ENEIDE: A High Quality Silver Standard Dataset for Named Entity Recognition and Linking in Historical Italian

Cristian Santini[a,c], **Sebastian Barzaghi**[b], **Paolo Sernani**[a], **Emanuele Frontoni**[a], **Laura Melosi**[a], **Mehwish Alam**[c]

[a]University of Macerata, Macerata, Italy
[b]University of Bologna, Bologna, Italy
[c]Telecom Paris, Polytechnic Institute of Paris, Palaiseau, France

sebastian.barzaghi2@unibo.it, mehwish.alam@telecom-paris.fr
{c.santini12, paolo.sernani, emanuele.frontoni, laura.melosi}@unimc.it

Abstract

This paper introduces ENEIDE (Extracting Named Entities from Italian Digital Editions), a silver standard dataset for Named Entity Recognition and Linking (NERL) in historical Italian texts. The corpus comprises 2,111 documents with over 8,000 entity annotations semi-automatically extracted from two scholarly digital editions: Digital Zibaldone, the philosophical diary of the Italian poet Giacomo Leopardi (1798–1837), and Aldo Moro Digitale, the complete works of the Italian politician Aldo Moro (1916–1978). Annotations cover multiple entity types (person, location, organization, literary work) linked to Wikidata identifiers, including NIL entities that cannot be mapped to the knowledge graph. To the best of our knowledge, ENEIDE represents the first multi-domain, publicly available NERL dataset for historical Italian with training, development, and test splits. We present a methodology for semi-automatic annotations extraction from manually curated scholarly digital editions, including quality control and annotation enhancement procedures. Baseline experiments using state-of-the-art models demonstrate the dataset's challenge for NERL and the gap between zero-shot approaches and fine-tuned models. The dataset's diachronic coverage spanning two centuries makes it particularly suitable for temporal entity disambiguation and cross-domain evaluation. ENEIDE is released under a CC BY-NC-SA 4.0 license.

Keywords: named entity recognition, entity linking, digital humanities, historical Italian, scholarly digital editions, Wikidata

1. Introduction

While the majority of Natural Language Processing (NLP) systems are trained and evaluated on contemporary born-digital texts, historical documents (including letters, printed books, and manuscripts) remain fundamental sources for scholars seeking to validate and enrich our understanding of historical events and cultural heritage. Despite their importance in the Digital Humanities field (DH), State-of-The-Art (SoTA) systems for Named Entity Recognition (NER) and Entity Linking (EL) (together NERL) demonstrate significant performance degradation (Ehrmann et al., 2023) when applied to historical documents. This degradation stems from multiple factors: the linguistic complexity and variation of historical texts, the presence of noise introduced by Optical Character Recognition (OCR), and the scarce representation of historical language varieties in modern web-crawled corpora used to train contemporary language models.

The development of robust historical NERL systems, which enable the detection of textual references to entities (such as persons, locations, and organizations) and their disambiguation through Knowledge Graphs (KGs) like Wikidata (Vrandečić and Krötzsch, 2014), critically depends on the availability of annotated historical corpora for training and evaluation. However, the creation of such resources faces substantial obstacles. First, the field lacks widely adopted community standards for annotation schemas and guidelines, limiting interoperability and reusability across projects (Ehrmann et al., 2023). Second, the time and cost required for expert human annotation create a significant bottleneck, leaving many historical language varieties severely under-resourced. Historical Italian exemplifies this challenge: currently, only one dataset, i.e., MHERCL (Graciotti et al., 2025), provides NERL annotations for historical Italian texts. However, this resource is domain-specific (music periodicals from the 20th century), contains only a test split, and therefore offers limited utility for training purposes.

In this landscape, Scholarly Digital Editions (SDEs) (Sahle, 2016) emerge as a promising yet underutilized resource. SDEs are digital resources that provide manually curated annotations of named entities in TEI/XML format, created by

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7126–7136
11–16 May 2026.

domain experts. They play a pivotal role in broadening the scope of NERL research in historical texts and enhancing the scalability of NLP applications. These resources contain manual annotations made by domain-experts for people, places, organizations, bibliographic resources, and other entities referenced in texts from different historical periods, typically disambiguated using Wikidata or other domain-agnostic authority files. Reusing such datasets can improve and verify the application of historical NERL systems due to the validity of the interpretative work of philologists and expert scholars (Valette, 2024). However, the systematic extraction and transformation of SDE annotations into reusable datasets remains an open methodological challenge.

This paper addresses this gap by presenting **ENEIDE** (*Extracting Named Entities from Italian Digital Editions*), a silver standard dataset for NERL in historical Italian, semi-automatically extracted from two Italian SDEs: DigitalZibaldone[1] (Stoyanova and Johnston, 2014), based on the philosophical diary of the Italian poet Giacomo Leopardi (1798–1837), and Aldo Moro Digitale[2] (Barzaghi et al., 2025), containing the works of the Italian politician Aldo Moro (1916–1978).

We refer to ENEIDE as a silver standard dataset to distinguish it from gold standard resources (Wissler et al., 2014), where all annotations are manually created and validated by expert annotators from scratch. Silver standard datasets are instead produced through semi-automatic methods (Rebholz-Schuhmann et al., 2010) — such as the extraction and transformation of pre-existing annotations, automated labeling, or distant supervision — followed by partial human validation. While silver standard resources may contain residual noise compared to fully manual gold annotations, they offer a scalable and cost-effective alternative for producing training data in low-resource scenarios, where gold standard annotation would be prohibitively expensive or time-consuming.

ENEIDE comprises 2,111 documents spanning two centuries and two distinct domains, with over 8,000 entity annotations across multiple types (person, location, organization, literary work) linked to Wikidata identifiers. To the best of our knowledge, ENEIDE represents the first multi-domain, publicly available dataset for NERL in historical Italian with training, development, and test splits.

This work offers three key contributions:

1. A discussion of the methodology for semi-automatic dataset extraction from SDEs, detailing sampling strategies, quality control procedures, and annotation enhancement techniques;
2. The description of ENEIDE, a multi-domain, diachronic corpus covering Italian philosophical, literary, political, and legal texts from the 19th and 20th centuries semi-automatically extracted from SDEs;
3. An empirical analysis of dataset quality and characteristics, including average token counts, entity distribution statistics, temporal coverage, and NIL entity analysis.

The remainder of this paper is structured as follows. Section 2 surveys related work on historical NERL datasets. Section 3 details our semi-automatic extraction pipeline and quality control procedures, while also offering dataset statistics. Section 4 provides baseline experiments and analysis demonstrating the dataset's utility. Section 5 discusses the dataset's reuse potential across different research scenarios. Section 6 summarizes our contributions and outlines future directions. Finally, Section 7 addresses current limitations and potential solutions for dataset improvement.

The dataset is released under a CC BY-NC-SA 4.0 license on Zenodo (Santini et al., 2025).[3]

2. Related Work

The challenges of adapting NERL approaches to documents from cultural heritage institutions have led to the creation of several annotated historical corpora. This section discusses the most relevant datasets containing literary and political texts, with particular attention to resources for Italian. A comprehensive survey of historical NERL datasets is available in (Ehrmann et al., 2023).

Literary and Multilingual Corpora A seminal work in NER for literary texts was presented by Bamman et al. (2019). Their dataset, *LitBank*, serves as a training and evaluation resource for multiple information extraction tasks: NER, coreference resolution, and event detection. The dataset comprises text samples from 100 English novels from Project Gutenberg. The annotations follow the ACE guidelines (Walker et al., 2006), including six coarse-grained entity categories (person, location, geo-political entity, facility, organization, and vehicle) and covering both common nouns and nested entities.

More recently, Romanello and Najem-Meyer (2024) presented a manually annotated dataset of named entity references in classical commentaries of Sophocles' *Ajax*. The dataset, called *AJMC*, was included in the HIPE-2022 shared task (Ehrmann

[1] digitalzibaldone.net
[2] aldomorodigitale.unibo.it
[3] doi.org/10.5281/zenodo.17407356

et al., 2022) and supports training and evaluation of systems for detecting general entities (persons, locations, and dates) and domain-specific entities (primary and secondary literary sources), along with their disambiguation using Wikidata. Entities are annotated according to 6 coarse-grained and 9 fine-grained classes, include nested entities at 1-level depth, and cover OCR-ed texts in three languages: English, French, and German. Similar to ENEIDE, entities are often referenced through abbreviations, such as "Cic." for "Cicero", a common practice in philological texts like classical commentaries and literary works such as Leopardi's Zibaldone.

Regarding literary entities specifically, one of the earliest datasets focusing on references to books, monographs, and essays is *LinkedBooks* (Colavizza and Romanello, 2017). Extracted from documents related to Venetian historiography in multiple languages (including Italian), this dataset contains numerous annotated citations from reference lists and footnotes, including abbreviated primary and secondary sources. While this resource was designed to train NER models capable of extracting and parsing literary work references, it does not include general entity types (persons, locations, organizations) and lacks Wikidata identifiers.

Italian Historical Corpora For Italian specifically, a pivotal contribution was made by Paccosi and Palmero Aprosio (2022), who describe *KIND*, a multi-domain NER dataset extracted from various text types, including news, literary texts, and political works. Like ENEIDE, this dataset includes excerpts from the digital edition of Aldo Moro's works. The resource contains semi-automatic annotations using three coarse-grained entity types: person, location, and organization. However, the annotations do not consider nested entities, and the dataset does not support entity disambiguation through Wikidata.

Most recently, Graciotti et al. (2025) introduced MHERCL, the first dataset providing NERL annotations for historical Italian and English texts. This resource is extracted from a domain-specific corpus of 19th-century music periodicals. Despite its rich tagset and substantial annotations of long-tail (i.e., unpopular) entities, MHERCL contains only a test split, limiting its utility for training purposes.

Positioning ENEIDE ENEIDE differs from these existing resources in several key aspects: (1) it is the first multi-domain corpus for NERL in historical Italian with training, development, and test splits; (2) it combines general entity types (person, location, organization) with domain-specific ones (literary works); and (3) it covers two distinct historical periods (19th and 20th centuries) and textual genres (philosophical-literary and political-legal), providing temporal broadness uncommon in existing Italian resources.

3. ENEIDE Dataset

The ENEIDE dataset spans multiple time periods and domains. It was semi-automatically extracted from two SDEs: Digital Zibaldone (DZ) and Aldo Moro Digitale (AMD). These sources cover two centuries (19th and 20th) and include various text types, making them suitable for testing EL systems on humanistic documents with complex contextual and historical features.

DZ is the digital edition of Giacomo Leopardi's *Zibaldone di pensieri*, containing over 4,500 pages of reflections on literature, history, and philosophy written between 1817 and 1832. The hypertextual structure of this work, composed of cross-references (internal links between notes) as well as external ones (bibliographic references, quotes, named entities, etc.) motivated digital scholars to re-mediate this literary work into a digital research platform which would enable users to dynamically mine the text's structural complexity through XML/TEI (Stoyanova, 2023).

The digital edition, collaboratively annotated by domain experts at Princeton University, serves as a valuable tool for assessing entity disambiguation methods for two main reasons. First, it includes thousands of expert-annotated references to people (`PER`), locations (`LOC`), and bibliographic works (`WORK`), linked when possible to Wikidata or VIAF. Second, Leopardi's notes offer a pre-digital example of an encyclopedic hypertext, comprising thousands of external references to historical figures, literary works, authors, and various facets of humanistic knowledge (Stoyanova, 2013).

The structure of this resource resembles that of the original work, with every fragment available in a single web page as an HTML document, ordered by its page and identified by a unique URI. For instance, accessing the URI `https://digitalzibaldone.net/node/p2721_1` allows users to view the first note on page 2721 of the Zibaldone. References to people, places, and bibliographic works are formatted as hyperlinks directing users to a unique identifier. The unique identifiers are based mostly on Wikidata or VIAF if a corresponding Wikidata entry does not exist. The content of the platform is publicly available under a CC-BY-NC-SA. For this project we used as source the HTML documents available inside the digital platform. A sample of an HTML file contained in the digital edition is available in Figure 1.

AMD presents the complete works of Aldo Moro, collecting political, legal, and journalistic texts from

```
<p xmlns="http://www.w3.org/1999/xhtml" xmlns:tei="http://www.tei-c.org/ns/1.0">
<div class="node" id="p2721_1"><div class="nodemeta" id="p2721_1_meta"></div><b class="para_num">[2721,1]</b>NBSP Anche il <a
 class="person" href="https://digitalzibaldone.net/node/Q518160">Gelli</a> confessava (ap. <a class="person"
 href="https://digitalzibaldone.net/node/Q3769747">Perticari</a>
<a href="/node/viaf34613848">Degli Scritt. del Trecento</a> l. 2. c. 13. p. 183.) che la lingua toscana
non era stata applicata alle scienze. (24. Maggio 1823.).</div></p>
```

Figure 1: Example of markup with entities annotated for a note in Digital Zibaldone.

the 1932 to 1978. Encoded in RDFa (Adida et al., 2007) with semantic web annotations, AMD links three entity types to Wikidata: person (PER), location (LOC), and organization (ORG). Figure 2 illustrates an example of how RDFa is used to inject semantic information in HTML elements representing mentioned entities in the text.

Both corpora present notable challenges for NERL systems: in Aldo Moro, entities may be referenced indirectly (e.g., "Pope" for "Pope Pius XII"), while in Leopardi's texts complex abbreviations (e.g., "Il." for "Iliad") and metonymic references are abundant. Moreover, some entities do not exist in Wikidata. This last category requires NIL prediction: the ability to identify entities that cannot be mapped to any KG entry.

3.1. Annotation Extraction and Sampling Strategy

We extracted entities using `Beautiful Soup`[4], a Python library for parsing HTML and XML documents, identifying annotated entities through HTML hyperlinks. For AMD, we first leveraged its public API[5] to download all works in HTML format. Then, annotations were extracted from `span` elements, classified into `PER`, `LOC` or `ORG` based on the `class` attribute. These elements were linked to potential Wikidata entities via the `owl:sameAs` property found in the `meta` elements of the HTML `head`.

In the case of DZ, entity annotations were located within `link` elements, categorized into `PER`, `LOC` or `WORK` based on regular markup patterns. For both editions, if a mention was not linked to Wikidata it was tagged as NIL. Given the ongoing development of both digital editions, the structure of the annotations is subject to change, necessitating adaptable pre-processing strategies in future research.

Each SDE required a different sampling approach due to their distinct structures. From DZ, we selected two sections of the Zibaldone: pages 1000–2001 and pages 2700–4000, written in 1821 and 1823 respectively, corresponding to the years of highest productivity for the poet. For AMD, we sampled the first paragraph of each document, an approach chosen because AMD documents vary significantly in length (from short letters to lengthy political speeches), and first paragraphs typically introduce the main topics and key entities.

To ensure quality and balance, we excluded texts whose length deviated significantly from the standard distribution (beyond 2 standard deviations) and retained only documents containing at least one entity. This process produced 1,050 items from DZ and 1,061 from AMD. We then created train, validation, and test splits using a 70/15/15 ratio with stratified sampling based on creation year to maintain similar chronological distributions across all splits. The source code used for extracting this dataset and the documentation for using ENEIDE is available on Github.[6]

3.2. Quality Control

Domain experts evaluated annotation quality using 100 randomly sampled documents from each dataset. The evaluation involved four expert scholars divided into two teams: one team with expertise in Italian literature (two PhD holders) and another in history (two PhD holders), coordinated by one of the authors to solve ambiguous cases. The evaluation followed these guidelines:

- Only annotate entities of type person, location, organization, and literary work that are referenced by proper names (i.e., exclude common nouns);
- Do not annotate nested entities; always annotate the longest surface form (e.g., *[Government of Italy]* rather than *[Government]* of *[Italy]*);
- Select the Wikidata identifier that best matches the entity in its historical and textual context;
- Consider figures of speech such as metonymy (e.g., literary works referred to by their creator's name).

Table 1 presents the results. DZ annotations achieved high quality with an F1 score of 95.6. AMD scored lower (79.8 F1), primarily due to missing annotations, as reflected in its recall of 70.6.

[4] pypi.org/project/beautifulsoup4
[5] aldomorodigitale.unibo.it/api/works
[6] github.com/sntcristian/ENEIDE

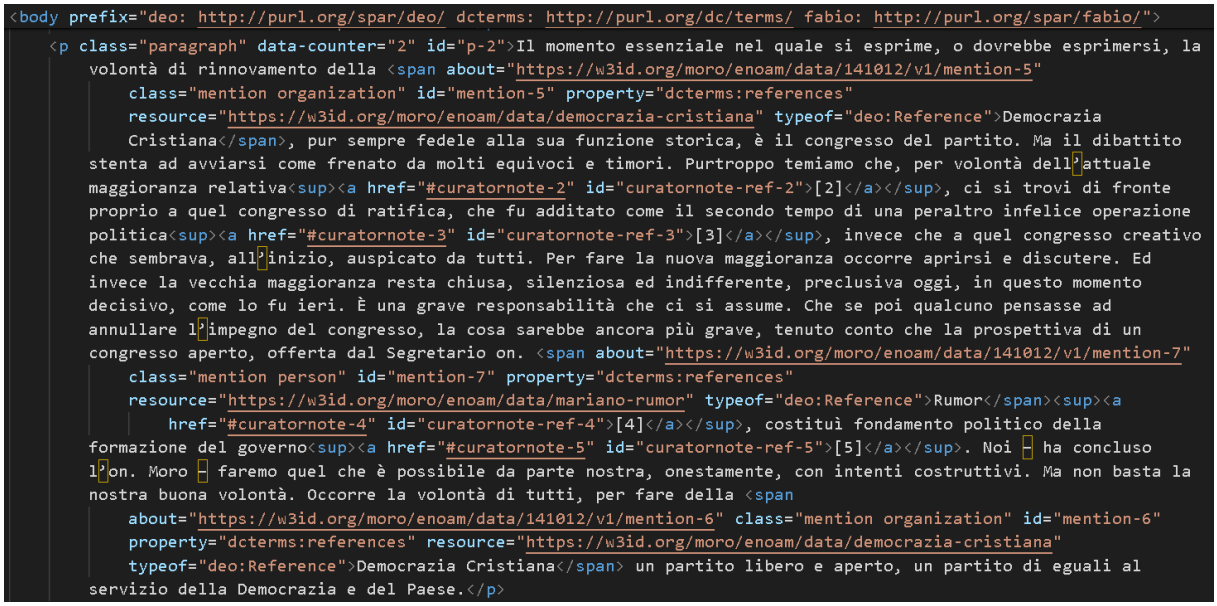

Figure 2: Example of entities annotated with HTML and RDFa in a document of the Edition of Aldo Moro's works.

Statistic	DZ	AMD
Annotations in samples	403	215
Wrong annotations	13	18
Missing annotations	10	64
Precision	96.8	91.6
Recall	94.4	70.6
F1	95.6	79.8

Table 1: Quality assessment statistics for DZ and AMD samples.

To improve AMD's annotation coverage, we developed a semi-automatic enhancement pipeline comprising four steps: (i) extracting frequent entity mentions with Wikidata IDs and validating them with domain experts; (ii) automatically annotating previously missed mentions of validated entities throughout the text using exact string matching; (iii) applying the Italian StanzaNLP NER model[7] to identify remaining unannotated entities; and (iv) conducting final expert validation of automatically added annotations. This approach increased annotation completeness while maintaining annotation quality through expert oversight at each stage.

3.3. Dataset Statistics

Tables 2 and 3 summarize key statistics for both corpora. Table 2 shows document counts and number of annotations, while Table 3 provides a fine-grained breakdown by entity type, reporting the number of NIL entities and unique Wikidata identifiers in each split. NIL entities constitute a marginal fraction of the total annotations: 7% in DZ and 2% in AMD. A notable difference between the two corpora is the entity overlap between training and test sets: 93.19% of unique entities in the DZ train and dev splits also appear in the test set, whereas AMD exhibits lower overlap (75.38%). This difference reflects the distinct natures of the two sources: DZ's philosophical reflections repeatedly reference a core set of classical authors and works, while AMD's political writings engage with a more diverse and time-specific set of contemporary figures and organizations. In terms of scale, ENEIDE is comparable to other medium-sized corpora for historical NERL, such as AJMC (Romanello and Najem-Meyer, 2024) and TopRes19th (Ardanuy et al., 2022).

4. Experiments

To establish baseline performance on ENEIDE and demonstrate its utility for evaluating NERL systems, we conducted experiments using SoTA models for both NER and EL tasks. All models were chosen for being open-source and versioned, thus allowing better reproducibility of our results.

4.1. Named Entity Recognition

For NER, we evaluated four architectures. First, we considered three instruction-tuned Large Language Models (LLMs) using zero-shot prompting: LLaMA 3.1-8B[8], a general-purpose multilingual

[7] stanfordnlp.github.io/stanza/ner_models.html

[8] meta-llama/Llama-3.1-8B-Instruct

	DZ	AMD
Documents (train)	735	743
Documents (dev)	157	159
Documents (test)	158	160
Tokens per Doc (train)	246.6	125.9
Tokens per Doc (dev)	253.2	119.3
Tokens per Doc (test)	248.7	119
Annotations (train)	2,935	2,766
Annotations (dev)	727	604
Annotations (test)	617	657
% Identifiers Overlap (train+dev vs test)	93.19	75.38

Table 2: Number of documents, average tokens, total annotations, and Wikidata entity overlap percentages in ENEIDE.

	DZ					AMD				
Split	**PER**	**LOC**	**WORK**	**NIL**	**IDs**	**PER**	**LOC**	**ORG**	**NIL**	**IDs**
train	1,661	488	786	182	623	759	940	1,067	64	583
dev	375	149	203	74	276	158	226	206	13	203
test	318	130	169	42	241	194	190	205	9	238

Table 3: Annotation counts by entity type (PER=person, LOC=location, WORK=literary work, ORG=organization), NIL entities, and unique Wikidata identifiers (IDs) in each split.

instruction-tuned model; Minerva-7B[9], an Italian-focused instruction-tuned model; and Ministral-8B[10], a multilingual instruction-tuned model from the Mistral family. Each model was prompted to identify and classify named entities according to the entity types present in each dataset (person, location, organization, and literary work) and to return them in a list formatted as JSON. A detailed overview of the prompts used is given in Appendix A.

In addition, to compare LLMs used in a zero-shot setting with a smaller model trained on ENEIDE, we decided to fine-tune a GLiNER model (Zaratiana et al., 2023) already pre-trained for universal NER on Italian[11] using the full ENEIDE training set (DZ and AMD).

Table 4 presents the results. GliNER (fine-tuned) achieved the best performance across both datasets, with F1 scores of 0.782 on DZ and 0.876 on AMD. The slightly lower performance on DZ reflects the challenges posed by references to `WORK` entities, where the model achieved an F1 score of 0.592 using a *strict match* criterion and 0.724 using *relaxed match* (at least one token overlap).

All the LLMs, despite having a larger number of parameters, exhibited very low precision and recall, suggesting difficulties in correctly interpreting instructions for NER when prompted with domain-specific and historical data. Overall, the moderate F1 scores across all zero-shot models highlight the difficulty of historical NER and the need for specialized approaches.

Dataset	Model	Precision	Recall	F1
DZ	LLaMA 3.1-8B	0.438	0.376	0.405
	Minerva-7B	0.531	0.028	0.052
	Ministral-8B	0.351	0.329	0.340
	GLiNER (fine-tuned)	**0.851**	**0.723**	**0.782**
AMD	LLaMA 3.1-8B	0.719	0.603	0.656
	Minerva-7B	0.659	0.138	0.228
	Ministral-8B	0.678	0.530	0.595
	GLiNER (fine-tuned)	**0.887**	**0.866**	**0.876**

Table 4: NER results on DZ and AMD test sets using zero-shot instruction-tuned LLMs.

4.2. Entity Linking

For EL, we evaluated three publicly available general-purpose models trained on Wikipedia: BLINK-ita[12] (Pozzi et al., 2023), an Italian adaptation of the BLINK model (Wu et al., 2020); mGENRE[13] (De Cao et al., 2022), a multilingual generative entity retrieval model; and BELA[14] (Plekhanov et al., 2023), a multilingual bi-

[9] sapienzanlp/Minerva-7B-instruct-v1.0
[10] mistralai/Ministral-8B-Instruct-2410
[11] huggingface.co/DeepMount00/universal_ner_ita
[12] github.com/rpo19/pozzi_aixia_2023
[13] github.com/facebookresearch/GENRE
[14] github.com/facebookresearch/BELA

encoder for EL. We evaluated all models in the entity disambiguation task, i.e. we used ground-truth entity spans from the test sets and evaluated the models' ability to correctly link them to Wikidata identifiers.

Table 5 presents the results. BELA achieved the best performance on DZ (0.598 accuracy), while mGENRE performed best on AMD (0.689 accuracy). The higher accuracy scores on AMD compared to DZ can be attributed to two factors: first, AMD entities are predominantly contemporary political figures and organizations that are well-represented in Wikipedia training data; second, DZ contains numerous classical and literary references that may be underrepresented or absent in standard Wikipedia-based training corpora. These results demonstrate that while general-purpose EL models achieve reasonable performance on recent texts containing factual information, they struggle with specialized literary and historical content as in DZ.

Dataset	Model	Accuracy
DZ	mGENRE	0.579
	BLINK-ita	0.502
	BELA	**0.598**
AMD	mGENRE	**0.689**
	BLINK-ita	0.643
	BELA	0.676

Table 5: Entity disambiguation results on DZ and AMD test sets using EL systems trained on Wikipedia.

5. Discussion

ENEIDE aims to be a valuable resource for NLP researchers in the Italian DH community. Since every text is annotated with temporal information related to its creation date, this dataset can be used effectively for training and evaluating algorithms for diachronic EL (Agarwal et al., 2018). Figures 3 and 4 show the temporal distribution of entities in both dataset partitions, highlighting their diachronic coverage. This distribution was obtained by querying Wikidata for the earliest recorded date of each entity in the corpus and plotting a frequency distribution of annotations by year.

The DZ partition (Figure 3) is particularly valuable for scholars focused on recognizing and linking literary entities and cultural references. By providing numerous annotations of classical works and literary authors across many historical periods (spanning from ancient Greece and Rome through the Renaissance to the early modern period) DZ serves as a challenging benchmark for NLP systems tailored to the humanistic domain. The temporal distribution reveals a concentration of entities from classical antiquity and the early modern period, reflecting Leopardi's philosophical interests and the foundational texts that influenced his thinking.

The AMD partition (Figure 4) enables improvement and verification of systems designed for detecting and disambiguating entities in the political domain. This partition contains rich annotations of historical figures and political organizations that shaped the 20th-century Italian history, with entity distributions concentrated in the mid-20th century, corresponding to Moro's active political career. The resource can therefore support historical studies that leverage EL systems to analyze the evolution of social networks and their relationship with political and historical dynamics (Nabiafjadi et al., 2021).

Beyond these domain-specific applications, ENEIDE's multi-domain nature makes it suitable for evaluating cross-domain generalization capabilities of NERL systems. The baseline experiments presented in Section 4 demonstrate that current SoTA models struggle with historical Italian texts, particularly in the literary domain, suggesting substantial room for improvement through specialized training or domain adaptation techniques. The dataset's inclusion of NIL entities further enables research on entity discovery and KG completion (Zhu et al., 2023), as these non-linkable entities represent gaps in existing knowledge resources that could be addressed through automated methods.

6. Conclusion

This paper presented ENEIDE, the first multi-domain, publicly available dataset for NERL in historical Italian with training, development, and test splits. Comprising 2,111 documents with over 8,000 entity annotations spanning two centuries and two distinct domains, ENEIDE addresses a critical gap in language resources for historical Italian NLP.

We presented a methodology for semi-automatic dataset extraction from SDEs, including quality control procedures and annotation enhancement techniques that can be potentially reproduced to other digital editions, given a certain degree of structural similarity with the ones that we treated in this work. Our baseline experiments demonstrate that current SoTA NERL systems struggle with historical Italian texts, particularly in the literary domain, with F1 scores ranging from 0.340 to 0.876 for NER and accuracy scores from 0.502 to 0.689 for EL. These results highlight both the challenge posed by historical texts and the value of ENEIDE as a benchmark for developing specialized NERL systems.

The dataset's diachronic coverage, spanning from classical antiquity to 20th-century Italy, makes

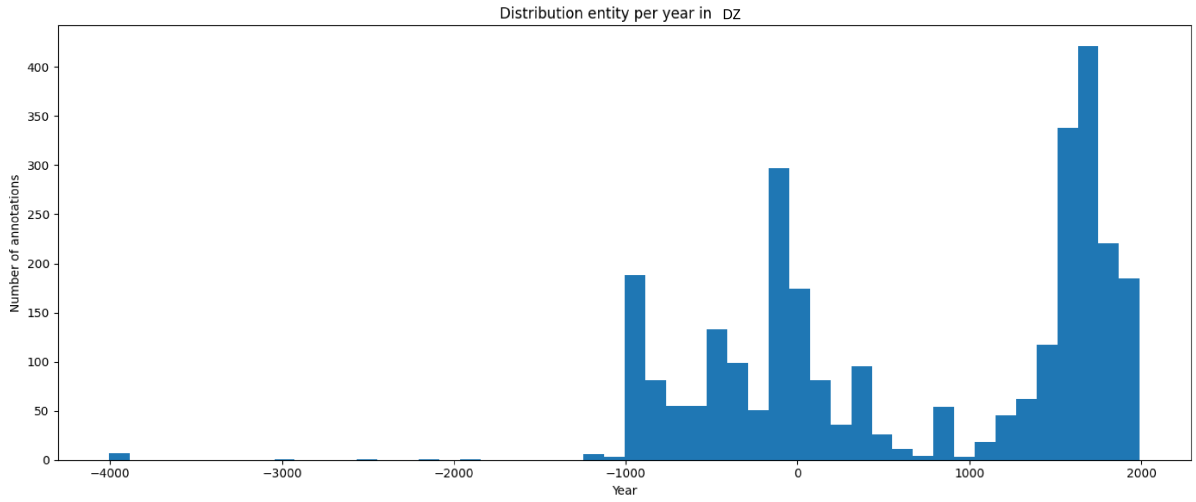

Figure 3: Temporal distribution of named entities in the DZ corpus, showing references spanning from classical antiquity to the early modern period.

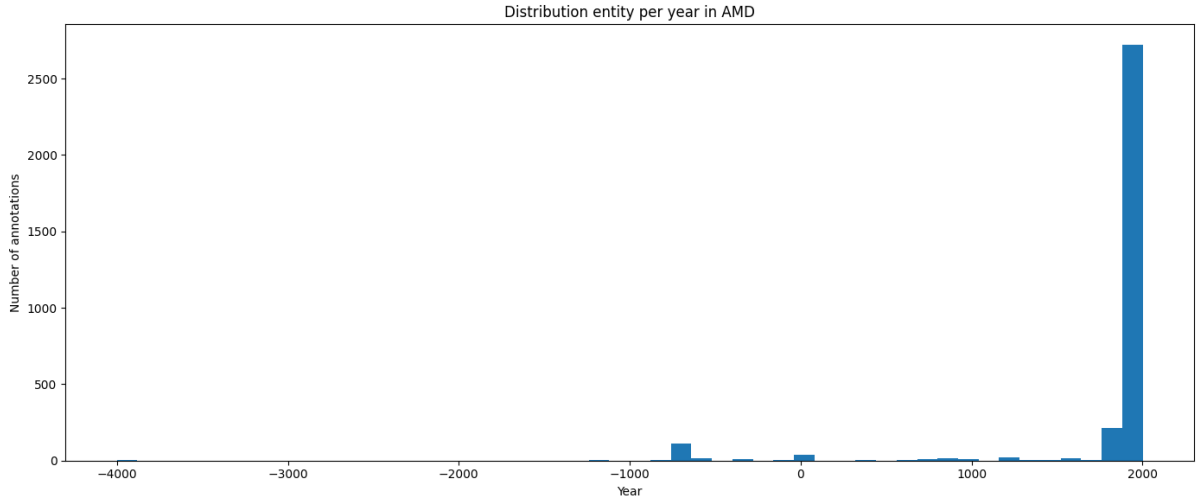

Figure 4: Temporal distribution of named entities in the AMD corpus, showing references concentrated in the 20th-century Italian history.

it particularly suitable for research on temporal entity disambiguation and cross-domain generalization. The inclusion of domain-specific entity types (literary works and organizations) alongside general types (persons and locations) further broadens its applicability to specialized NLP tasks in DH.

We release ENEIDE under a CC BY-NC-SA 4.0 license to encourage its adoption by the research community and hope it will stimulate further work on historical Italian NLP and the reuse of annotated resources in the DH as training and evaluation data for language technologies.

7. Limitations

Despite its contributions, ENEIDE has several limitations that should be considered when using this resource. First, due to its limited representativeness of historical Italian varieties, automatic systems trained exclusively on this resource may learn domain-specific biases if not carefully designed and tested for generalizability. The dataset covers only two authors and two specific domains (philosophical-literary and political-legal), which may not reflect the full linguistic and stylistic diversity of historical Italian texts.

Second, the annotation enhancement pipeline applied to AMD, while improving coverage, may have introduced some inconsistencies compared to the original expert-curated DZ annotations. Although the enhancement process included expert validation steps, the semi-automatic nature of the additions may result in subtle quality differences between the two partitions.

Finally, the dataset currently focuses on Italian

texts with limited multilingual coverage (some excerpts in French, Latin, and Greek appear in DZ). This limits its applicability for multilingual or cross-lingual NERL research.

Nonetheless, ENEIDE is an actively maintained resource, and future work may address these limitations by extending the dataset with additional authors, genres, and time periods to broaden its scope and enhance its reusability. We also plan to explore the integration of additional SDEs to increase domain diversity and temporal coverage.

Acknowledgments

We acknowledge the contribution of the development team of Aldo Moro Digitale (Barzaghi et al., 2025) and DigitalZibaldone (Stoyanova and Johnston, 2014) for providing the original HTML files from which ENEIDE was extracted. This publication is based upon work from COST Action CA24121 Knowledge Graphs in the Era of Large Language Models (KGELL), supported by COST (European Cooperation in Science and Technology, www.cost.eu)."

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Sonnet 4.5) to perform a proofreading of the initial draft. After using this tool, the authors reviewed and edited the content as needed to take full responsibility for the publication's content. No AI tool was used for designing, interpreting and analysing the research itself.

References

Ben Adida, Mark Birbeck, Shane McCarron, and Ivan Herman. 2007. Rdfa core 1.1.

Prabal Agarwal, Jannik Strötgen, Luciano Del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. dianed: Time-aware named entity disambiguation for diachronic corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 686–693.

Mariona Coll Ardanuy, David Beavan, Kaspar Beelen, Kasra Hosseini, Jon Lawrence, Katherine McDonough, Federico Nanni, Daniel van Strien, and Daniel CS Wilson. 2022. A dataset for toponym resolution in nineteenth-century english newspapers. *Journal of Open Humanities Data*, 8.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. 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)*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Sebastian Barzaghi, Alessio Palmero Aprosio, Francesco Paolucci, Francesca Tomasi, Sara Tonelli, Marialaura Vignocchi, and Fabio Vitali. 2025. The semantic digital edition of aldo moro's writings: A workflow supporting data sharing and replicability. *J. Comput. Cult. Herit.*

Giovanni Colavizza and Matteo Romanello. 2017. Annotated References in the Historiography on Venice: 19th–21st centuries. *Journal of Open Humanities Data*, 3:2–2.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Overview of hipe-2022: named entity recognition and linking in multilingual historical documents. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 423–446. Springer.

Arianna Graciotti, Leonardo Piano, Nicolas Lazzari, Enrico Daga, Rocco Tripodi, Valentina Presutti, and Livio Pompianu. 2025. Ke-mhisto: Towards a multilingual historical knowledge extraction benchmark for addressing the long-tail problem. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria.

Samira Nabiafjadi, Maryam Sharifzadeh, and Mostafa Ahmadvand. 2021. Social network analysis for identifying actors engaged in water governance: An endorheic basin case in the middle east. *Journal of Environmental Management*, 288:112376.

Teresa Paccosi and Alessio Palmero Aprosio. 2022. KIND: an Italian multi-domain dataset for named

entity recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 501–507, Marseille, France. European Language Resources Association.

Mikhail Plekhanov, Nora Kassner, Kashyap Popat, Louis Martin, Simone Merello, Borislav Kozlovskii, Frédéric A Dreyer, and Nicola Cancedda. 2023. Multilingual end to end entity linking. *arXiv preprint arXiv:2306.08896*.

Riccardo Pozzi, Riccardo Rubini, Christian Bernasconi, and Matteo Palmonari. 2023. Named entity recognition and linking for entity extraction from italian civil judgements. In *International Conference of the Italian Association for Artificial Intelligence*, pages 187–201. Springer.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M. van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Katrin Tomanek, Elena Beisswanger, and Udo Hahn. 2010. The CALBC silver standard corpus for biomedical named entities — a study in harmonizing the contributions from four independent named entity taggers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Matteo Romanello and Sven Najem-Meyer. 2024. A Named Entity-Annotated Corpus of 19th Century Classical Commentaries. *Journal of Open Humanities Data*, 10(1).

Patrick Sahle. 2016. What is a scholarly digital edition? *Digital scholarly editing: Theories and practices*, 1:19–39. Publisher: Open Book Publishers Cambridge.

Silvia Stoyanova. 2013. Fragmentary narrative and the formation of pre-digital scholarly hypertextuality: G. Leopardi's Zibaldone and its hypertext rendition. In *Proceedings of the 3rd Narrative and Hypertext Workshop*, NHT '13, pages 1–6, New York, NY, USA. Association for Computing Machinery.

Silvia Stoyanova. 2023. Working with the Digital Edition of Giacomo Leopardi's Zibaldone. *magazén*, 4(3):13.

Silvia Stoyanova and Ben Johnston. 2014. Remediating Giacomo Leopardi's Zibaldone: Hypertextual Semantic Networks in the Scholarly Archive. In *Proceedings of the Third AIUCD Annual Conference on Humanities and Their Methods in the Digital Ecosystem*, AIUCD '14, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Mathieu Valette. 2024. What does perspectivism mean? an ethical and methodological countercriticism. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 111–115.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus. Artwork Size: 1572864 KB Pages: 1572864 KB.

Lars Wissler, Mohammed Almashraee, Dagmar Monett Díaz, and Adrian Paschke. 2014. The gold standard in corpus annotation. *IEEE GSC*, 21.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. Gliner: Generalist model for named entity recognition using bidirectional transformer. *arXiv preprint arXiv:2311.08526*.

Fangwei Zhu, Jifan Yu, Hailong Jin, Lei Hou, Juanzi Li, and Zhifang Sui. 2023. Learn to not link: Exploring nil prediction in entity linking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10846–10860.

Language Resource References

Santini, Cristian and Barzaghi, Sebastian and Sernani, Paolo. 2025. *ENEIDE: A High Quality Silver Standard Dataset for Named Entity Recognition and Linking in Historical Italian*. Zenodo. PID https://doi.org/10.5281/zenodo.17407356.

A. Prompts for NER with Instruction-Tuned LLM

This appendix details the prompts used to perform NER with instruction-tuned LLMs, as presented in Section 4.1. Each LLM was prompted to identify and classify named entities according to the entity types present in each dataset (person, location, organization, and literary work) and to return them in a list formatted as JSON.

A.1. Prompt for DZ

System Prompt: You are a philologist with expert knowledge of Italian literature. Your task is to annotate references to Persons, Locations, and Literary Works within historical texts. You generate structured responses in JSON format. Do not write Python code.

User Prompt: Extract the references to named entities of type “person”, “location”, and “work” within the input text, taken from the collection “Zibaldone di pensieri” by the poet and philologist Giacomo Leopardi (1817–1832). Extract the entities in the response using a JSON format as in the provided example.

Example Input: “La Divina Commedia venne scritta da Dante a Firenze”.

Example Output:

```
[["Divina Commedia", "work"],
["Dante", "person"],
["Firenze", "location"]]
```

Input Text: `{Document}`

A.2. Prompt for AMD

System Prompt: You are a political scientist with expert knowledge of Italian politics. Your task is to annotate references to Persons, Locations, and Organizations within historical texts. You generate structured responses in JSON format. Do not write Python code.

User Prompt: Extract the references to named entities of type “person”, “location”, and “organization” within the input text, taken from the writings of the Italian politician Aldo Moro (1916–1978). Extract the entities in the response using a JSON format as in the provided example.

Example Input: “Nel luglio Togliatti si trasferì a Mosca dove partecipò al VI congresso dell’Internazionale Comunista.”

Example Output:

```
[["Togliatti", "person"],
["Mosca", "location"],
["Internazionale Comunista", "organization"]]
```

Input Text: `{Document}`