Integrating TEI Publication, Guided Exploration, and Vector Databases for Semantic Search in the Voci dall'Inferno Project

Angelo Mario Del Grosso, Elvira Mercatanti, Carla Congiu, Marina Riccucci
Cnr-Istituto di Linguistica Computazionale "A. Zampolli", Via Moruzzi, 1, Pisa, Italy
Università di Pisa, P.za Evangelista Torricelli, 2, Pisa, Italy
{angelomario.delgrosso, elviramercatanti}@cnr.it
c.congiu1@studenti.unipi.it, marina.riccucci@unipi.it

Abstract

This paper presents the *Voci dall'Inferno* digital environment, which integrates TEI-based encoding, web publication, and embedding-based semantic retrieval for testimonies by Italian Holocaust survivors. The corpus comprises written and oral sources encoded in XML-TEI through an ODD customization that documents provenance, structure, and interpretive layers. The web application, build on eXist-db, supports guided access, management, visualization, and exploratory analysis of encoded data. Within this infrastructure, we report a pilot semantic-retrieval study on references to Dante's *Divine Comedy*, using SentenceTransformers embeddings and a vector database to retrieve both literal and non-literal Dantean passages. Given the current corpus size, findings are interpreted as exploratory and method-oriented, and future large-scale validation will be conducted. We also address the ethical and legal constraints that shape access policies and long-term reuse in sensitive historical collections.

Keywords: Sentence similarity, web application, eXist-db, TEI-based digital archives, Divine Comedy

1. Introduction

Preserving, curating, and providing access to testimony archives, including both written and oral sources, is essential for historical, literary, and linguistic scholarship, as well as for ensuring that these stories remain available for research, education, and public memory.

Holocaust testimony archives remain an underused resource across a wide range of fields, including linguistics, oral history, sociology, and related disciplines. By preserving firsthand testimonies, these sources serve as repositories of collective memory and contribute to safeguarding intangible cultural heritage.

The lack of shared infrastructures and interoperability mechanisms, together with the legal and ethical complexity associated with access conditions and data governance, remains a major challenge (Calamai et al., 2021). These issues are particularly evident in the Italian context, where collections are frequently managed by individual researchers or small teams, and where institutions may lack trusted digital repositories and clear policies for archiving primary sources and scholarly analyses. As a result, historically and culturally significant testimonies risk being dispersed or even lost (Abete et al., 2025).

Testimony archives still struggle to be fully integrated into research-data management and open-science agendas, where long-term sustainability depends on shared protocols, interoperable infrastructures, and robust legal and ethical procedures (Calamai and Frontini, 2018). From this perspective, the (re)use of these sources is commonly associated with the adoption of the FAIR principles, which promote findability, accessibility, interoperability, and reusability as prerequisites for durable and responsible dissemination (Jong et al., 2018).

This paper builds on our contribution to the 2024 edition of the HTRes workshop (Anuradha et al., 2024), where we introduced the *Voci dall'Inferno* initiative and described the collection and curation of both oral and written resources, focusing mainly on non-literary testimony texts. We outlined the data-acquisition pipeline, annotation choices, quality-control procedures, and intended research and evaluation use cases (Del Grosso et al., 2024), with particular attention to investigating the presence and use of Dante's lexicon in the corpus. We also reported initial dataset statistics and observations, and discussed the limitations and future directions that motivate the present submission.

The scope of this paper is twofold but methodologically unified: (i) to document an interoperable TEI-based archival and publication framework for heterogeneous Holocaust testimonies, and (ii) to present a pilot retrieval component for Dantean intertextuality built on sentence embeddings and vector search. Our main contribution is therefore the end-to-end integration of data curation, encoding, publication, and computational exploration, with explicit attention to reproducibility, interpretability, and governance constraints (Mercatanti et al., 2025a).

In this context, the *Voci dall'Inferno* initiative[1] contributes to ongoing efforts to preserve and make testimonies accessible by providing a structured, interoperable digital archive spanning both written and oral sources, and by offering tools for explo-

[1]The GitHub repository for the project is available at https://github.com/CoPhi/voci-inferno/.

Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026, pages 1–11
11 May 2026.

ration and analysis that support long-term reuse under appropriate legal and ethical constraints (Mercatanti et al., 2025b).

The remainder of the paper is structured as follows. Section 2 reviews related work and situates our contribution within TEI-based initiatives and semantic retrieval approaches. Section 3 outlines the methodology and data-processing workflow adopted in the project. Section 4 describes the *Voci dall'Inferno* web application and its main functionalities for guided exploration and analysis. Section 5 presents our semantic search component based on embeddings and vector databases, focusing on the automatic retrieval of explicit and implicit Dantean echoes. Finally, we summarize current limitations and future directions in Section 6.

2. Related Work

Like other digital archives, testimony collections require harmonized metadata practices, robust workflows from digitization to preservation, and technical solutions that support controlled access while maintaining visibility and reusability (Lin et al., 2020).

The design of our application is informed by the broader ecosystem of tools for the production, dissemination, and exploration of TEI-encoded content (Bénière et al., 2024). In particular, frameworks such as TEI Publisher[2] provide configurable web interfaces for browsing, searching, and rendering TEI documents, often leveraging XQuery/XSLT pipelines and XML databases. Client-side approaches such as CETEIcean[3] enable in-browser transformation of TEI to HTML, supporting lightweight publication workflows and interactive presentation. Related projects and toolkits (e.g., TEI Boilerplate[4]) similarly aim to facilitate the rendering of TEI documents on the web with minimal infrastructure.

For digital editions with advanced navigation and reading interfaces, EVT (Edition Visualization Technology)[5] is a widely used environment for publishing scholarly editions and integrating multiple views (e.g., diplomatic and interpretative transcriptions, facsimiles, and apparatus).

In this comparison, TEI Publisher is used as a conceptual benchmark for publication patterns and user-facing functionalities rather than as our deployment stack. Our implementation is based on eXist-db and XQuery while remaining compatible with architectural principles shared by TEI-native publication environments.

Recent advances in semantic retrieval are largely driven by Transformer encoders derived from BERT (Devlin et al., 2018), which enable queries and text units to be mapped to dense vector representations and compared in an embedding space (Venkatesh Sharma et al., 2024). In this setting, retrieval can be implemented with bi-encoder architectures and optionally complemented by re-ranking components for improved precision. Sentence-level models (e.g., Sentence-BERT and related *SentenceTransformers* variants) (Mayil and Jeyalakshmi, 2023) are now commonly adopted to support semantic similarity and passage retrieval, especially when the goal is to capture paraphrases and non-literal correspondences rather than exact lexical matching (Zhou et al., 2023).

Operationally, embedding-based retrieval is typically supported by vector indexes and specialized databases providing approximate nearest-neighbor search, metadata filtering, and hybrid lexical–semantic retrieval. Among widely used infrastructures, Weaviate[6] and LlamaIndex[7] offer integrated environments for storing vectors alongside structured metadata and executing similarity queries at scale, and they can be combined with application-layer logic to expose results. In the *Voci dall'Inferno* project, these technologies served as methodological references for developing the testimony archive, combining TEI-based publication workflows with semantic retrieval.

3. Methodology

Our methodology integrates (i) philological and linguistic work to curate heterogeneous testimonies, (ii) TEI-based modeling and encoding to represent textual phenomena and preserve interpretability, (iii) digital modules for guided web access and visual analytics, and (iv) computational modules for semantic retrieval of Dantean echoes using BERT-based embeddings.

Data acquisition and transcription. The corpus is built from heterogeneous sources, including written documents (e.g., diaries, memoirs, manuscripts, and printed texts) and oral testimonies (audio or audiovisual interviews); in the current release, testimonies used in retrieval experiments are in Italian. Source selection is based on resources currently available to the team (an unpublished oral-interview collection for spoken testimonies and unpublished private and family collections for written materials) and on explicit inclusion criteria: documented provenance, sufficient metadata for contextualization, legal feasibility for research use and controlled publication, and transcription/encoding feasibility within the project workflow. Exclusion

[2] https://teipublisher.com/index.html
[3] https://github.com/TEIC/CETEIcean
[4] http://teiboilerplate.org/
[5] https://github.com/evt-project/evt-viewer/
[6] https://weaviate.io/
[7] https://www.llamaindex.ai/

Figure 1: Example of an XQuery function used in the web application

criteria include uncertain provenance, unresolved rights constraints, and insufficient textual or audio quality for reliable encoding. To reduce circularity in downstream analysis, selection is not based on the prior presence of Dantean references. Oral sources follow a transcription workflow that combines manual revision with semi-automatic support, using speech-to-text tools such as the CLARIN transcription portal (Draxler et al., 2024) to produce transcriptions that are faithful to the spoken interaction and suitable for linguistic analysis.

Domain-specific transcription language. To ensure consistency and machine processability, we formalize transcription conventions as a domain-specific language (DSL) (Bambaci and Boschetti, 2020). Oral testimonies contain a complex set of verbal and non-verbal phenomena (e.g., interruptions, unfinished segments, repetitions, false starts, pauses, background noise, prosodic cues; and, for audiovisual sources, gestures and body movements) that must be explicitly identified and encoded. We therefore define a controlled set of transcription markers and a parsing strategy grounded in a context-free grammar approach, in line with established transcription ecosystems such as CHAT (MacWhinney, 2019) and DT2 (Bois, 1991), enabling formal recognition of units such as speaker changes, gaps, unclear spans, and prosodic variations. This design facilitates readability for encoders and downstream extraction of structured information.

TEI-based encoding and customization. All testimonies are encoded in XML-TEI through a unified data model that supports both oral and written sources while preserving source-specific characteristics. The model supports multiple annotation layers, including document structure, source alignment, entities, events, relations, and interpretive tags. Encoding decisions aim to (i) retain stable links to primary sources (including alignment to audio segments or line-based references where applicable), (ii) capture document structure (divisions, segments, and anchors), and (iii) represent the authorial and analytical layers needed by the project, such as additions, deletions, named entities, events, places, and relations. In the current implementation, core documentary and linguistic layers are represented inline in TEI for transparency and editorial traceability, while computationally derived layers can be externalized as stand-off annotations when needed. In this setup, stand-off annotation is used to represent named entities, relations among mentioned individuals, cited events, and quotations through six dedicated TEI lists: `listPerson`, `listPlace`, `listOrg`, `listRelation`, `listEvent`, and `listBibl`. An ODD customization governs element and attribute usage, together with project-specific constraints, ensuring consistency, validation, and long-term maintainability (Mercatanti et al., 2025c).

Information extraction and linguistic analysis. On top of the TEI backbone, we extract structured

information for guided access and analysis, including witness-level metadata, testimony provenance, and corpus-derived annotations for maps, timelines, and exploratory statistics. In addition, we curate and visualize lexical information derived from linguistic analysis of the testimonies, with particular attention to Dantean lexicon, quotations, and allusions. This extraction layer is explicitly separated from semantic-retrieval testing: editorial annotations (explicit quotations, implicit quotations, allusions) define the ground truth, whereas retrieval experiments are executed as a distinct computational phase. This separation is intended to limit circularity and to make evaluation assumptions explicit.

Guided access and visual analytics in the web application. Publication and exploration are implemented through a web application built on eXist-db. XQuery functions retrieve TEI-encoded content and assemble HTML views that provide guided access to witnesses, testimonies, and analytical outputs. The interface integrates interactive components for reading and listening to encoded testimonies, and for exploring encoded phenomena, named entities, relationships, maps, timelines, and corpus-level statistics. The architecture is modular and extendable. Scalability across different collections depends on harmonized metadata quality, governance policies, and indexing strategies.

Embedding-based semantic retrieval of Dantean echoes. To complement editorially curated links and lexicon-driven analysis, we integrate an embedding-based semantic retrieval module that surfaces non-literal correspondences between testimony fragments and passages of the *Commedia*. Textual units (e.g., verses, terzine, or longer segments) are normalized and represented with sentence-level Transformer embeddings, then indexed in a vector database to enable nearest-neighbor queries. Retrieved candidates are returned with bibliographic coordinates and metadata (cantica, canto, verse locus) to support verification and scholarly use. This module supports both explicit quotations (typically characterized by high lexical overlap) and implicit quotations or allusions, where semantic proximity is more informative than string matching.

Ethical and legal considerations (data governance). Given the sensitivity of Holocaust-related testimonies and the heterogeneity of archival provenance, methodological choices are complemented by governance measures that regulate access, documentation, and reuse. We adopt a privacy- and rights-aware approach to dissemination, ensuring that publication and computational processing remain aligned with applicable legal and ethical constraints while preserving the research value of the archive.

4. The *Voci dall'Inferno* web application

The digital corpus is accessible through the *Voci dall'Inferno* web application, developed as an integrated environment for managing, presenting, exploring, and analyzing encoded testimonies (Mercatanti et al., 2025a). Built on eXist-db, the application combines the HTML templating framework with XQuery functions to process XML-TEI documents and generate HTML fragments assembled into end-user pages. This architecture separates presentation from data-processing logic, improves maintainability, supports incremental scaling, and contributes to long-term sustainability. The platform is therefore extendable across collections that adopt compatible TEI and metadata profiles.

For example, Figure 1 shows an XQuery function (`app:contaTestimonianzeArchivio()`) that returns the number of testimonies and their categories.

The web application provides several features for consulting and exploring heterogeneous data derived from encoding. The current *Voci dall'Inferno* corpus includes 25 testimonies from 20 witnesses and reflects substantial variation in both source type and testimonial profile, including accounts by people who experienced the Lager firsthand and by others who were never deported. Importantly, corpus inclusion is not conditioned by prior Dantean evidence; this design supports a less circular setting for retrieval analysis. The testimonies currently processed, encoded, and semantically analyzed are Italian-language materials, consistently normalized through the same preprocessing pipeline.

To support guided access and conceptual clarity, the project adopts a hierarchical taxonomy that groups witnesses according to their historical experience (Fig. 2). At the top level, the collection is organized under *Witnesses* and divided into *Deported witnesses* and *Non-deported witnesses*. The *Deported* branch is further articulated into *Jewish deportees* and *Non-Jewish deportees*, the latter including *Italian Military Internees* (IMI) and *Italian Civil Internees* (ICI). The *Non-deported witnesses* branch includes Jewish partisans, currently represented in the archive by the testimony of Emanuele Artom.

Although still subject to refinement, this taxonomy provides a coherent organizational framework for managing the corpus's heterogeneity while clearly identifying the provenance and historical context of each testimony.

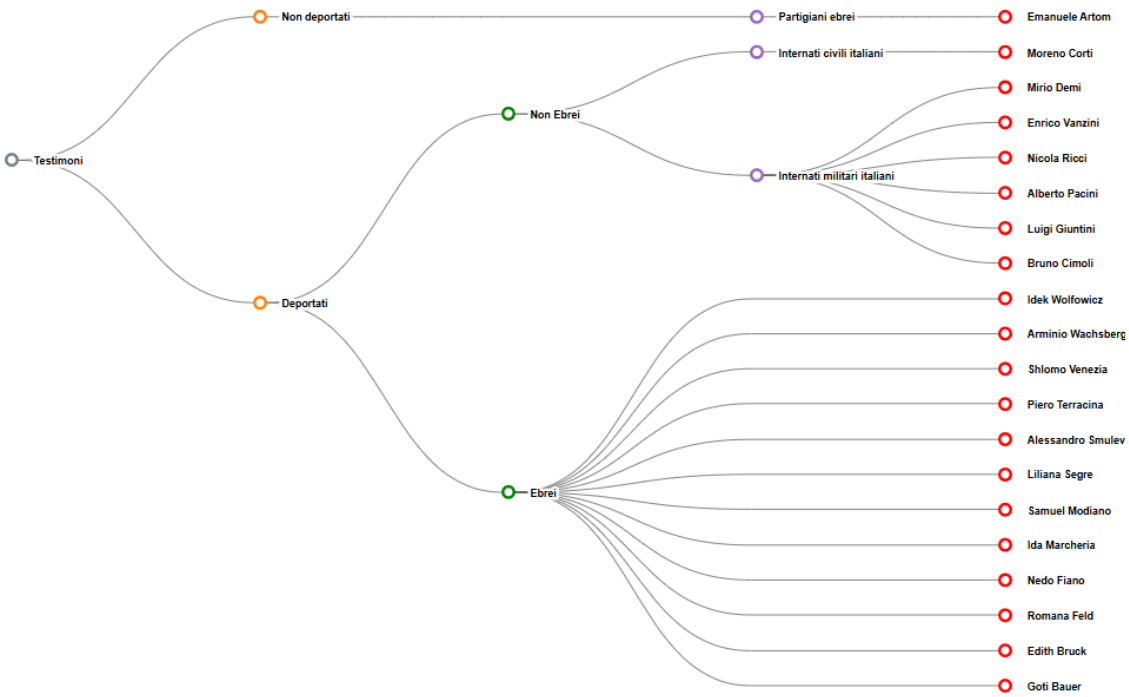

Figure 2: The *Voci dall'Inferno* project taxonomy of witnesses

After encoding, the application enables the extraction and visualization of heterogeneous data types, including written and spoken phenomena, named entities, maps, interpersonal relationships, timelines, and statistical analyses. Interactive visualizations are implemented with Highcharts, a JavaScript library that enhances accessibility and supports intuitive engagement with quantitative analyses.

The application is structured into nine main sections: Home, The Project, Voices, Search for a Witness, Dante, Statistics, Automatic Transcription, Events, and Bibliography (Fig. 3).

Figure 3: Homepage of the application

The core section, *Voices*, provides access to witnesses and their testimonies through an alphabetical navigation menu that allows users to filter results by the initial letter of a witness's surname and open individual profile pages.

For each witness, users are directed to a dedicated page displaying a brief biographical profile together with the list of encoded testimonies currently available for consultation. The page also provides visualizations to support analytical exploration of the encoded data, including (i) a directed, labeled graph of interpersonal relationships, (ii) two maps showing places mentioned and the witness's movements before, during, and after deportation (Fig. 4), and (iii) a timeline of the main events cited by the witness (Fig. 5).

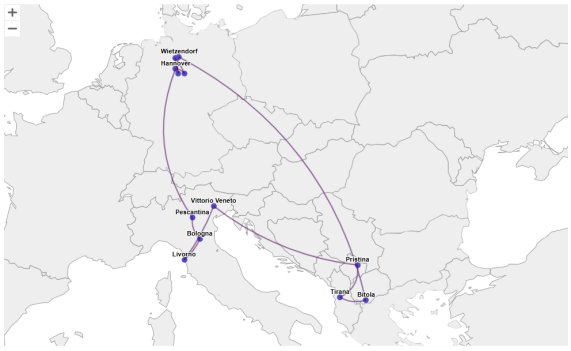

Figure 4: Map of the witness's movements

Upon selecting a testimony, users access a dedicated page presenting structured metadata extracted dynamically from the XML-TEI source through XQuery functions. The displayed information varies by testimony type (oral vs. written), enabling a differentiated representation of source-specific features. For oral testimonies, the interface

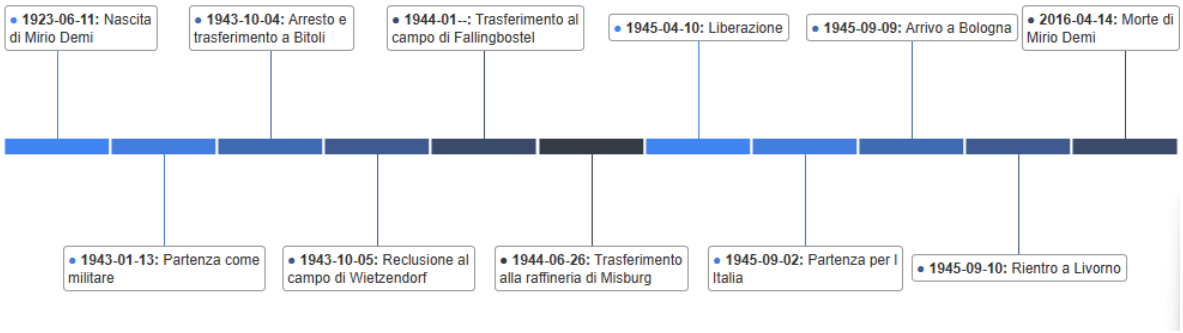

Figure 5: Timeline of the main events cited in the testimony

synchronizes the structured abstract (*regesto*), organized into segments, with the timeline, allowing users to read each segment and listen to the corresponding portion of the audio recording.

The transcription interface adapts to the specific type of resource. For written testimonies, an image-based mode displays the source image alongside its transcription (Fig. 6); alternatively, users can view the transcription alone, with a legend of encoded phenomena that can be interactively highlighted in the text. The same highlighting functionality is available for oral sources, with the legend dynamically tailored to the resource type (Fig. 7).

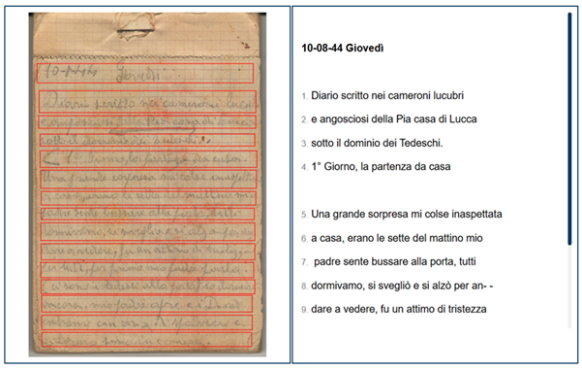

Figure 6: Alignment of the primary source image with its corresponding transcription

Each testimony is further enriched by a set of statistical visualizations, including charts on encoded phenomena, witness expression patterns, and named-entity distribution. Additionally, the interface provides an analysis of references to the *Divine Comedy*. Explicit and implicit quotations, allusions, and Dantean lexical references can be explored interactively: selecting a citation highlights the corresponding verses or passages of the poem referenced in the testimony.

The web application features two analytical sections, *Dante* and *Statistics*, dedicated to exploring the encoded corpus. The *Dante* section examines the presence and distribution of references to the *Divine Comedy* within the testimonies. It provides visualizations and quantitative summaries of explicit and implicit quotations, allusions, and lexical references identified during transcription and encoding. The analysis is twofold: it investigates both the typology of Dantean references and the witnesses who most frequently employ Dante's language to articulate the Lager experience. Out of 25 testimonies (produced by 20 witnesses), 10 contain references to the poem, totaling 61 occurrences: 15 explicit quotations, 4 implicit quotations, 7 allusions, and 35 terms from the *Comedy* (Fig. 8). Among the 19 encoded quotations, the majority refer to the *Inferno* (16 cases), with only limited references to the *Purgatorio* (2) and the *Paradiso* (1).

The *Statistics* section provides broader corpus-level analyses. It presents visualizations and summary data concerning the composition of the archive, including the distribution of testimonies by witness category and the provenance of the two main categories of sources, oral and written. The section also documents the overall extent of the encoded material: to date, 18 hours, 35 minutes, and 48 seconds of oral recordings and 395 pages of written sources have been transcribed and encoded.

5. Semantic search with embeddings and vector databases

To complement guided access to the corpus, we are developing *Dante Similarity Search*, a semantic retrieval module that enables users to identify potential Dantean echoes across testimonies, including non-literal correspondences, through embedding-based similarity and vector search.

Dante Similarity Search is currently implemented as a prototype web application designed to detect echoes of Dantean language in concentration-camp survivor testimonies by linking prose frag-

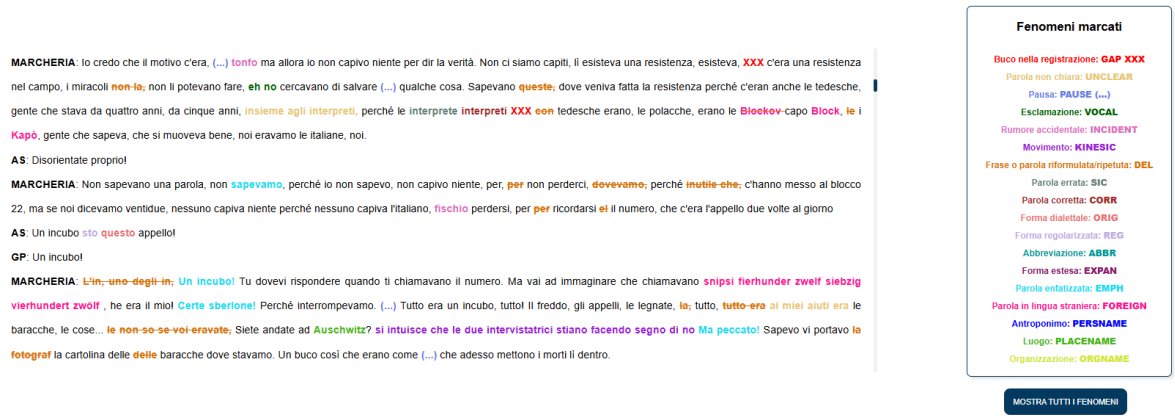

Figure 7: Transcription of an oral testimony with in-text highlighting of encoded phenomena

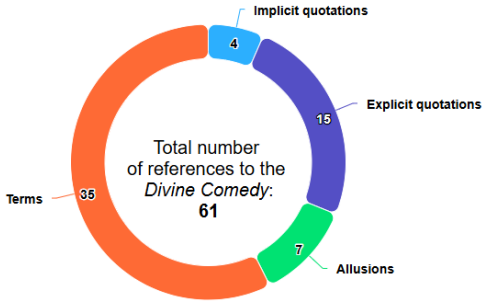

Figure 8: References to the *Divine Comedy*

ments to passages from the *Commedia*. Given a query, the system returns the closest candidates, ranked by similarity and enriched with metadata (*cantica*, *canto*, position), making results verifiable, citable, and suitable for subsequent human validation (Congiu et al., 2025) (Fig. 9). The approach goes beyond lexical overlap and targets semantic proximity that can surface non-literal correspondences.

The development of Dante Similarity Search draws inspiration from projects by William Mattingly[8] that aim to automatically identify biblical quotations in Latin. In Vulgata spaCy (2022)[9], the Clementine Vulgate is cleaned and organized into CSV format, and a spaCy pipeline is built by combining embeddings trained on the Patrologia Latina[10] (Bloom/floret) with two components: an EntityRuler to detect direct or partial quotations and a machine-learning model to detect quotations in context. A subsequent step links each occurrence to specific verses while addressing spelling and punctuation variability and incomplete quotations. Constraints such as requiring phrases of at least four words reduce false positives. With Weaviate Vulgate (2024), the approach is recast in terms of semantic similarity: verses are converted into vectors using SentenceTransformers (LaBSE[11]), indexed in Weaviate, and made accessible through a Streamlit web application that allows users to submit a text fragment and retrieve the most closely related verses.

The architecture integrates a vector-

[8] https://www.wjbmattingly.com/
[9] https://github.com/wjbmattingly/vulgata-spacy
[10] https://patristica.net/latina/
[11] https://huggingface.co/sentence-transformers/LaBSE

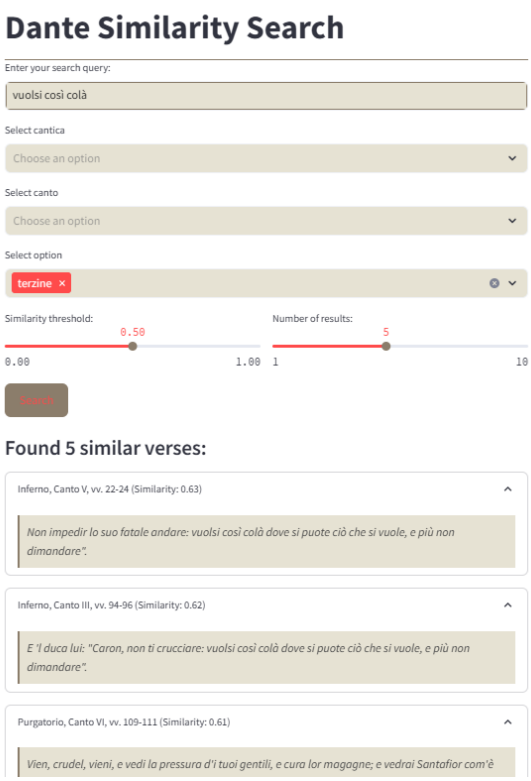

Figure 9: Dante Similarity Search

representation pipeline and a semantic retrieval module. Texts are normalized and transformed into contextual embeddings via *SentenceTransformers*[12]. We compared three candidate models in our pilot setting: `LaBSE`, `paraphrase-mpnet-base-v2`, and `all-mpnet-base-v2`. In our data, `LaBSE` produced weaker rankings for Dantean fragments, while `paraphrase-mpnet-base-v2` recovered some relevant passages but with lower ranking consistency. `all-mpnet-base-v2` provided the best overall trade-off in top-ranked relevance and ranking stability, and was therefore selected as the operational model. We note, however, that this remains a pragmatic choice for a pilot setup, and domain adaptation to historical/literary Italian remains a key next step. Vectors are then indexed in *Weaviate*[13], which supports efficient nearest-neighbor queries and structured metadata management needed to reconstruct the Dantean reference associated with each result. Interaction takes place through a *Streamlit*[14]-based interface, designed for entering queries and inspecting matches, including similarity scores and textual references. For testing purposes and to support external access and service sharing, the application can be exposed via *Cloudflare Tunnel*[15], simplifying deployment without requiring complex network configurations.

A central methodological component is the construction of datasets from the *Commedia*, transformed into collections of homogeneous, queryable textual units, each associated with bibliographic metadata. The choice of granularity directly affects interpretability and result quality: smaller units support precise anchoring but may be semantically fragile, whereas larger units introduce context and stability at the expense of precision. From this perspective, segmentation into *terzine* prioritizes semantic coherence, reduces ambiguity, and is effective when an echo is distributed across multiple lines; segmentation into single verses maximizes precision and is particularly suited to identifying explicit quotations, although it requires careful normalization in Python and filtering by metadata to limit spurious matches; segmentation into sentences offers a useful compromise, especially for paraphrases and reformulated echoes, preserving semantic relations that a single verse may not make explicit.

Overall, the workflow maps user input to the retrieval of the most similar Dantean candidates within a reproducible pipeline in which technical choices (model, textual units, search parameters)

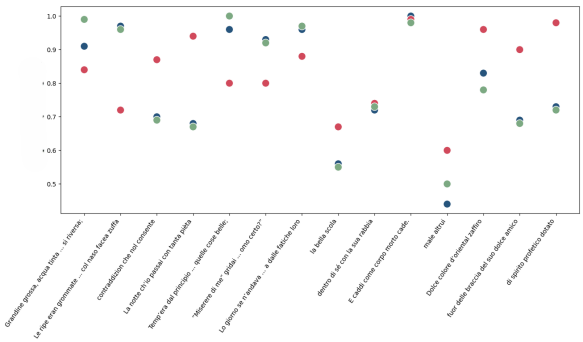

Figure 10: Similarity scores for a set of explicit Dante quotations identified in the testimonies. Each quotation is listed along the x-axis, while the y-axis represents a similarity score.

remain traceable and results interpretable. Given the current corpus size, we frame this component as a pilot study and avoid strong generalization claims. To support systematic validation, we will define an explicit protocol including *precision*, *recall*, and baseline comparisons. We will also set up experiments to distinguish explicit quotations, implicit quotations, and allusions. Editorial annotation and computational testing will be treated as separate phases to reduce circularity.

Indexing in Weaviate and inspection through Streamlit turn similarity into an operational tool: results become explorable and filterable, and can be critically evaluated thanks to metadata and textual coordinates. Quantitatively, explicit quotations are the most robust case: because testimony fragments often preserve literal portions of the *Commedia*, the system returns precise and reliable alignments (Fig. 10). Implicit quotations are more challenging: paraphrastic reformulations can preserve thematic resonance while substantially altering lexical form, making exact retrieval less stable even when top candidates remain semantically coherent with the input fragment.

A representative failure case concerns Nicola Ricci's diary quotation, "Si va verso la fame, si va verso il freddo, si va verso l'inferno," for which the system did not recover the expected match with *Inferno* III, 1–3 ("Per me si va ne la città dolente / per me si va ne l'etterno dolore / per me si va tra la perduta gente"). For a human reader, the allusion is immediate; for the model, it is an implicit intertextual signal that exceeds purely distributional similarity. This example clarifies a key limitation of the current setup and motivates the next step: fine-tuning on a *Commedia*-centered corpus enriched with commentaries, paraphrases, and curated intertextual links, in order to improve recognition of non-explicit echoes and reduce confusion between shallow lexical similarity and deeper semantic-literary correspondence.

[12] https://sbert.net/

[13] https://weaviate.io/

[14] https://streamlit.io/

[15] https://developers.cloudflare.com/cloudflare-one/networks/connectors/cloudflare-tunnel/

The output is not limited to on-screen consultation: retrieval results can be reintegrated into the encoding workflow in a structured form. Starting from a testimony fragment, the system stores the top candidates across different textual granularities (terzine, verses, and sentences), each with a similarity score and bibliographic coordinates (cantica, canto, verse range). This enrichment process is implemented as a Python routine over testimony XML files and is designed to remain auditable, reversible, and compatible with subsequent human validation. The enrichment procedure follows an explicit criterion: for each encoded quotation, the system searches across the three textual unit types (terzine, verses, and sentences) and, for each type, inserts the three closest results. The encoded documents are then enriched with the suggested references (Fig. 11).

Figure 11: For each encoded quotation, the system automatically searches terzine, verses, and sentences, adding the three closest matches with bibliographic references to the XML files in a structured format.

6. Conclusions and Further Directions

This paper presented an integrated framework for testimony curation, TEI encoding, web publication, and embedding-based semantic retrieval, together with a pilot study on Dantean echoes. The current results are promising but preliminary and should be interpreted as proof-of-concept evidence rather than as definitive performance claims. We delivered a first operational release of the *Voci dall'Inferno* web application, including dedicated views for exploring Dantean lexicon, quotations, and allusions, and we defined a reproducible workflow linking editorial annotation and computational analysis. Given the limited experimental scale, we prioritize methodological transparency and a clear separation between annotation and retrieval evaluation. Future work will focus on: (i) refining the transcription workflow by integrating more robust support for automatic recognition and DSL-based encoding across different types of primary sources; (ii) improving interpretability and methodological soundness by connecting the archive to major SSH research infrastructures, in particular the CLARIN-IT community; (iii) extending the system with additional visualization components (e.g., temporal and thematic facets) to better support scholarly workflows; (iv) expanding the benchmark and reporting full IR metrics; (v) measuring annotation reliability for explicit, implicit, and allusive categories; and (vi) evaluating multilingual and domain-adapted embedding models for historical and literary Italian.

7. Acknowledgments

The project has progressed primarily through the work of undergraduate and graduate students, mostly enrolled in the Digital Humanities and Italian Studies degree programs at the University of Pisa. To date, twenty-two students have completed their degree theses within the framework of *Voci dall'Inferno*. The initiative has also benefited from the support of several institutions and research infrastructures: the Department of Philology, Literature and Linguistics at the University of Pisa; the Institute for Computational Linguistics "A. Zampolli" of the National Research Council (CNR) in Pisa[16]; the CLARIN-IT research infrastructure[17]; the Collaborative and Cooperative Philology Lab (CoPhiLab)[18]; the CLARIN Knowledge Centre for Digital and Public Textual Scholarship (DiPText-KC)[19]; the Centro Interdipartimentale di Studi Ebraici (CISE)[20]; the Centro di Documentazione Ebraica Contemporanea (CDEC)[21].

7.1. Ethical considerations and limitations

The current version of the *Voci dall'Inferno* web application is accessible only to project members in a restricted environment. This choice reflects both the sensitive nature of the collection and the need to ensure responsible management of sources and content in compliance with data-protection requirements. The ethical and legal framework is particularly complex due to the heterogeneity of the corpus: while some testimonies derive from publicly accessible archives, others consist of unpublished materials from private family collections. Obtaining consent for research purposes is often challenging because many data subjects are difficult to contact due to age or because they are deceased. Furthermore, the testimonies span different historical periods, with some sources predating the entry into

[16] https://www.ilc.cnr.it/
[17] https://www.clarin-it.it/it
[18] https://cophilab.ilc.cnr.it/
[19] https://diptext-kc.clarin-it.it/
[20] https://cise.unipi.it/
[21] https://www.cdec.it/

force of the GDPR, which complicates the identification of lawful bases for processing.

An additional layer of complexity concerns the dual nature of the sources (oral and written), with oral testimonies presenting specific challenges for data governance. Audio recordings constitute a particularly sensitive category of research material, as they simultaneously function as historical documents, scientific sources, and biometric identifiers. This hybridity creates tensions between openness and protection that directly affect decisions about accessibility, reuse, and long-term sustainability within research infrastructures.

Recent collaboration with the ROADS project, which has defined a FAIR-by-design model for managing oral archives throughout the entire data life cycle (from data collection to publication and reuse), has enabled us to test and refine this approach on the *Voci dall'Inferno* archive. The ROADS framework provides a transferable model to guide the transition from project-based archives to FAIR, sustainable, and reusable research resources, ensuring compliance with data-protection requirements while respecting the sensitivity of the documented contexts. A key factor in the sustainability of the ROADS model is the involvement of legal experts embedded within participating institutions, who mediate between Open Science principles and data-protection constraints. Their contribution supports the legal robustness and transparency of research resources, highlighting how the long-term reuse of sensitive historical data depends not only on technical solutions but also on stable governance frameworks and legal accountability (Abete et al., 2026).

In practice, this work could inform a set of concrete strategies to be progressively implemented within the *Voci dall'Inferno* archive. These may include: (i) formalizing the distribution of responsibilities among stakeholders (e.g., via joint controllership agreements and the identification of a single contact point for data-subject requests), (ii) documenting provenance and consent status at the item level, while adopting a diligent-search approach for legacy materials collected before current standards, aimed at contacting data subjects or, where this is no longer possible, their potential heirs, (iii) introducing layered information and multi-level consent procedures in the case of newly collected data, allowing participants to make informed choices about different levels of access, dissemination, and reuse of their testimonies, and (iv) applying data-minimization measures (e.g., redacted public views and restricted access to particularly sensitive content), supported by logging mechanisms.

From this perspective, the model defined by the ROADS project represents a fundamental guideline for orienting future decisions concerning data management, consent documentation, and access policies. Its adoption could support the gradual transition from a project-based archive to a FAIR, sustainable, and reusable research resource, while ensuring compliance with data-protection requirements and respect for the sensitivity of the historical contexts represented in the corpus.

8. Bibliographical References

Giovanni Abete, Silvia Calamai, Sergio Canazza, Alessandro Casellato, Elvira Mercatanti, and Monica Monachini. 2026. La filiera legale di ROADS. Una proposta FAIR per archivi orali analogici. Technical report, Zenodo.

Giovanni Abete, Cesarina Vecchia, Silvia Calamai, Alessandro Casellato, Sergio Canazza, Elvira Mercatanti, Monica Monachini, Roberta Ottaviani, Giulia Zitelli Conti, and Giada Zuccolo. 2025. On the lifecycle of Italian oral archives: the ROADS project. In *La voce della grammatica. Nuove prospettive sull'interazione tra fonetica e morfologia, sintassi, lessico*. Associazione Italiana di Scienze della voce.

Isuri Anuradha, Martin Wynne, Francesca Frontini, and Alistair Plum, editors. 2024. *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Luigi Bambaci and Federico Boschetti. 2020. Encoding the Critical Apparatus by Domain Specific Languages: The Case of the Hebrew Book of Qohelet. In *La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica.*, Quaderni di Umanistica Digitale, Milano. Università Cattolica del Sacro Cuore.

Sarah Bénière, Floriane Chiffoleau, and Laurent Romary. 2024. TEI specifications for a sustainable management of digitized holocaust testimonies. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 10–17, Torino, Italia. ELRA and ICCL.

John W. Du Bois. 1991. Transcription design principles for spoken discourse research. *Pragmatics*, 1:71–106.

Silvia Calamai and Francesca Frontini. 2018. FAIR data principles and their application to speech and oral archives. *Journal of New Music Research*, (47):339–354.

Silvia Calamai, Stefania Scagliola, Fabio Ardolino, Christoph Draxler, Arjan van Hessen, and Henk van den Heuvel. 2021. Ravensbrück Interviews: How to Curate Legacy Data to Make it CLARIN Compliant. In *Selected Papers from the CLARIN Annual Conference 2021, virtual event, September 27-29, 2021*, volume 189 of *Linköping Electronic Conference Proceedings*, pages 1–9. Linköping University Electronic Press.

Carla Congiu, Angelo Mario Del Grosso, and Marina Riccucci. 2025. Verso l'implementazione di un sistema di riconoscimento diallusioni al lessico dantesco nelle testimonianze del Lager: il caso d'uso in project. In *Diversità, Equità e Inclusione: Sfide e Opportunità per l'Informatica Umanistica nell'Era dell'Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD2025*, Verona. AIUCD. Num Pages: 663.

Angelo Mario Del Grosso, Marina Riccucci, and Elvira Mercatanti. 2024. The Impact of Digital Editing on the Study of Holocaust Survivors' Testimonies in the context of project Project. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 1–9, Torino, Italia. ELRA and ICCL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*.

Christoph Draxler, Henk van den Heuvel, Arjan van Hessen, Pavel Ircing, and Jan Lehečka. 2024. Speech technology services for oral history research. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 38–43, Torino, Italia. ELRA and ICCL.

Franciska De Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dawei Lin, Jonathan Crabtree, Ingrid Dillo, Robert R. Downs, Rorie Edmunds, David Giaretta, Marisa De Giusti, Hervé L'Hours, Wim Hugo, Reyna Jenkyns, Varsha Khodiyar, Maryann E. Martone, Mustapha Mokrane, Vivek Navale, Jonathan Petters, Barbara Sierman, Dina V. Sokolova, Martina Stockhause, and John Westbrook. 2020. The TRUST Principles for digital repositories. *Scientific Data*, 7(1):144.

Brian MacWhinney. 2019. Chat manual.

V.Valli Mayil and T.Ratha Jeyalakshmi. 2023. Pretrained Sentence Embedding and Semantic Sentence Similarity Language Model for Text Classification in NLP. In *2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP)*, pages 1–5.

Elvira Mercatanti, Carla Congiu, Angelo Mario Del Grosso, and Marina Riccucci. 2025a. Voci dall'Inferno: a Web application to study and analyze the Lager testimonies. In *DH2025 - Building Access and Accessibility, Open Science to all Citizens*, Lisbon, Portugal. Zenodo.

Elvira Mercatanti, Angelo Mario Del Grosso, and Marina Riccucci. 2025b. Voci dall'Inferno: Dante per esprimere l'indicibile: Un'applicazione digitale per esplorare le testimonianze non letterarie dei sopravvissuti ai Lager. *Umanistica Digitale*, (20):527–562.

Elvira Mercatanti, Marina Riccucci, and Angelo Mario Del Grosso. 2025c. Voci dall'Inferno: a TEI-Based Digital Archive for finding Dante in Concentration Camp Testimonies. In *"New Territories". Text Encoding Initiative Conference and Members' Meeting 2025*, Kraków, Poland. Zenodo.

K. Venkatesh Sharma, Pramod Reddy Ayiluri, Rakesh Betala, P. Jagdish Kumar, and K. Shirisha Reddy. 2024. Enhancing query relevance: leveraging SBERT and cosine similarity for optimal information retrieval. *International Journal of Speech Technology*.

Ya Zhou, Ning Zhao, Guimin Huang, Nanxiao Deng, and Qingkai Guo. 2023. Sentences Similarity Model Based on Fusion of Semantic, Syntactic and Word Order Multi-Features. In *2023 4th International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*, pages 121–124.