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.

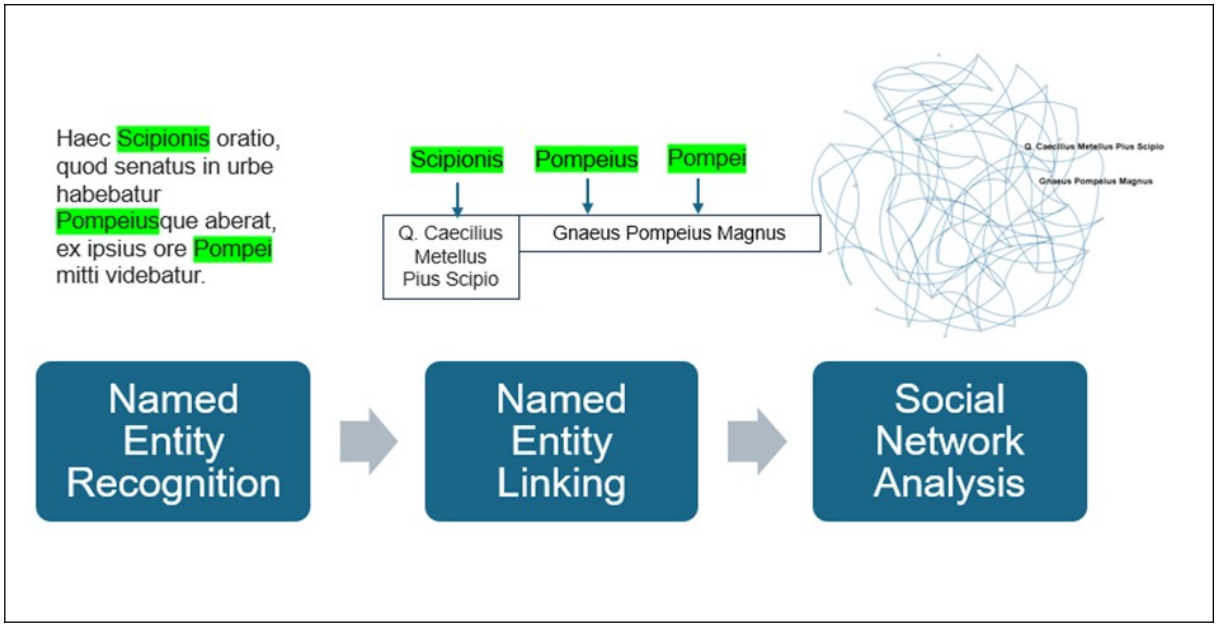

Fig. 1: Main steps of the *NIKAW* workflow.

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.

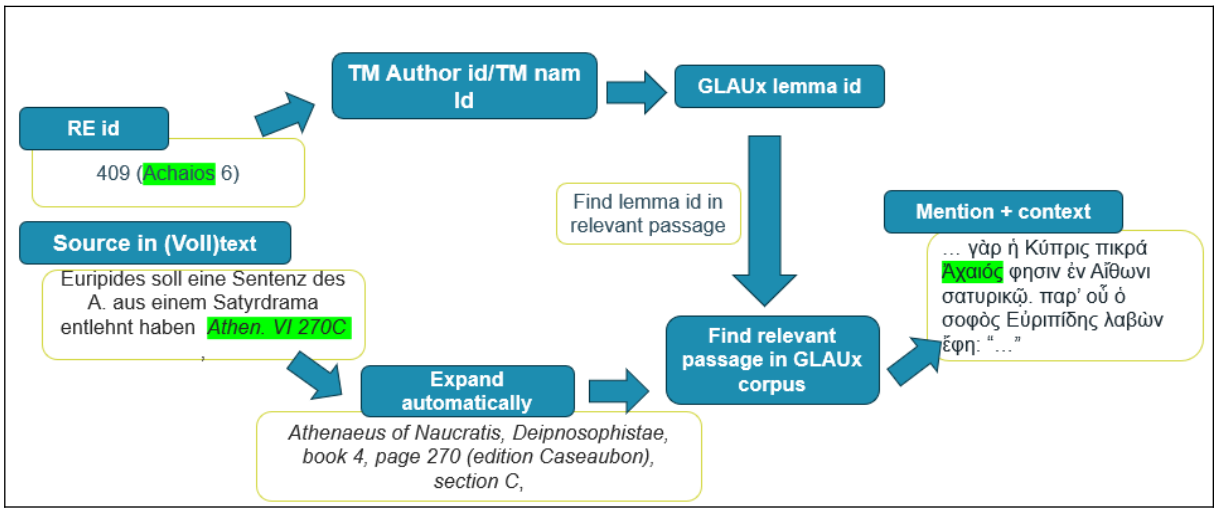

Fig. 3: Pipeline for the automatic creation of NEL training data.

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.

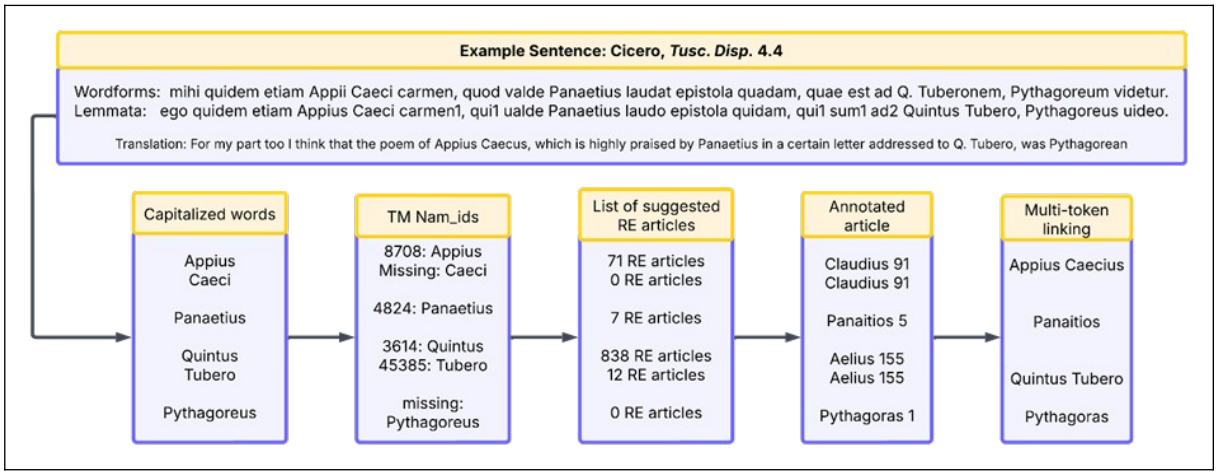

Fig. 4: Manual annotation process for the Plato case study.

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

Fig. 1: Main steps of the *NIKAW* workflow.

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.

Fig. 3: Pipeline for the automatic creation of NEL training data.

Fig. 4: Manual annotation process for the Plato case study.

Tab.1: Evaluation of the quality of the silver data.

Tab. 2: List of texts included in the Plato case-study.

Tab. 3: Results on ‘Plato case study data’.

Author Contact Information[64]

Margherita Fantoli
Assistant Professor
Faculty of Arts
KU Leuven
E-mail: margherita.fantoli@kuleuven.be

64