.

Heřmánková et al. (2022): P. Heřmánková / M. Horster / J. Prag, Digital Epigraphy in 2022: A Report from the Scoping Survey of the FAIR Epigraphy Project, online 2022, https://doi.org/10.5281/zenodo.6610696 (last access 11.07.2025).

Hyvönen (2020): E. Hyvönen, Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery, in: Semantic Web 11/3 (2020), 187–193, https://doi.org/10.3233/sw-190386 (last access 11.07.2025).

Kaffee et al. (2017): L.A. Kaffee / A. Piscopo / P. Vougiouklis / E. Simperl / L. Carr / L. Pintscher, A Glimpse into Babel: An Analysis of Multilinguality in Wikidata, Proceedings of the 13th International Symposium on Open Collaboration = OpenSym, New York 2017, 1–5.

Lorito (2018): R. Lorito, L'Epigrafia latina e i database online, in: La Biblioteca di Classico Contemporaneo 6 (2018), 335–346.

Middle (2024): S. Middle, Linked Ancient World Data: Implementation, Advantages, and Barriers, DCO 10/1 (2024), 1–49, https://doi.org/10.11588/dco.2024.10.104105 (last access 19.03.2026).

Möller et al. (2007): K. Möller / T. Heath / S. Handschuh / J. Domingue, Recipes for Semantic Web Dog Food – The ESWC and ISWC Metadata Projects, in: K. Aberer et al., The Semantic Web. Lecture Notes in Computer Science = ISWC ASWC 2007 vol. 4825, Berlin / Heidelberg 2007, 802–815.

Mora-Cantallops et al. (2019): M. Mora-Cantallops / S. Sánchez-Alonso / E. García-Barriocanal, A systematic literature review on Wikidata, in: Data Technologies and Applications 53/3, 250–268, https://doi.org/10.1108/DTA-12-2018-0110 (last access 11.07.2025).

Orlandi (2021): S. Orlande, Digital Projects in Epigraphy: Research Needs, Technical Possibilities, and Funding Problems, in: I. Velasquéz Soriano / D. Espinosa (eds.), Epigraphy in the Digital Age: Opportunities and challenges in the Recording, Analysis and Dissemination of Inscriptions, Oxford 2021, 1–8. https://doi.org/10.2307/j.ctv1xsm8s5.5 (last access 11.07.2025).

Pivetta (1997/1998): B Pivetta, Le gentes di Altino romana. Tesi di laurea, Relatore G. Cresci Marrone, Venice 1997/1998.

Thornton (2024): K. Thornton / K. Seals Nutt / A. Chen, Encoding Archaeological Data Models as Wikidata Schemas: Utilizing Shape Expressions to Structure Collaborative Linked Open Data for Digital Storytelling Within the International Dura-Europos Archive, in: The International Journal of Technology, Knowledge, and Society 21/1 (2024), 69–83, https://doi.org/10.18848/1832-3669/cgp/v21i01/69-83 (last access 11.07.2025).

Tupman (2021): C. Tupman, Where Can Our Inscriptions Take Us? Harnessing the Potential of Linked Open Data for Epigraphy, in: I. Velasquéz Soriano / D. Espinosa (eds.), Epigraphy in the Digital Age: Opportunities and challenges in the Recording, Analysis and Dissemination of Inscriptions, Oxford 2021, 115–128, https://doi.org/10.2307/j.ctv1xsm8s5.15 (last access 11.07.2025).

Zhao (2023): F. Zhao, A systematic review of Wikidata in Digital Humanities projects, in: Digital Scholarship in the Humanities 38/2 (2023), 852–874, https://doi.org/10.1093/llc/fqac083 (last access 11.07.2025).

Figure and Table References

Figg. 1–3: Anna Clara Maniero Azzolini.

Figg 4–9: https://www.wikidata.org (last access 11.07.2025).

Tab. 1: Anna Clara Maniero Azzolini.

Author Contact Information[54]

Anna Clara Maniero Azzolini
PhD Student
University of London, School of Advanced Study
Digital Humanities Research Hub
E-mail: anna.manieroazzolini@london.ac.uk

54 The rights pertaining to content, text, graphics, and images, unless otherwise noted, are reserved by the author. This contribution is licensed under CC BY-SA 4.0.

DIGITAL CLASSICS ONLINE

Digital Mapping of Toponyms in Paradoxographical Texts: The Case of the *Paradoxographus Florentinus*

Pietro Zaccaria, Monica Berti

Abstract: The article presents the first results of a digital project devoted to the study of ancient paradoxography, a literary tradition consisting of collections of strange-but-(supposedly-)true phenomena concerning the natural world and the human sphere. The article is structured in two parts. First, we describe the first steps towards the creation of a relational database of ancient paradoxography, in collaboration with *Trismegistos* (TM Paradoxography). Second, we focus on the possibilities and limitations of the digital study of the toponyms mentioned in ancient paradoxography, by discussing the digital annotation of the toponyms mentioned in the paradoxographical collection known as the *Paradoxographus Florentinus*. This case study shows that the digital mapping of toponyms can be a valuable research tool to better understand the structure of paradoxographical texts and, more generally, the geographical horizon of ancient paradoxography.*

1. Introduction

'Paradoxography' is a modern term which is used to indicate a group of texts produced in Greco-Roman antiquity (from the 3rd century BCE onwards) consisting of literary collections of strange-but-(supposedly-)true phenomena concerning the natural world and the human sphere (e.g., stones, bodies of water, plants, and animals with amazing characteristics; human beings with prodigious powers, peculiar customs of peoples, and curious historical anecdotes).[1]

Even if paradoxography cannot be understood as a strictly defined literary genre (cf. de Martini [2023], 47-58), it still seems possible to identify a set of 'paradoxographical' texts which share similar contents, structures, and methods. Seven paradoxographical collections have been directly preserved by the manuscript tradition. These include pseudo-Aristotle's *On Marvelous Things Heard*, Antigonus' *Collection of Marvelous Histories*, Apollonius' *Amazing Stories*, Phlegon of Tralles' *On Marvels*, and the anonymous collections known as the *Paradoxographi Florentinus*, *Vaticanus*, and *Palatinus*. Many other lost (Greek and, in a few cases, Latin) collections are known thanks to fragments preserved by later authors or, less often, ancient papyri.[2] Basically, these collections contain lists of phenomena

* This article is the result of the cooperation of the two authors. Pietro Zaccaria is mainly responsible for parts 1, 3, and 4; Monica Berti for part 2. We gratefully acknowledge the generous support of the Research Foundation Flanders (FWO) (Pietro Zaccaria: FWO 1203624N).

1 The term 'paradoxography' goes back to Westermann (1939), who famously entitled his edition Παραδοξόγραφοι. *Scriptores rerum mirabilium Graeci*. The term 'paradoxographer' (παραδοξογράφος) is first attested in Tz. *Chil.* 2,35,154. On ancient paradoxography, see Ziegler (1949); Giannini (1963); Giannini (1964); Jacob (1983); Sassi (1993); Schepens / Delcroix (1996); Wenskus / Daston (2000); Pajón Leyra (2011); Geus / King (2018); Greene (2019); Lightfoot (2021), 42–57; Rusten / Yu (2022); Pajón Leyra (2022); Yu (2023). See also the studies collected in Gerolemou (2018); Kazantzidis (2019); Schorn / Mayhew (2024); Zucker et al. (2024).

which defy logic but still (supposedly) belong to the real world, with each phenomenon being usually located in a specific place and related on the authority of a specific source.

Because of paradoxography's widespread tendency to mention places and authors, the study of paradoxographical collections can significantly benefit from digital approaches which can allow us to extract, annotate, and analyze the numerous Named Entities (i.e., proper names) mentioned in them. The present article aims to present the first results of a recently started digital project led by the authors and devoted to the digital study of ancient paradoxography. The article is structured in two main parts. First, we describe the first steps towards the creation of a relational database of ancient paradoxography. Second, we focus on the possibilities and limitations of the digital study of toponyms in ancient paradoxography, by examining the specific case of the paradoxographical collection known as the *Paradoxographus Florentinus*.

2. 'TM Paradoxography': a Relational Database of Ancient Paradoxographical Texts

As explained above, the study of ancient paradoxography – a treasure-trove of Named Entities (toponyms and ethnics, authors, book titles, historical and mythological figures, etc.) – can greatly benefit from the application of digital approaches. In order to apply digital methods, we have started setting up a digital database of ancient paradoxography (TM Paradoxography: now accessible at https://www.trismegistos.org/paradoxography/ [last access 17.03.2026]), in collaboration with the KU Leuven-based project *Trismegistos*[3] (*TM*),[4] specifically with Mark Depauw and Yanne Broux (KU Leuven).

A preliminary step towards the creation of the database has been preparing a digital corpus of ancient paradoxography.[5] Because of the specific problems posed by fragmentary texts,[6] we have chosen to limit ourselves, in a first phase, to the seven preserved paradoxographical collections. The relatively limited amount of tokens (ca. 30,000) makes it feasible to prepare a philologically reliable corpus – a time-consuming, but fundamental work phase.[7]

---

2 After Giannini (1965), new editions of all (extant and fragmentary) paradoxographers are in preparation for *FGrHist* IV E (Paradoxography and Antiquities), published both in print and in Brill's *Jacoby Online*. The paradoxographers of the imperial period and undated authors are published in Schorn (2022).

3 https://www.trismegistos.org/ (last access 20.03.2026).

4 Depauw / Gheldof (2014).

5 We gratefully acknowledge the help of Stef Janssens, who ably assisted us in creating the digital corpus in the framework of an internship (KU Leuven, 2023–2024) under the supervision of Pietro Zaccaria.

6 On digital approaches to historical fragmentary texts, see Berti (2021) and Berti (2026).

7 By "tokens" we mean words and punctuation marks. The paradoxographical texts openly available in digital resources are valuable, but are based on old editions. The website *Paradoxography* (https://sites.google.com/site/paradoxography/Home [last access 24.07.2025]) offers the Greek texts of the *Paradoxographus Florentinus* and the *Paradoxographus Vaticanus* based on Giannini (1965), as well as English translations of the *Paradoxographus Florentinus*, the *Paradoxographus Vaticanus*, Apollonius, and Antigonus. The Loeb edition of pseudo-Aristotle's *On Marvelous Things Heard* by Hett (1936) is openly available (https://archive.org/details/minorworks00arisuoft/page/n5/mode/2up?view=theater [last access 24.07.2025]), but of limited use from a philological point of view. As far as concerns data in a machine readable format, the Scaife Viewer of the Perseus Project allows to navigate the edition of the *Paradoxographus Florentinus* by Öhler (1913) (https://scaife.perseus.org/library/urn:cts:greekLit:tlg0580.tlg001/ [last access 24.07.2025]) and the edition of pseudo-Aristotle by Bekker (1831) (https://scaife.perseus.org/library/urn:cts:greekLit:tlg0086.tlg027/ [last access 24.07.2025]) on the basis of structured XML files. These online resources were last accessed on 24.07.2025.

For each collection, we have prepared a digital version of the text in a TXT format based on a recent and authoritative edition. Where available, we have extracted the texts from the XML files of the *Jacoby Online* project[8] and converted them into plain text files. The texts of the reference editions have been carefully checked and, if necessary, compared with other existing editions.[9] As a rule, we have not changed the critical text established by the editors of the reference editions, but we have corrected it in the cases of typos, missing words, and minor inconsistencies.[10] In particular, we have used capitals for all Named Entities (personal names, toponyms, ethnics, book titles, etc.) – including those cases in which such names were not capitalized in the reference editions – and removed all extratextual information except for paragraph numbers and critical signs.[11] In what follows, we briefly list the editions which have been used to create the digital corpus of TM Paradoxography:

- Pseudo-Aristotle, *On Marvelous Things Heard* (TLG {0086} = TM Author 6108 = TM AuthorWork 11181: 10,492 tokens): Reference edition: Giacomelli (2023) (not digitized).[12] This edition not only offers an improved critical text, but also presents a new order of the chapters. The TLG has the old edition by Bekker (1831) {0086.027}.

- Antigonus, *Collection of Marvelous Histories* (TLG {0568} = TM Author 71 = TM AuthorWork 11264: 7,342 tokens): Reference edition: Musso (1985) (not digitized), checked against Giannini (1965) (digitized in the TLG as {0568.001}) and the unpublished dissertation by Eleftheriou (2018)[13].

- Apollonius, *Amazing Stories* (TLG {0569} = TM Author 1635 = TM AuthorWork 5497: 2,603 tokens): Reference edition: Spittler (2022) (*FGrHist* 1672, published both in print and online in Brill's *Jacoby Online*: https://doi.org/10.1163/1873-5363_jciv_a1672 [last access 15.04.2026]). A TXT file has been created on the basis of the XML file downloaded from Brill's database. The text has then been compared with the print edition in *FGrHist* IV E.2[14] and with Giannini (1965) (digitized in the TLG as {0569.001}).

- Phlegon of Tralles, *On Marvels* (TLG {0585} = TM Author 684 = TM AuthorWork 6473: 5,538 tokens): Reference edition: Shannon-Henderson (2022) (*FGrHist* 1667, published both in print and online in Brill's *Jacoby Online*: https://doi.org/10.1163/1873-5363_jciv_a1667 [last access 15.04.2026]).[15] A TXT file has been created on the basis of the XML file downloaded from Brill's database. The text has then been compared with the print edition in *FGrHist* IV E.2[16] and with Stramaglia (2011) (digitized in the TLG as {0585.004}).[17]

---

8 https://scholarlyeditions.brill.com/bnjo/ (last access 24.07.2025).

9 Where available, critical reviews have also been taken into account: Braccini (2023) has been particularly useful.

10 A full list of textual changes is available at https://www.trismegistos.org/paradoxography/changes.php (last access 22.01.2026).

11 In Computational Linguistics and Digital Humanities, the so-called Named Entities (NEs) mean proper names denoting 'real-world objects', such as persons, places, and organizations: see Nouvel et al. (2016). For NEs in historical languages like ancient Greek and Latin, see Berti (2019), Berti (2024), and Berti (2025).

12 We thank Ciro Giacomelli for having provided us with a PDF version of the text.

13 Openly available at: https://theses.hal.science/tel-01835129/ (last access 24.07.2025).

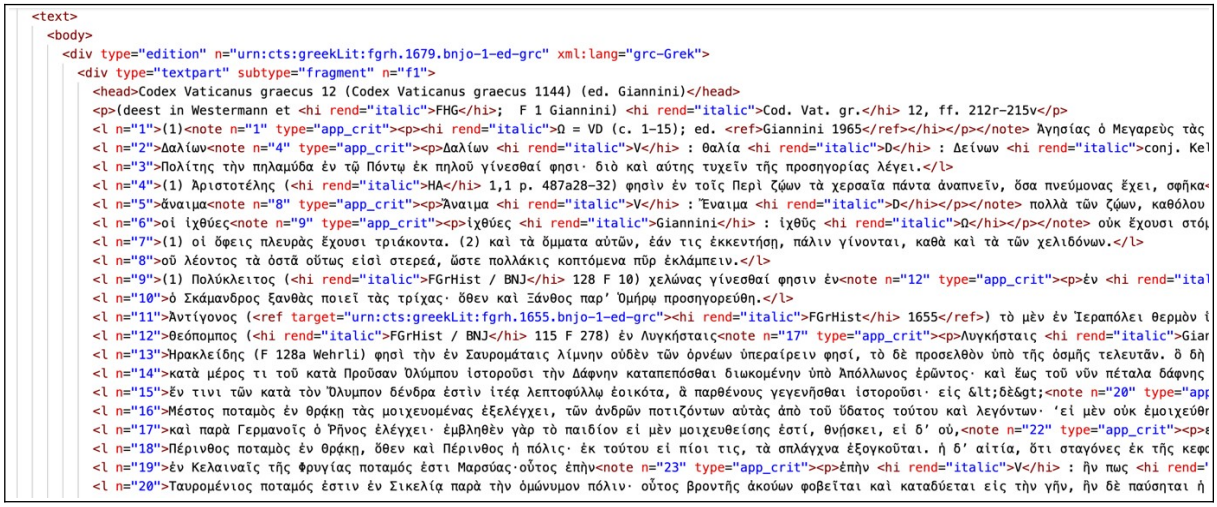
14 Schorn (2022), 387–519. In all cases in which a text was published both in print and online, comparing the two editions has allowed us to correct the text at various places. Although in principle identical, they differ in several places, with the print versions (which have been published after the online versions) usually (but not always) having the better text.

15 In Brill's *Jacoby Online*, Phlegon's paradoxographical collection is also included as *FGrHist*/*BNJ* 257 F 36 (Jacoby [1929], https://doi.org/10.1163/1873-5363_boj_a257 [last access 15.04.2026]; McInerney [2012], https://doi.org/10.1163/1873-5363_bnj_a257 [last access 15.04.2026]).

- *Paradoxographus Florentinus* (TLG {0580} = TM Author 8255 = TM AuthorWork 15191: 1,395 tokens): Reference edition: Greene (2022) (*FGrHist* 1672, published both in print and online in Brill's *Jacoby Online*: https://doi.org/10.1163/1873-5363_jciv_a1680 [last access 15.04.2026]). A TXT file has been created on the basis of the XML file downloaded from Brill's database. The text has been compared with the print edition in *FGrHist* IV E.2[18] and with Giannini (1965) (digitized in the TLG as {0580.001}).

- *Paradoxographus Vaticanus* (TLG {0582} = TM Author 8278 = TM AuthorWork 15763: 1,798 tokens): Reference edition: Sørensen (2022a) (*FGrHist* 1679, published both in print and online in Brill's *Jacoby Online*: https://doi.org/10.1163/1873-5363_jciv_a1679 [last access 15.04.2026]). A TXT file has been created on the basis of the XML file downloaded from Brill's database. The text has been compared with the print edition in *FGrHist* IV E.2[19] and with Giannini (1965) (digitized in the TLG as {0582.001}).

- *Paradoxographus Palatinus* (TLG {0581} = TM Author 8277 = TM AuthorWork 15762: 693 tokens): Reference edition: Sørensen (2022b) (*FGrHist* 1681, published both in print and online in Brill's *Jacoby Online*: https://doi.org/10.1163/1873-5363_jciv_a1681 [last access 15.04.2026]). A TXT file has been created on the basis of the XML file downloaded from Brill's database. The text has been compared with the print edition in *FGrHist* IV E.2[20] and with Giannini (1965) (digitized in the TLG as {0581.001}). A new, valuable edition of the text is offered by de Martini (2023) (unpublished dissertation openly available at: https://iris.unige.it/handle/11567/1128715 [last access 24.07.2025]).



```
<text>
  <body>
    <div type="edition" n="urn:cts:greekLit:fgrh.1679.bnjo-1-ed-grc" xml:lang="grc-Grek">
      <div type="textpart" subtype="fragment" n="f1">
        <head>Codex Vaticanus graecus 12 (Codex Vaticanus graecus 1144) (ed. Giannini)</head>
        <p>(deest in Westermann et <hi rend="italic">FHG</hi>;  F 1 Giannini) <hi rend="italic">Cod. Vat. gr.</hi> 12, ff. 212r–215v</p>
        <l n="1">(1)<note n="1" type="app_crit"><p><hi rend="italic">Ω = VD (c. 1–15); ed. <ref>Giannini 1965</ref></hi></p></note> Ἀγησίας ὁ Μεγαρεὺς τὰς
        <l n="2">Δαλίων<note n="4" type="app_crit"><p>Δαλίων <hi rend="italic">V</hi> : Θαλία <hi rend="italic">D</hi> : Δείνων <hi rend="italic">conj. Kel
        <l n="3">Πολίτης τὴν πηλαμύδα ἐν τῷ Πόντῳ ἐκ πηλοῦ γίνεσθαί φησι· διὸ καὶ αὐτης τυχεῖν τῆς προσηγορίας λέγει.</l>
        <l n="4">(1) Ἀριστοτέλης (<hi rend="italic">HA</hi> 1,1 p. 487a28–32) φησὶν ἐν τοῖς Περὶ ζῴων τὰ χερσαῖα πάντα ἀναπνεῖν, ὅσα πνεύμονας ἔχει, σφῆκα
        <l n="5">ἄναιμα<note n="8" type="app_crit"><p>Ἄναιμα <hi rend="italic">V</hi> : Ἔναιμα <hi rend="italic">D</hi></p></note> πολλὰ τῶν ζῴων, καθόλου
        <l n="6">οἱ ἰχθύες<note n="9" type="app_crit"><p>ἰχθύες <hi rend="italic">Giannini</hi> : ἰχθῦς <hi rend="italic">Ω</hi></p></note> οὐκ ἔχουσι στό
        <l n="7">(1) οἱ ὄφεις πλευρὰς ἔχουσι τριάκοντα. (2) καὶ τὰ ὄμματα αὐτῶν, ἐάν τις ἐκκεντήσῃ, πάλιν γίνονται, καθὰ καὶ τὰ τῶν χελιδόνων.</l>
        <l n="8">οὗ λέοντος τὰ ὀστᾶ οὕτως εἰσὶ στερεά, ὥστε πολλάκις κοπτόμενα πῦρ ἐκλάμπειν.</l>
        <l n="9">(1) Πολύκλειτος (<hi rend="italic">FGrHist / BNJ</hi> 128 F 10) χελώνας γίνεσθαί φησιν ἐν<note n="12" type="app_crit"><p>ἐν <hi rend="ital
        <l n="10">ὁ Σκάμανδρος ξανθὰς ποιεῖ τὰς τρίχας· ὅθεν καὶ Ξάνθος παρ' Ὁμήρῳ προσηγορεύθη.</l>
        <l n="11">Ἀντίγονος (<ref target="urn:cts:greekLit:fgrh.1655.bnjo-1-ed-grc"><hi rend="italic">FGrHist</hi> 1655</ref>) τὸ μὲν ἐν Ἱεραπόλει θερμὸν ἱ
        <l n="12">Θεόπομπος (<hi rend="italic">FGrHist / BNJ</hi> 115 F 278) ἐν Λυγκήσταις<note n="17" type="app_crit"><p>Λυγκήσταις <hi rend="italic">Giar
        <l n="13">Ἡρακλείδης (F 128a Wehrli) φησὶ τὴν ἐν Σαυρομάταις λίμνην οὐδὲν τῶν ὀρνέων ὑπεραίρειν φησί, τὸ δὲ προσελθὸν ὑπὸ τῆς ὁσμῆς τελευτᾶν. ὃ δὴ
        <l n="14">κατὰ μέρος τι τοῦ κατὰ Προῦσαν Ὀλύμπου ἱστοροῦσι τὴν Δάφνην καταπεπόσθαι διωκομένην ὑπὸ Ἀπόλλωνος ἐρῶντος· καὶ ἕως τοῦ νῦν πέταλα δάφνης
        <l n="15">ἔν τινι τῶν κατὰ τὸν Ὄλυμπον δένδρα ἐστὶν ἰτέᾳ λεπτοφύλλῳ ἐοικότα, ἃ παρθένους γεγενῆσθαι ἱστοροῦσι· εἰς &amp;lt;δὲ&amp;gt;<note n="20" type="ap
        <l n="16">Μέστος ποταμὸς ἐν θράκῃ τὰς μοιχευομένας ἐξελέγχει, τῶν ἀνδρῶν ποτιζόντων αὐτὰς ἀπὸ τοῦ ὕδατος τούτου καὶ λεγόντων· 'εἰ μὲν οὐκ ἐμοιχεύθη
        <l n="17">καὶ παρὰ Γερμανοῖς ὁ Ῥῆνος ἐλέγχει· ἐμβληθὲν γὰρ τὸ παιδίον εἰ μὲν μοιχευθείσης ἐστί, θνήσκει, εἰ δ' οὐ,<note n="22" type="app_crit"><p>
        <l n="18">Πέρινθος ποταμὸς ἐν θράκῃ, ὅθεν καὶ Πέρινθος ἡ πόλις· ἐκ τούτου εἰ πίοι τις, τὰ σπλάγχνα ἐξογκοῦται. ἡ δ' αἰτία, ὅτι σταγόνες ἐκ τῆς κεφ
        <l n="19">ἐν Κελαιναῖς τῆς Φρυγίας ποταμός ἐστι Μαρσύας·οὗτος ἐπὴν<note n="23" type="app_crit"><p>ἐπὴν <hi rend="italic">V</hi> : ἦν πως <hi rend="
        <l n="20">Ταυρομένιος ποταμός ἐστιν ἐν Σικελίᾳ παρὰ τὴν ὁμώνυμον πόλιν· οὗτος βροντῆς ἀκούων φοβεῖται καὶ καταδύεται εἰς τὴν γῆν, ἣν δὲ παύσηται ἡ
```

Fig. 1: Extract from the XML file of the *Paradoxographus Vaticanus* (*FGrHist* 1679) downloaded from the *Jacoby Online* project.

16 Schorn (2022), 9–338.

17 The *TLG* also has the edition by Giannini (1965), 170–218: {0585.001}.

18 Schorn (2022), 633–785.

19 Schorn (2022), 579–632.

20 Schorn (2022), 787–831.

The TXT files were converted into CSV files to preserve their internal structure in chapters and paragraphs, with each token on a separate row, including punctuation, and with sequential numbers to keep the order of the tokens in each text (see fig. 1–3). The CSV files were then imported in a FileMaker Database, set up by Mark Depauw, connected to the *Trismegistos* environment.

```
1|1|Ἀγησίας ὁ Μεγαρεὺς τὰς γεράνους φησίν, ὅταν ἐκ τῆς θράκης ἀπαίρειν μέλλωσιν, ὑπὸ μιᾶς περιρραίνεσθαι κύκλῳ πάσας· εἶθ' ὅταν βοήσῃ ἐκείνη, τὰς μὲν ἐξαίρειν καθαπερεὶ κελευστοῦ
1|2|ὅταν δὲ τὸ πέλαγος διαπεραιῶνται, δύο μὲν ἐκτείνειν τὰς πτέρυγας, τὴν δὲ γινομένην ὑπόκοπον ἐπὶ τούτων ἐφίζουσαν ἀναπαύεσθαι.
2||Δαλίων φησίν ἐν τῇ πρώτῃ τῶν Αἰθιοπικῶν ἐν τῇ Αἰθιοπίᾳ θηρίον γίνεσθαι κροκόττην καλούμενον· τοῦτο ἐρχόμενον πρὸς τὰς ἐπαύλεις κατακούειν τῶν λαλουμένων, καὶ μάλιστα τὰ ὀνόματ
3||Πολίτης τὴν πηλαμύδα ἐν τῷ Πόντῳ ἐκ πηλοῦ γίνεσθαί φησι· διὸ καὶ ταύτης τυχεῖν τῆς προσηγορίας λέγει.
4|1|Ἀριστοτέλης φησὶν ἐν τοῖς Περὶ ζῴων τὰ χεραῖα πάντα ἀναπνεῖν, ὅσα πνεύμονας ἔχει, σφῆκα δὲ καὶ μέλισσαν οὐκ ἀναπνεῖν.
4|2|ὅσα τε κύστιν ἔχει, πάντα καὶ κοιλίαν· οὐχ ὅσα δὲ κοιλίαν καὶ κύστιν.
5||ἄναιμα πολλὰ τῶν ζῴων, καθόλου δὲ ὅσα πλείω πόδας ἔχουσι τῶν τεσσάρων.
6||οἱ ἰχθύες οὐκ ἔχουσι στόμαχον· διό, ἐὰν διώκηται ὁ ἐλάττων ὑπὸ μείζονος, ἄγει τὴν κοιλίαν ὑπὸ τὸ στόμα.
7|1|οἱ ὄφεις πλευρὰς ἔχουσι τριάκοντα.
7|2|καὶ τὰ ὄμματα αὐτῶν, ἐάν τις ἐκκεντήσῃ, πάλιν γίνονται, καθὰ καὶ τὰ τῶν χελιδόνων.
8||τοῦ λέοντος τὰ ὀστᾶ οὕτως εἰσὶ στερεά, ὥστε πολλάκις κοπτόμενα πῦρ ἐκλάμπειν.
9|1|Πολύκλειτος χελώνας γίνεσθαί φησιν ἐν τῷ Γάγγῃ, ὧν τὸ χελώνιον μεδίμνους χωρεῖν πέντε.
9|2|ὁ Ἀγαθαρχίδης δὲ τοῖς χελωνίοις χρῆσθαι πλείοις ὡς ὁροφώμασι τῶν καλυβῶν.
10||ὁ Σκάμανδρος ξανθὰς ποιεῖ τὰς τρίχας· ὅθεν καὶ Ξάνθος παρ' Ὁμήρῳ προσηγορεύθη.
11||Ἀντίγονος τὸ μὲν ἐν Ἱεραπόλει θερμὸν ὕδωρ πάντα ἀπολιθοῦν φησι, καὶ αὐτὸ δὲ πήσσεσθαι καὶ λίθον γίνεσθαι.
12||Θεόπομπος ἐν Λυγκήσταις φησίν εἶναι ὕδωρ ὀξύ, ὃ τοὺς πίνοντας μεθύσκει.
13||Ἡρακλείδης [φησὶ] τὴν ἐν Σαυρομάταις λίμνην οὐδὲν τῶν ὀρνέων ὑπεραίρειν φησί, τὸ δὲ προσελθὸν ὑπὸ τῆς ὁσμῆς τελευτᾶν. ὃ δὴ καὶ περὶ τὴν ἄορνιν κατὰ τὴν Ἰταλίαν δοκεῖ γίνεσθαι
14||κατὰ μέρος τι τοῦ κατὰ Προῦσαν Ὀλύμπου ἱστοροῦσι τὴν Δάφνην καταπεπόσθαι διωκομένην ὑπὸ Ἀπόλλωνος ἐρῶντος· καὶ ἕως τοῦ νῦν πέταλα δάφνης ἐν τοῖς λίθοις ἀναμεμιγμένα εὑρίσκεσθ
15||ἔν τινι τῶν κατὰ τὸν Ὄλυμπον δένδρα ἐστὶν ἰτέᾳ λεπτοφύλλῳ ἐοικότα, ἃ παρθένους γεγενῆσθαι ἱστοροῦσι· εἰς <δὲ> δένδρα ταύτας ἀμειφθῆναι τὸν Βορρᾶν φευγούσας ἐρῶντα. καὶ νῦν ἔτ
16||Μέστος ποταμὸς ἐν θράκῃ τὰς μοιχευομένας ἐξελέγχει, τῶν ἀνδρῶν ποτιζόντων αὐτὰς ἀπὸ τοῦ ὕδατος τούτου καὶ λεγόντων· “εἰ μὲν οὐκ ἐμοιχεύθης, ἄρρεν τέκοις, εἰ δ' οὖν, θῆλυ”.
17||καὶ παρὰ Γερμανοῖς ὁ Ῥῆνος ἐλέγχει· ἐμβληθὲν γὰρ τὸ παιδίον εἰ μὲν μοιχευθείσης ἐστί, θνῄσκει, εἰ δ' οὔ, ζῇ.
18||Πέρινθος ποταμὸς ἐν θράκῃ, ὅθεν καὶ Πέρινθος ἡ πόλις· ἐκ τούτου εἰ πίοι τις, τὰ σπλάγχνα ἐξογκοῦται. ἡ δ' αἰτία, ὅτι σταγόνες ἐκ τῆς κεφαλῆς Γοργόνος ἐν τούτῳ ἐρρύησαν βασταζ
19||ἐν Κελαιναῖς τῆς Φρυγίας ποταμός ἐστι Μαρσύας· οὗτος ἐπὴν πως αὐλοῦ ἀκούσῃ, βομβεῖ μέγα, ἣν δὲ κιθάρας, μετὰ σιγῆς ῥεῖ, ἀποπνιγέντος ἐν αὐτῷ Μαρσύου τοῦ αὐλητοῦ.
20||Ταυρομένιος ποταμός ἐστιν ἐν Σικελίᾳ παρὰ τὴν ὁμώνυμον πόλιν· οὗτος βροντῆς ἀκούων φοβεῖται καὶ καταδύεται εἰς τὴν γῆν, ἣν δὲ παύσηται ἡ βροντή, πάλιν ἄνεισιν ἐκ τῆς γῆς καθά
```

Fig. 2: TXT file of the *Paradoxographus Vaticanus* (1–20) structured with pipes indicating chapters and paragraphs.

```
1,1,1,"Ἀγησίας"
1,1,2,"ὁ"
1,1,3,"Μεγαρεὺς"
1,1,4,"τὰς"
1,1,5,"γεράνους"
1,1,6,"φησίν"
1,1,7,","
1,1,8,"ὅταν"
1,1,9,"ἐκ"
1,1,10,"τῆς"
1,1,11,"θράκης"
1,1,12,"ἀπαίρειν"
1,1,13,"μέλλωσιν"
1,1,14,","
1,1,15,"ὑπὸ"
1,1,16,"μιᾶς"
1,1,17,"περιρραίνεσθαι"
```

Fig. 3: Extract of the CSV file of the *Paradoxographus Vaticanus* with each token on a separate row and sequential numbers.

Within this environment, we annotate all Named Entities with existing identifiers in TM Geo, TM Author, TM AuthorWork, TM Real, and TM God. This environment parses each token and proposes matches for entities already present in the database of *Trismegistos*. We have checked and, whenever necessary, corrected these matches. We have then annotated the other tokens not yet present in *Trismegistos* and, in the case of multi-token entities, we have related single tokens in order to represent real entities. As shown in fig. 4, for example, the toponym Potniai, mentioned in *Paradoxographus Florentinus* 1, can be linked to TM Geo 52457 (Potniai [Tachi], https://www.trismegistos.org/place/52457 [last access 20.03.2026]) (fig. 5), which is linked to the toponym Potniai in the gazetteer *Pleiades* (https://pleiades.stoa.org/places/541070 [last access 20.03.2026]) (fig. 6).

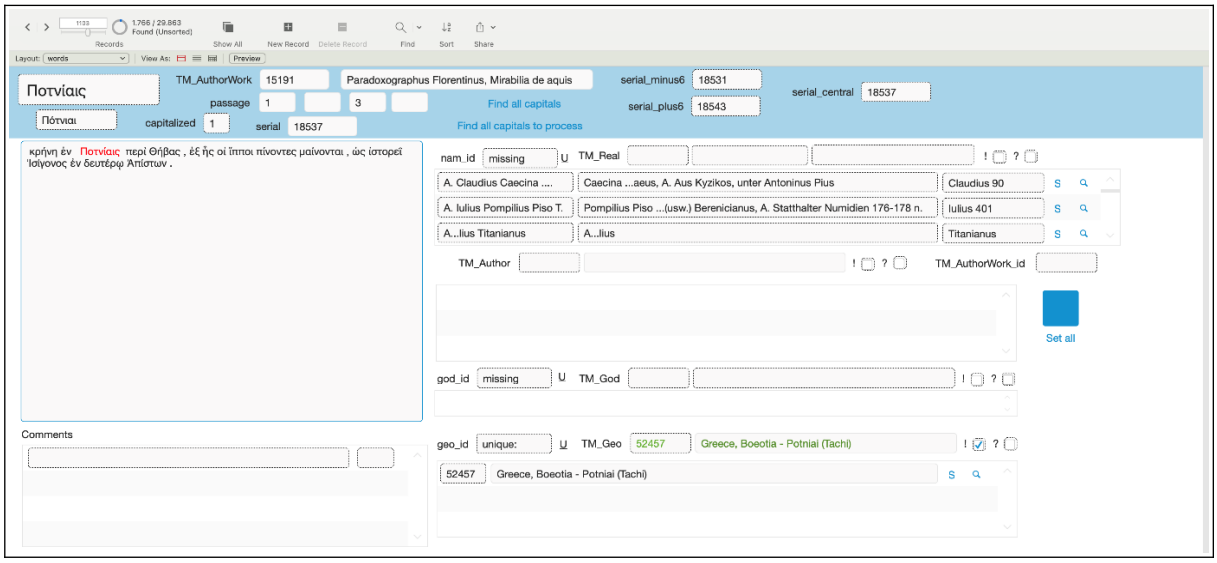


Fig. 4: Annotation of the toponym Potniai (TM Geo 52457) in *Paradoxographus Florentinus* 1.

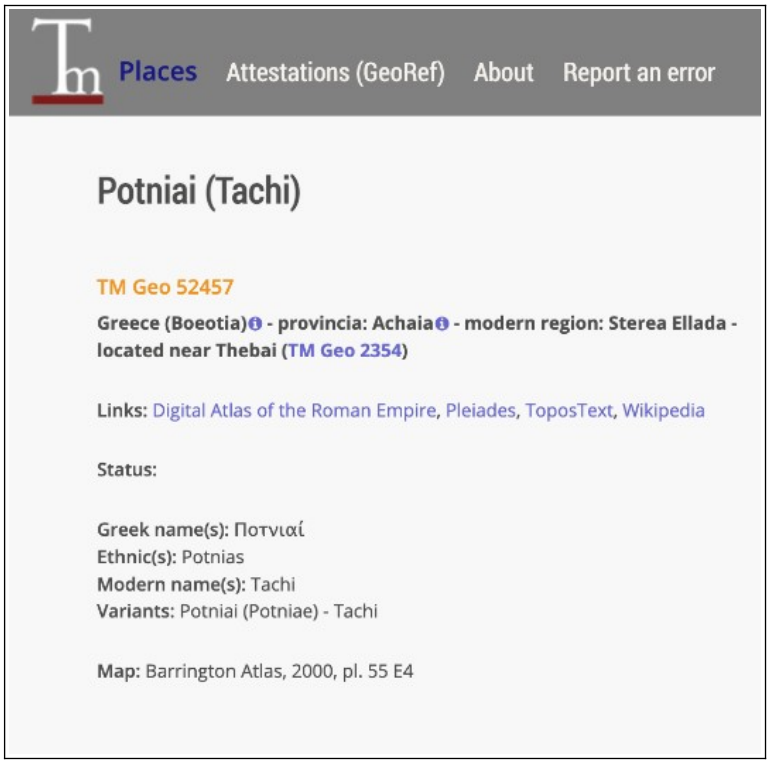


Fig. 5: TM Geo 52457 (https://www.*Trismegistos*.org/place/52457 [last access 24.07.2025]).

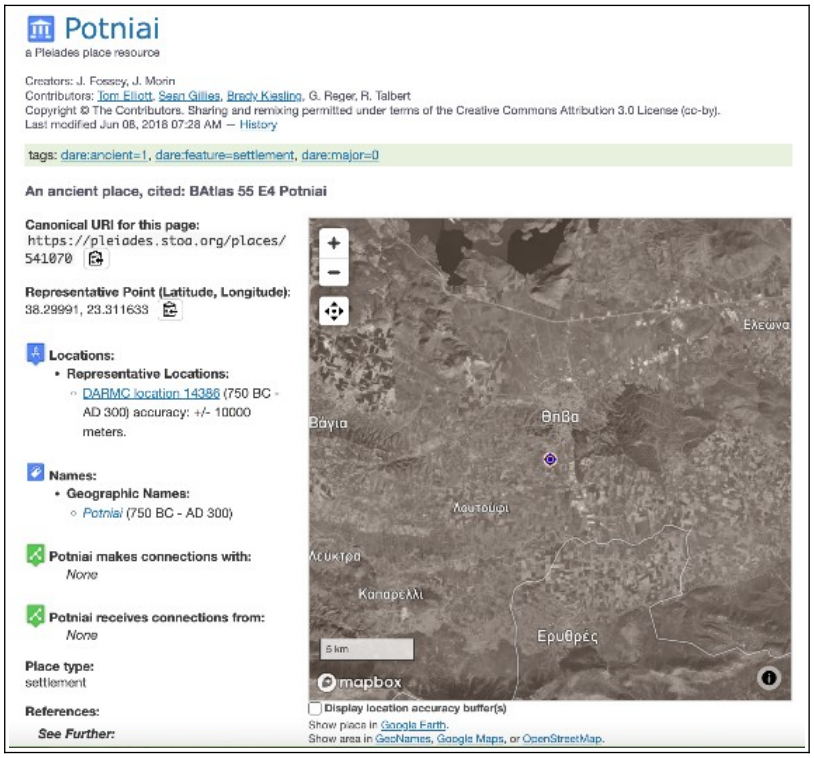


Fig. 6: Potniai in *Pleiades* (https://*Pleiades*.stoa.org/places/541070 [last access 24.07.2025]).

This environment also allows us to use a second annotation layer in those cases in which different Named Entities are associated to refer to one real entity. In *Paradoxographus Florentinus* 1, for example, Potniai is mentioned in the textual string ἐν Ποτνίαις περὶ Θήβας ("in Potniai near Thebes"). At a first level, Ποτνίαις has been linked, as noted above, to TM Geo 52457 (Potniai [Tachi]: https://www.*Trismegistos*.org/place/52457 [last access 24.07.2025]) while Θήβας has been annotated as TM Geo 2354 (Thebai: https://www.*Trismegistos*.org/place/2354 [last access 24.07.2025]). At a second level, however, the tokens περί and Θήβας can be included with Ποτνίαις, as shown in fig. 7.

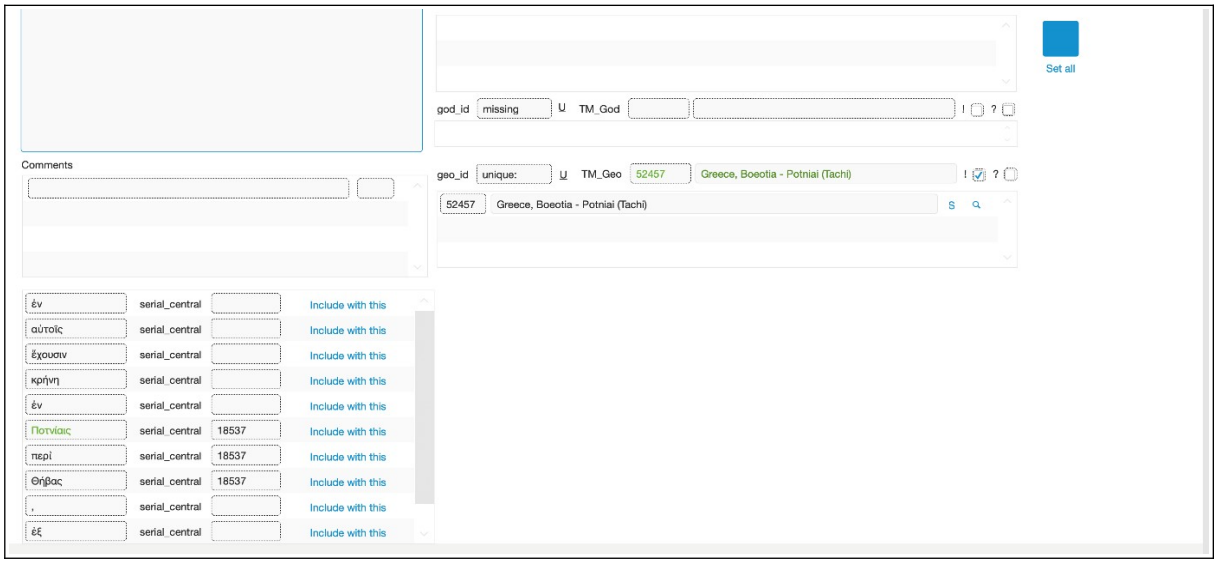


Fig. 7: Second layer annotation of Ποτνίαις περὶ Θήβας (*Paradoxographus Florentinus* 1).

Data annotated in *Trismegistos* can be exported to visualize and extract information on annotated Named Entities like personal names and toponyms. These entities, that have been disambiguated in the annotation phase and provided with unique identifiers of authority lists (e.g., TM Geo and *Pleiades*), constitute a first dataset that will be used to annotate other paradoxographical texts. Data in ancient

Greek – which is in most cases still lacking in a digital format – is fundamental for further analyses on the richness and variety of the language of ancient paradoxography for onomastics and historical geography.

This relational database of ancient paradoxography, whose completion is envisaged for 2026 and which is openly accessible, will allow scholars to apply digital approaches to paradoxographical collections on the basis of a complete and reliable textual basis, and to export linguistic data and metadata to be used in larger projects. The following section explores the application of a digital approach which can greatly enhance our understanding of ancient paradoxography, namely, the digital annotation and mapping of the toponyms mentioned in ancient paradoxographical texts.

3. Toponyms in Ancient Paradoxography: the *Paradoxographus Florentinus*

The location of phenomena is quintessential to paradoxography, as the marvels described by paradoxographers are usually located in specific places.[21] As Jacob aptly notes:

> "Le merveilleux résulte de l'association d'un phénomène avec un *topos* précis, ce rapport n'étant ni explicité ni expliqué. L'anecdote paradoxographique évoque souvent des différences minimes observables dans un lieu particulier et le distinguant de l'espace environnant […]."[22]

In some cases, paradoxographers even seem to have modified their sources in order to locate the described phenomena in specific places[23] – which has been defined as paradoxography's "geographical isolation or singularization".[24] By linking their marvels to precise places, paradoxographers could also firmly anchor them to the real world and thereby enhance the credibility of their accounts.[25] Moreover, geography could serve as a way for a paradoxographer to structure his material.[26] Based on the (probably partially corrupted) title transmitted by the *Suda* (κ 227 Adler, s.v. Καλλίμαχος: Θαυμάτων τῶν εἰς ἅπασαν τὴν γῆν κατὰ τόπους ὄντων συναγωγή, *Collection of Marvels from Every Land Arranged According to Places*), it is usually assumed that Callimachus' collection – traditionally regarded as the first paradoxographical text – was arranged geographically.[27] Theopompus' *Thaumasia* – whether the

21 On geography and paradoxography, cf. Jacob (1983), 134–135; Stern (2008), 439–440; Pajón Leyra (2011), 33–35; Dueck (2012), 64–67; Geus (2016); Geus / King (2018); Eleftheriou (2018), 91–96 (esp. on Antigonus); Rusten / Yu (2022). Yu (2023), 265 speaks of paradoxography's tendency to tie marvels "to specific *lieux de memoire*, that is, landmarks vested with cultural significance".

22 Jacob (1983), 135.

23 A famous example is the mobility of cattle horns: while Aristotle (*Hist. An.* 3,9,517a27-30) locates the phenomenon "in Phrygia and elsewhere" (ἐν Φρυγίᾳ εἰσὶ βόες καὶ ἄλλοθι), Antigonus (Mir. 75), citing Aristotle, places the phenomenon in Phrygia alone (ἐν Φρυγίᾳ δὲ βοῦς εἶναι, οἳ κινοῦσι τὰ κέρατα). Cf. Jacob (1983), 134–135; Schepens / Delcroix (1996), 392–393; Schepens / Schorn (2010), 407; Lightfoot (2021), 77–78. Another interesting example is provided by the deer in Achaea (ps.-Arist. Mir. 5, to be compared with Arist. *HA* 8(9),611b8-20), see Giacomelli (2024), 245–246.

24 As aptly put by Schepens / Delcroix (1996), 392; Schepens / Schorn (2010), 407. Cf. also Giacomelli (2024), 235–236.

25 García Teijeiro / Molinos Tejada (1994), 276; Stern (2008), 439–440; Shannon-Henderson (2013), 4–5; Geus (2016), 244; Geus / King (2018), § 3; Nichols (2018), 11–12.

26 On geography as a possible systematic arrangement for paradoxography, see Schepens / Delcroix (1996), 394–398; Pajón Leyra (2011), 33–35. Paradoxographical collections could follow a topical, alphabetical, or bibliographical order. Different principles of classifications could of course be combined: see Pajón Leyra (2011), 33–40. By contrast, Yu (2023), 275 considers Greek paradoxography as deprived of any "apparent hodological, cartographic, or conceptual order".

27 For the fragments, see F 407–411 Pfeiffer = Giannini (1965), 15–20. Cf. Ziegler (1949), 1140–1141; Giannini (1964), 105–109; Pajón Leyra (2011), 33–34; Geus / King (2018), § 3; Lightfoot (2021), 47.

title refers to book 8 of the *Philippica* or to a later collection of marvels related by Theopompus throughout the *Philippica*[28] – also seems to have followed a geographical pattern,[29] while various collections focused on specific regions, such as Nymphodorus of Syracuse's *On the Wonders in Sicily* (*FGrHist*/*BNJ*[2] 572 F 1–2) and Polemon of Ilium's *On Marvelous Rivers in Sicily* (T 2 Giannini). And even if none of the preserved collections seem to use geography as their primary principle of organization,[30] it has been argued that specific sections of some collections may follow a geographical or even a periegetic or hodological order.[31] Studying the geographical location of the phenomena mentioned in the collections can therefore enhance our understanding of the structure of the preserved collections, and shed light on the geographical horizon of paradoxography.[32] Despite the growing scholarly awareness of the role played by geography in ancient paradoxography, however, this aspect still needs to be systematically explored.

Our understanding of the geographical element of paradoxography can be significantly enhanced by the digital annotation and mapping of the places mentioned or alluded to in the preserved collections. Admittedly, the very nature of paradoxography makes it difficult to effectively visualize its perception of space. For paradoxographers usually do not offer geographical descriptions (like, e.g., Pliny the Elder) nor a narrative embedded in historical space (like, e.g., Herodotus), but mostly limit themselves to mention isolated toponyms or ethnic names, drawn from sources belonging to different periods. As Daniela Dueck aptly puts it, "[p]aradoxographies […] emphasized natural phenomena with geographical relevance, not an orderly spatial or linear description of foreign places."[33]

Nonetheless, it can still be useful to chart paradoxographical toponyms and geographical references onto modern maps by linking ancient toponyms to real places. This operation notoriously brings with it significant methodological challenges.[34] However, in so far as they are "thought of and approached as *part of* an interpretative process and not its end result",[35] such maps can be useful research tools. In this respect, digital methods and data are particularly relevant for two main reasons. 1) We need more data in ancient languages to be extracted from ancient sources and collected in authority lists in order to structure, annotate, and analyze further texts. By data in authority lists, we mean inflected and lem-

28 See Zaccaria (2024), 135, with references.

29 Apollon. Mir. 1 = *FGrHist*/*BNJ* 115 F 67b: Θεόπομπος ἐν ταῖς Ἱστορίαις ἐπιτρέχων τὰ κατὰ τόπους θαυμάσια. In Mir. 13, Apollonius also cites a paradoxical story that was originally found ἐν τῷ Κατὰ τόπους μυθικῷ, but the name of the author is unfortunately lost in a lacuna and cannot be recovered with any certainty. The *Suda* entry ε 3930 Adler, s.v. Ἔφιππος, which seems to concern Ephorus of Cumae (*FGrHist*/*BNJ* 70 T 1) rather than Ephippus of Olynthus (*FGrHist*/*BNJ* 126 T 1 = T 6 Ravazzolo), mentions the title Παραδόξων τῶν ἑκασταχοῦ βιβλία ιε΄.

30 Cf. Yu (2023), 275: "Notably, the *Shan hai jing*'s recurring concern for the careful elaboration and measurement of travel routes and circuits is entirely absent in Greek paradoxography, which leads the reader from one place to the next in no apparent hodological, cartographic, or conceptual order. … The driving impetus of Greek paradoxography, in the final equation, revolves around local rather than global knowledge."

31 See Geus (2016), 248–256 (with regard to *Par. Vat.* 47–56); Geus / King (2018), § 3 and Pajón Leyra (2024) (with regard to the section of pseudo-Aristotle's *On Marvelous Things Heard* concerning the Western Mediterranean: Mir. 78–121); Zaccaria (2024), 133–134 (with regard to pseudo-Aristotle's section concerning the Eastern Mediterranean: Mir. 123–138).

32 Of course, paradoxical reports were also a common *topos* in ancient geographical literature, with geographical treatises sometimes containing paradoxographical sections. Book 6 of Protagoras' *Geometry of the Inhabited World*, for example, focused on marvels: see Giannini (1965), 220 = *FGrHist* 2044 T 1. Cf. Giannini (1964), 130.

33 Dueck (2012), 67.

34 The study of ancient geographers and ancient representations of space has greatly advanced in recent years. See, e.g., Rathmann (2007); Talbert (2012); Roller (2015); González Ponce et al. (2016); Bianchetti et al. (2016); Roller (2019); Castro-Páez / Cruz Andreotti (2020); Shipley (2024). For approaches combining textual and digital methodologies, see, e.g., Barker et al. (2016b). For a digital project on annotating toponyms and representing them on modern maps, see the *Digital Periegesis* (https://www.periegesis.org/ [last access 24.07.2025]) with Barker et al. (2023).

35 Barker et al. (2016a), 18. Cf. also Barker et al. (2023), 142.

matized forms in ancient Greek and Latin with stable identifiers and metadata (e.g., Συρακουσῶν, Συράκουσαι, TM Geo 2210, *Pleiades* 462503, settlement).[36] Paradoxographical texts are rich in toponyms whose forms are still missing in authority lists and whose inclusion is therefore important for increasing the data that can be used for linguistic, philological, and historiographical analyses. 2) These forms allow us to trace the history of places keeping track of the linguistic variety of their names and descriptions over the centuries. Moreover and whenever possible, geographical coordinates enable us to visualize ancient places on modern maps for experiments with distant reading approaches, as we will see in the following paragraphs.[37]

The digital annotation and mapping of toponyms should of course be accompanied by close textual analysis. Paradoxographers used earlier sources, which they sometimes modified or misunderstood; moreover, the names of places and peoples are often corrupted in the manuscript tradition, which makes their identification all the more difficult. The toponyms we deal with are the result of a multi-layered chain of transmission, which poses us in front of delicate methodological challenges.[38]

In what follows, we discuss the potential value of and some methodological challenges posed by the digital annotation and mapping of paradoxographical toponyms by focusing on the specific case of the *Paradoxographus Florentinus*.

The *Paradoxographus Florentinus* (*TLG* {0580} = TM Author 8255 = TM AuthorWork 15191) is an anonymous collection of marvels carrying the title Κρῆναι καὶ λίμναι καὶ πηγαὶ καὶ ποταμοὶ ὅσοι θαυμάσιά τινα ἐν αὐτοῖς ἔχουσιν (*Springs, lakes, streams, and rivers which have some amazing qualities in them*).[39] The preserved manuscript witnesses seem to derive from Flor. Laur. Plut. 56.1 (= F; often dated to the 13th–14th century, but perhaps as early as the 12th century[40]). All the 43 reports included by the *Paradoxographus Florentinus* deal with the amazing characteristics of different bodies of water (springs, lakes, streams, and – less frequently – rivers).[41] As usual in paradoxography, the collection regularly mentions not only its sources,[42] but also the geographical location of the described phenomena. The compiler's dates are unknown. Based on the dates of the sources mentioned in the collection, he has been traditionally dated to the 1st or 2nd century CE.[43] Taking into account not only the compiler's sources, but also his language, Öhler proposed to place the composition of the collection

36 For a rich collection of metadata, see the example of the ancient city of Syracuse in Sicily in the gazetteer Pleiades under the corresponding URI: https://pleiades.stoa.org/places/462503 (last access 24.07.2025). On Pleiades, see Elliott / Gillies (2009).

37 When ancient places are still locatable, geographical coordinates (latitude and longitude) are used to position them on a map. See the example of ancient Syracuse mentioned in the previous footnote with the coordinates 37.070078971, 15.2833356581. See Barker et al. (2024).

38 See especially Giacomelli (2024) (with regard to toponyms in pseudo-Aristotle's *On Marvelous Things Heard*).

39 The following discussion of the *Paradoxographus Florentinus* is based on Zaccaria (forthcoming). Besides the valuable commentaries by Öhler (1913) and Greene (2022), previous studies dealing with the collection include Ziegler (1949), 1161–1162; Giannini (1964), 135–136; Schepens / Delcroix (1996), 426; Geus / King (2018), § 4; Pajón Leyra (2011), 162–163.

40 See Giacomelli (2021), 350 n. 140; Greene (2022), 646 n. 2.

41 In general, the collection seems to follow a roughly thematic structure, with c. 1–27 focusing on springs and c. 28–43 dealing with lakes and ponds – though this distinction is not rigidly followed throughout the text.

42 These include: Isigonus' *Unbelievable Things* (c. 1, 2, 8, 9, 11–14, 21, 27, 36, 40, 43); pseudo-Aristotle's *On Marvelous Things Heard* (c. 7, 10, 29, 30, where "Aristotle" is mentioned as a source: see Giacomelli (2021), 350–355; Greene (2022), 650); Aristotle (c. 19); Ctesias of Cnidus (c. 3 and 17); Theopompus of Chios (c. 15 and 20); Hellanicus of Lesbos (c. 16); Amometus (c. 18); Heraclides Ponticus (c. 22); Herodotus (c. 23 and – implicitly – 32); Ariston of Ceos (c. 25); Hieronymus of Cardia (c. 33); Pythermus (c. 34); and Phaethon (c. 35). Other reports contain references to unnamed sources (φασίν/λέγουσιν, "they say": c. 5, 6, 24, 26, 32) or local informants (c. 41).

43 See Giannini (1964), 135–136; Schepens / Delcroix (1996), 426.

between 80 and 100 CE.[44] Öhler's arguments, however, were rejected by Ziegler, who argued that the collection may have been compiled in the imperial or late antique period.[45] The last editor, Greene, tends to date the text to the late 1st or 2nd century CE, but cautiously admits that "no evidence forbids a later date".[46]

As already mentioned in section 2, the reference edition of the *Paradoxographus Florentinus* adopted in TM Paradoxography is Greene 2022 (*FGrHist* 1672). A TXT file has been created on the basis of the XML file downloaded from Brill's database (https://doi.org/10.1163/1873-5363_jciv_a1680 [last access 15.04.2026]), which has been compared with the print edition of the text published in *FGrHist* IV E.2[47] and with Giannini 1965 (cataloged in the TLG as {0580.001}).[48] Greene's edition is accurate: only one accent has been corrected (6: τι σύστημά ἐν > τι σύστημα ἐν). Despite the limited length of the text, comparison with Giannini's edition clearly demonstrates the impact of editorial decisions on texts containing many textual problems, which often concern proper names.[49]

In several places, the Greek forms of Named Entities accepted by Greene, including toponyms, are different from those in Giannini's edition:

- 15: ἐν Χρωψὶ τῆς Θρᾴκης Greene (F) : ἐν Κίγχρωψι τῆς Θρᾴκης Giannini (ex Antig. Mir. 141: ἐν †κιγχρωψωσιν† τοῖς Θρᾳξίν).
- 20: ἐν Λυγκήστῳ Greene (λυγκίστω F) : ἐν Λυγκήσταις Giannini (ex Antig. Mir. 164)
- 24: Προιτίδας Greene (ex Vitr. 8,3,21) : Προιτίδος Giannini (F)
- 42: Λυχνίς Greene (F) : Λυχνῖτις Giannini (ex ps.-Scymn. 429, Diod. 16,8,1)

Cruces are used to mark different Named Entities as corrupted:

- 35: Φαέθων Greene (F) : †Φαέθων† Giannini
- 36: †Μυκλαίαν† Greene (F) : Ἀμυκλαίαν Giannini (ex Plin. nat. 8,104)
- 43: Τάλα Greene (F) : †Τάλα† Giannini

Capitals are not regularly used by Giannini to indicate Named Entities such as book titles and affiliations to philosophical schools:

- 1: Ἀπίστων Greene : ἀπίστων Giannini
- 8: Ἀπίστων Greene : ἀπίστων Giannini
- 25: Περιπατητικὸς Greene : περιπατητικὸς Giannini
- 43: Ἀπίστων Greene : ἀπίστων Giannini

44 Öhler (1913), 22, 148–150, 162–163.

45 Ziegler (1949), 1162.

46 Greene (2022), 649. Cf. also Geus / King (2018), § 4; Giacomelli (2021), 350–351 n. 141.

47 Schorn (2022), 633–785.

48 Previous editions include Westermann (1839), 183–191; Ideler (1841), 184–189; Landi (1895), 532–538; Öhler (1913), with a still valuable commentary. A Spanish translation with notes is provided by Gómez Espelosín (1996), 253–261.

49 Besides those concerning Named Entities, the following textual differences have been identified: 24: τῆς Ἀρκαδίας Greene : δὲ τῆς Ἀρκαδίας Giannini; 24: μὴ ποτὶ Greene : μὴτ' ἐπὶ Giannini; 24: ἐκτὸς ἰόντα Greene : ἐντὸς ἐόντα Giannini; 24: ἔκρυψεν Greene : ἔκοψεν Giannini; 24: αἳ γὰρ Greene : †αγαρ† Giannini; 33: οὔτ' ἰχθῦς Greene : οὔτε ἰχθῦς Giannini.

Moreover, Giannini does not print the title preceding the work in manuscript F (Κρῆναι καὶ λίμναι καὶ πηγαὶ καὶ ποταμοὶ ὅσοι θαυμάσιά τινα ἐν αὐτοῖς ἔχουσιν), apparently on the assumption that it should be considered a description of the work's content rather than an authorial title.

On the basis of the TXT file of Greene's edition of the *Paradoxographus Florentinus* – as described above in section 2 – a CSV file was generated and imported in a FileMaker Database connected to the *Trismegistos* environment. In the Database, all Named Entities have been digitally annotated with existing identifiers in TM Geo, TM Author, TM AuthorWork, TM Real, and TM God. In order to map the marvels mentioned in the collection, the TXT file has also been uploaded in the semantic annotation platform *Recogito*[50] in which toponyms and ethnic names pointing to the marvels' geographical locations have been annotated by linking them to existing *Pleiades*[51] identifiers.[52] Since annotation in *Recogito* was specifically meant to map the geographical distribution of marvels, only one geographical reference for each marvel has been annotated (with the exception of c. 34, where the same marvel involves the river Strymon and the city of Apollonia) (see fig. 8).

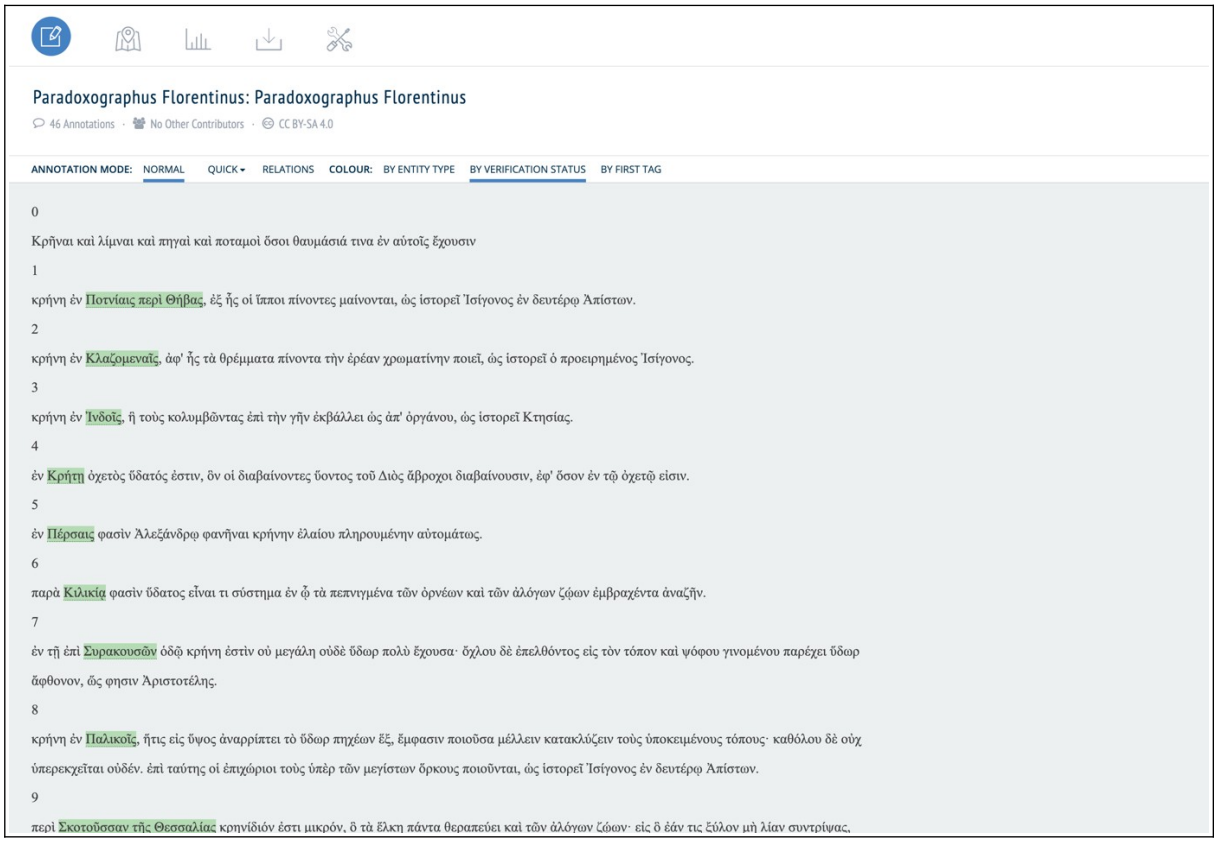


Fig. 8: The annotation of the text in *Recogito*.

The following table contains the results of the digital annotation of the toponyms associated with the marvels mentioned in the *Paradoxographus Florentinus*. In the column "Pleiades", only the identifiers which have been used within the *Recogito* environment have been added, with other possible identifiers within squared brackets. Chapters marked with an asterisk are discussed in the "Textual Notes" below.

As will become clear from the commentary on the individual passages, comparison with parallel sources is often crucial to correctly understand the geographical references in the collection. In c. 18, for example, comparison with a parallel passage in Antigonus' collection allows us to understand the

50 https://recogito.pelagios.org/ (last access 24.07.2025).

51 https://Pleiades.stoa.org/ (last access 24.07.2025).

52 On the annotation platform *Recogito*, see Vitale et. al (2021).

ambiguous location "in Arabia" as a reference to that part of Eastern Egypt between the Nile and the Red Sea which was called Arabia. Similarly, the "Bosporos" in c. 35 likely refers not to the Thracian, but to the Cimmerian Bosporos, as suggested by comparison with Dionysius of Byzantium and Strabo.

However, comparison with parallel sources also poses methodological challenges in those cases in which the compiler seems to have modified or simply misunderstood his source. As a general rule, we have tried to follow as much as possible the geography proposed by the compiler. For example, we have located the marvelous spring in c. 29, with the compiler, "near Carthage", even though his source most likely located it within the territory controlled by the Carthaginians in western Sicily, near Agrigentum. Similarly, we have located the lake mentioned in c. 32, with the compiler, near Abdera, even though his ultimate source (Herodotus) places it near Pistyrus. However, in those cases in which the compiler vaguely locates phenomena which can be located more precisely thanks to parallel sources, we have annotated those phenomena based on the parallel sources (in so far as their locations are compatible with those provided by the compiler). So, the spring of oil vaguely placed by the compiler "among the Persians" in c. 5, has been located, with other sources, along the course of the Oxos river (Amu Darya), while the lake described in c. 33, located by the compiler "in the land of the Nabataians in Arabia", has been identified with the Dead Sea.

Chapter	Greek form	TM Geo	*Pleiades*	Region
1	ἐν Ποτνίαις περὶ Θήβας	Potniai (Tachi): 52457 Thebai: 2354	Potniai: 541070	Boeotia
2	ἐν Κλαζομεναῖς	Klazomenai: 34304	Klazomenai (earlier): 550650 [Klazomenai (later): 550651]	Ionia
3	ἐν Ἰνδοῖς	India: 903	India: 50004	India
4	ἐν Κρήτῃ	Creta: 527	Creta (island): 589748	Crete
5*	ἐν Πέρσαις	Persis: 1704	Oxus (river): 59969 [Persis/Pars: 922698]	Bactria
6	παρὰ Κιλικίᾳ	Cilicia: 526	Cilicia: 658440 [Cilicia: 981514] [Cilicia: 628957]	Cilicia
7	ἐν τῇ ἐπὶ Συρακουσῶν ὁδῷ	Syrakousai (Siracusa): 2210	Syracusae/Syrakousai: 462503	Sicily
8	ἐν Παλικοῖς	Palike: 38477	Palikoi/Palicorum Stagna: 462408 [Palike: 465970]	Sicily
9*	περὶ Σκοτοῦσσαν τῆς Θεσσαλίας	Skotoussa (Hagia Triada): 33399 Thessalia: 2393	Skotoussa (Thessaly): 541107	Thessaly
10	ἐν Λούσοις τῆς	Lousoi: 37599	Lousoi: 570438	Arcadia

	Ἀρκαδίας	Arcadia: 286		
11	ἐν Ἀθαμᾶσι	Athamania: 361	Athamania: 540676	Thessaly/Epirus
12*	παρὰ Κλειτορίοις	Kleitor: 13605	Kleitor: 570359	Arcadia
13	ἐν Ἰταλίᾳ, ἐν τῷ Ῥεατίνῳ ἀγρῷ κρήνην … Μέντην ὀνομαζομένην	Italy: 932 Reate (Rieti): 10929 Mente: 64862	Reate: 413283	Italy (Samnium)
14	ἐγγὺς Κόσης	Cosa (Ansedonia): 32049	Cosa: 413107	Italy (Etruria)
15*	ἐν Χρωψὶ τῆς Θρᾴκης	Chropes: 64854 Thracia: 2414	Thracia: 501638 (tribe cannot be identified)	Thrace
16	περὶ Μαγνησίαν τὴν ἐπὶ Σιπύλου	Magnesia (Manisa): 3575 Sipylos (Manisa Dağı): 60948	Magnesia ad Sipylum: 550706	Lydia
17	ἐν Αἰθιοπίᾳ	Aethiopia: 51	Aethiopia: 39274	Aethiopia
18*	ἐν Ἀραβίᾳ … Ἴσιδος κρήνη	Eileithyiopolis (El-Kab): 611 Arabia: 10930 Isidos Krene: 64863	Eileithyiaspolis: 786020 [Arabia (region in Egypt): 756537]	Egypt
19	Ἄμμωνος κρήνην	Heliou Krene (Ain el-Gubah): 11135	Solis Fons: 716637	Egypt
20	ἐν Λυγκήστῳ	Lyncestis: 33751	Lynkos: 481903	Macedonia
21	ἐν Συκαμίναις πόλει	Sykaminos (Tell el-Samak): 7302	Tel Shiqmona: 678404	Phoenicia
22	ἐν Σαυρομάταις	Sarmatai: 7069	Sarmatia: 825371 [Sarmatae: 226752]	Sarmatia
23*	ἐν Μακροβίοις Αἰθίοψι	Aethiopia: 51	Aethiopia: 39274	Aethiopia
24*	ἐν Κλειτορίοις τῆς Ἀρκαδίας	Kleitor: 13605 Arcadia: 286	Kleitor: 570359	Arcadia
25	ἐν τῇ Κίῳ	Keos (Kea): 1032	Keos (island): 570348	Aegean islands
26	ἐν δὲ Σούσοις τῆς Περσίδος	Sousa: 3634 Persis: 1704	Susa/Seleucia ad Eulaeum: 912936	Susiana

			[Persis/Pars: 922698]	
27	ἐν δὲ Ἀλλιφάνῳ τῆς Ἰταλίας	Allifae (Alife): 14443 Italy: 932	Al(l)ifae: 432658	Italy (Samnium)
28	Ἄουερνός … λίμνη ἐν Ἰταλίᾳ περὶ Κούμας	Avernus Lacus (Lago d'Averno): 42488 Italy: 932 Cumae (Cuma): 14437	Avernus (lake): 432712	Italy (Campania)
29*	κατὰ Καρχηδόνα	Carthago: 484	Carthago: 314921	Africa
30	περὶ Γέλαν τῆς Σικελίας … λίμνη Σίλλα καλουμένη	Gela: 702 Sicilia: 2132 Silla: 64864	Gela: 462214 [Silla Limne: 465992, unlocated]	Sicily
31*	παρὰ τὸν Ἠριδανὸν ποταμὸν … κατὰ τὰς Ἠλεκτρίδας νήσους	Padus (Po): 42384 Elektrides: 61107	Padus/Eridanus (river): 393469	Italy
32*	τὴν κατὰ Ἄβδηρα λίμνην Κύστειρον καλουμένην	Abdera: 14 Kysteiros: 64865	Abdera: 501323	Thrace
33*	ἐν τῇ Ναβαταίων χώρᾳ τῶν Ἀράβων	Nabataea: 1413 Dead Sea: 17357	Mortuum Mare/Asphaltitis Limne: 697709 [Nabataea (region): 29677]	Nabataea
34	εἰς τὰς τοῦ Στρυμόνος ποταμοῦ δίνας … ἐν τῇ περὶ Ἀπολλωνίαν λίμνῃ	Strymon (Struma): 11828 Apollonia (Pojani): 3603	Strymon (river): 501629 Apollonia: 481728	Thrace Illyria
35*	τὸν ἐν Βοσπόρῳ ποταμὸν	Tanais (Don): 38285 Bosporus Cimmerius: 452	Tanais (river): 825398 [Cimmerius Bosp(h)orus: 854675]	Cimmerian Bosporus
36*	περὶ δὲ Ταρρακίναν τῆς Ἰταλίας … λίμνην … †Μυκλαίαν†	Tarracina (Terracina): 32551 Italy: 932 Amyclae: 63781	Tarracina(e)/Anxur: 433143 [Amyclae: 63781, unlocated]	Italy (Latium)

	καλουμένην	Fundanus Lacus (Lago di Fondi): 62976	[Fundanus L.: 432854]	
37.1	ἐπὶ τῆς ἐν Ἰταλίᾳ λίμνης καλουμένης μὲν Βηνάκου	Italy: 932 Benacus Lacus (Lake Garda): 43641	Benacus (lake): 383587	Italy
37.2	ἐν ἑτέρᾳ λίμνῃ τῆς Ἰταλίας Κουτιλίᾳ καλουμένῃ	Italy: 932 Cutiliae Lacus (Laghetto di Paterno): 66083	Cutiliensis (lake): 413115	Italy (Samnium)
38	λάκκος Οὐαδίμωνος καλουμένη λίμνη οὐ μεγάλη ἐν Ἰταλίᾳ	Vadimonis (Laghetto di Bassano): 61100 Italy: 932	Vadimonis (lake): 413370	Italy (Etruria)
39*	ἡ κατὰ Σάρδεις λίμνη καλουμένη δὲ Κολόη	Sardeis (Sart): 2090 Gygaia (Marmara Gölü): 60400	[Sardis/Hyde?: 550867] Gygaia/Koloe/Talaimenis (lake): 550556	Lydia
40	τὸ δὲ κατὰ τὴν Σουσιανὴν ὕδωρ … Μηδείας	Sousiane: 11936 Medeias Hydor: 64866	Elymais/Susiana: 912843	Susiana
41.1	ἐν Ἰταλίᾳ λίμνη Σάβατος καλουμένη	Italy: 932 Sabatinus Lacus (Lago di Bracciano): 52813	Sabatinus lacus: 413290	Italy (Latium)
41.2	περὶ τοῦ Κιμίνου λάκκου ἐν Ἰταλίᾳ	Ciminius: 41807 Italy: 932	Ciminius (lake): 413080	Italy (Etruria)
42	ἡ ἐν Μακεδονίᾳ λίμνη καλεῖται μὲν Λυχνίς	Macedonia: 1279 Lychnidus Lacus: 42523	Lychnidus L.: 481901	Macedonia
43*	ἐν Λυδίᾳ ἔστι λίμνη Τάλα μὲν καλουμένη	Lydia: 1269 Gygaia (Marmara Gölü): 60400	Gygaia/Koloe/Talaimenis (lake): 550556	Lydia

Tab. 1: Annotated places in the *Paradoxographus Florentinus*.

Based on the links to the gazetteer *Pleiades* annotated in the *Recogito* environment (see fig. 10), the following map has been generated:

Fig. 9: *Recogito* map with all the places annotated in the *Paradoxographus Florentinus.*

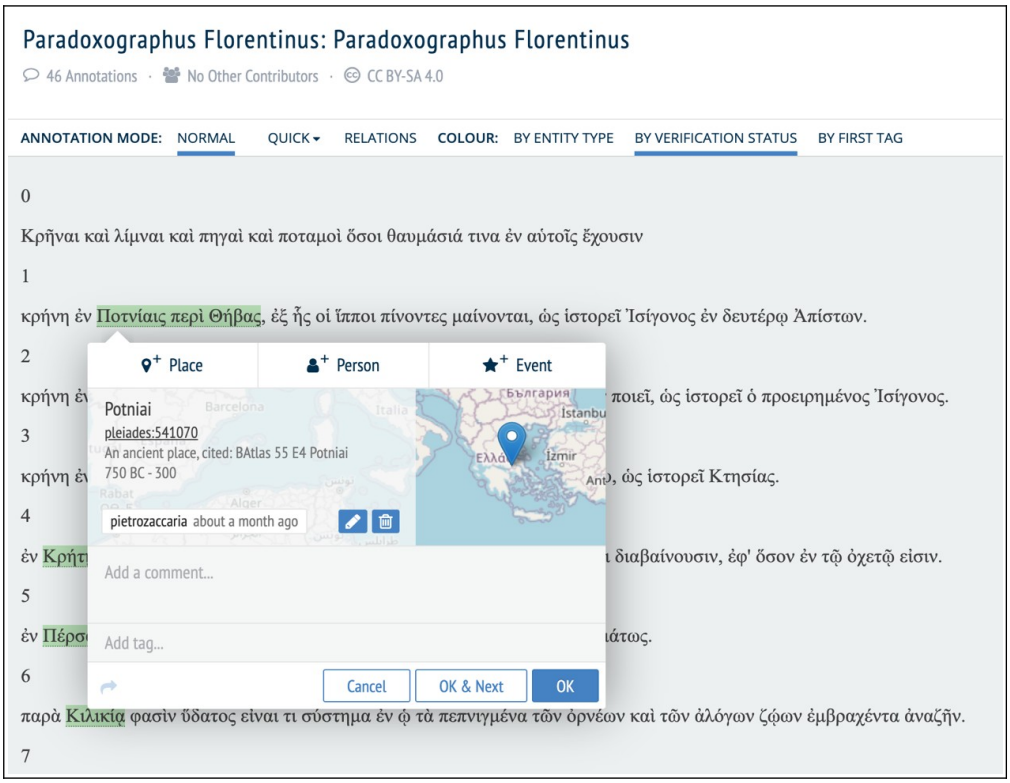


Fig. 10: A detail of the *Recogito* map with the location of Potniai near Thebes (*Paradoxographus Florentinus* 1).

Various interesting elements emerge from the above table and map. In what follows, we limit ourselves to some general considerations. The first is that the compiler always explicitly locates the described phenomena.[53] In general, he does not follow a geographical pattern, apparently jumping from one place to the other without a clear geographical order. Nonetheless, some sequences of chapters do focus on specific areas, which suggests that geography may serve as a structuring principle for some sections of the *Paradoxographus Florentinus*:[54]

- 7–8: Sicily
- 9–12: Greece (Thessaly, Arcadia, Thessaly/Epirus, Arcadia)
- 13–14: Italy (Samnium and Etruria)
- 17–18–19: Aethiopia and Egypt
- 27–31: Italy (Samnium, Campania, Po River) and Sicily; the sequence is interrupted by c. 29 ("near Carthage")
- 36–37.1–37.2–38: Italy (Latium, Garda Lake, Samnium, Etruria)
- 41.1–41.2: Italy (Latium and Etruria)

As persuasively shown by Greene, however, geography seems to work as a principle of organization only on a secondary level.[55] For the geographical sequences identified above seem to be part of source-based sequences. For example, c. 7–8 correspond to ps.-Arist. Mir. 56–57 (in the same order),[56] while c. 13–14 are both based on Isigonus of Nicaea (*FGrHist* 1659 F 8–9), which suggests that these reports were already combined in the same way in the compiler's sources.[57] C. 27–31 concern Italy and Sicily, with the exception of c. 29, which is located "near Carthage"; however, as argued below, this seems to be a misunderstanding on the part of the compiler of an original reference to that part of Sicily controlled by the Carthaginians. This suggests that this sequence is also primarily source-based and that the compiler's source (perhaps Isigonus, mentioned in c. 27 = *FGrHist* 1659 F 11, who may have cited pseudo-Aristotle, who offers parallels for c. 28–31)[58] already grouped the phenomena geographically.[59] Similarly, c. 36–38 may belong to a larger sequence (36–40) based, once again, on Isigonus (mentioned in c. 36 = *FGrHist* 1659 F 12) and perhaps ultimately derived from Varro.[60] As suggested by Greene, even c. 17–18–19, which focus on Aethiopia and Egypt, may be a source-based sequence, since they all have parallels in chapters of Antigonus' collection appearing close to each other, though this hypothesis remains more speculative.[61] Remarkable is also the fact that c. 12 and 24 refer

53 See already Greene (2022), 656.

54 See Greene (2022), 657, with some differences.

55 Greene (2022), 657–662, who identifies sequences based on source and, secondarily, on geography, topic, or other associative links.

56 But note that c. 8 cites Isigonus, *FGrHist* 1659 F 2 as a source.

57 The entire section 6–14 may derive from Isigonus, who may have "ordered his own work with an eye to both geography and theme": see Greene (2022), 659.

58 *Par. Flor.* 28 = ps.-Arist. Mir. 102; *Par. Flor.* 29 = ps.-Arist. Mir. 113; *Par. Flor.* 30 = ps.-Arist. Mir. 112; *Par. Flor.* 31 = ps.-Arist. Mir. 81: see Greene (2022), 659.

59 Cf. Greene (2022), 659–660.

60 See Greene (2022), 660.

61 *Par. Flor.* 17 = Antig. Mir. 145; *Par. Flor.* 18 = Antig. Mir. 149; *Par. Flor.* 19 = Antig. Mir. 144: see Greene (2022), 661–662, who discusses the larger sequence 15–23.

to the same phenomenon in the same location (probably on the basis of different sources). By contrast, c. 39 and 43, which probably refer to the same lake, were apparently considered by the compiler (who uses different toponyms) as two discrete entities (see "Textual Notes").

Something interesting also emerges when looking at the places mentioned by the compiler from a bird's eye view. Somewhat unexpectedly, the large majority of the marvels do *not* concern places located at the edges of the known world, but, broadly speaking, in the Mediterranean area and adjacent regions: mainland Greece and its northern parts (1, 9, 10, 11, 12, 15, 20, 24, 32, 34, 42), Crete and the Aegean islands (4, 25), Asia Minor (2, 6, 16, 39, 43), Sicily (7, 8, 30), Italy (13, 14, 27, 28, 31, 36, 37.1, 37.2, 38, 41.1, 41.2),[62] Egypt (18, 19), Africa (29), Phoenicia (21), and Nabataea (33). Some reports concern regions of the Persian empire in central Asia: Susiana (26, 40) and Bactria (5). Only a few items regard regions at the edges of the known world: Sarmatia (22), the Tanais river in the Bosporus Cimmerius (35), Aethiopia (17, 23), and India (3), with the reports concerning Aethiopia and India deriving from Ctesias and Herodotus (3 = Ctesias, *FGrHist*/*BNJ* 688 F 45sβ; 17 = Ctesias, *FGrHist*/*BNJ* 688 F 1lβ; 23 = Hdt. 3,23).[63] If this (as we suspect) is a general tendency of such collections, it suggests that paradoxography did not primarily concern incredible things placed in remote and almost legendary lands (where other ancient sources admittedly tend to place *mirabilia*),[64] but phenomena that happen somewhat unexpectedly but that still belong to the world which Greek readers were – to different degrees – familiar with (at least, through literature). Commenting on the *Paradoxographus Florentinus*, Greene rightly observes that:

> "While the compilation touches on areas in nearly every corner of the ancient world, few of the phenomena are located beyond the ambit of the Roman empire. […] Many of the marvels may be redolent of the exotic to a second-century reader in the central areas of Greece and Rome, but most are not unreachable."[65]

Not by chance, paradoxography emerges in the third century BCE, in a period when – thanks to, among other factors, Alexander's conquests – the geographical horizon of the Greeks had significantly expanded. Comparing the geographical horizon of Theophrastus' *History of Plants* with that of Aristotle's *History of Animals*, Stevens writes as follows:

> "In Theophrastus' *HP* we see a restructuring of intellectual geography which mirrors the contemporary political and cultural restructuring of the Mediterranean and Near East. Spatially and conceptually, Egypt, central Asia, and India are now more closely and neutrally linked to Greece and the Mediterranean, functioning within Theophrastus' narrative to a greater extent as parallels for parts of the Greek world instead of as monolithic or stereotyped opposites. Rather than the potentially unbounded environments for deadly and exotic animals represented in Aristotle's work, these regions appear in *HP* as territorial entities, connected parts of a world which could now be visited, inhabited, and compared by Greeks. In these very processes of comparison and

62 Great attention is paid to Italy, probably because of a Latin intermediate source, with the toponym "Italy" recurring nine times to locate places from Campania to Northern Italy (13, 27, 28, 36, 37.1, 37.2, 38, 41.1, 41.2). Italy, Sicily, and Carthage represent the western "border" of the *Paradoxographus Florentinus*' geographical horizon: cf. Greene (2022), 657.

63 As is to be expected, geographical references to remote regions are usually more vague than those used to locate marvels happening in countries closer to the Mediterranean world (3: ἐν Ἰνδοῖς; 5: ἐν Πέρσαις; 17: ἐν Αἰθιοπίᾳ; 22: ἐν Σαυρομάταις; 23: ἐν Μακροβίοις Αἰθίοψι; 35: τὸν ἐν Βοσπόρῳ ποταμόν). The only more specific references to remote places mentioned by the compiler are to Susiana (26: ἐν δὲ Σούσοις τῆς Περσίδος; 40: τὸ δὲ κατὰ τὴν Σουσιανὴν ὕδωρ … Μηδείας).

64 As famously put by Plin. nat. 7,21: *praecipue India Aethiopumque tractus miraculis scatent* ("India and parts of Ethiopia especially teem with marvels", transl. Rackham (1942), 519). See Dueck (2012), 64–67; Shipley (2024), 21–22.

65 Greene (2022), 657.

> connection, I would suggest, lies a final aspect of Theophrastus' intellectual geography which can be classed as Hellenistic."[66]

Mutatis mutandis, and although paradoxographical collections are based on a vast array of both pre-Hellenistic and Hellenistic sources and cannot be ascribed systematic geographical views anchored in specific historical circumstances, Stevens' remarks may also apply to the general geographical horizon of paradoxography, where nature serves as a source of wonder from the Mediterranean world up to India and Aethiopia.

Textual Notes:

5: ἐν Πέρσαις φασὶν Ἀλεξάνδρῳ φανῆναι κρήνην ἐλαίου πληρουμένην αὐτομάτως ("They say that in India [*read*: Persia] a spring filling with oil spontaneously appeared to Alexander").[67] Based on this passage alone, one would locate this episode in Persia ("among the Persians"). However, the same episode is also related by several historians of Alexander, who locate this spring along the course of the Oxos river (Amu Darya = TM Geo 42266 = *Pleiades* 59969), which formed the boundary between Bactria and Sogdiana.[68] *Paradoxographus Florentinus*' ἐν Πέρσαις apparently refers not to the region of Persia, but to the domain of the Persians. Although the specific location of the episode is not provided by the *Paradoxographus Florentinus*, in the *Recogito* environment we have annotated ἐν Πέρσαις as a reference to the Oxos river.

9: περὶ Σκοτοῦσσαν τῆς Θεσσαλίας κρηνίδιόν ἐστι μικρόν, ὃ τὰ ἕλκη πάντα θεραπεύει καὶ τῶν ἀλόγων ζώων· εἰς ὃ ἐάν τις ξύλον μὴ λίαν συντρίψας, ἀλλὰ σχίσας ἐμβάλῃ, ἀποκαθίσταται· οὕτως κολλῶδες ἔχει τὸ ὕδωρ, ὥς φησιν Ἰσίγονος ("Near Skotussa in Thessalia there is a small little spring that heals all wounds, even those of beasts of burden. If someone casts a branch into it, one that has not been entirely shattered but split, it is restored. So glutinous is its water, as Isigonos says") (= Isigonus, *FGrHist* 1659 F 5). While there is no doubt that chapter 9 is located by our compiler in Thessalian Scotoussa (TM Geo 33399 = *Pleiades* 541107; cf. also ps.-Arist. Mir. 125: ἐν δὲ Σκοτούσαις τῆς Θετταλίας), it is possible that the ultimate source of this report, Theopompus of Chios, located the phenomenon in Thracian Scotoussa (Skotoussa [Sidirokastro]: TM Geo 34021 = *Pleiades* 491722).[69] Based on the geographical notation of the *Paradoxographus Florentinus*, in the *Recogito* environment we have annotated περὶ Σκοτοῦσσαν τῆς Θεσσαλίας as a reference to the Thessalian city.

12: See below on c. 24.

15: Θεόπομπος ἱστορεῖ κρήνην ἐν Χρωψὶ τῆς Θρᾴκης, ἐξ ἧς τοὺς λουσαμένους παραχρῆμα μεταλλάσσειν ("Theopompos records that there is a spring among the Chropsi in Thrake; those who have bathed in it immediately perish") (= Theopomp., *FGrHist*/*BNJ* 115 F 270c). The name of this Thracian tribe is variously reported in a number of parallel sources:

- Ps.-Arist. Mir. 129: ἐν δὲ †Κύκλωψι† τοῖς Θρᾳξί.[70]
- Antig. Mir. 141: ἐν †κιγχρωψωσιν† τοῖς Θρᾳξίν.[71]

66 Stevens (2016), 148.

67 All translations of the *Paradoxographus Florentinus* are from Greene (2022).

68 See Plu. Alex. 57,4–5; Curt. 7,10,13–14; Arr. An. 4,15,7–8; cf. also Strab. 11,11,5; Ath. 2,42f. Cf. Öhler (1913), 64–65; Greene (2022), 677–682. On the identification of the Oxos river, see Greene (2022), 677 n. 119.

69 See Zaccaria (2024), 97–101. On this report, see also Öhler (1913), 71; Greene (2022), 688–693.

70 Ed. Giacomelli (2023), 162.

71 Ed. Musso (1985), 62–63; cf. also Eleftheriou (2018), 198.

- Plin. nat. 31,27: *in Thracia apud Cychros*.[72]
- Vitr. 8,3,15: *Chrobsi Thracia*.[73]
- *Par. Vat*. 38 simply locates the phenomenon in Thrace (ἐν Θρᾴκῃ).[74]

The form Χρωψί, although probably incorrect, should not be emended, since comparison with Vitruvius' *Chrobsi* suggests that the manuscript reading reflects "what seems to have been a later, if incorrect, understanding of the tribe's name".[75] Unfortunately, none of these ethnic names can be identified with a known Thracian tribe.[76] The flexibility of the digital environment, however, allows us to link different forms to one and the same unique identifier (in this case, TM Geo 64854). Since it is impossible to locate this mysterious tribe, we simply linked this chapter to the ancient region of Thrace in the *Recogito* environment (*Pleiades*: 501638). This annotation is not as precise as one would wish, but it still gives a correct idea of the general location of the described wonder.

18: ἐν Ἀραβίᾳ ἔστιν Ἴσιδος κρήνη, ἥτις κοτύλης οἴνου ἐμβληθείσης κίρναται καὶ πρὸς τὴν πόσιν εὔκρατος γίνεται, ὥς φησιν Ἀμώμητος ("In Arabia there is a fountain of Isis that becomes mixed when a cup of wine is cast into it, and it becomes well-mixed for drinking, as Amometos says") (= Amometus, *FGrHist*/*BNJ* 645 F 1b). This "Fountain of Isis" is located by the *Paradoxographus Florentinus* "in Arabia" (ἐν Ἀραβίᾳ). The precise location is provided by a parallel passage preserved by Antigonus (Mir. 149 = Amometus, *FGrHist*/*BNJ* 645 F 1a), which locates the spring "in Arabia in the *polis* of Leucothea" (TM Geo 611 = *Pleiades* 786020).[77] Based on this parallel passage, we are able to identify the Arabia mentioned by the *Paradoxographus Florentinus* with that part of Egypt between the Nile and the Red Sea (TM Geo 10930 = *Pleiades* 756537).[78]

23: Ἡρόδοτος ἐν Μακροβίοις Αἰθίοψι κρήνην ἱστορεῖ, ἀφ' ἧς τοὺς λουσαμένους λιπαίνεσθαι ("Herodotos reports that among the long-lived Aithiopians there is a spring that anoints those who bathe in it"). As explicitly stated by the compiler, this report is based on Herodotus.[79] More specifically, it comes from Herodotus' digression on Aethiopia (3,23), in which the historian reports that the effects of this marvelous spring could explain the origin of the epithet μακρόβιοι of the Aethiopians living in Libya, on the coast of the southern sea.[80] The *Paradoxographus Florentinus* does not explicitly comment on the spring's link to longevity, but the location of this phenomenon ἐν Μακροβίοις Αἰθίοψι may allude to this,[81] even though Herodotus also refers to this people as the "long-lived Aethiopians" (3,17; 21,3; 3,97). Other accounts of the spring – all ultimately going back to Herodotus – also locate

72 Ed. Mayhoff (1897), 10.

73 Ed. Callebat (1973), 18.

74 Ed. Sørensen (2022a), 584.

75 See Greene (2022), 701, approved by Braccini (2023).

76 See Öhler (1913), 80–83; Greene (2022), 701–704; Giacomelli (2023), 317; Zaccaria (2024), 111–113.

77 Antig. Mir. 149: κατὰ δὲ τὴν Ἀραβίαν ἐν πόλει Λευκοθέᾳ Ἀμώμητόν φησιν γράφειν, τὸν πραγματευθέντα τὸν ἐκ Μέμφεως ἀνάπλουν, εἰς τὴν καλουμένην Ἴσιδος κρήνην ἄν τις οἴνου ἐπιχέῃ κοτύλην, διότι γίγνεται τὸ ποτὸν εὔκρατον ("(Callimachus) says that Amometos, the author of Sailing up from Memphis writes that, in Arabia, in the city of Leukothea, if one pours a measure of wine into the so-called 'spring of Isis', the drink becomes well mixed"; transl. D'Hautcourt (2008), F 1a). Cf. Öhler (1913), 87; Greene (2022), 708–710.

78 Cf. D'Hautcourt (2008), on T 1a; Greene (2022), 708–710.

79 Cf. Öhler (1913), 92–93; Greene (2022), 719–722.

80 Hdt. 3,17: ἐπὶ τοὺς μακροβίους Αἰθίοπας, οἰκημένους δὲ Λιβύης ἐπὶ τῇ νοτίῃ θαλάσσῃ ("against the 'long-lived' Ethiopians, who dwelt on the Libyan coast of the southern sea"); 3,23: τὸ δὲ ὕδωρ τοῦτο εἴ σφί ἐστι ἀληθέως οἷόν τι λέγεται, διὰ τοῦτο ἂν εἶεν, τούτῳ τὰ πάντα χρεώμενοι, μακρόβιοι ("if this water be truly such as they say, it is likely that their constant use of it makes the people long-lived"). Transl. Godley (1921), 25, 31.

81 Greene (2022), 720 n. 369.

it in Aethiopia (Vitr. 8,3,8; Plin. nat. 31,17 = Thphr. F 214D FHS&G; Isid. Orig. 13,13,2) or among the "long-lived Aethiopians" (Mela 3,85–88; Solin. 30,9–11). Like the spring mentioned in c. 17,[82] the spring in c. 23 has thus been linked to the region of Aethiopia (*Pleiades*: 39274). The two springs, although sharing the same general location, are to be understood as discrete.[83]

24: ἐν Κλειτορίοις τῆς Ἀρκαδίας κρήνην φασὶν εἶναι, ἀφ' ἧς τοὺς πίνοντας μισεῖν τὸν οἶνον […] ("In Kleitor in Arkadia they say that there is [a] spring, and that those who drink from it hate wine […]"). The spring in Kleitor in Arcadia has already been described by the compiler in c. 12 (on the authority of Isigonus, *FGrHist* 1659 F 7: παρὰ Κλειτορίοις ὁ αὐτός φησιν εἶναι κρήνην, ἧς ὅταν τις τοῦ ὕδατος πίῃ, τοῦ οἴνου τὴν ὀσμὴν οὐ φέρει, "the same author says that in Kleitor there is a spring. Whenever someone drinks its water, he is not able to bear the smell of wine"), even though the two reports are not identical and likely derive from different sources.[84] Both locations (ἐν Κλειτορίοις τῆς Ἀρκαδίας – παρὰ Κλειτορίοις) have been annotated as *Pleiades* 570359.

29: Ἀριστοτέλης ἱστορεῖ κατὰ Καρχηδόνα κρήνην εἶναι ἐλαίου προσηνεστέραν· ἂν δὲ μή τις ἁγνὸς προσίῃ, ἐκλείπειν αὐτήν ("Aristotle reports that near Carthage there is a spring softer than oil. If someone who is not pure approaches the spring, it disappears") (= ps.-Arist. 113). While other sources locate this marvelous spring near Agrigentum,[85] the *Paradoxographus Florentinus* locates it "near / in the region of Carthage" (κατὰ Καρχηδόνα). This geographical notation seems to stem from the incorrect reading of the corresponding passage in pseudo-Aristotle's *On Marvelous Things Heard* – cited by our compiler as "Aristotle" – which locates this marvel "within the territory controlled by the Carthaginians" (113: ἐν δὲ τῇ ἐπικρατείᾳ τῶν Καρχηδονίων).[86] While pseudo-Aristotle meant western Sicily, near Agrigentum (note that this marvel belongs to a sequence of Sicilian *paradoxa*: ps.-Arist. Mir. 111–115), the *Paradoxographus Florentinus* (or his intermediate source) misleadingly took it as a reference to the region "near Carthage". The same location is also provided by Vitruvius (8,3,8: *Carthagini fons*). We have decided not to correct the geographical picture suggested by the compiler – however misleading – and to locate this marvel where the compiler locates it: near Carthage (TM Geo 484 = *Pleiades* 314921).

31: παρὰ τὸν Ἠριδανὸν ποταμὸν ἔστι λίμνη κατὰ τὰς Ἠλεκτρίδας νήσους, ὕδωρ ἔχουσα θερμόν, ὀσμὴν δὲ βαρεῖαν, ἀφ' ἧς οὐδὲν ζῷον γεύεται ("Along the Eridanos River there is a lake near the Elektridai Islands that has warm water and an oppressive smell, a lake whose water no living creature tastes"). This mephitic lake is located by the compiler along the Eridanus river, not far from the mythical Electrides islands, usually located in the upper Adriatic Sea opposite the mouth of the Po river.

---

82 *Par. Flor.* 17: Κτησίας δὲ ἐν Αἰθιοπίᾳ κρήνην ἱστορεῖ τῷ χρώματι κιννάβαρι παραπλησίαν· τοὺς δὲ πίνοντας ἀπ' αὐτῆς παραλλάττειν τὴν διάνοιαν, ὥστε καὶ τὰ κρυφίως πεπραγμένα ὁμολογεῖν ("Ktesias records that there is a spring in Aithiopia like cinnabar in color. Those who drink from this spring lose their minds, with the result that they also confess their hidden deeds").

83 Greene (2022), 722. The two springs are confused in Asheri et al. (1990), 239.

84 See Greene (2022), 697–700, 728–731. Cf. also Öhler (1913), 74–80.

85 See Plin. nat. 35,179: *in Sicilia Agragantino fonte*; Diosc. Mat. med. 1,73: κατὰ τὴν Ἀκραγαντίνων χώρων τῆς Σικελίας; Sol. 5,22: *in lacu Agrigentino*. Cf. Öhler (1913), 99–103; Greene (2022), 739–742.

86 Ps.-Arist. Mir. 113: ἐν δὲ τῇ ἐπικρατείᾳ τῶν Καρχηδονίων φασὶν ὄρος εἶναι ὃ καλεῖται †Οὐράνιον†, παντοδαπῆς μὲν ὕλης γέμον, πολλοῖς δὲ διαπεποικιλμένον ἄνθεσιν, ὥστε τοὺς συνεχεῖς τόπους ἐπὶ πολὺ μεταλαμβάνοντας τῆς εὐωδίας αὐτῶν ἡδίστην τινὰ τοῖς ὁδοιποροῦσι προσβάλλειν τὴν ἀναπνοήν. πρὸς δὴ τοῦτον τὸν τόπον κρήνην ἐλαίου φασὶν εἶναι, τὴν δὲ ὀσμὴν ἔχειν τῆς κέδρου τοῖς ἀποπτίσμασιν ὁμοίαν. δεῖν δέ φασι τὸν προσιόντα πρὸς αὐτὴν ἁγνὸν εἶναι, καὶ τούτου γινομένου πλεῖον ἀναβλύειν αὐτὴν τὸ ἔλαιον, ὥστε ἀσφαλῶς ἀρύεσθαι ("In the empire of the Carthaginians they say that there is a mountain called Uranium, full of every kind of timber, and made beautiful by many-coloured flowers, so that a succession of places sharing the sweet scent over a large district gives a most delightful air to travellers. At this place they say that there is a spring of oil, which has a scent like the cuttings of cedar. But he who approaches it must be pure, and when this is the case the oil bubbles up more than before, so that it can be safely drawn off"; transl. Hett (1936), 293. Cf. Ath. 2,42f = Thphr. F 214A FHS&G: ἐν τῇ Καρχηδονίων δὲ ἐπικρατείᾳ.

Since these mythical islands cannot be located, we have linked this marvel to the *Pleiades* identifier for the Po river, with which the Eridanus river can be identified. Although the location and identification of the Eridanus varied throughout antiquity, the identification with the Po in our passage is supported by the parallel passage in ps.-Arist. Mir. 81, which places the Electrides, the Eridanus, and the mephitic lake in the Adriatic gulf (ἐν τῷ μυχῷ τοῦ Ἀδρίου).[87]

32: τὴν κατὰ Ἄβδηρα λίμνην Κύστειρον καλουμένην φασὶ τὸ Ξέρξου στράτευμα πῖνον ἀναξηρᾶναι ("They say that there is a lake called Kysteiron near Abdera which Xerxes' army drank dry"). As it stands, this sentence looks like the factual report of a curious historical episode concerning a lake called Kysteiron near Abdera. However, this is in fact a misleading summary of information concerning the march of Xerxes' army through Thrace provided by Herodotus (7,108–109):[88]

> ἔχεται δὲ ταύτης Θασίων πόλις Στρύμη, διὰ δέ σφεων τοῦ μέσου Λίσος ποταμὸς διαρρέει, ὃς τότε οὐκ ἀντέσχε τὸ ὕδωρ παρέχων τῷ Ξέρξεω στρατῷ ἀλλ' ἐπέλιπε. […] διαβὰς δὲ τοῦ Λίσου ποταμοῦ τὸ ῥέεθρον ἀπεξηρασμένον πόλιας Ἑλληνίδας τάσδε παραμείβετο, Μαρώνειαν, Δίκαιαν, Ἄβδηρα. ταύτας τε δὴ παρεξήιε καὶ κατὰ ταύτας λίμνας ὀνομαστὰς τάσδε, Μαρωνείης μὲν μεταξὺ καὶ Στρύμης κειμένην Ἰσμαρίδα, κατὰ δὲ Δίκαιαν Βιστονίδα, ἐς τὴν ποταμοὶ δύο ἐσιεῖσι τὸ ὕδωρ, Τραῦσός τε καὶ Κόμψατος. κατὰ δὲ Ἄβδηρα λίμνην μὲν οὐδεμίαν ἐοῦσαν ὀνομαστὴν παραμείψατο Ξέρξης, ποταμὸν δὲ Νέστον ῥέοντα ἐς θάλασσαν. μετὰ δὲ ταύτας τὰς χώρας Θασίων τὰς ἠπειρώτιδας πόλις παρήιε, τῶν ἐν μιῇ λίμνη ἐοῦσα τυγχάνει ὡσεὶ τριήκοντα σταδίων μάλιστά κῃ τὴν περίοδον, ἰχθυώδης τε καὶ κάρτα ἁλμυρή· ταύτην τὰ ὑποζύγια μοῦνα ἀρδόμενα ἀνεξήρηνε. τῇ δὲ πόλι ταύτῃ οὔνομά ἐστι Πίστυρος.[89]

> "Next to it [*sc*. Mesambria] is a Thasian town, Stryme; between them runs the river Lisus, which now could not furnish water enough for Xerxes' army, but was exhausted. […] Having crossed the bed (then dried up) of the river Lisus he passed by the Greek cities of Maronea, Dicaea, and Abdera. Past these he went, and past certain lakes of repute near to them, the Ismarid lake that lies between Maronea and Stryme, and near Dicaea the Bistonian lake, into which the rivers Travus and Compsantus disembogue. Near Abdera Xerxes passed no lake of repute, but crossed the river Nestus where it flows into the sea. From these regions he passed by the cities of the mainland, one whereof has near it a lake of about thirty furlongs in circuit, full of fish and very salt; this was drained dry by no more than the watering of the beasts of burden. This town is called Pistyrus."[90]

87 See Greene (2022), 745–746. Cf. also Öhler (1913), 106–107.

88 Cf. Greene (2022), 748–750.

89 Ed. Wilson (2015), 633–634.

90 Transl. Godley (1922), 413–415.

The short report provided by the *Paradoxographus Florentinus* combines various elements which are ultimately based on Herodotus' account:

Paradoxographus Florentinus	**Herodotus**
τὴν κατὰ Ἄβδηρα λίμνην	κατὰ δὲ Ἄβδηρα λίμνην
Κύστειρον[91]	Πίστυρος [Πίστυρος Cr : Πύστιρος ADV]
τὸ Ξέρξου στράτευμα	τῷ Ξέρξεω στρατῷ
πῖνον	ἐπέλιπε
ἀναξηρᾶναι	ἀπεξηρασμένον … ἀνεξήρηνε

Tab. 2: Comparison between *Par. Flor.* 32 and Hdt. 7,108–109.

However, the two accounts significantly differ. While the *Paradoxographus Florentinus* claims that a lake called Kysteiron near Abdera was drunk dry by Xerxes' army, Herodotus reports (1) that the army drank dry the river Lisus; (2) that Xerxes passed *no lake of repute* near Abdera; and (3) that the beasts of burden drank dry an unnamed lake near the town of Pistyrus. Probably drawing on an intermediate source (note the compiler's use of φασί),[92] the *Paradoxographus Florentinus* seems to have created a new historical anecdote by combining elements derived from Herodotus' account. The question now arises of how to represent the report of the *Paradoxographus Florentinus* on a map. One may locate the lake Kysteiron (TM Geo 52835) near Abdera (TM Geo 14 = *Pleiades* 501323), based on the *Paradoxographus Florentinus*, or near Pistyrus (TM Geo 52835 = *Pleiades* 501569),[93] based on the Herodotean account. Since our aim is to represent, as far as possible, the geographical horizon of the compiler, we have eventually decided to locate the lake where the compiler locates it: near Abdera.

33: Ἱερώνυμος ἱστόρησεν ἐν τῇ Ναβαταίων χώρᾳ τῶν Ἀράβων εἶναι λίμνην πικράν, ἐν ᾗ οὔτ' ἰχθῦς οὔτε ἄλλο τι τῶν ἐνύδρων ζῴων γίνεσθαι· ἀσφάλτου δὲ πλίνθους ἐξ αὐτῆς αἴρεσθαι ὑπὸ τῶν ἐπιχωρίων ("Hieronymos reports that in the land of the Nabataians in Arabia there is a bitter lake in which no fish nor any other aquatic creature lives, but bricks of asphalt are collected from it by the local inhabitants") (= Hieronymus of Cardia, *FGrHist*/*BNJ* 154 F 5). The lake described here is the Dead Sea (TM Geo 17357 = *Pleiades* 697709), known in antiquity as the Asphaltitis or Asphaltites lake.[94] Even though the *Paradoxographus Florentinus* does not explicitly identify the lake, we have linked this item to the Dead Sea, since its location fits with the geographical notation provided by the compiler.

91 A lake called Κύστειρον is otherwise unattested and seems to be based on Herodotus' Πίστυρος (which, however, is the name of a city). According to Greene (2022), 749, "Κύστειρον is otherwise unattested, though it likely is the product of a conflation of κύστις ('bladder, pouch') and Πίστυρος. Mistake or not, a name based upon κύστις, which is regularly used of drinking pouches, befits the lake in light of its fate." However, it should be noted that a city called Kystiros (unlocated) is attested: see Hansen (2004), 1250, no. 1033. Macan (1908), 140 identifies this Kystiros with Herodotus' Πίστυρος.

92 See Greene (2022), 749.

93 On the uncertain identification and location of Pistyrus, see Loukopoulou (2004), 866–867; Vannicelli et al. (2017), 424, with references. There was also an *emporion* called Pistiros (TM Geo 38392), see *SEG* 43 (1993) no. 486; St. Byz. π 162, s.v. Πίστιρος, with Billerbeck / Neumann-Hartmann (2016), 73 n. 229.

94 See Öhler (1913), 108–109; Greene (2022), 750–752.

35: Φαέθων φησὶ τὸν ἐν Βοσπόρῳ ποταμὸν οὕτως εἶναι ψυχρόν, ὥστε μηδὲν τῶν ζῴων ὑπομένειν αὐτοῦ τὴν ψυχρότητα ("Phaethon says that the river in the Bosporos is so frigid that no living creature abides its extreme cold"). Commenting on this passage, Greene says:

> "A prohibitively cold river at the Bosporos is otherwise unattested, though the use of the definite article suggests that the compiler refers to a specific and well-known river. Throughout his exile poems Ovid complains of a number of rivers in the area near Tomis on the central-western shore of the Black Sea that are very cold or freeze in winter, though one would expect that the compiler has a river closer to the Bosporos in mind (e.g. the Alibeyköy or Kağıthane rivers, which feed into the Bosporos inlet known as the Golden Horn). Another possibility is that the report ultimately stems from an account that relates the freezing of either the Bosporos strait itself or the Golden Horn. Both were known to freeze during particularly harsh winters, and ancient and medieval authors found their freezing worthy of commemoration."[95]

A convincing solution has been recently suggested by Braccini, who argued that this passage makes perfect sense if we identify the Bosporus mentioned by the *Paradoxographus Florentinus* not with the Thracian Bosporus (TM Geo 61002 = *Pleiades* 520977), but with the Cimmerian Bosporus (TM Geo 452 = *Pleiades* 854675), in which case the frigid river can be identified with the Tanais (TM Geo 38285 = *Pleiades* 825398).[96] See Dion. Byz. Anaplus Bospori 2: τὸ δὲ πέρας ποταμὸς ὁ Τάναϊς, ὅρος τῶν δυεῖν ἠπείρων, ἀνατέλλων ἐκ τῆς διὰ κρυμὸν ἀοικήτου ("the river Tanais forms its limit, the boundary between two continents, and springing up from an area, uninhabited on account of the icy cold")[97]; cf. also Strab. 2,5,26: διὰ ψῦχος ἀοικήτου ("uninhabited because of the cold"); 11,2,2 (with regard to the Tanais): τοῦ δ᾽ ὑπὲρ τῶν ἐκβολῶν ὀλίγον τὸ γνώριμόν ἐστι διὰ τὰ ψύχη καὶ τὰς ἀπορίας τῆς χώρας ("but little of the part that is beyond its outlets is known to us, because of the coldness and the poverty of the country"[98]). Accordingly, we have annotated the textual string τὸν ἐν Βοσπόρῳ ποταμὸν as a reference to the Tanais river (TM Geo 38285 = *Pleiades* 825398).

36: περὶ δὲ Ταρρακίναν τῆς Ἰταλίας φησὶν Ἰσίγονος λίμνην εἶναι †Μυκλαίαν† καλουμένην καὶ παρ᾽ αὐτῇ πόλιν ἔρημον, ἧς τοὺς ἐνοικοῦντας στερηθῆναι τῆς πόλεως διὰ τὸ πλῆθος τῶν ὕδρων ("Isigonos says that around Tarrakine in Italy there is a lake called the Amyklaia. And around it there is a deserted city, whose inhabitants were robbed of their city by water snakes") (= Isigonus, *FGrHist* 1659 F 12). This marvel can be firmly located on the basis of the reference to the city of Tarracina (TM Geo 32551 = *Pleiades* 433143). Thanks to a parallel account offered by Pliny (nat. 3,59), moreover, we can reasonably consider the corrupted form †Μυκλαίαν† as a reference to Amyclae/Amynclae, a city located in Latium between Tarracina and Caieta (Gaeta) (TM Geo 63781 = *Pleiades* 63781, unlocated): *dein flumen Aufentum, supra quod Tarracina oppidum lingua Volscorum Anxur dictum, et ubi fuere Amyclae sive Amynclae a serpentibus deletae, dein locus Speluncae, lacus Fundanus, Caieta Portus etc*. ("Then comes the river Aufentum, above which is the town of Tarracina, called Anxur in the dialect of the Volsci, and the site of Amyclae, or Amynclae, the town destroyed by serpents, then the place called the Grottoes, Lake Fundanus, the port of Gaeta";[99] cf. also nat. 8,104: *in Italia Amynclas a serpentibus deletas*; Sol. 2,32; Serv. aen. 10,564). The Μυκλαία λίμνη mentioned by the *Paradoxographus Florentinus* is therefore probably to be identified with the Lago di Fondi (lacus Fundanus) (TM Geo 62976 = *Pleiades* 432854).[100] Μυκλαίαν has been corrected to <Ἀ>μυκλαίαν by Landi and Giannini

95 Greene (2022), 755.

96 Braccini (2023).

97 Transl. Nicholson / Russell (2024), 828.

98 Transl. Jones (1928), 193. Cf. Billerbeck (2023), 109–110.

99 Transl. Rackham (1942), 45.

100 Greene (2022), 755–757. Cf. also Öhler (1913), 111–112.

on the basis of Pliny's parallel passages.[101] However, since Pliny's manuscripts at 8,104 also have the variants *Minclas* and *Mynclas*,[102] Greene refrained from emending the transmitted text of the *Paradoxographus Florentinus*.[103]

39: See below on c. 43.

43: ἐν Λυδίᾳ ἔστι λίμνη Τάλα μὲν καλουμένη, ἱερὰ δὲ οὖσα νυμφῶν, ἣ φέρει καλάμων πλῆθος καὶ μέσον αὐτῶν ἕνα, ὃν βασιλέα προσαγορεύουσιν οἱ ἐπιχώριοι. θυσίας δὲ καὶ ἑορτὰς ἐπιτελοῦντες ἐνιαυσίους ἐξιλάσκονται· τούτων δὲ ἐπιτελουμένων, ἐπειδὰν ἐπὶ τῆς ἠιόνος κτύπος συμφωνίας γένηται, πάντες οἱ κάλαμοι χορεύουσι καὶ ὁ βασιλεὺς σὺν αὐτοῖς χορεύων παραγίνεται ἐπὶ τὴν ἠιόνα· οἱ δὲ ἐπιχώριοι ταινίαις αὐτὸν καταστέψαντες ἀποπέμπουσιν, εὐχόμενοι καὶ εἰς τὸ ἐπιὸν αὐτόν τε καὶ ἑαυτοὺς παραγενέσθαι ὡς εὐετηρίας ὄντι σημείῳ, ὡς ἱστορεῖ Ἰσίγονος ἐν δευτέρῳ Ἀπίστων ("In Lydia there is a lake called Tala, sacred to the nymphs, which carries a great number of reeds. In their midst there is one that the local inhabitants call the "king". They propitiate it by holding annual sacrifices and festivals. When these are held, whenever the sound of music occurs on the shore, all the reeds dance and the king, dancing along with them, comes to the shore. And the local inhabitants send it away after they have crowned it with fillets, praying that it will come to them again next year, since it is a sign of prosperity, as Isigonos records in the second book of *Unbelievable things*") (= Isigonus, *FGrHist* 1659 F 3). The Τάλα lake is probably identical with the lake Κολόη mentioned in c. 39.1 (ἡ κατὰ Σάρδεις λίμνη καλουμένη δὲ Κολόη πλῆθος μὲν ὄψου πάμπολυ τρέφει· ἔχει δὲ καὶ αὐτὴ νήσους οἰκουμένας πρὸς ἀπάτην· ἐπινήχονται γὰρ καὶ τῇ τῶν ἀνέμων πνοῇ συμμετοικοῦσι, "A lake near Sardis called the Koloë produces a great amount of food, but it also has islands that only seem settled. For they swim upon the lake and move along with the blast of the wind"), though the compiler understands them to be discrete, since he uses different names.[104] Both chapters have therefore been linked to the same TM Geo (60400) and *Pleiades* (550556) identifiers.

4. Conclusion

The article has described the first steps towards the creation of a relational database of ancient paradoxography, one which can allow us to extract, annotate, and analyse the numerous Named Entities mentioned in paradoxographical texts. The database will not only serve as a tool to apply digital approaches to ancient paradoxography, it will also allow scholars to export reliable linguistic data and metadata to be used in larger projects.

Moreover, the article has discussed the digital annotation in the *Trismegistos* and *Recogito* environments of the toponyms mentioned in the so-called *Paradoxographus Florentinus*. Despite the historical and philological problems posed by paradoxographical toponyms and the methodological issues raised by the visualization of toponyms mentioned by ancient texts on a modern map, the case study of the *Paradoxographus Florentinus* shows that the digital mapping of toponyms can be a valuable research tool to better understand the structure of paradoxographical texts and, more generally, the geographical horizon of ancient paradoxography.

101 Landi (1895), 536; Giannini (1965), 326–327.

102 Mayhoff (1909), 114; Ernout (1952), 60.

103 Greene (2022), 755–756.

104 See Greene (2022), 763, 772–774, with references. Cf. also Öhler (1913), 117–122.

References

Asheri et al. (1990): D. Asheri / S. M. Medaglia / A. Fraschetti, Erodoto, Le Storie, Volume III, Libro III, La Persia, Introduzione e commento di D. Asheri, Testo critico di S. M. Medaglia, Traduzione di A. Fraschetti, Milano 1990.

Barker et al. (2016a): E. Barker / S. Bouzarovski / L. Isaksen, Introduction: Creating New Worlds out of Old Texts, in: E. Barker / S. Bouzarovski / C. Pelling / L. Isaksen (eds.), New Worlds from Old Texts: Revisiting Ancient Space and Place, Oxford 2016, 1–21.

Barker et al. (2016b): E. Barker / S. Bouzarovski / C. Pelling / L. Isaksen (eds.), New Worlds from Old Texts: Revisiting Ancient Space and Place, Oxford 2016.

Barker et al. (2023): E. Barker / K. Konstantinidou / B. Kiesling / A. Foka, Journeying through Space and Time with Pausanias's Description of Greece, in: Literary Geographies 9/1 (2023), 124–160.

Barker et al. (2024): E. Barker / C. Palladino / S. Gordin, Digital Approaches to Investigating Space and Place in Classical Studies, in: The Classical Review 74/1 (2024), 1–19, https://doi.org/10.1017/S0009840X23002858 (last access 24.03.2026).

Bekker (1831): I. Bekker, Aristotelis opera, vol. 2, Berlin 1831.

Berti (2019): M. Berti, Named Entity Annotation for Ancient Greek with INCEpTION, in: K. Simov / M. Eskevich (eds.), Proceedings of CLARIN Annual Conference 2019, Leipzig 2019, 1–4.

Berti (2021): M. Berti, Digital Editions of Historical Fragmentary Texts, Heidelberg 2021, https://doi.org/10.11588/propylaeum.898 (last access 15.04.2026).

Berti (2024): M. Berti, Digital Canons and Catalogs of Fragmentary Literature, in: F. Neuerburg / T. Tsiampokalos / P. Wozniczka (eds.), Fragmente einer fragmentierten Welt. Zur Problematik des Umgangs mit Fragmenten in der gegenwärtigen klassisch-philologischen Forschung, Berlin 2024, 217–236, https://doi.org/10.1515/9783111508788-009 (last access 24.03.2026).

Berti (2025): M. Berti, Linked Ancient Greek and Latin (LAGL) and Wikidata: Structuring and Reusing Data of Classical Literature (discussion paper): in Journal of Open Humanities Data 11/72, 2025, 1–12, https://doi.org/10.5334/johd.423 (last access 15.04.2026).

Berti (2026): M. Berti, Annotating the Ancient World. Critical Annotations and Digital Editions: in P. d'Hoine / D. Kohler / W. Decock (eds.), Charting the Future of Historical Humanities, Turnhout 2026, 21–46.

Bianchetti et al. (2016): S. Bianchetti / M. R. Cataudella / H.-J. Gehrke (eds.), Brill's Companion to Ancient Geography. The Inhabited World in Greek and Roman Tradition, Leiden / Boston 2016.

Billerbeck (2023): M. Billerbeck, Dionysios von Byzanz, Anaplus Bospori, Die Fahrt auf dem Bosporos, Einleitung, Text, Übersetzung und Kommentar, Basel 2023.

Billerbeck / Neumann-Hartmann (2016): M. Billerbeck / A. Neumann-Hartmann, Stephani Byzantii Ethnica, Volumen IV: Π–Y, Berlin / Boston 2016.

Braccini (2023): T. Braccini, Review of S. Schorn (ed.), Die Fragmente der Griechischen Historiker Continued IV E: Paradoxography and Antiquities. Fascicle 2. Paradoxographers of the Imperial Period and Undated Authors [Nos. 1667–1693], Leiden / Boston 2022, in: BMCR 2023.10.10.

Callebat (1973): L. Callebat, Vitruve. De l'architecture. Livre VIII. Texte établi traduit et commenté par L. C., Paris 1973.

Castro-Páez / Cruz Andreotti (2020): E. Castro-Páez / G. Cruz Andreotti (eds.), Geografía y cartografía de la Antigüedad al Renacimiento. Estudios en honor de Francesco Prontera, Alcalá de Henares 2020.

de Martini (2023): A. de Martini, Il cosiddetto Paradoxographus Palatinus. Edizione critica, traduzione e commento filologico e interpretativo, Diss. Università degli Studi di Genova 2023.

Depauw / Gheldof (2014): M. Depauw / T. Gheldof, Trismegistos. An Interdisciplinary Platform for Ancient World Texts and Related Information, in: Ł. Bolikowski / V. Casarosa / P. Goodale / N. Houssos / P. Manghi / J. Schirrwagen (eds.), Theory and Practice of Digital Libraries – TPDL 2013 Selected Workshops (Communications in Computer and Information Science 416), Cham 2014, 40–52, https://doi.org/10.1007/978-3-319-08425-1_5 (last access 15.04.2026).

D'Hautcourt (2008): A. D'Hautcourt, Amometos (645), in: Jacoby Online. Brill's New Jacoby, Part III, edited by I. Worthington, Leiden 2008, https://doi.org/10.1163/1873-5363_bnj_a645 (last access 15.04.2026).

Dueck (2012): D. Dueck, Geography in Classical Antiquity, with a chapter by K. Brodersen, Cambridge 2012.

Eleftheriou (2018): D. Eleftheriou, Pseudo-Antigonos de Carystos. Collection d'histoires curieuses, 1. Introduction – édition – traduction, Diss. Université Paris Nanterre 2018.

Elliott / Gillies (2009): T. Elliott / S. Gillies, Digital Geography and Classics, in: DHQ 3/1 (2009), https://www.digitalhumanities.org/dhq/vol/3/1/000031/000031.html (last access 24.07.2025).

Ernout (1952): A. Ernout, Pline l'Ancien. Histoire naturelle, livre VIII, Paris 1952.

García Teijeiro / Molinos Tejada (1994): M. García Teijeiro / M. T. Molinos Tejada, Paradoxographie et religion, Kernos 7 (1994), 273–285.

Gerolemou (2018): M. Gerolemou (ed.), Recognizing Miracles in Antiquity and Beyond, Berlin / Boston 2018.

Geus (2016): K. Geus, Paradoxography and Geography in Antiquity. Some Thoughts About the Paradoxographus Vaticanus, in: F. J. Gonzáles Ponce / F. J. Gómez Espelosín / A. L. Chávez Reino (eds.), La letra y la carta. Descripción verbal y representación gráfica en los diseños terrestres grecolatinos. Estudios en honor de Pietro Janni, Sevilla 2016, 243–257.

Geus / King (2018): K. Geus / C. G. King, Paradoxography, in: P. T. Keyser / J. Scarborough (eds.), Oxford Handbook of Science and Medicine in the Classical World, Oxford 2018, https://doi.org/10.1093/oxfordhb/9780199734146.013.20 (last access 15.04.2026).

Giacomelli (2021): C. Giacomelli, Ps.-Aristotele, De mirabilibus auscultationibus. Indagini sulla storia della tradizione e ricezione del testo, Berlin 2021.

Giacomelli (2023): C. Giacomelli, Pseudo-Aristotele, De mirabilibus auscultationibus. Edizione critica, traduzione e commento filologico, Roma 2023.

Giacomelli (2024): C. Giacomelli, Suspicious Toponyms in the De mirabilibus auscultationibus: Textual Problems, "Forgeries," and Methodological Issues, in: S. Schorn / R. Mayhew (eds.), Historiography and Mythography in the Aristotelian Mirabilia, London / New York 2024, 234–257.

Giannini (1963): A. Giannini, Studi sulla paradossografia greca. I. Da Omero a Callimaco: motivi e forme del meraviglioso, in: Rendiconti / Istituto Lombardo, Accademia di Scienze e Lettere, Classe di Lettere, Scienze morali e storiche 97 (1963), 247–266.

Giannini (1964): A. Giannini, Studi sulla paradossografia greca. II. Da Callimaco all'età imperiale: la letteratura paradossografica, Acme 17 (1964), 99–140.

Giannini (1965): A. Giannini, Paradoxographorum Graecorum Reliquiae. Milano 1965.

Godely (1921): A.D. Godley, Herodotus, II, Books III–IV, London / Cambridge (MA) 1921.

Godley (1922): A.D. Godley, Herodotus, III, Books V–VII, London / Cambridge (MA) 1922.

Gómez Espelosín (1996): F. J. Gómez Espelosín, Paradoxógrafos Griegos. Rareza y maravillas, Madrid 1996.

Gonzáles Ponce et al. (2016): F. J. Gonzáles Ponce / F. J. Gómez Espelosín / A.L. Chávez Reino (eds.), La letra y la carta. Descripción verbal y representación gráfica en los diseños terrestres grecolatinos. Estudios en honor de Pietro Janni, Sevilla 2016.

Greene (2019): R. Greene, A Most Amazing Conversation: The Social Contexts of Wonder-Telling and the Development of Paradoxography, NECJ 46 (2019), 28–45.

Greene (2022): R. Greene, 1680. Paradoxographus Florentinus, in: S. Schorn (ed.), Die Fragmente der Griechischen Historiker Continued IV E: Paradoxography and Antiquities. Fascicle 2. Paradoxographers of the Imperial Period and Undated Authors [Nos. 1667–1693], 633–785, Leiden 2022 (online edition [2018], https://doi.org/10.1163/1873-5363_jciv_a1680 (last access 15.04.2026).

Hansen (2004): M. H. Hansen, Unlocated, in: M. H. Hansen / T. H. Nielsen (eds.), An Inventory of Archaic and Classical Poleis, Oxford 2004, 1250.

Hett (1936): W. S. Hett, Aristotle. Minor Works, London / Cambridge (MA) 1936.

Ideler (1841): J. L. Ideler, Physici et Medici Graeci Minores. Volumen I, Berlin 1841.

Jacob (1983): C. Jacob, De l'art de compiler à la fabrication du merveilleux. Sur la paradoxographie grecque, Lalies 2 (1983), 121–140.

Jacoby (1929): F. Jacoby, Die Fragmente der Griechischen Historiker. Zweiter Teil: Zeitgeschichte. B: Spezialgeschichten, Autobiographien und Memoiren. Zeittafeln, Berlin 1929.

Jones (1928): H. L. Jones, Strabo. Geography. Books 10–12, Cambridge (MA) / London 1928.

Kazantzidis (2019): G. Kazantzidis (ed.), Medicine and Paradoxography in the Ancient World, Berlin 2019.

Landi (1895): C. Landi, Opuscula de fontibus mirabilibus, de Nilo etc. ex cod. Laur. 56, 1 descripta, SIFC 3 (1895), 531–548.

Lightfoot (2021): J. Lightfoot, Wonder and the Marvellous from Homer to the Hellenistic World, Cambridge 2021.

Loukopoulou (2004): L. Loukopoulou, Thrace from Strymon to Nestos, in: M. H. Hansen / T. H. Nielsen (eds.), An Inventory of Archaic and Classical Poleis, Oxford 2004, 854–869.

Macan (1908): R. W. Macan, Herodotus. The Seventh, Eight, & Ninth Books, Vol. I, Part I, London 1908.

Mayhoff (1909): C. Mayhoff, C. Plini Secundi Naturalis Historiae libri XXXVII, vol. II, libri VII–XV, Stuttgart 1909.

McInerney (2012): J. McInerney, Phlegon of Tralles (257), in: I. Worthington (ed.), Jacoby Online. Brill's New Jacoby, Part II, Leiden 2012, https://doi.org/10.1163/1873-5363_bnj_a257 (last access 15.04.2026).

Musso (1985): O. Musso, Antigonus Carystius. Rerum mirabilium collectio, Napoli 1985.

Nichols (2018): A. Nichols, Ctesias' Indica and the Origins of Paradoxography, in: M. Gerolemou (ed.), Recognizing Miracles in Antiquity and Beyond, Berlin / Boston 2018, 3–16.

Nicholson / Russell (2024): O. Nicholson / T. Russell, Dionysios of Byzantion, in: D.G.J. Shipley (ed.), Geographers of the Ancient Greek World: Selected Texts in Translation, Volume II, Cambridge 2024, 820–852.

Nouvel et al. (2016): D. Nouvel / M. Ehrmann / S. Rosset, Named Entities for Computational Linguistics, London / Hoboken (NJ) 2016.

Öhler (1913): H. Öhler, Paradoxographi Florentini anonymi opusculum de aquis mirabilibus, ad fidem codicum manu scriptorium editum commentario instructum, Tübingen 1913.

Pajón Leyra (2011): I. Pajón Leyra, Entre ciencia y maravilla. El género literario de la paradoxografía griega, Zaragoza 2011.

Pajón Leyra (2022): I. Pajón Leyra, Mythography and Paradoxography, in: R. Scott Smith / S.M. Trzaskoma (eds.), The Oxford Handbook of Greek and Roman Mythography, New York 2022, 396–408.

Pajón Leyra (2024): I. Pajón Leyra, Islands and Their Marvels as Structural Principle in the So-Called Historiographical Section of the De mirabilibus auscultationibus, in: S. Schorn / R. Mayhew (eds.), Historiography and Mythography in the Aristotelian Mirabilia, London / New York 2024, 10–31.

Pfeiffer (1949): R. Pfeiffer, Callimachus, Volumen I, Fragmenta, Oxford 1949.

Rackham (1942): H. Rackham, Pliny. Natural History. Books 3–7, Cambridge (MA) / London 1942.

Rathmann (2007): M. Rathmann (ed.), Wahrnehmung und Erfassung geographischer Räume in der Antike, Mainz 2007.

Roller (2015): D. W. Roller, Ancient Geography: The Discovery of the World in Classical Greece and Rome, London / New York 2015.

Roller (2019): D. W. Roller (ed.), New Directions in the Study of Ancient Geography, University Park (PA) 2019.

Rusten / Yu (2022): J. S. Rusten / K. W. Yu, Paradoxography, in: Oxford Classical Dictionary 2022, https://doi.org/10.1093/acrefore/9780199381135.013.4728 (last access 15.04.2026).

Sassi (1993): M. M. Sassi, Mirabilia, in: G. Cambiano / L. Canfora / D. Lanza (eds.), Lo spazio letterario della Grecia antica, Volume I, La produzione e la circolazione del testo, Tomo II, L'ellenismo, Rome 1993, 449–468.

Schepens / Delcroix (1996): G. Schepens / K. Delcroix, Ancient Paradoxography: Origin, Evolution, Production and Reception, in O. Pecere / A. Stramaglia (eds.), La letteratura di consumo nel mondo greco-latino, Cassino 1996, 373–460.

Schepens / Schorn (2010): G. Schepens / S. Schorn, Verkürzungen in und von Historiographie in klassischer und hellenistischer Zeit, in: M. Horster / C. Reitz (eds.), Condensing Texts – Condensed Texts, Stuttgart 2010, 395–433.

Schorn (2022): S. Schorn (ed.), Die Fragmente der Griechischen Historiker Continued IV E: Paradoxography and Antiquities. Fascicle 2. Paradoxographers of the Imperial Period and Undated Authors [Nos. 1667–1693], Leiden / Boston 2022.

Schorn / Mayhew (2024): S. Schorn / R. Mayhew (eds.), Historiography and Mythography in the Aristotelian Mirabilia, London / New York 2024.

Shannon-Henderson (2013): K. E. Shannon-Henderson, Authenticating the Marvellous: 'Mirabilia' in Pliny the Younger, Tacitus, and Suetonius, in: Working Papers on Nervan, Trajanic and Hadrianic Literature 1/9 (2013), 1–26.

Shannon-Henderson (2022): K. E. Shannon-Henderson, 1667. Phlegon of Tralleis, in: S. Schorn (ed.), Die Fragmente der Griechischen Historiker Continued IV E: Paradoxography and Antiquities. Fascicle 2. Paradoxographers of the Imperial Period and Undated Authors [Nos. 1667–1693], Leiden / Boston 2022, 9–338. (online edition [2019], https://doi.org/10.1163/1873-5363_jciv_a1667 [last access 15.04.2026]).

Shipley (2024): D. G. J. Shipley (ed.), Geographers of the Ancient Greek World: Selected Texts in Translation, 2 Vols., Cambridge 2024.

Sørensen (2022a): S.L. Sørensen, 1679. Paradoxographus Vaticanus, in: S. Schorn (ed.), Die Fragmente der Griechischen Historiker Continued IV E: Paradoxography and Antiquities. Fascicle 2. Paradoxographers of the Imperial Period and Undated Authors [Nos. 1667–1693], Leiden / Boston 2022, 579–632 (online edition [2020], https://doi.org/10.1163/1873-5363_jciv_a1679 [last access 15.04.2026]).

Sørensen (2022b): S.L. Sørensen, 1681. Paradoxographus Palatinus, in: S. Schorn (ed.), Die Fragmente der Griechischen Historiker Continued IV E: Paradoxography and Antiquities. Fascicle 2. Paradoxographers of the Imperial Period and Undated Authors [Nos. 1667–1693], Leiden / Boston 2022, 787–831 (online edition [2017], https://doi.org/10.1163/1873-5363_jciv_a1681 [last access 15.04.2026]).

Spittler (2022): J. Spittler, 1672. Apollonios, in: S. Schorn (ed.), Die Fragmente der Griechischen Historiker Continued IV E: Paradoxography and Antiquities. Fascicle 2. Paradoxographers of the Imperial Period and Undated Authors [Nos. 1667–1693], Leiden / Boston 2022, 387–519. (online edition [2016], https://doi.org/10.1163/1873-5363_jciv_a1672 [last access 15.04.2026]).

Stern (2008): J. Stern, Paradoxographus Vaticanus, in: S. Heilen et al. (eds.), In Pursuit of Wissenschaft. Festschrift für W.M. Calder III zum 75. Geburtstag, Hildesheim et al. 2008, 437–466.

Stevens (2016): K. Stevens, From Herodotus to a 'Hellenistic' World? The Eastern Geographies of Aristotle and Theophrastus, in: E. Barker / S. Bouzarovski / C. Pelling / L. Isaksen (eds.), New Worlds from Old Texts: Revisiting Ancient Space and Place, Oxford 2016, 121–152.

Stramaglia (2011): A. Stramaglia, Phlegon Trallianus Opuscula de rebus mirabilibus et de longaevis, Berlin 2011.

Talbert (2012): R. J. A. Talbert (ed.), Ancient Perspectives: Maps and Their Place in Mesopotamia, Egypt, Greece and Rome, Chicago / London 2012.

Vannicelli et al. (2017): P. Vannicelli / A. Corcella / G. Nenci, Erodoto. Le Storie. Volume VII. Libro VII. Serse e Leonida, a cura di P. Vannicelli, Testo critico di A. Corcella, Traduzione di G. Nenci, Milano 2017.

Vitale et al. (2021): V. Vitale / P. de Soto / R. Simon / E. Barker / L. Isaksen / R. Kahn, Pelagios – Connecting Histories of Place. Part I: Methods and Tools, in: International Journal of Humanities and Arts Computing 15.1, 2021, 5–32, https://doi.org/10.3366/ijhac.2021.0260 (last access 15.04.2026).

Wenskus / Daston (2000): O. Wenskus / L. Daston, Paradoxographoi, in: DNP 9 (2000), 309–314.

Westermann (1839): A. Westermann, Παραδοξόγραφοι. Scriptores rerum mirabilium Graeci. Braunschweig / London 1839.

Wilson (2015): N. G. Wilson, Herodoti Historiae, Oxford 2015.

Yu (2023): K. W. Yu, Textualizing Wonders: Ancient Greek Paradoxography in Comparative Perspective, in: G. W. Most / M. Puett (eds.), After Wisdom: Sapiential Traditions and Ancient Scholarship in Comparative Perspective, Leiden / Boston 2023, 251–283.

Zaccaria (2024): P. Zaccaria, Pseudo-Aristotle, De mirabilibus auscultationibus 122–138 and Theopompus' Philippica, in: S. Schorn / R. Mayhew (eds.), Historiography and Mythography in the Aristotelian Mirabilia, London / New York 2024, 86–146.

Zaccaria (forthcoming): P. Zaccaria, Paradoxographus Florentinus, in: Trends in Classics – Greek and Roman Humanities Encyclopedia – Historiography, forthcoming.

Ziegler (1949): K. Ziegler, Paradoxographoi, in: RE XVIII/3, 1949, 1137–1166.

Zucker et al. (2024): A. Zucker / R. Mayhew / O. Hellmann (eds.), The Aristotelian Mirabilia and Early Peripatetic Natural Science, London / New York 2024.

Figure and Table References

Author Contact Information[105]

Dr. Pietro Zaccaria
KU Leuven, Ancient History
Blijde-Inkomststraat 21 – box 3309
3000 Leuven
E-mail: pietro.zaccaria@kuleuven.be

PD Dr. Monica Berti
Universität Leipzig
Lehrstuhl für Alte Geschichte
Beethovenstraße 15
04107 Leipzig
E-mail: monica.berti@uni-leipzig.de

DIGITAL CLASSICS ONLINE

More than Names? Challenges and Opportunities for Ancient Named Entity Recognition

Chiara Palladino

Abstract: This paper focuses on the conceptual challenges of Named Entity Recognition and Classification for Ancient Greek and Latin texts. It examines the shifting definitions of 'name' and 'named entity', their changes over time, the overlaps and differences between them, and shows how their use is often flawed by implicit assumptions on naming mechanisms in language and culture. It then offers examples of ancient place-naming practices that may challenge these assumptions, highlighting the limitations of current vocabularies and standards, and pointing to the need for a domain-specific approach to the problem of Named Entities in ancient languages.

Named Entities and Digital Classics

The ability to recognise, identify and analyse information from proper names is a fundamental aspect of humanistic research. In Classical Philology, names of persons and places convey essential information: personal names can lead to an enhanced understanding of social dynamics, character usage in narratives, or family networks. Place names can be extracted and plotted on maps to provide context and to visualise the material spaces a document describes, or they can help better conceptualise travel and movement. Citations of works and authors can provide insight into textual transmission and the circulation of ideas. Referencing obscure names to entries in dictionaries and encyclopaedias can provide an essential aid to reading, teaching, and learning about an ancient document. Furthermore, the indexing and association of textual information with external metadata, such as findspots, editors and collections can support more systematic research into material artefacts such as papyri, coins and inscriptions.

The automatic extraction and classification of names is conducted through Named Entity Recognition (NER), a basic task of Natural Language Processing (NLP). At the most basic level, NER consists of the parsing of a text to extract relevant strings, which represent names. The importance of NER has long been recognised in Computer Science, as Named Entities are more or less static strings in language and can be used as anchors to improve on various text processing tasks, such as morphosyntactic and part-of-speech tagging. The recognition and indexing of names, however, is also a fundamental part of Classical Philology. Organising, cataloguing, and indexing book titles, author names, and place information constitute one of the chief occupations of the philologist since antiquity.

In recent years, the revolution of Transformers and Large Language Models has had a significant impact on the computational study of the ancient world: the emphasis has been very much on linguistic analysis, handwriting recognition, and translation, but there is increasing attention towards extraction

and classification tasks.[1] Various NER models are currently available for Latin, Ancient Greek, and other premodern languages, with significantly improved performances on existing benchmarks.[2]

However, these methods face structural challenges that are typical of the computational processing of ancient documents. Some of the problems depend on the externalities of available technologies, which may include outdated edition formatting, noisy OCR output, unstable orthographies or spellings, or lack of unified Unicode standards.[3] More specifically, performance issues in NER models have been associated with the inability to extract outliers like names in foreign languages, or with inconsistencies in the detection of string boundaries in the case of multi-word names. Significant performance drops happen regularly when such models are tested on out-of-domain texts, which indicates overfitting in training and fine-tuning: this problem is related to the paucity of available high-quality annotated datasets and reliable guidelines.[4]

Finally, the lack of domain-specific tagsets is a challenge to research in this area: the labels used to classify entities according to semantic typologies are predominantly adapted from modern NLP tagsets, which are often used in domains that have very little in common with ancient documents. Considerable progress has been made in the domain of citation of ancient authors and works,[5] which can be annotated and extracted through carefully designed guidelines. Palladino et al. proposed an experimental tagset designed specifically for the annotation in Latin and Greek, with the intent of ensuring a basic level of interoperability across annotated corpora.[6] However, there is no agreed-upon strategy for handling named entities as structured information in ancient texts: in a sector where the lack of annotated datasets is an increasingly important problem, reconciliation strategies are often employed to ensure interoperability and data exchange across projects and to support fair evaluation of NER outputs.[7]

This situation reveals a fundamental lack of consensus on the definition of *ancient named entities*, and what criteria can be used to recognise them. In fact, what exactly constitutes a Named Entity, and to what extent it overlaps with the grammatical notion of name, is far from established. In this paper, we will more closely examine these competing notions and their ties to linguistic and cultural definitions. We will mainly consider toponyms in Ancient Greek and Roman sources, as a case study where cultural mechanisms of naming and transmission introduce a layer of complexity that helps question our preconceived notions of what constitutes a name, as well as problematising the assumptions underpinning extraction and classification tasks.

What is a Name?

The definition of what constitutes a name is obviously complex, and it has been studied in an impressive number of fields, from philosophy to logic, to linguistics, to anthropology. In fact, the defining criteria of what constitutes a name in language are very nuanced and full of potential contradictions.

We usually refer to names in language as something distinct and different from common nouns, and with a certain degree of specificity. In grammar, some criteria are commonly considered as defining proper names: capitalisation, lack of lexical meaning/semantic emptiness, morphosyntactic regularity

1 Sommerschield et al. (2023).

2 Beersmans et al. (2023).

3 Ehrmann et al. (2024).

4 Palladino / Yousef (forthcoming).

5 Romanello / Najem-Meyer (2022); Berti (2023).

6 Palladino et al. (2024).

7 Palladino / Yousef (2024).

(lack of inflection and determiners), untranslatability, and absence in traditional dictionaries. In general, these criteria point to the seemingly unique status of names as static strings of language that do not have a stable lexicographical definition like common nouns.

This definition is usually enriched by an idea that derives from the philosophy of language, the definition of name as a *rigid designator* that points to a *referent*. The concept owes considerable debt to Kripke:[8] differently from a common noun, a name is considered to point unambiguously to a uniquely identifiable object, the referent. For instance, while the noun 'city' indicates a class of objects, 'Athens' unambiguously denotes an individual instance of that class, that is, a specific thing in the world.

The concept of Named Entity is obviously much more recent. Officially, the idea of 'Entity Expression' emerged during the MUC-6 Evaluation Campaign in 1995, even though already in 1991 Lisa F. Rau designed a method to extract company names from texts, a task commonly understood as part of NER today.[9] The chief idea was to extract text strings that could be identified as entities of interest, and further disambiguate them using labels, or tags, that identified their type. In 1995, entities of interest were grouped as 'Entity Name Expressions', identified as persons, organisations and locations, 'Numeric Expressions' and 'Time Expressions'. Events, relationships, and coreference resolution were progressively added to the task, although today they are often considered more advanced applications. The benefits of extracting and classifying Named Entities were numerous, from a better understanding of context to improvements in other areas of language processing, like morphosyntactic parsing or translation.

However, what exactly constitutes a Named Entity and how to recognise it are questions that have shifted significantly over the years, and changes have occurred in almost every major evaluation campaign since 1995.[10] Named Entities never received an authoritative linguistic definition, contributing to the mistaken assumption that they may overlap with proper names: since the very inception, however, it is evident that the idea of Named Entity was broader and more flexible, but also vaguer.

The most stable defining criterion for a Named Entity seems to be the idea of rigid designator in the Kripkean sense.[11] This is also the assumption behind Entity Linking, which is based on the idea that a Named Entity extracted from a text can be associated with the corresponding referent, as it appears recorded in an ontology or authority list, through the operation of *semantic annotation*. However, the continuous redefinition of the domains of interest for NER, alongside a considerable expansion in label vocabularies, led to a certain differentiation in what the designator represents, increasingly problematising the Kripkean notion.

The relationship between *name*, *entity*, and *referent* is not a stable one. In its original formulation, anything could potentially be considered a referent: 'gold' or 'whale', for example, are specific instances of more general classes of objects, but they still designate something with specific properties and characteristics without being proper names in the grammatical sense.[12]

The question, therefore, becomes one of granularity. In their seminal work on NER, Nouvel et al. show that, while the idea of denoting something unique is somehow common to all definitions, the mechan-

8 Kripke (1996).

9 Marrero et al. (2013).

10 See Appendix 5 in Nouvel et al. (2016) for a non-exhaustive list.

11 Note, however, that Nadeau / Sekine (2007) referred to potentially *one or many* rigid designators.

12 Marrero et al. (2013). Additional problems relate to the notion of coreference, where pronouns, synonyms, and other discourse elements may also be associated with unique referents. For the sake of space, we will follow recent trends in NER that consider coreference as a separate task, and we will limit the present discussion to names.

isms through which this is achieved are heterogeneous.[13] This is one of the most important differences between the grammatical definition of name and the concept of Named Entity. Descriptive expressions are often considered NEs, provided that they point to a referent: for example, the string 'the daughter of Augustus' contains one proper name (Augustus), but has another identifiable and stable referent, Julia the Elder. Therefore, the *entity* it refers to is different from the *name* it contains (or there are two entities, but only one name). In other cases, a name in the grammatical sense may not be an entity at all: in the sentence 'Athens is a place-name', 'Athens' is not an identifiable city but *the name of the name* Athens, opposed to 'the name Alexandria' or 'the name Rome'. Thus, whether there is a referent or not depends very much on the definition of what constitutes a 'real object'.

Other categories are inconsistently grouped under the definition of name, Named Entity, or neither: patronymics, ancestry names, nicknames, political roles tend to have limited importance in traditional NLP, but are extremely important in literary studies and are sufficiently stable in their specific context that they may function as rigid designators. In situations of coreference, different names may point to the same referent, but be perceived as having different meanings or even point to culturally different objects: famously, this is the case of Frege's puzzle, where the 'morning star' and the 'evening star' have the same referent (the Planet Venus), but differ in their meaning to the point of being perceived as different versions of Venus (the planet as it appears in different locations in the sky a few months apart) and the phrase 'the morning star is the evening star' is not a tautology but an expression contributing new knowledge.

Fort et al. define Named Entities based on the concepts of *referential unicity* (contextual unicity of the referent), *referential autonomy* (should be sufficient to identify the referent), *denominational stability* (more regular and less numerous than other noun phrases), and *referential relativity* (the referent is considered relatively to a domain model).[14] The notion of referential autonomy, as discussed above, does not necessarily mean that a Named Entity is a name, but rather that the text string is sufficient, without additional context, to identify a specific object. On the other hand, referential unicity and relativity imply a consideration of *domain specificity*. The extraction and classification of Named Entities must be useful for a goal, rather than being a text processing exercise: therefore, the application of the task should be preceded by the delimitation of the elements of interest for the domain under investigation. The idea of Named Entity becomes, therefore, increasingly domain-dependent: as different knowledge domains may have different ideas of what constitutes a string of interest, it becomes a more flexible concept than the grammatical notion of proper name.

Following this trend, tags and vocabularies to classify Named Entities have grown exponentially to accommodate increasingly diverse domains, leading to tagsets of more than 200 labels and to strategies of annotation that divide Named Entities into many different components, aiming to achieve very fine semantic granularity.[15] Another approach to the problem has been suggested using nested entities, that is, the construction of named entities made up of multiple names in a hierarchical 'nested' structure. Strategies of this complexity become very close to ontologies, as they require multiple levels of internal ramifications.

On the other hand, the advent of Large Language Models and Generative AI has highlighted the lack of consistency in the development of datasets. Machine Learning, and particularly language models, require well-defined benchmarks to achieve reliable results in training and evaluation. Therefore, recent research has emphasised the need for harmonisation across labelling systems, in order to decrease

13 Nouvel et al. (2016).

14 Fort et al. (2009).

15 Sekine et al. (2002).

redundancy and ambiguity in datasets.[16] Thus, NLP is divided between these two demands, and this becomes more impactful in cases where the quantity and availability of data is limited.

There seems to be a tendency to trust the annotator's common sense to know what constitutes a Named Entity in each context: ultimately, the individual's judgement is considered the best criterion. This, however, is a problematic assumption, based on the idea that the concept of name and naming mechanisms are somehow intuitive and almost cultural universals. Not only this is demonstrably untrue, but it is especially dangerous for contexts that are situated far away in time and space from the annotator. The need for consistent datasets requires a careful definition of what constitutes a Named Entity, and the instability of this concept does not provide easy solutions.

In the section that follows, we will focus on ancient place-names, or toponyms, to illustrate how naming practices pose crucial challenges to the definitions described above. We focus on place-names specifically, not only to keep within the space allotted by this paper, but also because place-naming is a very good example of a practice with cultural specificity, as the context of creation of a name affects its manifestations in documents. There is growing scholarship on personal naming practices in the ancient world, but the study of toponymic practices and cultural transmission of place-names is often limited to etymology or individual case studies.[17] Place-naming as a cultural practice, while abundantly investigated in other fields, is still a novel area of investigation for Classical Antiquity.

Ancient Names, Modern Named Entities

Names and Place-Naming Practices

The operation of naming something obviously bears a degree of intentionality, even though the original meaning of a name may be lost in time. Place-naming is a social act common to all human cultures. It is impossible to detach place-naming from the human experience of place-making, that is, the process of understanding, conceptualising, and communicating spatial knowledge. Place-names stabilise features in the environment around which human activity is organised, but also function as linguistic tools to store and transmit geographical information.[18] Far from being a simple top-down imposition of a conventional model of reality, place-names are intimately connected with the material world they define, but also with local knowledge and memory.[19]

The standard grammatical definition of proper name is problematic for ancient, indigenous, and non-Western languages, and sometimes it directly contradicts the dynamics of place-naming. Criteria such as lack of inflection or determinants, as well as untranslatability or exclusion from dictionaries, are generally unsatisfactory. Likewise, dictionaries have an inconsistent approach to the inclusion of proper names, particularly place-name derivatives such as ethnonyms and demonyms.

It hardly needs mention that capitalisation is an unreliable mechanism to distinguish proper names in ancient texts. Capitalisation did not consolidate in Greek and Latin scripts as an orthographic convention to mark proper names until the Middle Ages, and even then, it was used inconsistently. This means that, in the manuscript tradition, capitalisation is not the prevalent convention to mark names, which are often not distinguished at all from the rest of the text. However, since we work mainly with modern critical editions, the editorial intermediation has already set out a text where certain distinc-

16 Palladino / Yousef (2024).

17 On personal names, see the recent contributions of De La Escosura Balbás et al. (2024); Bonnet (2024). On toponymy, notable etymological studies are Georgacas (1959); McDonald (1958).

18 Eades (2017).

19 Palladino (2023).

tions were made by typographic conventions or conscious choices that are almost never documented in the critical apparatus.

This is most visible in cases where geographic features originate from, or are included in, a toponym. One of the most basic place-naming mechanisms is using a notable feature to name an area, blurring the boundary between naming and descriptive function. While the assumption is that the name is going to progressively lose its lexical meaning and become associated with the referent, this is not always the case. Furthermore, what is valid for modern perceptions is not necessarily similar in ancient contexts. This leads to some notable inconsistencies in the editorial tradition of proper names.

One example of this is the Isthmus of Corinth. This toponym is often indicated in texts through the common noun 'ἰσθμός', which appears capitalised in modern editions to distinguish it from other straits located elsewhere (with notable inconsistencies, where even within the same edition the word is occasionally present in lowercase when referred to the Corinthian isthmus: see for example Paus. 2,1,3; Diod. 12,59,1).[20] This, however, is purely a modern convention based on modern associations with a historically important location: the correct identification of the isthmus does not rely on the word itself, but on the notion that the geographical context of the passage is known to the reader.

The case of the Pillars of Heracles ('στῆλαι Ἡράκλειαι') is the opposite: the common noun is never capitalised in modern editions even though it indicates the toponym (which is, however, regularly capitalised in translations). Two important exceptions are Philostr. Ap. 4,47 and one passage of the *Geography* of Strabo (15,1,6), where we find 'Στῆλαι'.[21]

The Acropolis of Athens is generally considered a place-name today: however, the term 'ἀκρόπολις' simply denotes a spatially identifiable part of a polis, equivalent to the common noun 'citadel'. In fact, the acropolis of Thebes has its own name, Cadmea. A similar case is 'τὰ μακρὰ τείχη', which appears always lowercase (e.g. Xen. Hell. 2,3,11; 4,4,18; Thuk. 1,107; 5,26; etc.),[22] but is consistently capitalised in translations because it indicates a precise object in the mind of a modern reader, the Long Walls of Athens.

These cases are problematic, because they only function as rigid designators within a shared spatial framework (such as the city of Athens), and the uniqueness of their referent is purely based on a geographical characteristic. They most definitely were *places*, but their usage does not fit the strict linguistic definition of *name*, as they are used with their original lexical meaning.

The idea of lack of lexical meaning implies that a place-name does not carry information about the object being designated: the 'White House' does not designate a house that is white, while a noun generally gives information about the type of object being described ('house'). In the case of toponyms, this goes hand in hand with the phenomenon of grammaticalisation, where a toponym with a specific etymology and clear origin progressively loses that association, becoming devoid of meaning.

The fact that toponyms grammaticalise, however, is not sufficient to consider them meaningless. This is especially problematic for indigenous toponymic systems, where place-naming practices harness the whole descriptive power of language and toponyms are designed to provide information about the places they refer to.[23] Fields like critical toponymy recognise place-names and place-naming as cultural practice to consider within both connotative and denotative functions, being at the same time util-

20 See for example: Spiro (1903); Jones / Wycherley (1955); Musti (1986); Fischer / Vogel (1888–1906); Oldfather (1933–1967).

21 Meineke (1877); Kayser (1870).

22 We cite here the authoritative Oxford editions, but various other available editions have been consulted: Marchant (1900); Jones / Powell (1942).

23 Keith Basso demonstrates this powerful mechanism in Western Apache place-names, which include examples like 'Green Rocks Side By Side Jut Down Into Water', or 'Gray Willows Curve Around A Bend'; Basso (1996), 23.

itarian and symbolic.[24] In the ancient world, the strong relationship between the geography of a place and the meanings associated to it by its inhabitants are highlighted by Strabo in his proem to the *Geography* (2,5,17): in addition to geographic features, the traditions and values associated to them by people are imposed, but end up becoming almost natural constituents of a place, although (differently from physical characteristics) they are subject to change through time.

Ancient toponyms may be lexically meaningless in modern languages, but they may not have been in their original usage. The expression 'συμπληγάδες (πέτραι)' (Apollod. 1,9,22) illustrates the origin of a place-name through this explanatory/descriptive mechanism. In the passage, Phineus describes the dangers of the route to the Argonauts, including the 'clashing rocks', an area continuously tormented by the wind and the sea. The expression is used to describe a feature of the landscape in a meaningful way. 'Symplegades', however, is also the actual toponym of the area, undoubtedly attested as such in ancient literature (and even capitalised in the English translation of this passage).[25] This ambiguity reflects the complicated mechanisms of place-naming, where the boundary between toponym and meaning is not always clear-cut.

Whether through actual etymology or through transmission practices, ancient toponyms bear the traces of this process of landscape conceptualisation: for example, the names of the minor seas of the Mediterranean were linked to the surrounding areas or islands ('Tyrrhenian', 'Phoenician'), environmental characteristics (e.g. 'Pontos Euxeinos', 'hospitable', vs. 'Pontos Axeinos', 'inhospitable'), local mythologies, or even navigational information, such as the dangerous character of an area ('Skyllaion') or the local winds to which it was exposed ('Zephyrios Limen').[26] Although many toponyms progressively lose association with their original etymologies, this is often counterbalanced by the consolidation of folk etymologies reflecting dynamics of degrammaticalisation: far from demonstrating lack of meaning, in the ancient world the attestation of different traditions on the origins of toponyms show the strategic importance attached to places. So, the many competing origin stories that associated the Aegean Sea with various local heroes and traditions reflect a desire by various communities (notably, but not exclusively Athens) to claim a strategic natural affinity with that space.[27]

The use and understanding of toponyms require a shared cultural background, but also some knowledge of local spatial configurations. Thus, what Ancient Greeks and Romans considered as place-names may have different answers. In some ways, this is a translation problem: there is no perfect cultural equivalent for a word or expression across different languages, and this extends to naming practices. What defines a name is culturally specific, rather than a universal: therefore, different cognitive and linguistic mechanisms may point to different definitions of what a name and its components are.

Place-Names as Named Entities

This descriptive and additive nature makes toponyms an interesting challenge for Named Entity Recognition. As we have seen above, Named Entities are defined under broader terms than names, and they rely on more flexible and domain-specific criteria to define strings of interest. The individuation of a string of interest relies on the ability to recognise it as a designator for a unique object in the world: this idea, however, relies on some level of shared knowledge.

One of the key assumptions behind Named Entities is that there is knowledge that the referent exists, but this is not always the case for the ancient world, where much contextual information is vague or lost to us, and we cannot always be sure whether we can associate a word in a document with a spe-

24 Choo (2023).

25 See for instance the famous Frazer translation (1921), and more recent translations in English by Hard (1997) and Smith / Trzaskoma (2007). See also the most recent authoritative Italian translation by Sarpi / Ciani (1996).

26 For these examples, see especially Morton (2001), 70ff.

27 Ceccarelli (2012).

cific individual or place. This does not mean that they did not originally exist or were not considered real.

Furthermore, place-names are, by definition, distributed and indefinite:[28] while it is quite easy to identify the physical boundaries of the individual named 'Augustus', it is extremely difficult to precisely define where 'Roma' begins or ends, and those boundaries change through time and context. Defining what makes anything a 'place' is a complicated matter. For example, defining a place on the basis of geographical location or cartographic coordinates is deeply problematic for the ancient world. Places may also change their physical location or configurations, sometimes significantly, in the course of a very long chronological span.[29] Thus, the concrete things denoted by a place-name could be identified in very different ways, depending on context: the referent of a place-name and its characteristics are not necessarily stable.

As we have seen above, Named Entities tend to include descriptive naming mechanisms ('the daughter of Augustus') under their definition. However, because place-naming is by definition descriptive and relational, it is much harder to understand when a string of text represents a descriptive place-name and when it is a simple noun phrase, especially when there is no contextual information or any evidence of consolidated usage patterns. Because different languages have very different place-naming practices, virtually any sort of geographical description could be considered a place-name, as long as it designates a specific and identifiable feature in a unique way and its usage can be demonstrated across a variety of contexts.[30] In other words, in many cases it is very difficult to understand whether a *spatial entity* is also a *named entity*.

This represents one of the biggest challenges for the automated extraction of place-names, the detection of entity boundaries: in many cases, it is very difficult to establish where a place-name begins and ends. The most common example is instances where a proper name in the conventional sense appears with descriptive attributes or common nouns, denoting a geographical feature, such as 'Νεῖλος ποταμός' or 'Κάσιος ὄρος', or have the function of further specifying a place in implicit contraposition with a different one, as in the case of 'κάτω Αἴγυπτος', '*mare Superum*', or '*Germania Inferior*'. These cases also challenge conventional morphosyntactic annotation, because different strategies are adopted to mark proper names and their constituents.[31]

More complex expressions like 'Ἡράκλειαι στῆλαι αἱ ἐν τῇ Εὐρώπῃ' (Ps. Skyl. 1) or 'Ἰβηρικὴ παραλία' (Strab. 3,4,16) certainly denote *spatial entities*, because they can be associated with a specific geographic area that is precisely recognizable and cannot be confused with any other. However, the problem is to understand whether they are perceived as *place-names*. In these cases, the attestation of regular usage in the tradition is the only discriminating criterion that can be adopted. Even so, some instances may put into question our own lack of evidence and our familiarity with the context. Ancient ethnographic practices, for example, often use relational expressions to denote specific subgroups: the Ethiopians are a particularly interesting example, where the regularity of usage to indicate distinct geographic areas based on relative location ('ὑπὲρ Συήνης', 'πρὸς/ὑπὲρ Αἰγύπτῳ', 'ὑπὲρ Μαύρους', etc.) suggests that these were fixed expressions, indicating different subgroups often marked in modern commentaries as 'Eastern' and 'Western' Ethiopians.[32]

---

28 Eades (2017).

29 Georgacas (1959).

30 Clearly, personal names are not totally devoid of the issue. Strings like 'the winner of the 60th Olympiad' or 'the wife of Augustus' nephew' have their own share of problems. In the case of place-names, however, the question has a cognitive implication: these expressions may not have been considered as noun phrases, but as actual *place-names*, in the same way as 'Rome' or 'Athens'.

31 Because of the strong relation between toponym and landscape terms, this is also a translation problem, as languages differ wildly in the ways in which they define landscape features such as 'mountain' or 'sea'.

In computational processing, boundaries need to be clearly defined and cannot be vague. The definition of boundaries may also affect subsequent tasks, such as the classification via semantic vocabularies. The inclusion or exclusion of descriptive words within the name may mean a change in label: cases such as 'city of the Gaditanes' ('Γαδιτανῶν πόλις'), 'pillars of Heracles' ('Ἡράκλειαι στῆλαι'), 'temple of Juno' ('*Junonis templum*'), 'altar of Apollo' ('βωμὸς Ἀπόλλωνος') are clear instances of this, where the *name* appearing in the text actually differs from the type of the *entity* being talked about.

The task of semantic labelling is peculiar to the treatment of Named Entities in the computational space, as it supports further disambiguation and processing, and it helps define extraction patterns. However, the classification of an entity is never completely devoid of the surrounding context.

Geographical features like constellations, winds, and compass points tend to be used differently in ancient and modern traditions. Wind names, for example, may be used in the ancient world to indicate actual winds or to express cardinal directions metonymically ('towards Zephyrus'), but winds are also regularly personified, calling into question the entire usage of capitalisation in modern editions.[33] Similarly, ancient Mediterranean cultures have a tendency to personify (or, more precisely, deify) rivers, as representations of important seasonal phenomena (the Nile and the Egyptian god Hapi) or mythical entities (the Styx in the Greek tradition). Spatial features or natural forces are often associated with superhuman beings of various kinds, but at least in part retain their material characteristics, to the point where it becomes extremely tricky to classify them using modern definitions: this difficulty is an indication of the challenges behind the idea of a uniquely identifiable referent, when in practice, the referent for the name 'Styx' is a dualistic object.

A linguistically interesting example is provided by phenomena of metonymy, that is, the practice of using the name of a place to refer to its inhabitants ('France won the World Cup') or vice-versa: the former is considered typical of modern languages, while the latter is much more common in Ancient Greek toponymic practices.[34] So much so that the 'people-for-place' mechanism is evident in the etymology of certain Greek toponyms, which derive from the ethnonym of their inhabitants. The most famous case is Delphi, where the toponym 'Δελφοί' is actually the name of the local population. However, 'Δελφοί' in its original ethnonymic sense did not fall out of use, and it is attested in texts to indicate either the Delphian people or the place of Delphi.[35] While this distinction can sometimes be clarified from context and some occurrences are used unambiguously for the people or the place, the decision is often left to the scholar's interpretation. This phenomenon, however, is a core practice of ancient place-naming, and it may well be that even an ancient author would not be able to tell the difference – and it may not have mattered.

Conclusions

A substantial part of the operation of recognising and isolating Named Entities relies on the existence of an implicit common interpretive framework, where there is fundamental agreement on what constitutes a distinct piece of information. In the case of ancient documents, however, the uncertainty of the surrounding context makes this assumption fragile. The nuances involved in the formation of ancient toponyms, ethnonyms, epithets, and other named features like winds and personified natural forces, challenge existing definitions at many levels: historical, interpretive, linguistic, and operational. Thus,

32 See Agathem. 2,7; Hdt. 7,70,1; Paus. 1,33,4; 1,33,5; 6,26,2; Strab. 1,2,28; 2,3,8; 17,1,53.

33 Shipley (2021).

34 Poibeau (2006).

35 Kron et al. (2019).

ancient place-names force us to contextualise our definitions as the product of a cultural and cognitive framework, and help us problematise ideas of 'name' and 'named entity'.

Clearly, one does not need to reconstruct a toponym's etymology or history to recognise it as a toponym. However, computational processing requires a remarkable degree of precision in the definition of what constitutes information of interest, and clearly defined criteria are paramount to the creation of guidelines and datasets. For this reason, the study of how an ancient community conceived naming practices and the perceptions attached to those names provides useful insights and supports a better definition of the matter under investigation.

On the other hand, when it comes to the concrete study of a text, it is important to emphasise that the simple extraction of names is hardly enough. To support serious scholarly analysis of ancient place information, it is necessary to further link toponyms to location data and gazetteer information through Entity Linking,[36] but also to combine this information with the analysis of morphosyntax, sentiments, and ideas associated with such occurrences.

Amidst all these challenges, the more flexible conceptual nature of Named Entities presents an opportunity, rather than a limitation. The idea of NE does not rely on rigid conceptual boundaries but focuses on domain-specific definitions within the context of application. Therefore, it provides new ways to look at entities as structured information in texts. We have more freedom to consider named entities through the framework of culturally situated practices, such as naming and spatial conceptualisation. Thus, rather than Named Entity Recognition, it would be more productive to think in terms of Information Extraction within specific knowledge domains.[37]

Leveraging on domain knowledge, therefore, is fundamental for the definition of methods of computational analysis. The challenge is to design vocabularies and standards that take the cultural and linguistic specificity of ancient naming practices into account.[38] This scholarly endeavour is, at its core, philological, but goes hand in hand with the opportunities of automated processing and digital treatment of ancient documents, in order to allow for in-depth investigation to a bigger and more ambitious scale.

36 Beersmans et al. (2024).

37 For example, Romanello and Najiem Meyer (2022) leverage on the notion of 'Knowledge Entity' readapted from Zhang et al. (2021) for the extraction of citations of Classical works.

38 Broux (2015).

Sources

Online Sources

Open Greek and Latin, Perseus Digital Library, Scaife Viewer, https://scaife.perseus.org/ (last access 04.09.2025).

ΛΟΓΕΙΟΝ, https://logeion.uchicago.edu/ (last access 04.09.2025).

TLG, Thesaurus Linguae Graecae, https://stephanus.tlg.uci.edu/ (last access 04.09.2025).

Editions

Fischer / Vogel (1888-1906): K. T. Fischer (post I. Bekker & L. Dindorf) and F. Vogel, Diodori bibliotheca historica, 5 vols., 3rd edn., Leipzig 1888–1906 (repr. 1964), retrieved from: http://stephanus.tlg.uci.edu/Iris/Cite?0060:001:1447586 (last access 28.07.2025).

Frazer (1921): James G. Frazer (ed.), Apollodorus, The Library, Volume I: Books 1–3.9, Cambridge (Mass.) 1921.

Hard (1997): Robin Hard (ed.), Apollodorus, The Library of Greek Mythology, Oxford / New York 1997.

Jones / Powell (1942): H. S. Jones / J. E. Powell, Thucydidis Historiae, 2 vols., Oxford 1942 (repr. 1963).

Jones / Wycherley (1955): W. H. S. Jones / R. E. Wycherley, Pausanias' Description of Greece, 5 vols, Cambridge (Mass.) / London 1955.

Kayser (1870): C. L. Kayser, Flavii Philostrati opera, vol. 1, Leipzig 1870 (repr. Hildesheim 1964), retrieved from: http://stephanus.tlg.uci.edu/Iris/Cite?0638:001:312314 (last access 28.07.2025).

Marchant (1900): E. C. Marchant, Xenophontis Opera Omnia, vol. 1, Oxford 1900 (repr. 1968).

Meineke (1877): A. Meineke, Strabonis geographica, 3 vols., Leipzig 1877, retrieved from: http://stephanus.tlg.uci.edu/Iris/Cite?0099:001:1841753 (last access 28.07.2025).

Musti (1986): Pausania, Guida della Grecia. Libro II. La Corinzia e l'Argolide. Testo e traduzione a cura di D. Musti, Milano 1986 (repr. 2008).

Oldfather (1933–1967): C. H. Oldfather, Diodorus Siculus. Library of History, Cambridge (Mass.) / London 1933–1967.

Sarpi / Ciani (1996): Apollodoro. I miti Greci (Biblioteca). A cura di P. Sarpi. Traduzione di M. G. Ciani, Milano 1996 (repr. 2013).

Smith / Trzaskoma (2007): Apollodorus' Library and Hyginus' Fabulae. Two Handbooks of Greek Mythology. Translated, with Introductions, by R. Scott Smith and Stephen M. Trzaskoma, Indianapolis / Cambridge 2007.

Spiro (1903): F. Spiro, Pausaniae Graeciae descriptio, 3 vols., Leipzig: Teubner, 1903 (repr. 1:1967), retrieved from: http://stephanus.tlg.uci.edu/Iris/Cite?0525:001:209605 (last access 28.07.2025).

References

Basso (1996): K. H. Basso, Wisdom Sits in Places: Landscape and Language Among the Western Apache, Albuquerque (NM) 1996.

Beersmans et al. (2023): M. Beersmans / E. de Graaf / T. Van de Cruys / M. Fantoli. Training and Evaluation of Named Entity Recognition Models for Classical Latin, in: Proceedings of the Ancient Language Processing Workshop (ALP 2023), Shoumen (Bulgaria) 2023.

Beersmans et al. (2024): M. Beersmans / A. Keersmaekers / E. De Graaf / T. Van De Cruys / M. Depauw / M. Fantoli, "Gotta catch 'em all!": Retrieving people in Ancient Greek texts combining transformer models and domain knowledge, in: Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), Bangkok / online 2024, 152–164.

Berti (2023): M. Berti, Named Entity Recognition for a Text-Based Catalog of Ancient Greek Authors and Works, online 2023, https://zenodo.org/records/8108058 (last access 04.09.2025).

Bonnet (2024): C. Bonnet / R. Häussler (transl.), The Names of the Gods in Ancient Mediterranean Religions, Cambridge 2024.

Broux (2015): Y. Broux, Graeco-Egyptian Naming Practices: A Network Perspective, Greek, Roman, and Byzantine Studies 55 (2015), 706–720.

Ceccarelli (2012): P. Ceccarelli, Naming the Aegean Sea, Mediterranean Historical Review 27 (2012), 25–49.

Choo (2023): S. Choo, Assessing the Validity of Critical Toponymy Perspectives for Understanding Human Perception of Places: An Analytical Framework, in: G. O'Reilly (ed.), Place Naming, Identities and Geography, Cham 2023, 29–50.

De La Escosura Balbás et al. (2024): C. de la Escosura Balbás / A. Kurilic / G. E. Rallo (eds.), Name and Identity. Selected Studies on Ancient Anthroponymy through the Mediterranean, Oxford 2024.

Eades (2017): G. L. Eades, The Geography of Names: Indigenous to Post-Foundational, London / New York 2017.

Ehrmann et al. (2024): M. Ehrmann / A. Hamdi / E. L. Pontes / M. Romanello / A. Doucet, Named Entity Recognition and Classification in Historical Documents: A Survey, ACM Computing Surveys 56 (2024), 1–47.

Fort et al. (2009): K. Fort / M. Ehrmann / A. Nazarenko, Towards a Methodology for Named Entities Annotation, in: Proceedings of the Third Linguistic Annotation Workshop on – ACL-IJCNLP '09, Suntec (Singapore) 2009, 142–145.

Georgacas (1959): D. J. Georgacas, A Contribution to Study of Greek Toponymy, Names. A Journal of Onomastics 7/2 (1959), 65–83.

Kripke (1996): S. A. Kripke, Naming and necessity, Oxford / Cambridge 1996.

Kron et al. (2019): C. Kron / W. L. Little / J. C. Wolfe / P. Ajaka / B. O. Allen / C. Brown / M.-C. H. de Marneffe / M. Elsner / M. D. Grioni / B. D. Joseph / A. B. Kessler / H. F. Young, What's In a Name? Issues in Named Entity Recognition, Paper to be presented at Annual Meeting of American Name Society, New York 2019, https://bpb-us-w2.wpmucdn.com/u.osu.edu/dist/4/27964/files/2016/01/ANSabstract-FINAL-2kf9qiy.pdf (last access 28.07.2025).

Marrero et al. (2013): M. Marrero / J. Urbano / S. Sánchez-Cuadrado / J. Morato / J. M. Gómez-Berbís, Named Entity Recognition: Fallacies, Challenges and Opportunities, Computer Standards & Interfaces 35 (2013), 482–489.

McDonald (1958): W. A. McDonald, Early Greek Attitudes toward Environment As Indicated in the Place-Names, Names. A Journal of Onomastics 6 (1958), 208–216.

Morton (2001): J. Morton, The Role of the Physical Environment in Ancient Greek Seafaring, Leiden / Boston 2001.

Nadeau / Sekine (2007): D. Nadeau / S. Sekine, A Survey of Named Entity Recognition and Classification, Lingvisticae Investigationes 30 (2007), 3–26.

Nouvel et al. (2016): D. Nouvel / M. Ehrmann / S. Rosset, Named Entities for Computational Linguistics, London / Hoboken 2016.

Palladino (2023): C. Palladino, Not the Same Landscape. Rediscussing Digital Approaches to Spatial Knowledge Systems, in: C. Palladino / G. Bodard (eds.), Can't Touch This: Digital Approaches to Materiality in Cultural Heritage, London 2023.

Palladino et al. (2024): C. Palladino / M. Fantoli / E. de Graaf / M. Berti / M. Romanello / T. Yousef / M. Beersmans / T. Gheldof / L. Soffiantini / E. Litta Modignani Picozzi, Experience and Challenges with Named Entities – Workshop at DHBenelux 2024, Leuven / online 2024, https://zenodo.org/records/11366870 (last access 04.09.2025).

Palladino / Yousef (2024): C. Palladino / T. Yousef, Development of Robust NER Models and Named Entity Tagsets for Ancient Greek, in: R. Sprugnoli / M. Passarotti (eds.), Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024, Torino 2024, 89–97.

Palladino / Yousef (forthcoming): C. Palladino / T. Yousef, Named Entity Recognition in Classical Languages: Two Approaches, in: E. De Graaf / A. Keersmaekers / S. Stopponi / S. Peels-Matthey (eds.), Computational Approaches to Ancient Greek and Latin, forthcoming.

Poibeau (2006): T. Poibeau, Dealing with Metonymic Readings of Named Entities, in: Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci 2006), Vancouver 2006, 1962–1968.

Romanello / Najem-Meyer (2022): M. Romanello / S. Najem-Meyer, Guidelines for the Annotation of Named Entities in the Domain of Classics, online 2022, https://doi.org/10.5281/zenodo.6368101 (last access 04.09.2025).

Sekine et al. (2002): S. Sekine / K. Sudo / C. Nobata, Extended Named Entity Hierarchy, in: M. González Rodríguez / C. P. Suarez Araujo (ed.), Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02), Las Palmas 2002.

Shipley (2021): D. G. J. Shipley, Sun, Sea, and Sky: On Translating Directions (and Other Terms) in the Greek Geographers, in: E. Boutsikas / S. C. McCluskey / J. Steele (ed.), Advancing Cultural Astronomy: Studies In Honour of Clive Ruggles, Cham 2021, 105–136.

Sommerschield et al. (2023): T. Sommerschield / Y. Assael / J. Pavlopoulos / V. Stefanak / A. Senior / C. Dyer / J. Bodel / J. Prag / I. Androutsopoulos / N. de Freitas, Machine Learning for Ancient Languages: A Survey, Computational Linguistics 49 (2023), 703–747.

Zhang et al. (2021): C. Zhang / P. Mayr / W. Lu / Y. Zhang, Extraction and Evaluation of Knowledge Entities from Scientific Documents, Journal of Data and Information Science 6 (2021), 1–5.

Author Contact Information[39]

Dr. Chiara Palladino
Assistant Professor
Department of Classics and Ancient History
Durham University
38 N Bailey
Durham, DH1 3EU
E-mail: chiara.palladino@durham.ac.uk

39

DIGITAL CLASSICS ONLINE

The *NIKAW* Project: An Infrastructure of Texts, Entities and Language Models to Study the Circulation of Knowledge in the Ancient World

Margherita Fantoli, Marijke Beersmans, Jens Bürger, Evelien de Graaf, Mark Depauw, Alek Keersmaekers, Bart Thijs, Tim Van de Cruys, Toon Van Hal

Abstract: This paper presents the foundational work of the interdisciplinary project *NIKAW* (*Networks of Ideas and Knowledge in the Ancient World*), which aims to analyse social networks in ancient Greek and Latin texts through mentions of historical figures. As a critical first step, we address the challenge of Named Entity Recognition (NER) for these languages by leveraging transformer-based models enriched with domain-specific knowledge. Our experiments highlight data sparsity and annotation inconsistencies as key bottlenecks for model performance. In the second phase, we introduce a pipeline for Named Entity Linking (NEL), utilizing the *Wikisource* edition of the *Pauly-Wissowa Encyclopedy* as a knowledge base. We detail the creation of silver-standard (automatically annotated) and gold-standard (human-verified) training datasets, and report preliminary results from fine-tuning the BLINK model for NEL.

Section 1: Introduction

"μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος"[1]: the opening line of the Iliad immediately immerses the reader in the dense universe of (here mythological) people which characterize Ancient Greek and Latin literature. For readers of classical works, every text introduces a rich array of characters: not only heroes like Achilles but also gods, generals, warriors, rulers, citizens, slaves, philosophers, and intellectuals. Particularly texts belonging to the literary tradition, a diverse set of non-documentary texts, ranging from epic to historiography, philosophy, oratory, moral treatises etc., often refer to prominent figures of the ancient world. The people mentioned include both fictional and historical individuals who appear, in some cases repeatedly, in the textual and material evidence that has survived from antiquity.

In the project *NIKAW* (*Networks of Ideas and Knowledge in the Ancient World*), we aim to represent the vast array of people mentioned in ancient literature as a network, capturing the interconnectedness of individuals in the literary landscape. By analysing how this network evolves or shifts in response to different parameters – such as the authors' origin, date, or religious views included in the corpus – we seek to evaluate whether the network reflects well-documented cultural transformations studied by classical scholarship.

Between the current state of text mining capabilities for ancient languages and the realization of this ambitious goal lies a long path fraught with highly challenging obstacles. During the project design phase, we developed a pipeline structured around three key steps: Named Entity Recognition (NER),

1 Hom. Il. 1,1.

Named Entity Disambiguation or Linking (NED/NEL), and Social Network Analysis (SNA). In the NER phase, we aim to train a model capable of identifying named entities in Ancient Greek and Latin texts, with a particular focus on accurately labelling mentions of people. In the NEL phase, our goal is to disambiguate these mentions and link them to an existing knowledge base. Finally, in the SNA phase, we intend to construct a network of citations using the disambiguated mentions. The overall visualization of the pipeline is provided in fig. 1.

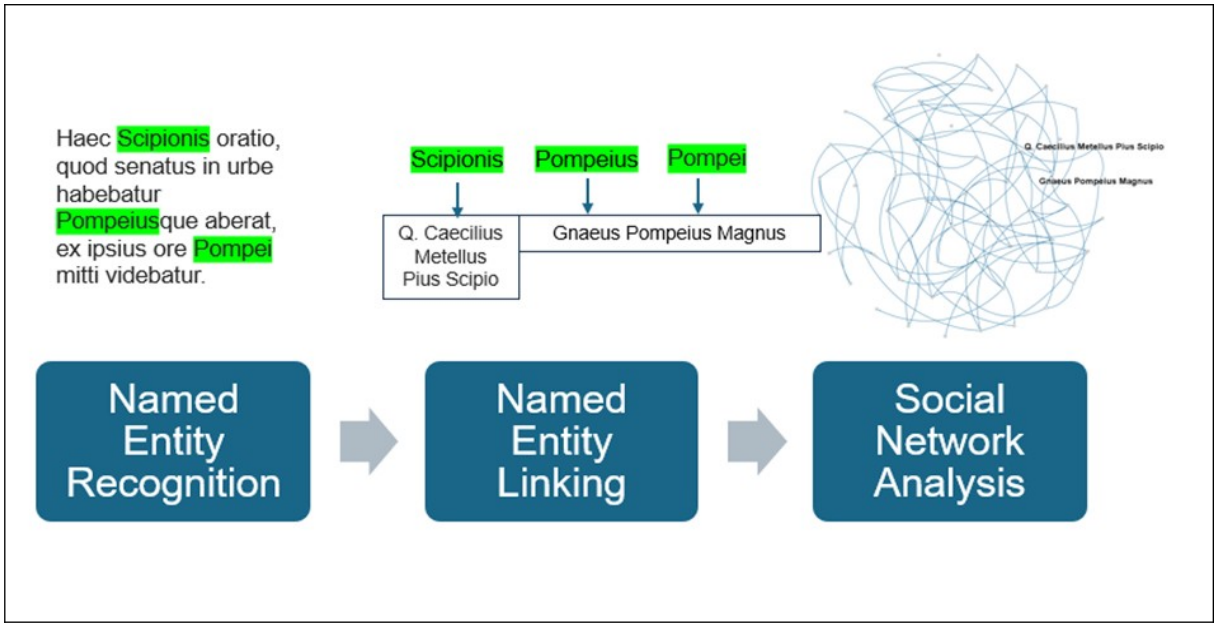


Fig. 1: Main steps of the *NIKAW* workflow.

With this paper, we concentrate on two key aspects of the work with named entities, namely:

- To what extent can we, in the current state of art, rely on automation for the task of processing named entities in classical studies, and what are the necessary steps to move forward along this path? How can we balance the ambition of automated processing and the need for high-quality data?

- The annotation of named entities in texts is a task that crosses boundaries between linguistic and historical annotation of texts. This requires combining tools and approaches typical of the Natural Language Processing (NLP) domain with domain-specific, content-oriented resources developed within the community of Ancient Studies. What are the possibilities to enable this combination, and how can we streamline the process?

Concretely, this paper discusses the first two steps of the workflow, as developed in the first two years of the project: after presenting the corpora on which we rely (Section 2: Data and Digital Infrastructure), we discuss Named Entity Recognition and Named Entity Linking (Section 3 and 4 respectively), while Section 5 (Ongoing and Future Projects) outlines the projects which have originated from the current research but are still in an early stage.

Section 2: Data and Digital Infrastructure

The primary research question driving this project is to understand whether the advent of Christianity caused a significant shift in the references to individuals in texts. To address this question, the first step is to compile a representative corpus of Ancient Greek and Latin texts that reflect this transformation. To ensure the reusability of the data and the replicability of our findings, we have opted for open-access corpora.

For the Greek texts, we rely on the *GLAUx* corpus, curated by project member Alek Keersmaekers.[2] This corpus, one of the largest open-access resources available, contains 20 million tokens and has been automatically lemmatized and morpho-syntactically annotated.[3]

The corpus spans twelve centuries, with the latest texts belonging to the 4th century AD. The *GLAUx* interface (developed in close cooperation with *Trismegistos*+) is currently browsable at https://glaux.be/search.php (last access 29.08.2025).[4] The corpus has also been connected with the *Trismegistos* database (*TM*).[5] Apart from information on all texts recorded for antiquity before 800 CE (*TM Texts*), *TM* also contains authority files for ancient authors known through direct and indirect attestations (*TM Authors*), as well as personal and place names and the variant forms in which they are attested (*TM People* and *TM Places*). The text-ids of the *GLAUx* corpus have been matched with the *TM AuthorWork* ids. Moreover, for people, *TM NamVar*, a list of name variants, has been linked to the lemmas of the *GLAUx* corpus, so that it is possible to directly retrieve all the passages where a certain name (or name variant) is attested.

The *TM*+ team has recently developed an algorithm capable of extracting and expanding references to ancient works in modern scholarship. These references are then connected to entries in *TM AuthorWork* and *TM Authors*, allowing us to locate corresponding works in the *GLAUx* corpus. While the process of identifying specific chapters or paragraphs within a work is still being refined, this pipeline has proven invaluable for the Named Entity Linking (NEL) phase, as discussed in Section 4.

For the Latin corpus, the landscape is more fragmented. Our goal was to identify a lemmatized, open-access corpus covering both classical antiquity and the early centuries of late antiquity. At this stage, we are working with two corpora. The first is the *LASLA* corpus, a manually lemmatized and morpho-syntactically tagged collection of Latin classical literature, openly available on Dataverse and linked to the *LiLa Knowledge Base*.[6] The onomastic and topographical data of the *LASLA* corpus has been integrated into the *TM* database, with places and names annotated using *TM* identifiers. While *LASLA* is an excellent resource, its diachronic range is more limited than required for the *NIKAW* project.

To address this limitation, we also use the *Corpus latin antiquité et antiquité tardive lemmatisé*[7], hereafter referred to as the *Corpus Latin*. This corpus, automatically annotated using the Pie-Extended *LASLA* model,[8] is the most extensive open-access Latin corpus with linguistic annotation. Currently, we combine both corpora: where available, we use the manually annotated *LASLA* corpus, and for texts not included in *LASLA*, we rely on the *Corpus Latin*, importing relevant texts into the *TM*+ custom made relational database. However, this approach presents challenges, such as the need for auto-

2 Keersmaekers (2021).

3 At the time of writing, also the *Opera Graeca Adnotata* has been released (Celano [2024]), but it was not available when the project started.

4 Keersmaekers et al. (2024).

5 https://www.trismegistos.org/ (last access 25.03.2026).

6 Fantoli et al. (2022).

7 Clérice (2020).

8 https://github.com/chartes/deucalion-model-lasla (last access 29.08.2025).

matic sentence-splitting and the persistence of errors due to the process of text recognition from printed editions. While we manually corrected these errors for small-scale experiments, large-scale preprocessing may require additional, partially automated efforts.

The *TM+* custom-made relational database is currently developed in FileMaker, a user-friendly tool that integrates seamlessly with the various modules of the *TM* infrastructure and the corpora we use. We employ FileMaker for manual text annotation (see Section 4), while Python scripts handle NLP tasks using tables exported from FileMaker. All annotated datasets are shared in open formats (e.g., CSV, TSV) to promote transparency and reuse.

Section 3: Named Entity Recognition

After identifying the corpus, the following natural step to undertake is to identify mentions of people in the corpus, which is a subtask of the Named Entity Recognition effort. While for contemporary sources NER models achieve highly satisfactory results, the accuracy when applied to historical texts is still lagging behind, due to lack of annotated data, noisy input and language change.[9] Noise is generally low in Latin and Ancient Greek editions due to the high quality of digitization, but tokenization errors in the Corpus Latin can still impact named entity recognition. For instance, words that were hyphenated in the edition are split into two tokens, which prevents the correct detection as proper nouns. In the case of the *NIKAW* project, the lack of annotated data for training models has proved to be the most significant issue. Moreover, as explained below, while for Latin most of the available training data for the NER task come from a single project, which results in a general consistency of the annotation choices, for Ancient Greek matters are complicated by a lack of consistency across the different datasets in terms of categories used, handling of ambiguous cases etc.

In our initial NER experiment,[10] we focused on Latin and compared the performance of three models (two transformer-based *LatinBERT* models and a shallow Conditional Random Field [CRF] model) on the only existing dataset for Classical Latin, annotated within the *Herodotos* project.[11] We excluded the Latin portions of a multilingual medieval charter dataset[12] due to linguistic and entity type differences from classical Latin. We benchmarked our three models against two existing models for Classical Latin: a neural BiLSTM-CRF entity recognizer, trained on classical Latin as part of the *Herodotos* project,[13] and LatinCy, a SpaCy pipeline for Latin backed by the multilingual BERT architecture[14] and fine-tuned for NER on a custom dataset combining *Herodotos* project data and Latin UD treebanks.[15] The goal of the paper was to evaluate whether *LatinBERT*,[16] which had not yet been fine-tuned for NER, could outperform existing models. This was motivated by the growing use of transformer models for various NLP tasks in classical languages,[17] including NER.[18] The results showed that this approach allowed us to achieve significant improvement over existing models, both on in-domain and

9 Ehrrmann et al. (2023).

10 See Beersmans et al. (2023) for more details.

11 Erdmann et al. (2016) and (2019).

12 Torres Aguilar (2022).

13 Erdmann et al. (2016).

14 Devlin et al. (2018).

15 Burns (2023).

16 Bamman / Burns (2020).

17 Sommerschield et al. (2023).

18 Yousef et al. (2023).

out-of-domain data. We tested the models on newly annotated texts of the *LASLA* corpus (Tacitus, *Historiae*, book 1; Cicero, *Orationes Philippicae*, I; the first three of Juvenal's *Saturae*),[19] in order to see whether the results were robust in the context of slight changes in the annotation style. While confirming the fact that *BERT* models outperformed the other models, this experiment showed a drop in the quality of the prediction: for instance, for people, the category in which we are most interested, the performance dropped from an F1 score of 0.92 to 0.85 (when looking at the annotation of the full entity, and not of the single tokens that compose it), and yet this was a less dramatic drop than for the other detected categories (places and groups). Such a difference highlights the strong influence of annotation consistency on the performance of the models. In particular, we identified the following aspects as causing most of the prediction errors:

- Boundary detections: multitoken entities are a regular source of errors. For instance, the sequence Cetrius Seuerus Subrius Dexter Pompeius Longinus (Tac. Hist. 1,31) contains 3 person entities: Cetrius Severus, Subrius Dexter, and Pompeius Longinus. These were predicted as Cetrius Seuerus Subrius and Dexter Pompeius Longinus by one *BERT* model and as Cetrius Seuerus Subrius Dexter and Pompeius Longinus by the other *BERT* model.

- Foreign names and names following a Greek declension were rarely tagged (e.g. Penelope, Aristotelen).

- Ambiguous entities: ambiguous tokens that occur both as entity and non-entity are frequently considered non-entities (e.g. *Oriens*, *Pax*, *Fides*…).

Hence, for our follow-up experiment,[20] which focused on Ancient Greek, we modified our approach. First, we concentrated on the category of people, being the primary focus within the *NIKAW* project. Second, we combined linguistic information with existing gazetteers to address the limitations of the models. Unlike Latin, Ancient Greek lacks a single, dedicated dataset for NER, such as the *Herodotos* dataset. However, we were able to make use of four distinct datasets from various projects that included named entities: the *Odyssey*,[21] the EpiDoc XML of the *Deipnosophistae* of Athenaeus of Naucratis, retrieved from the Perseus digital library, the *STEP* Bible corpus available on GitHub,[22] which contains the full Ancient Greek New Testament and Pausanias' *Periegesis Hellados* from the *Periegesis* project.[23] A significant portion of our work involved harmonizing the annotations across these datasets to enable their joint use for model training. Additionally, we annotated a randomly selected sample of sentences from the GLAUx corpus to create an 'out-of-domain' test set, minimizing genre, time, and author biases. We compared the performance of four transformer models (*Ancient Greek BERT*, *ELECTRA*, *GrεBerta*, and *UGARIT*)[24] on the NER task. While *Ancient Greek BERT* and *UGARIT* performed similarly overall, *Ancient Greek BERT* showed a slight advantage in identifying people versus a miscellaneous category. We therefore selected *Ancient Greek BERT* for the subsequent experiments. To enhance the model's performance, we integrated domain knowledge, a strategy

19 The annotated texts are available at https://github.com/NER-AncientLanguages/Ner-Latin-RANLP (last access 29.08.2025).

20 For the details, see Beersmans et al. (2024).

21 Pelagios (2021).

22 STE (2023).

23 Foka et al. (2021).

24 See Singh et al. (2021) for *Ancient Greek Bert*, Mercelis/Keersmaekers (2022) for *ELECTRA*, Riemenschneider / Frank (2023) for *GrεBerta* and Palladino / Yousef (2024) for *UGARIT*.

proven effective for low-resource languages,[25] and previously applied to classical languages.[26] Specifically, we utilized the *TM Gazetteers*: *NamVar*, which includes personal names and their variants, and *TM GeoVar*, which contains spelling and linguistic variants of placenames from ancient texts. By incorporating information on whether a capitalized word appeared in *NamVar* but not in *GeoVar*, we improved the model's performance, achieving an F1 score of 0.9 on the out-of-domain test set.

To address the challenge of identifying multi-token entities, we leveraged syntactic information from the *GLAUx* corpus. This involved expanding entities to include capitalized words syntactically dependent on tokens annotated as PERS. For example, in the expression "περὶ Ἡρώδου τοῦ Ἀθηναίου" (Philostr. soph. 2,1,15: "Concerning Herodes the Athenian"), the multitoken entity "Ἡρώδου τοῦ Ἀθηναίου" can be recognized in this manner, because "τοῦ Ἀθηναίου" is syntactically dependent on "Ἡρώδου". This approach significantly improved the recall of multi-token entities. These experiments demonstrated that combining transformer models with domain and linguistic knowledge is highly effective for mining Ancient Greek texts. Despite these positive results, error analysis revealed that annotation choices, particularly for ambiguous categories such as book or honorific titles (e.g., Φαραώ), continue to impact model performance.

Our work on both Ancient Greek and Latin NER highlighted the critical limitation of insufficient annotated data and inconsistencies across existing datasets. To address this, we co-initiated a collaborative effort within the scholarly community to develop shared guidelines for named entity annotation.[27] Several *NIKAW* members are actively contributing to this initiative, underscoring the importance of collaborative infrastructure for achieving robust results in large-scale experiments.

Section 4: Named Entity Linking

Named Entity Linking (NEL) is the task of disambiguating named entities mentioned in a text by associating them with entries in a knowledge base.[28] It involves two key steps: candidate generation, which identifies all possible entities that could match the mention, and candidate ranking, which evaluates and orders these candidates based on their likelihood of being the correct match. Additionally, the prediction of unlinkable entities can be incorporated into this process.[29] This task mirrors the mental reasoning of a reader who, upon encountering a name like 'Alexander', must determine which specific individual (among those they know) is most likely being referenced. To achieve this, readers – and NEL systems – often rely on external resources, such as *Wikidata* or contextual commentaries, leveraging both external and contextual knowledge to make accurate decisions.

In the domain of classical studies, NEL experiments remain relatively rare. A few digital datasets have been manually created, where entities mentioned in texts are disambiguated using identifiers: the *Patristic Text Archive*,[30] the *STEP Bible Project*,[31] the *Odyssey* annotated by Josh Kemp,[32] *Trismegistos*

25 Fetahu et al. (2022).

26 See for instance the work of Broux / Depauw (2015) and Berti et al. (2019).

27 Palladino et al. (2024).

28 For a general overview on Named Entity Linking, cf. Ji et al. (2022).

29 For a more detailed overview of the subtasks involved, cf. Sevgili et al. (2022) and Shen et al. (2015).

30 https://pta.bbaw.de/en/ (last access 29.08.2025).

31 https://www.stepbible.org/ (last access 29.08.2025).

32 Kemp (2021).

People,[33] and the *Greek Fragmentary Tragedians Online*.[34] Monica Berti's work has focused on developing semi-automatic pipelines for entity annotation within several projects, such as the *Linked Ancient Greek and Latin* project[35], the *Digital Athenaeus* project[36], and the *Digital Harpocration*[37]. The only attempt to fully automate the process occurred during the HIPE 2022 shared task,[38] where mentions of entities in classical commentaries were linked to *Wikidata* as part of the *Ajax Multi-Commentary* project.[39] Overall, the results of the NEL task were relatively low (e.g. recall reached at most 0.39), highlighting the challenges posed by current resources. These datasets and experiments make use of a variety of knowledge bases, including *Wikipedia*, *Wikidata*, project-specific identifier sets, and domain-specific resources such as the *Lexicon of Greek Personal Names*[40]. In the *NIKAW* project, we first addressed the problem of selecting and processing a knowledge base (Section 4.1). Next, we created training data using different strategies (Section 4.2), and we are now exploring the performance of various NEL models (Section 4.3).

Section 4.1: Creating a Knowledge Base

Identifying a knowledge base which could support the disambiguation of all people mentioned in Ancient Greek and Latin texts was no easy task. Historically, *Wikipedia*-derived knowledge bases (such as *DB Pedia* of *Wikidata*) have been used for NEL, to the point that a specific term exists for describing the process of mapping entities to *Wikipedia* ('Wikification').[41] However, despite the advantages of using such large resources, several shortcomings, which might particularly affect the results and evaluation of NEL, have been highlighted, such as the presence of duplicated or conflated entities.[42] Moreover, it is difficult to assess the extent of coverage of classical antiquity. Several initiatives aim to enrich the information on Greek and Roman antiquity on *Wikipedia*, for instance, the *WikiProject Classical Greece and Rome*,[43] or the 2023 datathon aiming at introducing *WikiData* entries related to publications in Classical Philology,[44] but they represent ongoing work, and are not specifically focusing on prosopographical information on people of the Ancient World. To provide an example of the potential lack of coverage, when we look at the entries in *Paulys Realencyclopädie der classischen Altertumswissenschaft*, which will be discussed intensively later, for the name Abaskantos, 8 different people are listed, of which only two are present in *WikiData*. Of these two, one has an entry only in seven languages,[45] while the other appears only in the Portuguese version,[46] which also flags the question of what version of *Wikipedia* should be used for retrieving the textual descriptions of the entities used for the NEL task.

---

33 Broux/Depauw (2015).

34 Antonopulos (2023).

35 https://www.lagl.org/ (last access 29.08.2025).

36 https://www.digitalathenaeus.org/ (last access 29.08.2025).

37 https://www.lagl.org/tools/harpocration/index.php?what=urn:cts:greekLit:tlg0533 (last access 29.08.2025).

38 Ehrmann et al. (2022).

39 https://mromanello.github.io/ajax-multi-commentary/ (last access 29.08.2025).

40 https://www.lgpn.ox.ac.uk/ (last access 29.08.2025).

41 See Mihalcea / Csomai (2007), or Shnayderman et al. (2019)

42 Pellizzari di San Girolamo (2023).

43 https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Classical_Greece_and_Rome (last access 29.08.2025).

44 For more information see https://diptext-kc.clarin-it.it/knowledge/knowledge-pills/introduction-to-wikidata-and-the-importation-of-bibliographic-elements-from-zotero/ (last access 29.08.2025).

45 https://www.wikidata.org/wiki/Q305622 (last access 05.01.2026).

46 https://pt.wikipedia.org/wiki/Abascanto_de_Cef%C3%Adsia (last access 29.08.2025).

Based on this lack of certainty about the feasibility of using *Wikipedia*, we decided to investigate the possibility of using a domain-specific knowledge base. As already mentioned, several resources and protocols provide unique identifiers for different kinds of entities (Greek names/persons, places, texts, passages of texts),[47] but to the best of our knowledge, no system targets specifically people. One structured knowledge base was set up by Matteo Romanello and is available on GitHub,[48] but focuses on texts and, as a consequence, on authors, a small subset of the total number of people that are mentioned in the corpus. We decided henceforth to thoroughly investigate two potential knowledge bases for the disambiguation of people, with a very different genesis.[49]

Initially, we considered an initiative with a comparable goal to the *NIKAW* project, the *ToposText* resource. The *ToposText* website offers a substantial corpus of English translations of Ancient Greek and Latin texts, enriched with manual annotations of various entities, including places, people, monuments etc. While the primary focus of *ToposTexts* is on geographical locations, in particular on Greek geographical locations,[50] their documentation indicates that they have also annotated a wide range of other entities, providing a classification system (e.g. 'animal', 'female', 'group', 'datable event') and assigning unique identifiers when possible, using various resources such as *WikiData* or *Trismegistos* places.[51] This approach appeared to align well with our requirements for a knowledge base, as it was grounded in a corpus similar to the one we aimed to analyze and already offered a structured classification and unique identifiers for entities. However, it appeared rather clearly that the list of entities was not constituted with the goal of designing a consistent catalogue, and that there were some tangible mistakes in the labelling of the entities, while the classification itself could result in inconsistent choices. For example, at the time of our investigation, Sappho[52] and nymphs[53] were incorrectly labelled as male, while constellations and stars were categorized under the 'astronomic' class – yet planets were classified as 'places'. These inconsistencies made it difficult to discern the underlying criteria for classification. Although some issues have been corrected over time, the dataset's current state is not fully documented, which undermines its usability for our scope.

The second resource we considered was the above-mentioned *Paulys Realencyclopädie der classischen Altertumswissenschaft*, whose publication was started in 1890 by Georg Wissowa and completed in 1980, building on a previous version published between 1837 and 1864 by August Friederich Pauly. A monumental work to which many of the most prominent classical philologists contributed, it contains approximately 100,000 entries on antiquity-related topics. All the entries are currently being digitized on the German *Wikisource* (we refer to the *Wikisource* version of the *Paulys Realencyclopädie* as *RE*). All printed double pages of the lexicon, around 27,600, are available as scans, and 65,704 articles are open, hence the full text of the *RE* entry is available (*Volltext* henceforth). The remaining articles cannot be entirely transcribed yet because they are still under copyright based on the year of death of their author, but they are regularly added to the resource as soon as the copyright expires. In addition, for all the entries, the register of keywords (*Stichwörter*) is available, meaning the list of the entries with a very short description (a few words) of its content (*Kurztext*). As an example, fig. 2 shows the different components of the entry for the freedman Hiberus (Hiberus 2).

---

47 For texts and passages of texts, cf. the Canonical Text Services protocol, see for instance: https://web.archive.org/web/20211130011501/http://www.homermultitext.org/hmt-doc/cite/cts-urn-overview.html (last access 29.08.2025).

48 https://github.com/mromanello/hucitlib (last access 29.08.2025).

49 An extended version of the comparison is available in de Graaf et al. (2024).

50 https://topostext.org/the-project (last access 29.08.2025).

51 https://topostext.org/people/ (last access 29.08.2025).

52 https://topostext.org/people/483 (last access 29.08.2025).

53 https://topostext.org/people/159 (last access 29.08.2025).

Fig. 2: The *RE* entry for Hiberus 2. "Hiberus 2" is the *Stichwort*, while "Kaiserlicher Freigelassener des Tiberius" (box on the right) is the *Kurztext*. The *Volltext* is the paragraph providing information on this person.

The *RE* was integrated into the *TM* database and preprocessed in a semi-automated manner, with manual verification conducted by the *TM+* team. This process involved identifying entries that described individuals (totalling 48,750 entries) and reconstructing the full names of the persons mentioned. Additionally, as will be discussed in Section 4.2, each *RE* name was linked to its corresponding *NamVar* in *TM+*. This enabled the specific subset of the *RE* to function as a knowledge base for our purposes.

We compared the coverage of individuals in *ToposText* and the *RE* by evaluating how often the correct match was included in the list of potential candidates generated through a fuzzy match between a person's name extracted from a text and the two knowledge bases (i.e. by relying on the surface form of the entity). To achieve this, we annotated a sample of Latin texts with both *ToposText* and *RE* identifiers for the individuals mentioned and assessed the results. The experiment demonstrated that the *RE* was more suitable for this task, as the total number of unlinkable mentions was significantly lower when using the *RE* compared to *ToposText*. Therefore, in the rest of our work, we decided to proceed with the *RE*.

This work also revealed certain limitations of the *RE* as a knowledge base, albeit affecting only a minority of cases. These limitations are partly a natural consequence of the *RE*'s origins as a printed work, whose publication spanned several years. Beyond the issue of missing entries, we encountered instances where multiple entries corresponded to the same entity (e.g. Hadrianus 1 = Aelius 64), as well as cases where a single entry referred to multiple entities (Phorbas 1, referring to multiple heroes).

Section 4.2: Creating Training Data

Existing digital datasets for NEL in classical texts do not use the *RE* as a source for identifiers. Although Rollinger employed *RE* identifiers to disambiguate individuals in Cicero's social network,[54] his dataset remains unavailable in a digital format. Consequently, we had to begin our work from scratch – particularly for Greek texts, since only Latin texts were annotated in our *ToposText*/*RE* experiment. Manual entity disambiguation is a labour-intensive and time-consuming process. To streamline our efforts, we implemented two complementary strategies:

- Automated training data generation by integrating the *TM+* infrastructure with the *GLAUx* corpus.
- A small-scale case study combined with manual annotation to produce high-quality training data.

We classify the data generated through automation as 'silver data' – structured but unverified – whereas manually annotated data constitutes 'gold-standard' reference material.

As noted earlier, the *GLAUx* corpus is linked to the *TM NameVariants* database, with each variant mapped to its corresponding *GLAUx* lemma. This connection enables us to extract all passages containing a specific name variant (e.g., every instance where "Thucydides" appears). Additionally, *RE* entries for individuals are connected to their respective name variants in the database. For example, The *RE* entry Iulius 131, referring to Gaius Iulius Caesar, has been linked to the *NamVar* and *Nam ID*s of each of the *tria nomina* (i.e. *NamVar ID* 69567 and *Nam ID* 9067 for Gaius, etc.). This integration creates a direct pathway from an *RE* entry about a person to every textual occurrence of their name.

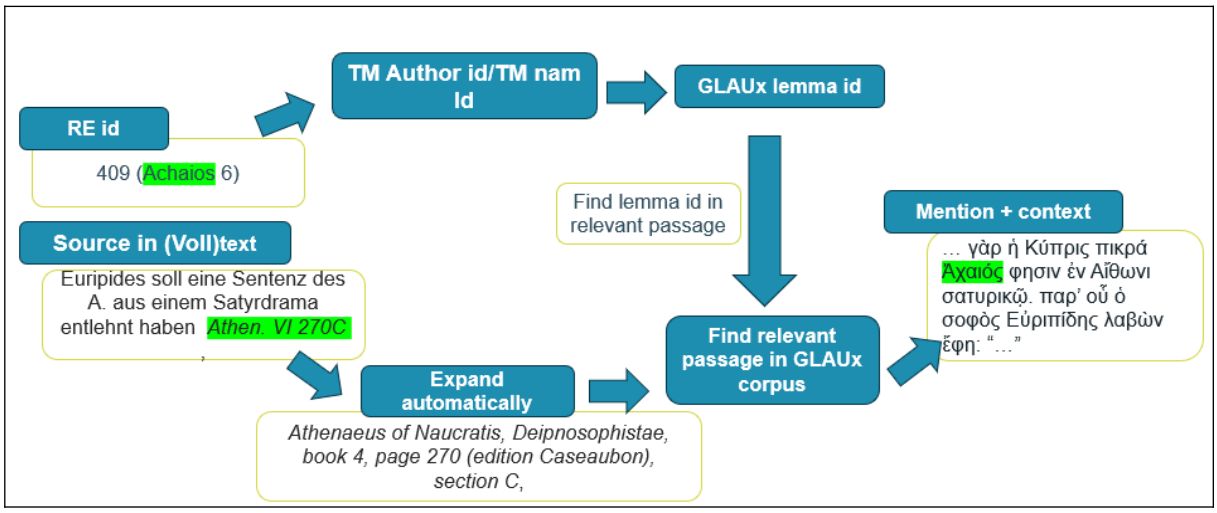


Fig. 3: Pipeline for the automatic creation of NEL training data.

In cases where the *TM+* algorithm successfully located and disambiguated a personal name within retrieved passages, identification proved largely accurate. However, the pipeline encountered several limitations. Challenges in matching RE-referenced passages to the *GLAUx* corpus stemmed from three primary issues: (1) inconsistencies between the reference systems used in the *RE* and *GLAUx*, (2) errors of the *TM+* algorithm, and (3) instances where cited passages contained no explicit mention of the person named in the *RE* entry.

To address this issue, we expanded the search to the entire text (e.g., beyond the specific chapter or paragraph indicated by the algorithm). If the name variant appeared fewer than 50 times in the text, we assumed that these passages referred to the correct individual, while if more than 50 mentions were present, we assumed that the chances were high that other individuals with the same name occurred in

54 Rollinger (2014).

the text. The threshold of 50 occurrences was set arbitrarily as an initial benchmark for evaluating the pipeline's performance: in order to assess the impact of this decision on the quality of the silver data, we conducted a manual evaluation of the pipeline on a subset of retrieved mentions. In the future, by testing different thresholds, we want to assess how the amount of wrongly labeled entries in the silver training data impacts the performance of the model. Tab. 1 wants to give a quantitative answer to the following questions:

- Is the full reference extracted from the *RE Volltext* in its entirety (first row of tab. 1)? The extraction is successful 55 times out of 80, half of which result in the retrieval of the correct passage.
- Does the *GLAUx* ID of the retrieved mention match the reference from the *RE* (second row)? Half of the cases yield to the identification of the correct mention in the text: unsurprisingly, this happens mostly when the correct passage has been found, and only 8 times when the extended the search to the full text.
- Is the established link correct, i.e. is the individual identified in the *RE* article the individual mentioned in the text (third row)? This happens 54 times out of 80 (which is a positive result), and mainly when the passage is correctly identified, even though also the extended search has yielded some correct data (16 correctly linked mentions).

	Passage found	**Passage not found**	**Total (of 80)**
Reference fully extracted	28	27	55
Correct mention identified	35	8	43
Correct link established	38	16	54

Tab. 1: Evaluation of the quality of the silver data.

In total, we processed 22,764 entries, producing 120,761 automatically extracted references to texts, of these, 16,664 were found to be referring to texts in the *GLAUx* corpus. 4,322 were precisely located in their respective texts, while for the other references we expanded the search to the full text: we retained the mentions if the name occurred less than 50 times in the text, while we discarded the reference if the name occurred 50 times or more in the text. After removing duplicates, we ended up with 13,964 mention-entity pairs. Despite the risk of introducing errors into the training data, we retained the entire set of passages. However, only the mentions that were precisely located were used for the evaluation of the NEL system, as will be described in Section 4.3.

The second strategy involved creating a manually annotated gold dataset. Given the overarching objective of the *NIKAW* project – applying SNA to named entities – one of the two PhD researchers began working on a smaller-scale case study. This case study allowed us to test the SNA methodology on a restricted corpus. Specifically, the case study focused on annotating mentions co-occurring with Plato in texts where Plato is mentioned a significant number of times. The texts included in this case study are detailed in tab. 2.

Author	Work	Period	Christian Work
Greek [GLAUx corpus]			
[Plato]	*Epistulae*	BCE	☐
Aristoteles	*Metaphysica*	BCE	☐
Dionysius Halicarnassensis	*De Demosthenis dictione*	BCE	☐
Strabo	*Geographica*	BCE	☐
Plutarchus	*Quaestiones convivales*	CE	☐
Claudius Aelianus	*Varia historia*	CE	☐
Clemens Alexandrinus	*Stromata*	CE	☑
Origenes	*Contra Celsum*	CE	☑
Diogenes Laertius	*Vitae philosophorum*	CE	☐
Galenus	*De placitis Hippocratis et Platonis*	CE	☐
Latin [LASLA / *Corpus Latin*]			
Cicero	*Tusculanae Disputationes*	BCE	☐
Tertullian	*De Anima*	CE	☑
Seneca	*Ad Lucilium Epistulae Morales*	CE	☐
Lactantius	*Divinarum Institutionum*	CE	☑
Apuleius	*Pro Se De Magia Liber*	CE	☐

Tab. 2: List of texts included in the Plato case-study.

The annotation process was conducted using the *TM+* custom-made FileMaker interface. Fig. 4 provides a schematic overview of the annotation workflow. Through its connection with *TM Names*, capitalized words are automatically assigned a set of potential *RE* articles, which are then manually selected by the annotator. This process also ensures that multi-token entities are fully annotated. The annotation was carried out collaboratively by a PhD student and a member of the *TM+* team, which simultaneously contributed to the enrichment of the *TM* database.

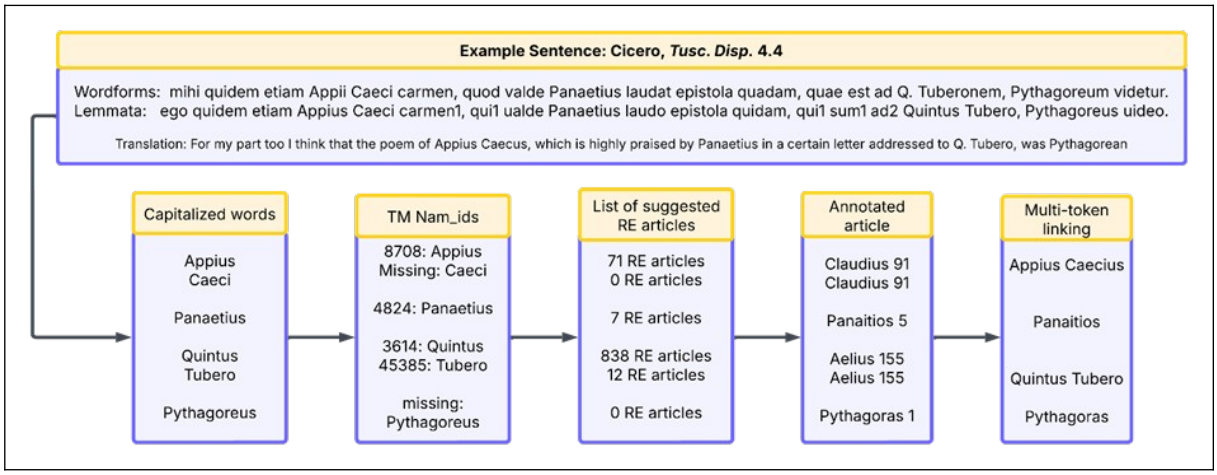


Fig. 4: Manual annotation process for the Plato case study.

The process revealed several conceptual challenges that are relevant to the project as a whole. These include disagreements among scholars regarding the identification of individuals mentioned in texts, as well as variations in textual sources, which are often particularly pronounced for infrequent names. Additionally, the use of the *RE* introduced specific challenges, such as the lack of a standardized practice for selecting the keyword for articles involving multi-token names (e.g., the Roman *tria nomina* system). To these difficulties, we must add the general limitations of the *RE* as a knowledge base, as outlined in Section 4.1. Finally, several steps in the pipeline linking the *GLAUx* corpus to the broader FileMaker infrastructure are performed (semi-)automatically, which inevitably introduces errors. These include issues with the lemmatization and capitalization system of the *GLAUx* corpus, as well as gaps in the association between *TM Names* and *RE* entries for certain individuals.

To illustrate the effort involved in completing the annotation process, we present statistics from the Greek portion of the corpus. The manual annotation of person entities required approximately 230 hours of work (shared between the two annotators), resulting in roughly 35,000 annotated mentions corresponding to about 6,000 distinct entities. These annotated datasets will be made publicly available upon finalization of the project.

Section 4.3: Training a Named Entity Linking System

Automating Named Entity Linking or Disambiguation remains a notoriously challenging NLP task, even for contemporary sources. Various approaches have been developed over time, with early efforts primarily focused on linking textual corpora to *Wikipedia* entries using hand-crafted features to capture contextual information.[55] Mihalcea and Csomai adapted machine learning methods from word sense disambiguation to the Wikification task for concepts, leveraging *Wikipedia*'s hyperlinked structure.[56] Subsequent machine learning approaches were further developed by Milne and Witten, Ratinov et al., and Rao et al.[57] More recent approaches rely on neural NEL models, employing deep learning to model relationships between textual information and knowledge bases. These methods create embeddings to represent both textual mentions and knowledge base entities. For instance, Yamada et al. proposed Wiki2Vec, which jointly maps words and entities to the same embedding space.[58]

For the *NIKAW* project, we face two unique challenges: first, because we prioritize domain-specific resources, we are not using Wikipedia as our knowledge base, which prevents us from building upon ex-

55 Bunescu / Pasça (2006;) Cucerzan (2007).

56 Mihalcea / Csomai (2007).

57 Milne/Witten (2008), Ratinov et al. (2011), and Rao et al. (2013).

58 Yamada et al. (2016).

isting Wikification methods; second, our knowledge base lacks structured elements (such as hyperlinks or categorized entries), requiring us to rely primarily on unstructured textual descriptions of entities. Consequently, we adopted a domain-independent approach, testing models that only require textual entity descriptions in the knowledge base. Additionally, we are currently focusing solely on disambiguating pre-identified entity mentions rather than implementing end-to-end systems that simultaneously perform NER and NEL.

Contemporary neural approaches for NEL use pretrained language models (like *BERT* and *RoBERTa*) fine-tuned for NEL tasks. Currently, we evaluated one of these architectures, *BLINK*,[59] and we plan to test a second one, as specified in the Conclusions. *BLINK* follows a traditional two-step process:

- Candidate generation: Creates a list of potential candidates by encoding mentions and entity descriptions using transformer models and retaining the closest embeddings.
- Candidate ranking: Concatenates mentions with descriptions to learn a joint representation of candidates and their mentions and rank candidates accordingly.

Conceptually, *BLINK*'s discriminative approach resembles human reasoning (selecting from existing options).

Since the original *BLINK* implementation was English-only, we chose two multilingual transformers that include Ancient Greek as new potential base transformers. In fact, we have only worked with Ancient Greek datasets at the moment. The first one is *UGARIT_grc_alignment* (below '*UGARIT*'),[60] already tested for the Ancient Greek NER. The second is *PhilBerta*,[61] a model trained from scratch on a high-quality dataset of classical Latin, Ancient Greek and English texts about antiquity. Furthermore, we experimented with two scenarios; in the first, the knowledge base contains the *Volltext* where available (below: _voll), in the second only the *Kurztext* was used for all entities in the KB (below: _kurz). The top k entities to retrieve was set to 64.

Tab. 3 shows the results of the NEL on a held-out test dataset of the silver data, only including those for which the full passage was found.

Model	Bi-encoder recall@top64	Cross-encoder accuracy	Overall accuracy
UGARIT_kurz	76,87	6.53	5,02
UGARIT_long	89,72	76,12	68,38
Philberta_kurz	84,80	3,81	3.24
Philberta_long	88,94	47,23	42,01

Tab. 3: Results on 'Plato case study data'.

Results show that there is still a large room for improvement. In the Conclusions, we outline what strategies we are currently undertaking.

59 Wu et al. 2020.

60 Yousef et al. (2022).

61 Riemenschneider / Frank (2023).

Conclusions and Future Work

Reflecting on the past two years of research, our work has provided preliminary answers to two core questions: the reliability of automation in processing named entities for classical studies and the effective integration of NLP tools with historical resources. While we developed automated pipelines for NER and NEL, manual annotation by experts often proved more reliable – and sometimes even more efficient – than training and correcting models. A persistent challenge is defining the acceptable margin of error in annotations for meaningful cultural analysis. Our hybrid approach combines automation with (semi)manual work, but a significant gap remains between NLP's potential for modern languages and its current results for ancient ones. For now, full automation is unfeasible, necessitating further annotation efforts and dataset standardization.

Our project benefits from unique collaborations – such as with the teams of *TM+*, *Trismegistos*, *GLAUx*, and *LASLA* – but broader progress requires interlinking resources (following models like Monica Berti's annotation projects, or the *LiLa Knowledge Base* for linguistic resources)[62], stronger communication between linguists and historians, and expanded open-access datasets (e.g., fully integrating the *RE* with *Wikidata*). Looking ahead, we are exploring synthetic training data generation using a lightweight multilingual model (trained on 435M+ Latin and Greek tokens) to produce annotated, authority-aligned sentences, reducing manual effort. For NEL, we are testing approaches like surface-form matching with *Trismegistos* aliases, prior probability ranking (using *RE* metadata, such as the length of the *RE* entries as an indication of their importace), and historical consistency filters (e.g. comparing the date of the text and the birth date of the person who is supposed to be mentioned in a passage). We are also evaluating generative models like *GENRE*[63], which may address knowledge-base gaps by directly producing entity names instead of only selecting those available in the knowledge base.

62 Passarotti et al (2020).

63 De Cao et al. (2020) and (2021).

References

Aguilar (2022): S. T. Aguilar, Multilingual named entity recognition for medieval charters using stacked embeddings and BERT-based models, in: Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, Marseille 2022, 119–128.

Antonopoulos et al. (2023): A. Antonopoulos / S. Chronopoulos / N. Ntaliakouras / P. Taktikou / A. Psomiadou / I. Markelis, Developing a Database for the Greek Fragmentary Tragedians, Digital Classics Online 9 (2023), 15–29, https://doi.org/10.11588/DCO.2023.9.95214 (last access 29.08.2025).

Bamman / Burns (2020): D. Bamman / P. J. Burns, Latin BERT: A contextual language model for classical philology, arXiv preprint (2020), https://arxiv.org/abs/2009.10053 (last access 29.08.2025).

Beersmans et al. (2023): M. Beersmans / E. de Graaf / T. Van de Cruys / M. Fantoli, Training and evaluation of named entity recognition models for classical Latin, in: A. Anderson et al. (ed.), Proceedings of the ancient language processing workshop, Shoumen 2023, 1–12, https://aclanthology.org/2023.alp-1.1/ (last access 29.08.2025).

Beersmans et al. (2024): M. Beersmans / A. Keersmaekers / E. de Graaf / T. Van de Cruys / M. Depauw / M. Fantoli, "Gotta catch 'em all!": Retrieving People in Ancient Greek Texts Combining Transformer Models and Domain Knowledge, in: Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), Bangkok 2024, 152–164.

Berti et al. (2019): M. Berti / K. Simov / M. Eskevich, Named Entity Annotation for Ancient Greek with INCEpTION, in: Proceedings of CLARIN Annual Conference 2019, Leipzig 2019, 1–4.

Broux / Depauw (2015): Y. Broux / M. Depauw, Developing Onomastic Gazetteers and Prosopographies for the Ancient World Through Named Entity Recognition and Graph Visualization: Some Examples from Trismegistos People, in: L. M. Aiello / D. McFarland (ed.), Social Informatics, Cham 2015, 304–313, https://doi.org/10.1007/978-3-319-15168-7_38 (last access 29.08.2025).

Burns (2023): P. J. Burns, Latincy: Synthetic trained pipelines for Latin NLP, arXiv preprint (2023). arXiv:2305.04365.

Bunescu / Paşca (2006): R. Bunescu / M. Paşca, Using Encyclopedic Knowledge for Named Entity Disambiguation, in: D. McCarthy / S. Wintner (ed.), 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento 2006, 9–16, https://aclanthology.org/E06-1002/ (last access 29.08.2025).

Celano (2024): G. G. A. Celano, Opera graeca adnotata: Building a 34M+ Token Multilayer Corpus for Ancient Greek, arXiv preprint (2024), https://arxiv.org/abs/2404.00739 (last access 29.08.2025).

Clérice (2021): T. Clérice, Corpus Latin antiquité et antiquité tardive lemmatisé (Version 0.1.3) [Computer software], Zenodo (2021), https://doi.org/10.5281/zenodo.4337145 (last access 29.08.2025).

Cucerzan (2007): S. Cucerzan, Large-Scale Named Entity Disambiguation Based on Wikipedia Data, in: J. Eisner (ed.), Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague 2007, 708–716, https://aclanthology.org/D07-1074/ (last access 29.08.2025).

De Cao et al. (2020): N. De Cao / G. Izacard / S. Riedel / F. Petroni, Autoregressive Entity Retrieval (Version 3), arXiv (2020), https://doi.org/10.48550/ARXIV.2010.00904 (last access 29.08.2025).

De Cao et al. (2021): N. De Cao / L. Wu / K. Popat / M. Artetxe / N. Goyal / M. Plekhanov / L. Zettlemoyer / N. Cancedda / S. Riedel / F. Petroni, Multilingual Autoregressive Entity Linking (Version 1), arXiv (2021), https://doi.org/10.48550/ARXIV.2103.12528 (last access 29.08.2025).

de Graaf et al. (2024): E. de Graaf / M. Depauw / M. Fantoli, "Nescio Carneades iste qui fuerit": Evaluation of Knowledge Bases for Named Entity Linking for Latin Texts, in: The First Workshop on Data-driven Approaches to Ancient Languages, Ghent 2024, 1–11.

Devlin et al. (2018): J. Devlin / M.-W. Chang / K. Lee / K. Toutanova, BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, arXiv (2018), abs/1810.04805.

Ehrmann et al. (2021): M. Ehrmann / A. Hamdi / E. L. Pontes / M. Romanello / A. Doucet, Named Entity Recognition and Classification on Historical Documents: A Survey, arXiv (2021), http://arxiv.org/abs/2109.11406 (last access 29.08.2025).

Ehrmann et al. (2022): M. Ehrmann / M. Romanello / S. Najem-Meyer / A. Doucet / S. Clematide, Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents, in: A. Barrón-Cedeño et al. (ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Cham 2022, 423–446, https://doi.org/10.1007/978-3-031-13643-6_26 (last access 29.08.2025).

Erdmann et al. (2016): A. Erdmann / C. Brown / B. Joseph / M. Janse / P. Ajaka / M. Elsner / M.-C. de Marneffe, Challenges and Solutions for Latin Named Entity Recognition, in: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), Osaka 2016, 85–93.

Erdmann et al. (2019): A. Erdmann / D. J. Wrisley / B. Allen / C. Brown / S. Cohen-Bodénès / M. Elsner / Y. Feng / B. Joseph / B. Joyeux-Prunel / M.-C. de Marneffe, Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis 2019, 2223–2234.

Fantoli et al. (2022): M. Fantoli / M. Passarotti / F. Mambrini / G. Moretti / P. Ruffolo, Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin, in: T. Declerck et al. (ed.), Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference, Marseille 2022, 26–34, https://aclanthology.org/2022.ldl-1.4/ (last access 29.08.2025).

Fetahu et al. (2022): B. Fetahu / A. Fang / O. Rokhlenko / S. Malmasi, Dynamic Gazetteer Integration in Multilingual Models for Cross-Lingual and Cross-Domain Named Entity Recognition, in: M. Carpuat et al. (ed.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle 2022, 2777–2790, https://doi.org/10.18653/v1/2022.naacl-main.200 (last access 29.08.2025).

Foka et al. (2021): A. Foka / D. A. McMeekin / K. Konstantinidou / N. Mostofian / E. Barker / O. C. Demiroglu / E. Chiew / B. Kiesling / L. Talatas, Mapping Ancient Heritage Narratives with Digital Tools, London 2021, 55–65.

Ji et al. (2022): S. Ji / S. Pan / E. Cambria / P. Marttinen / P. S. Yu, A Survey on Knowledge Graphs: Representation, Acquisition, and Applications, IEEE Transactions on Neural Networks and Learning Systems 33/2 (2022), 494–514, https://doi.org/10.1109/TNNLS.2021.3070843 (last access 29.08.2025).

Keersmaekers (2021): A. Keersmaekers, The GLAUx Corpus: Methodological Issues in Designing a Long-Term, Diverse, Multi-Layered Corpus of Ancient Greek, Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021 (2021), 39–50, https://doi.org/10.18653/v1/2021.lchange-1.6 (last access 29.08.2025).

Keersmaekers et al. (2024): A. Keersmaekers / F. Pietowski / T. Van Hal / M. Depauw, The Browser-Based GLAUx Treebank Infrastructure: Framework, Functionality, and Future, Cybernetics and Information Technologies 24/4 (2024), 164–174, https://doi.org/10.2478/cait-2024-0041 (last access 29.08.2025).

Kemp (2021): J. Kemp, Beyond Translation: Building Better Greek Scholars, Pelagios (2021), https://medium.com/pelagios/beyond-translation-building-better-greek-scholars-561ab331a1bc (last access 10.07.2024).

Mercelis / Keersmaekers (2022): W. Mercelis / A. Keersmaekers, Electra-grc (2022), https://huggingface.co/mercelisw/electra-grc (last access 10.07.2024).

Mihalcea / Csomai (2007): R. Mihalcea / A. Csomai, Wikify! Linking Documents to Encyclopedic Knowledge, Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (2007), 233–242, https://doi.org/10.1145/1321440.1321475 (last access 29.08.2025).

Milne / Witten (2008): D. Milne / I. H. Witten, Learning to Link with Wikipedia, Proceedings of the 17th ACM Conference on Information and Knowledge Management (2008), 509–518, https://doi.org/10.1145/1458082.1458150 (last access 29.08.2025).

Oliveira et al. (2021): I. L. Oliveira / R. Fileto / R. Speck / L. P. F. Garcia / D. Moussallem / J. Lehmann, Towards Holistic Entity Linking: Survey and Directions, Information Systems 95 (2021), https://doi.org/10.1016/j.is.2020.101624 (last access 29.08.2025).

Palladino / Yousef (2024): C. Palladino / T. Yousef, Development of Robust NER Models and Named Entity Tagsets for Ancient Greek, in: R. Sprugnoli / M. Passarotti (ed.), Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024, Paris 2024, 89–97, https://aclanthology.org/2024.lt4hala-1.11/ (last access 29.08.2025).

Palladino et al. (2024): C. Palladino / M. Fantoli / E. de Graaf / M. Berti / M. Romanello / T. Yousef / M. Beersmans / T. Gheldof / L. Soffiantini / E. Litta Modignani Picozzi, Experience and Challenges with Named Entities – Workshop at DHBenelux 2024: Named Entity Annotation Guidelines and Tutorials, Zenodo (2024), https://doi.org/10.5281/ZENODO.11366870 (last access 29.08.2025).

Passarotti et al. (2020): M. Passarotti / F. Mambrini / G. Franzini / F. M. Cecchini / E. Litta / G. Moretti / P. Ruffolo / R. Sprugnoli, Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin, Studi e Saggi Linguistici 58/1 (2020), https://doi.org/10.4454/ssl.v58i1.277 (last access 29.08.2025).

Pelagios (2021): Pelagios, Beyond Translation: Building Better Greek Scholars, Medium (2021), https://medium.com/pelagios/beyond-translation-building-better-greek-scholars-561ab331a1bc (last access 10.07.2024).

Pellizzari di San Girolamo (2023): C. C. Pellizzari di San Girolamo, Conflations and Duplications in Wikidata Items: Causes, Detection, Solutions, and Issues, Wikidata@ISWC (2023), https://api.semanticscholar.org/CorpusID:265381505 (last access 29.08.2025).

Rao et al. (2013): D. Rao / P. McNamee / M. Dredze, Entity Linking: Finding Extracted Entities in a Knowledge Base, in: T. Poibeau et al. (ed.), Multi-source, Multilingual Information Extraction and Summarization, Berlin 2013, 93–115, https://doi.org/10.1007/978-3-642-28569-1_5 (last access 29.08.2025).

Ratinov et al. (2011): L. Ratinov / D. Roth / D. Downey / M. Anderson, Local and Global Algorithms for Disambiguation to Wikipedia, in: D. Lin et al. (ed.), Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland 2011, 1375–1384, https://aclanthology.org/P11-1138/ (last access 29.08.2025).

Riemenschneider / Frank (2023): F. Riemenschneider / A. Frank, Exploring Large Language Models for Classical Philology, in: A. Rogers et al. (ed.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto 2023, 15181–15199, https://doi.org/10.18653/v1/2023.acl-long.846 (last access 29.08.2025).

Rollinger (2014): C. Rollinger, Amicitia sanctissime colenda. Freundschaft und soziale Netzwerke in der späten Republik, Heidelberg 2014.

Sevgili et al. (2022): Ö. Sevgili / A. Shelmanov / M. Arkhipov / A. Panchenko / C. Biemann, Neural Entity Linking: A Survey of Models Based on Deep Learning, Semantic Web 13/3 (2022), 527–570, https://doi.org/10.3233/SW-222986 (last access 29.08.2025).

Shen et al. (2015): W. Shen / J. Wang / J. Han, Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions, IEEE Transactions on Knowledge and Data Engineering 27 (2015), 443–460, https://doi.org/10.1109/TKDE.2014.2327028 (last access 29.08.2025).

Shnayderman et al. (2019): I. Shnayderman / L. Ein-Dor / Y. Mass / A. Halfon / B. Sznajder / A. Spector / Y. Katz / D. Sheinwald / R. Aharonov / N. Slonim, Fast End-to-End Wikification (Version 1), arXiv (2019), https://doi.org/10.48550/ARXIV.1908.06785 (last access 29.08.2025).

Singh et al. (2021): P. Singh / G. Rutten / E. Lefever, A Pilot Study for BERT Language Modelling and Morphological Analysis for Ancient and Medieval Greek, Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2021) (2021), 128–137.

Sommerschield et al. (2023): T. Sommerschield / Y. Assael / J. Pavlopoulos / V. Stefanak / A. Senior / C. Dyer / J. Bodel / J. Prag / I. Androutsopoulos / N. de Freitas, Machine Learning for Ancient Languages: A Survey, Computational Linguistics (2023), 1–44, https://doi.org/10.1162/coli_a_00481 (last access 29.08.2025).

STEPBible (2023): STEPBible, STEPBible-Data, GitHub (2023), https://github.com/STEPBible/STEPBible-Data (last access 10.07.2024).

Wu et al. (2020): L. Wu / F. Petroni / M. Josifoski / S. Riedel / L. Zettlemoyer, Scalable Zero-shot Entity Linking with Dense Entity Retrieval, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020), 6397–640, https://doi.org/10.18653/v1/2020.emnlp-main.519 (last access 29.08.2025).

Yousef et al. (2023): T. Yousef / C. Palladino / G. Heyer / S. Jänicke, Named Entity Annotation Projection Applied to Classical Languages, in: Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Dubrovnik 2023, 175–182.

Yousef et al. (2022): T. Yousef / C. Palladino / F. Shamsian / A. d'Orange Ferreira / M. Ferreira dos Reis, An Automatic Model and Gold Standard for Translation Alignment of Ancient Greek, Proceedings of the Thirteenth Language Resources and Evaluation Conference (2022), 5894–5905, https://aclanthology.org/2022.lrec-1.634 (last access 29.08.2025).

Figure and Table References

Author Contact Information[64]

Margherita Fantoli
Assistant Professor
Faculty of Arts
KU Leuven
E-mail: margherita.fantoli@kuleuven.be

64

DIGITAL CLASSICS ONLINE

Detecting Eastern and Western Names in the Latin Corpus of the *SERICA* Project – With Special Regard to the *Confucius Sinarum Philosophus* (1687) as a Case Study

Andrea Balbo, Elisa Della Calce

Abstract: This paper aims to describe the role of ICT within the *SERICA* (***S**ino-**E**uropean **R**eligious **I**ntersections in **C**entral **A**sia. Interactive Texts and Intelligent Networks*) Project, especially by focusing on the corpus of Latin texts we are progressively building. Particular attention will be paid to the annotation of Named Entities (NEs) through *Recogito*[1] from a very peculiar 17th century Latin text, entitled *Confucius Sinarum Philosophus* (*CSP*) and edited by the Jesuits Prospero Intorcetta, Christian Herdtrich, François de Rougemont, and Philippe Couplet in 1687. Despite including the translation of three Confucian texts (*Daxue*, *Zhongyong*, *Lunyu*), the *CSP* contains various references to Graeco-Hellenistic and Roman literature, and this comes as unsurprising since the Jesuit *Ratio Studiorum* (1599) was indebted to pagan classical authors. Yet the reception of ancient Latin literature can be further investigated by resorting to digital technologies. The annotation and the extraction of NEs allow in fact to take into account an extensive amount of data and to establish a first mapping concerning the impact of classical antiquity on the *CSP*, so as to detect which authors were mentioned more often and to reflect on their pattern of distribution within the work.

1. The SERICA Project[2]

The *SERICA* (***S**ino-**E**uropean **R**eligious **I**ntersections in **C**entral **A**sia. Interactive Texts and Intelligent Networks*) Project has contributed to bringing to scholars' attention Eurasian studies or Silk Road(s) studies that are counted among the liveliest and most attractive subjects in global history. This specific research field has just started receiving in Italy the attention it deserves, although the wide-ranging perspective and the complexity involved in such a complicated network, which spans at least two millennia and more than ten thousand kilometres, have not always been taken into due account. These studies have often uncovered the role of travelling merchants and missionaries, who from Late Antiquity onwards started tracking the earliest cultural path which joined East Asia and Europe in a structural and permanent interchange, which reached its peak in the 16th and 18th centuries, with the production of a huge number of texts and the creation of multilingual libraries, especially in China. These works aimed at circulating to the Far East elements of classical European education, and at the same time communicated to Europe the first widespread information on geography, customs and cultural

1 https://recogito.pelagios.org/ (last access 20.03.2026).

2 This paper has been adapted from a presentation given at the International Workshop *Nomina Omina. Detecting and Preserving Ancient Greek and Latin Proper Names in the Age of Artificial Intelligence (Leipzig University, June 27–29 2024)*. Paragraphs 1–2 are by Andrea Balbo and paragraphs 3–4 are by Elisa Della Calce. Our thanks go in particular to Monica Berti, for organizing this Workshop and for her precious advice (especially on the 'semantic annotation'), to Philip Barras and Simone Mollea for their thorough reading and useful suggestions. It goes without saying that only the authors are responsible for any remaining mistakes.

bases of the civilizations of East Asia. At the Western end of the Silk Road(s), in Anatolia and in the Levant, the entangled interactions of the Western and Eastern Christianities and the Muslim world in the Middle Ages have also received sustained – though uneven – scholarly attention. New approaches are now underlining the dialogue and controversy within what can be regarded as a pervasive Mediterranean 'culture of disputation'. A reassessment of these texts and other sources bearing witness to Christian-Muslim interactions (for example ethnographic works) appears badly needed, so as to shed light on their intended audience, wider circulation and later reception.

This multifaceted process of communication and cultural exchange has been the subject of many historical studies, always on the basis of a small selection of documents and almost never from the point of view of the linguistic intersection between the civilizations involved, nor from the perspective of Eastern contacts with Europe and the West. Our scope is precisely to bring together an interdisciplinary team of experts and at the same time to promote further national or international collaborations underlining the connections with other similar projects. In this regard, while from a strictly administrative standpoint the team is mainly composed by scholars belonging to the Universities of Pisa and Turin, a constellation of partner universities or institutions is also involved and cooperations are already established with Siena, Venice, Naples 'L'Orientale', Bologna-Ravenna, Seoul, and Beijing. The research team includes historians, philologists, and computer scientists.[3] While the role of philologists and specialists of the various Oriental languages is indispensable, as they guarantee a precise and rigorous interpretation and contextualization of all the scrutinized texts, the presence of information scientists and the creation of collaborative virtual environments is a further added value of this project, which promotes the convergence of technologies and applications between different domains. In this sense, the present project has brought together and unified research in the philological, linguistic and historical fields, which, despite having reached very high levels of refinement, still remain distinct and separate and too often do not create a dialogue between them.

By selecting some of the most significant case studies within a vast period (400 BC–1700 AD), we have been investigating how books and manuscripts reached Europe and were sometimes further disseminated through translations. Such text circulation attests not only to diplomatic-commercial relations, but also to an osmotic interaction between different agencies, faiths, social systems and cultural universes that moved along these routes. Considering "material things as entry points into history",[4] the project's digital platform and virtual mapping of the historical-geographical paths either mentioned in these books or actually covered by their circulation, seek to highlight, in its broadest sense, the uninterrupted contact between Asia and Europe. Choosing a long historical period and a wide geographical area permits the adoption of a trans-disciplinary agenda, which promotes a rethinking of traditional methodologies and has the aim of achieving a non-Eurocentric reading of the past: both these aspects allow us to reflect on and answer crucial questions for contemporary society.

Consequently, the novelty in our approach can be summarized as a twofold one: it presents and makes available new documents, often unpublished or scarcely known, for most of them have never been translated into Italian or any modern language. Secondly, this has been made possible also thanks to the pivotal role played by the ICT, which permits the evaluation and interpretation of the data by combining them into a united, transversal and dynamic corpus, capable of exploiting the potential of information technologies. New diachronic perspectives of comparative studies have thus been opened; in particular, the strict cooperation with information scientists permits the development of a digital library and a new corpus of texts and documents from medieval and early modern times concerning the Middle and the Far East. By proposing a fresh new approach to this kind of literature, we wish to place the study of East-West trajectory into the wider scientific discussion about Eurasian cultural exchanges, in particular considering how Eastern sources dealt with and were received in the West. Fur-

3 See https://serica.unipi.it/il-team (last access 29.08.2025).

4 Ulrich et al. (2015), XI.

thermore, it is difficult to find platforms that combine inter-linguistic and philological studies with historical reconstructions and a clear visual rendering.

One of the objectives of the *SERICA* Project is the development of a digital library on the website,[5] which aims to include texts regarding the Central Asian routes between China and Europe. Many of these texts were written in Latin by Jesuit missionaries who travelled to China between the 16th and the 18th centuries and drew up several works to spread in Europe Chinese history and culture, along with Confucian thought.

In this contribution, I intend to describe the digital side of the project, especially by focusing on the results achieved and its future perspectives (§2), while Elisa Della Calce aims to show a semantic annotation experiment (§3–§4) – which was started during an Erasmus Training stay at Leipzig University in 2024 (22.04–03.05.24) – by using *Recogito* on the *Confucius Sinarum Philosophus* (henceforth *CSP*).

The *CSP* is one of the crucial and longest texts that we planned to include in *SERICA*'s database: it was printed in Paris in 1687 and edited by the Jesuits Prospero Intorcetta, Christian Herdtrich, François de Rougemont, and Philippe Couplet, with the aim of spreading Confucianism in Europe. The Jesuits resorted to Latin as a medium to make Confucianism more understandable to Western cultivated readers. The *CSP* contains a prefatory epistle, dedicated to Louis XIV; an introduction (*Proëmialis Declaratio*); a life of Confucius; the Latin translation of three Confucian Books, namely *Daxue* (The Great Learning), *Zhongyong* (The Doctrine of the Mean), *Lunyu* (The Analects), and some chronological tables related to Chinese monarchy.[6]

2. The Role of ICT within the *SERICA* Project

Thanks to collaboration with private companies, which strengthens the links between the university and the local area, once fully realized, the website will feature a highly usable and effective user interface, and will be structured according to different phases of design prototyping (from Wireframing to Graphic Design), including the definition of general categories of users (Proto-Person); the construction of Use cases and Users stories; the description of navigation paths (User journeys); and the analysis of the effectiveness of the proposal through interviews and user testing. Once completed, it will feature a digital library equipped with innovative semantic web and artificial intelligence tools and a diachronic array of interactive maps that make use of geolocalisation tools. Data have been collected into an interactive database, with visual mapping of the texts and objects documenting mutual exchanges, mainly those from East to West. The final result will present a navigable interactive and 'talking' diachronic map of the various routes, where images and historical details will contribute to reconstruct a context as precise as possible of the knowledge that a traveller might have had.

This digital library is being realized using suites that include functions such as uploading texts and metadata, semantic enrichment, and data management. Such a knowledge base will represent a concrete model (and the first one in Italy) to display interactive data of historical relationship between different Eastern countries and Europe (for a similar project see the CHCD at Boston University). In addition, the book and other textual materials have become an integral part of the platform. Particular attention is being paid to the processes of linguistic analysis (through Natural Language Processing techniques), so as to normalize the texts and extract the greatest possible amount of structured information to be used as input for the advanced queries and analyses of artificial intelligence. A context of particular interest in the research area of the project is multilingualism, i.e. having texts written in different

5 https://serica.unipi.it/ (last access 29.08.2025).

6 The text of *CSP* is currently available in a PDF format on https://archive.org/ (last access 29.08.2025). Text quotations from this work have been adapted to modern typographic criteria (e.g. & → *et*; <j> → <i>).

languages – among which ancient Greek, Latin, Chinese, Arabic – which must be normalized and adapted to a common knowledge model in order to be able to create machine learning models with the greatest possible quantity of homogeneous data so as to guarantee greater accuracy. Another important aspect is the temporal diversity of the origin of the sources, which means that NLP analysis might be particularly difficult, since normally the available tools only consider the current versions of a language and not its older or classical versions. In particular, its hybrid nature (digital library-interactive map-database) makes it a pilot model in the Italian research panorama of Asian language studies, which can be connected to an ever-growing network of information, both national and international, cooperating with projects currently already active or in progress. As already suggested, the creation of an information framework structured through semantic and thematic paths and by means of a virtual guide based on web ontologies and deep learning is central to the success of the project.

The final goal is to create a virtual place that can be used not only by experts for academic purposes, but also by a wider audience for educational or divulgation purposes. Although the project is carried out in a scientific manner by experts from different academic sectors, one can also envisage the potential didactic impact, involving many students in the study and the use of these digital tools. Interdisciplinary teaching also increases student learning and encourages an intertwined approach at school and at the academy. In particular, we are proud to mention the fact that the project permitted a temporary enrolling of at least twenty young researchers, some of whom still undergraduates, or at the early stage of their professional academic career.

Uploading documents on *SERICA*'s website enables users to carry out targeted lexical searches (see fig. 1) by accessing texts that previously were digitized and spread only in a non-searchable format. This process involves several steps. To begin with, the document, depending on how it looks or has been digitized, is transcribed or processed by an OCR software. Afterwards, the document is subjected to linguistic-typographical correction and normalization in order to remove any mistakes. At this stage, as far as Latin texts are concerned, the abbreviations are replaced by their corresponding words (e.g. & > *et*), the grapheme <j> is transcribed as <i>, and all accents on vowels are removed (e.g. *Proëmialis Declaratio*, p. IX: *eò etiam magis, quòd placuit* > *eo etiam magis, quod placuit*) so as to facilitate an exact lexical search on the database. Once this step is completed, the text is transferred into XML format, tagged according to the TEI standard, and uploaded on the site.

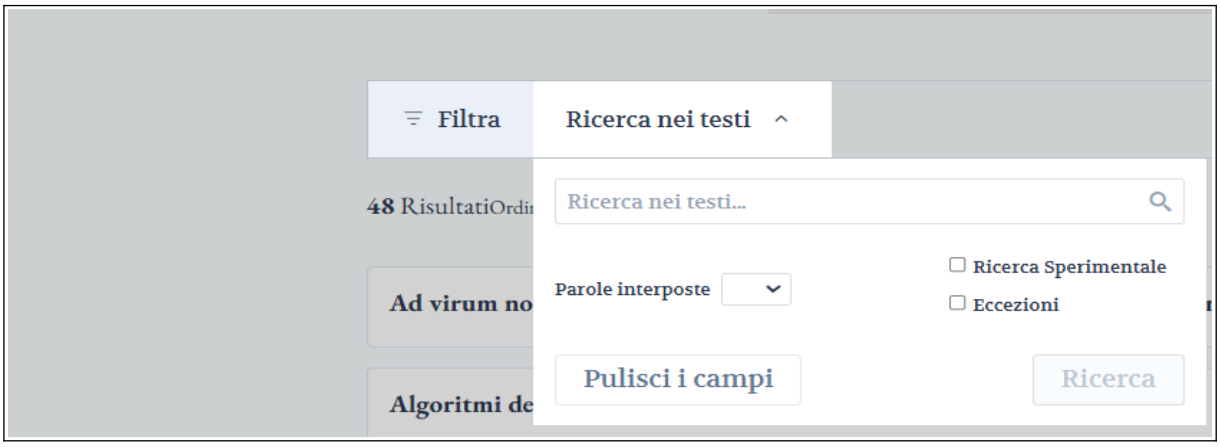


Fig. 1: The search mask of *SERICA*'s website.

3. *CSP* and Annotation through *Recogito*

This section focuses on the relationship between my literary interests in the *CSP* and the semantic Annotation process by means of *Recogito*, "an open-source, free and online semantic annotation tool developed by the Pelagios Network".[7]

My starting point was to question what impact this annotation and extraction process may have on my research goals, which particularly concern the influence of Latin authors in the *CSP*. Despite being composed in 17th century, this work includes several echoes of and quotations from pagan Latin literature, especially from Cicero. The Jesuit education was based on the *Ratio Studiorum* (1599), which put the emphasis on readings suitable to Catholic morality in the Counter-Reformation era. In this regard, a substantial number of Graeco-Roman classics was recommended to be read and learned, among which Cicero was a model for language and style.[8]

In light of this, the annotation of Named Entities (NEs)[9] or, more precisely, of proper personal names (henceforth PPNs) enabled me to consider a larger amount of data and create a mapping of the influence of Western ancient classics in the *CSP*, in order to investigate *which* authors and characters were most quoted from pagan Classical Literature. It follows that I privileged the extraction and analysis of PPNs over those of place names, yet limiting myself to some general considerations on applying *Recogito*'s gazetteers in a text as complex as the *CSP*. Since the annotation has not yet been concluded, the results I intend to share are only related to the first sections of the *CSP* (the *EP* = *Epistula praefatoria* and the *PD* = *Proëmialis Declaratio*).[10]

First of all, since automatic annotation via Herodotus Latin NER in *Recogito* proved unsuccessful, I proceeded with a manual annotation of the Latin text,[11] also in the belief that, due to the particular status of the *CSP*'s structure and content, a 'close reading approach' can contribute to a first text analysis.[12]

---

7 Del Rio Riande / Vitale (2020), 2. In addition to this, cf. also Simon et al. (2017) and Gregory (2021).

8 Cf. e.g. on this Balbo (2020), Balbo (2022), 112–117, Della Calce/Mollea (2023), 46–48, and Della Calce (2024), 279–282 (with bibliography).

9 For a general overview of the linguistic annotation, cf. at least Montemagni (2023), 162–176, Berti (2019), who in particular dealt with the annotation of ancient Greek, and Berti (2023), 316–318. Cf. also Yousef et al. (2023), who instead focused on "a processing pipeline to transfer NE annotations from a text in modern languages to parallel texts in classical or low-resourced languages" (175). On the annotation as one of the Pelagios Network's "core Activities", cf. https://pelagios.org/activities/annotation/ (last access 01.07.2025).

10 This paper refers to my use of *Recogito* mostly during the period of my training stay in Leipzig and of the *Nomina Omina* Workshop. The data and the number of annotations extracted, however, have been revised and updated in view of these proceedings and analyzed only with respect to the PPNs.

11 As far as Latin and NER are concerned, cf. at least e.g. Beersmans et al. (2023), 1–2, also with reference to Ehrmann et al. (2021) as well as to earlier bibliography, such as Erdmann et al. (2016) and (2019): "for modern high-resource languages, generic NER off-the-shelf solutions, focusing mainly on identifying locations, organizations and people, can produce highly accurate annotations. For historical languages, even prolific ones like Latin, the task remains a challenge, in part due to a lack of annotated corpora and tools (Ehrmann et al., 2021)". For a short overview on NER algorithms provided by *Recogito*, cf. del Rio Riande/Vitale (2020), 5 and Berti (2019), 1, n. 3: "*Recogito* (https://recogito.pelagios.org/ [last access 20.03.2026]) provides automatic NER tagging for historical data using Stanford CoreNLP (English, French, German and Spanish), experimental Latin NER with the Herodotus Latin NER plugin […] and experimental Hebrew NER with the Kima NER plugin (https://geo-kima.org [last access 01.07.2025])".

12 Cf. on this Montemagni (2023), 176–177, on the wake of Moretti (2005)'s dichotomy "Close Reading – Distant Reading": "l'annotazione linguistica può potenziare le funzionalità di ricerca di tipo Close Reading permettendo le ricerche per lemma, invece che per forma, oppure ricerche mediante schemi che combinano variamente informazione (morfo)sintattica e lessicale. Gli stessi testi diventano esplorabili mediante tecniche di Distant Reading, che permettono l'estrazione di generalizzazioni a partire dai testi".

The *CSP*, in fact, is a very articulated work which was exclusively composed in Latin. The Jesuits transliterated all Chinese names into Latin characters, but their initial plan was to juxtapose the Chinese and the Latin. Yet the Chinese text was in the end excluded, as N. Dew (2009) and others have pointed out:

> "Originally, the plan had been to include the original Chinese texts with the translations, but this aim had to be abandoned because of the practical difficulties involved in printing the Chinese – although in some passages a remnant of this intention survived, in the form of superscript numerals that would have guided the reader from each Latin word to the corresponding Chinese character".[13]

What is more, the Jesuits' transliteration of Chinese characters normally diverges from modern transcription: for instance, the grapheme <ç> does not always appear transliterated in the same way, as is the case with *Heu çie* > Houji[14] and *ço xi*> Zuoshi.[15] Therefore, since I have no specific sinological background, I have classified the extracted data so as to juxtapose the Jesuit transcription of Chinese names and that which is transliterated into modern Chinese, according to T. Meynard's edition of the *CSP*, which includes the English translation of the text from the *EP* to the first Confucian Book (*The Great Learning*).

For all names extracted I have also provided the link to repositories such as *Wikidata*, in order to give a unique identifier for each name and then distinguish it from other similar names that refer to different entities.[16] As far as the Western names are concerned, I preferred reporting both the original transcription and the corresponding lemma (e.g. *Ciceronem* → *Cicero*).

3.1. What Criteria for the Annotation?

Since the *CSP* is a very peculiar text at a thematic and structural level, it is extremely important to establish some strict criteria for this annotation. To begin with, I annotated only the PPNs which were explicitly cited. I therefore took into account historical or legendary names, while I did not consider names related to divinities, honorific attributes or titles, dynasties, religious or philosophical congregations and ethnonyms. In line with this, I excluded antonomasias, such as *Philosophus* to indicate *Confucius*, and personifications.

Despite focusing on the PPNs and, particularly, on the Graeco-Hellenistic and Roman ones, I limit myself, in passing, to some remarks on the annotation of place names through *Recogito* and its available gazetteers, which I quote below:

- ***HistoGIS*** – A GIS repository for historical temporalized spatial data by the Austrian Centre for Digital Humanities
- ***Pleiades*** – *Pleiades Gazetteer of the Ancient World*
- ***CHGIS*** – *China Historical GIS*
- ***DPP Places*** – Places from the Digitizing Patterns of Power project

13 Dew (2009), 210. Cf. also Della Calce (2024), 287–288.

14 https://www.wikidata.org/wiki/Q1207613 (last access 01.07.2025). According to transcription by Meynard (2011), 189.

15 https://www.wikidata.org/wiki/Q230192 (last access 01.07.2025). According to transcription by Meynard (2011), 102.

16 This is all the more important the more the name Thomas, who corresponds to Han Lin (*PD*, p. CIX) according to Meynard (2011), 227, can be confused with the name of the saint and Doctor of the Church, Thomas Aquinas (*PD*, p. LXIX and XCIII).

- ***DARE*** – *Digital Atlas of the Roman Empire*
- ***MoEML*** – *Map of Early Modern London*
- ***HGIS de las Indias*** – *Historical-Geographic Information System for Spanish America (1701-1808)*
- ***GeoNames*** – A subset of *GeoNames* populated places, countries and first-level administrative divisions
- ***Kima*** – *Kima Historical Gazetteer – place names in the Hebrew script*.

Regardless of the fact that I did not select the gazetteers which were unsuitable to the context of composition of the *CSP* (e.g. *MoEML* and *HGIS de las Indias*), they do not always convey the exact geographical information which the *CSP* provides at historical and geo-political levels. For example, in the *EP* there are 11 occurrences of place names, that is, *Gallia*/*Galliae* (6 occurrences), *Sina* (2 occurrences), *Africa*, *Asia*, *Europa* (1 occurrence respectively). However, would it be preferable to resort to the gazetteer *Pleiades*, thereby preserving the Latin uses and tradition, or to other gazetteers such as *Geonames*, which enable us to select only the modern 'France'? By the same token, the gazetteer *Pleiades* does not annotate place names as *Asia* in the *EP*, as it designates only a geographical limited portion of this continent in the ancient world. It is then preferable to use other gazetteers, such as *Geonames*, for place names relating to modern regions. With regard to China, progress has been achieved in representing the printed Western Maps of China to 1735, as is the case with the *Regnum Chinae* edited in 2022 by Marco Caboara. And this could be used as an important resource to refine any geolocalization activity, by integrating, for instance, further notes and references which could be useful in describing how better to interpret place names in their historical context.

Also regarding the analysis of the PPNs, other interesting incongruities emerge:

- the name attributed to the Buddha is variously expressed (*Xe*, *Xe Kia*, *Xaca*, *Foe*, according to Jesuit transcription),[17]
- when Jesuits allude to *Doctores Sinenses* or those who have a double name, according to Western and Eastern traditions (especially Jesuit missionaries, *Li Mateo* corresponds to *Matthaeus Riccius*), it is better to explicit both names to disambiguate (e.g. Xu Guangqi = Paul Siu),[18]
- some mythical characters are inconsistently transliterated, as is the case with *Aedipus* (*PD*, p. LIX) and *Oedipus* (*PD*, p. XVIII),
- *Recogito* does not allow any continuous annotation when the original phrasing seems to be interrupted (*PD*, p. CIX: *Matthaei scilicet Riccii*; *PD*, p. CX, *Iulius vero Aleni*).

Nonetheless, *Recogito* enabled me to download the annotations as spreadsheet data in CSV format. This means that I could further analyze the data tables, especially those relating to Graeco-Hellenistic and Roman names, in order to provide a more precise overview of their use and frequency.

---

17 *PD*, pp. XXVII–XXVIII: *ex hac* [*scil*. *Mo-ye*] *natus est illi filius,* Xe *primum, sive* Xe Kia *dictus (quo etiam nomine tota Bonziorum colluvies ac superstitio significatur; Japonii tamen corrupto vocabulo Sinico* Xaca *fecere) deinde, cum trigesimum attigit aetatis annum,* Foe *nominatus*. Furthermore, the reference to the so-called brothers Cheng (*PD*, pp. XXXIV–XXXVI, along with Meynard 2011, 127, n. 58 and 130, n. 9) could not be separated so as to distinguish two different figures and *Wikidata* resource was adapted accordingly (*s.v*. Cheng brothers, https://www.wikidata.org/wiki/Q48880876 [last access 01.07.2025]).

18 https://www.wikidata.org/wiki/Q420427 (last access 01.07.2025). Cf. *infra* n. 20.

5 PPNs have been extracted from the *EP*, namely 3 occurrences of Confucius and 1 of Louis XIV and Philippe Couplet respectively. This implies that Couplet was interested in highlighting only the names that were crucial both to emphasize the main character of the work (*Confucius*) and to accomplish the rhetorical aims related to his dedication by mentioning himself as editor and Louis XIV as recipient.

By contrast, the data extracted from the *PD* are definitely more, amounting to 516 annotations of PPNs. 76 occurrences are relating to *Confucius*[19] and 326 occurrences to Eastern/Middle Eastern names, not only transliterated Chinese names, but also names from Buddhist and Hindu as well as Assyrian and Jewish tradition.[20] As I said before, I am not going to dwell on this latter category, given my different research purposes. In this sense, Graeco-Hellenistic and Roman PPNs amount to 33 occurrences, as is evident from the following table (Tab. 1), which summarizes the remaining data (114 occurrences).

Category	PPNs (Occurrences)	PPNs (Characters)
1. Pagan Graeco-Hellenistic and Roman culture	33	16
2. Christian culture and tradition from Antiquity to Middle Ages	25	9
3. Western names related to Modern Age (from the Second Half of 15th century to 17th century approximately)[21]	56	15

Tab. 1: Categories of Western PPNs.

The first category reflects the highest number in terms of referents, thus contributing further to confirming the Jesuits' knowledge of several pagan characters and their interest in Western ancient literature. While showing the highest number of occurrences, the third category has fewer referents than the first, since Matteo Ricci is the most mentioned character (37 occurrences), which is unsurprising when one considers Ricci's crucial role in applying the Jesuit method of 'accommodation'.[22] Accordingly, the proportion between Jesuit and non-Jesuit characters shows the prevalence of the former over the latter in terms of occurrences (50 vs. 6) and referents (10 vs. 5).

19 These occurrences consist of the Latin name *Confucius*, the name *Cum-çu* (corresponding to Kongzi: cf. Meynard [2011], 175) and the reference to the Jesuit missionary Giulio Aleni (1582–1649, https://www.wikidata.org/wiki/Q2707504 [last access 01.07.2025]) as *Occidentis advena Confucius* (*Si lai cum çu*). Also the occurrence *Confucius* in the capture at *PD*, p. XLII has been included.

20 The name *Confucius* (see *supra*, n. 19) as well as Matteo Ricci's Chinese name (*Li Mateo*) have been excluded. Those who are referred to *Doctores Sinenses* or Chinese people converted to Christianism (Ignatius, Leo, Luke, Mathias, Michael, Paul, Peter, Philip, Thomas, cf. *PD*, p. CV and CIX, along with Meynard [2011], 222–223 and 227 for the corresponding Chinese names) have been included. By the same token, the name *Cham Colaus*, in the capture at *PD*, p. XLII, has been included. He corresponds to Colaus Zhang, *scil*. Zhang Juzheng (Meynard [2011], 141), who lived between 1525 and 1582, https://www.wikidata.org/wiki/Q197234 (last access 01.07.2025).

21 In the case of discontinuous syntagms (*Matthaei scilicet Riccii*; *Iulius vero Aleni*) two occurrences have been included in the total amount.

22 Cf. Mungello (1985), 44–73 and Catto (2014).

In the second category, pride of place goes to Lactantius (7 occurrences),[23] followed by Jerome (6 occurrences) and Augustine of Hippo (4 occurrences),[24] whereas Cicero (5 occurrences) is the most mentioned author in the category of Graeco-Hellenistic and Roman names. Then, in order of importance, there are 3 occurrences of Oedipus, Orpheus, Trismegistus, Plato respectively, 2 occurrences of Aristotle, Epictetus, Pythagoras, Seneca, Socrates respectively, and 1 occurrence of Aeneas, Hesiod, Homer, Plutarch, Solon, Varro respectively. This statistical survey allows us to draw some further remarks, starting from the central role played by Cicero. He is, in fact, the only figure of pagan antiquity who is called by different names (2 occurrences of *Cicero* and 3 occurrences of *Tullius*), thereby showing how the Jesuit authors were familiar with him. Not unsurprisingly the *Ratio Studiorum* (1599) attributed a crucial role to Ciceronian works as far as rhetoric, style and philosophy are concerned.

Furthermore, it is also worth considering that *Oedipus*, *Orpheus* and *Trismegistus* recur more than others. On the one hand, *Orpheus* and *Trismegistus* are commonly connected with the notion of *prisca theologia*. In this regard, D. E. Mungello, referring to *China illustrata* (1667), composed by another Jesuit, Athanasius Kircher, has pointed out:

> "Kircher saw certain vestiges of Christianity in China. […] For Kircher, the strongest religious influences on China were the Egyptian and Greek pagan religions. Kircher's position is understandable only in terms of his groundings in the Hermetic tradition of Christian apologetics which argued that certain pagan texts contained vestiges of Christianity. The primary texts of this tradition were ascribed to Hermes Trismegistus, Orpheus and Pythagoras and the tradition is referred to variously as Hermetism or *prisca theologia* (Ancient Theology)".[25]

On the other hand, *Oedipus*, who solved the famous Sphinx's riddle, is associated with two Chinese rulers at *PD*, p. XVIII and LIX, who are therefore regarded as *Aedipi*.

> *Lineolis et quidem paucis tota res constat: Nos eas proxime hic depingemus, unaque declarabimus, quoties et quomodo variatae figuras novas, et quasi nova rerum significata conficiant. Annis mille et octingentis Monarchia steterat, cum tandem Oedipus apparuit, Regulus, inquam,* Ven vam*: hic lineolis octo octies inter se mutuo commutatis conatus est, octo rerum principalium mutuas vicissitudines exponere.*
> "Everything in the world is classified according to a set of drawings of short lines. We shall draw these lines below and also demonstrate how they form every day new figures with almost new significations. The monarchy had stood for one thousand eight hundred years when at last a kind of Oedipus appeared: the minor king Wen Wang came to explain the permutations of the eight principles by combining and recombining the eight trigrams (groups of three lines) eight times."
> *PD*, p. XVIII; transl. Meynard (2011), 101.

23 Cf. von Collani (1990), 41–42: "Ph. Couplet cites several times Aurelius Augustine (354–430), Jerome (347–420), Thomas Aquinas (1225–1274) and finally relatively often the early Latin Father Lactantius (ca. 250/260–after 317). Although not precisely stated, it is easily seen that Ph. Couplet takes the quotations from the *opus maximum* by Lactantius *De divinis institutionibus libri VII.* Lactantius regards the true religion being supported by the revelations of paganism, for instance by the prophecies of the sibyls and Hermes Trismegistos, who predicted the future true Redeemer".

24 Also St. Paul and Thomas Aquinas (2 occurrences respectively) as well as Basil of Caesarea, John the Evangelist, John Chrysostom (1 occurrence respectively) are included in this category.

25 Mungello (1985), 136–137.

> Ye kim*, sive eum qui de mutationibus inscribitur; quippe cuius author est idem qui gentis Sinicae Fundator* Fo hi*. Recte tu quidem: At, amabo te, quid tandem libri fuit, cuius Authorem* Fo hi *praedicas? Figurae aenigmaticae quatuor et sexaginta, sive lineolae 384 partim continuae partim interruptae et praeterea nihil. Bene habet. At si aenigmaticae, ergo perobscurae; si tam obscurae, ergo aedipo fuit opus qui lucem afferret.* […] *Aedipi fuerunt magnus ille Princeps et quasi conditor* Cheu *Familiae tertiae Imperialis* Ven vam *dictus, nec non eiusdem filius* Cheu cum*. Hi solverunt aenigma, et figuras interpretati sunt.*
> "It is entitled the *Yijing*, or the *Book of Changes*, written by Fuxi, the Founder of the Chinese race. You may well ask: 'But, do tell, what is in this book whose author you claim to be Fuxi? It is nothing more than sixty-four mysterious figures, or 384 lines, some being continuous, others broken'. That is all very well. But these figures are mysterious to the point of being utterly obscure, so obscure that it would take another Oedipus to elucidate them. […] Wen Wang, this great Prince who can almost be considered the founder of the Third Dynasty Zhou, and his son Zhou Gong, were the true Oedipuses, in that they solved the riddle and interpreted the figures."
> *PD*, p. LIX; transl. Meynard (2011), 163–164.

It is also interesting to observe that *Sinicus noster Epictetus* corresponds to *Confucius* at *PD*, pp. XIII–XIV.[26] In this case, a Greek name conveys a Chinese one, so as to emphasize analogies between two philosophical traditions which actually diverged from a chronological and geographical viewpoint. Furthermore, *Confucius* as *Sinicus Epictetus* is compared to Socrates, Plato, Seneca, and Plutarch, who are cited in the plural form (*et vero in Europa illa, ubi iam Socrates, et Platones, ubi Senecae, Plutarchi prope viluerunt, an speremus fieri posse ut plausum referat Sinicus noster Epictetus?*). This can be regarded not only as a simple literary embellishment, but also as a rhetorical strategy to emphasize the relevance of their spread across Europe.[27] By the same token, the famous Athenian legislator Solon is compared to the ancient legendary Chinese rulers Yao and Shun. For this reason he is mentioned in the plural form: *illi* […] *veri Solones gentis Sinicae* (*PD*, p. XV).

4. Annotating Personal Names in the *EP* and *PD*. Some Final Remarks

Although only a part of the *CSP* has been analyzed, the annotation and extraction of PPNs may be a useful instrument for more detailed analyses in Jesuit Latin texts. As for Graeco-Hellenistic and Roman personal names, to which particular attention has been paid, it is true that they are a minority compared to Eastern names. However, it is equally true that, by combining digital tools with traditional methods of analysis, it was possible to reflect on these names in a de-contextualized form and, at the same time, to analyze them further in specific passages, by creating interesting intersections of linguistic, literary, and thematic viewpoints. In other words, close and distant reading appeared complementary both to carry out a manual annotation of names and to analyze an amount of data from a comparative perspective in terms of referents and occurrences.[28]

26 On *Sinicus Epictetus* cf. Tommasi (2020), 80.

27 Consequently, for consistency with this passage, I also included the occurrence *Platones* (*PD*, p. LVIII: *Platones subinde aliquos audire te credas, aliosve Philosophos haudquaquam male sentientes de Deo*).

28 Cf. *supra*, §3.

In this sense, two main patterns of citation of these Graeco-Hellenistic and Roman personal names emerged as basically functional to:

1. compare Western and Eastern intellectual figures (Solon, Orpheus, Oedipus), at times by resorting to a 'catalogue', as is the case with *PD*, p. LXXVII, which relates to "the early knowledge and worship of God"[29]: *apud Graecos Socrates, Pythagoras, Plato, Epictetus; et apud Latinos, Varro, Tullius, Seneca aliique Philosophi de Deo multa recte senserunt, atque scripserunt*.

2. introduce a source, especially by reporting quotations or other authors' thought, as is the case with Cicero at *PD*, p. LXXXIX, where the phrase *quot enim, teste Tullio, hominum linguae, tot nomina Deorum* derived from Cicero's *De natura deorum* (1,84)[30] or, to give another example, with Lactantius.[31]

To conclude, among some future prospects, we would especially like to focus on collecting more data, by extending the analysis to the entire *CSP*, in order to understand fully how these digital techniques can contribute to building a bridge between Eastern and Western cultures in the same way as Latin used to be considered a vehicular language by Jesuit Fathers.

29 Meynard (2011), 183.

30 Cf. Della Calce (2024), 290–292.

31 Cf. e.g. *PD*, p. LVIII: *et quemadmodum Lactantius dicebat a nullo Ciceronem, quam ab ipso Cicerone vehementius posse refutari, ita nec hos novatores a nullis certius, quam a seipsis refutari posse*; *PD*, p. LXXXIII: *et de illis quidem aliarum gentium superstitionibus, deque ipsarum prisca religione constat ex primis litterarum cuiusque gentis monumentis ex Orpheo, inquam, et primis Poëtarum Hesiodo, Homero, et c. qui et ipsi (uti idem Lactantius ait) multo ante natum Philosophiae nomen fuerunt, et habiti sunt sapientes, et tamen tam inepta de Deo Deorumque generationibus figmenta et fabulas protulerunt*.

References

Balbo (2020): A. Balbo, Classics, Latin and Greek Authors in the Proemialis Declaratio of Confucius Sinarum Philosophus (1687), Itineraria 19 (2020), 153–172.

Balbo (2022): A. Balbo, The Epistula praefatoria of the Confucius Sinarum Philosophus: A Rhetorical Analysis in Search of Cicero and Seneca, in: A. Balbo / J. Ahn / K. Kim (eds.), Eastern and Western Civilizations. Searching for a Respublica Romanosinica, Berlin / Boston 2022, 111–130.

Beersmans et al. (2023): M. Beersmans / E. de Graaf / T. Van de Cruys / M. Fantoli, Training and Evaluation of Named Entity Recognition Models for Classical Latin, in: A. Anderson / S. Gordin / B. Li / Y. Liu / M. C. Passarotti (eds.), Proceedings of the Ancient Language Processing Workshop, Shoumen 2023, 1–12.

Berti (2019): M. Berti, Named Entity Annotation for Ancient Greek with INCEpTION, in: K. Simov / M. Eskevich (eds.), CLARIN Annual Conference Proceedings, Leipzig 2019, 1–4.

Berti (2023): M. Berti, L'antichità greco-romana e le tecnologie digitali, in: F. Ciotti (ed.), Digital Humanities. Metodi, strumenti, saperi, Roma 2023, 312–324.

Caboara (2022): M. Caboara, Regnum Chinae: The Printed Western Maps of China to 1735, Leiden, 2022.

Catto (2014): M. Catto, Il mito gesuitico della Cina, in: F. D'Arelli, Matteo Ricci. L'altro e diverso mondo della Cina, Milano 2014, 9–27.

Della Calce (2024): E. Della Calce, Cicerone nel Confucius Sinarum Philosophus (1687). Discussione di alcuni casi emblematici tra filosofia, religione e morale, Ciceroniana online 8/1 (2024), 277–308.

Della Calce / Mollea (2023): E. Della Calce / S. Mollea, Seneca in Jesuit Thought: the Case of Antonio Possevino and His Secondary Sources (Muret, Ortino, Perera, Erasmus, and Others), Latinitas Series Nova 11 (2023), 45–69.

Del Rio Riande / Vitale (2020): G. del Rio Riande / V. Vitale, Recogito-in-a-Box: From Annotation to Digital Edition, Modern Languages Open (2020), 1–13.

Dew (2009): N. Dew, Orientalism in Louis XIV's France, Oxford / New York 2009.

Ehrmann et al. (2021): M. Ehrmann / A. Hamdi / E. Linhares Pontes / M. Romanello / A. Doucet, Named Entity Recognition and Classification on Historical Documents: A Survey, arXiv:2109.11406 [cs.CL] – ACM Computing Surveys 56/2 (2023), 1–47.

Erdmann et al. (2016): A. Erdmann / C. Brown / B. Joseph / M. Janse / P. Ajaka / M. Elsner / M.-C. de Marneffe, Challenges and Solutions for Latin Named Entity Recognition, in: E. Hinrichs / M. Hinrichs / T. Trippel (eds.), Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), Osaka 2016, 85–93.

Erdmann et al. (2019): A. Erdmann / D. J. Wrisley / B. Allen / C. Brown / S. Cohen-Bodénès / M. Elsner / Y. Feng / B. Joseph / B. Joyeux-Prunel / M.-C. de Marneffe, Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities, in: J. Burstein / C. Doran / T. Solorio (eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis (MN.) 2019, 2223–2234.

Gregory (2021): I. Gregory, Barker, Elton, Leif Isaksen, Rebecca Kahn, Rainer Simon, Valeria Vitale, and the Pelagios Network, Project Creators. Recogito: Semantic Annotation without the Pointy Brackets. Other, Renaissance and Reformation 44/3 (2021), 243–249.

Meynard (2011): T. Meynard S. J. (ed.), Confucius Sinarum Philosophus (1687). The First Translation of the Confucian Classics, Roma 2011.

Montemagni (2023): S. Montemagni, Trattamento automatico del linguaggio e Digital Humanities: metodi e strumenti, sfide, in: F. Ciotti (ed.), Digital Humanities. Metodi, strumenti, saperi, Roma 2023, 160–177.

Moretti (2005): F. Moretti, La letteratura vista da lontano, Torino 2005.

Mungello (1985): D. Mungello, Curious Land: Jesuit Accommodation and the Origins of Sinology, Stuttgart 1985.

Simon et al. (2017): R. Simon / E. Barker / L. Isaksen / P. de Soto Cañamares, Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2, Journal of Map & Geography Libraries 13/1 (2017), 111–132.

Tommasi (2020): C. O. Tommasi, Epictetus in the Forbidden City. Accommodation and Resilience in Matteo Ricci's "Twenty-five Paragraphs", Itineraria 19 (2020), 73–104.

Ulrich et al. (2015): L. T. Ulrich / I. Gaskell / S. J. Schechner / S. A. Carter, with Photographs by Samantha S. B. van Gerbig, Tangible Things: Making History through Objects, New York 2015.

von Collani (1990): C. von Collani, Philippe Couplet's Missionary Attitude Towards the Chinese in Confucius Sinarum Philosophus, in: J. Heyndrickx, (ed.) Philippe Couplet, S. J. (1623–1693). The Man Who Brought China to Europe, Nettetal 1990, 37–54.

Yousef et al. (2023): T. Yousef / C. Palladino / G. Heyer / S. Jänicke, Named Entity Annotation Projection Applied to Classical Languages, in: S. Degaetano-Ortlieb / A. Kazantseva / N. Reiter / S. Szpakowicz (eds.), Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Dubrovnik 2023, 175–182.

Figure and Table References

Fig. 1: The search mask of *SERICA*'s website.

Tab. 1: Categories of Western PPNs.

Author Contact Information[32]

Prof. Andrea Balbo
Università di Torino
E-mail: andrea.balbo@unito.it

Dr. Elisa Della Calce
Università di Torino
E-mail: elisa.dellacalce@unito.it

32