

Augmenting Historical Research through AI: A Case Study on Norcia (Italy) 1859

Sara Alimenti^{1,*}, Carla Binucci², Giulio Biondi^{1,*}, Emilio Di Giacomo², Fabrizio Grosso^{1,*}, Marco Legittimo^{1,*}, Giuseppe Liotta², Regina Lupi³ and Paola Venuti¹

¹Ce.Di.Pa - University of Perugia, Perugia, Italy

²Department of Engineering - University of Perugia, Perugia, Italy

³Department of Political Sciences - University of Perugia, Perugia, Italy

Abstract

This paper presents the objectives, methodology, and preliminary results of an interdisciplinary project on the cultural heritage debate in 19th-century Italy, part of an ongoing initiative at the Research Centre on the Digital Transformation of Cultural Heritage (CeDiPa), University of Perugia. The case study focuses on the transition from the Papal States to the Kingdom of Italy and the choices (preserve, renovate, or modify) made after the earthquake that almost completely destroyed the town of Norcia in Valnerina in August 1859. Our source, the daily newspaper “La Gazzetta dell’Umbria” (1860–1870), was OCR-transcribed, then processed with different LLMs to extract key entities into a knowledge graph, giving researchers a formalised information base that can be queried and traced back to the original documentary evidence. Domain expertise remains central, both in identifying significant patterns within large amounts of data and in monitoring the extraction and disambiguation process. The proposed infrastructure enables a virtuous circle between historical research and computational sciences, suggesting new historiographical hypotheses and raising unexpected questions.

Keywords

Knowledge Graph, Digital History, Document Processing, LLM-based Knowledge Extraction, Semantic Network Analysis, Graph Analysis

1. Introduction

Large digitised archives of historical newspapers hold information that is, in practice, inaccessible at scale: reconstructing the network of actors, places, and concepts behind a decades-long public debate requires cross-referencing more pages than a single scholar can process by hand. Large Language Models (LLMs) can automate a substantial part of this reconstruction, extracting structured knowledge from unstructured historical text at a scale manual analysis cannot match, while still requiring the historian’s expertise to validate results against the sources.

This preliminary work of an ongoing interdisciplinary project between historians and computer scientists at CeDiPa, University of Perugia, has two aims. First, we devise a pipeline based on fully open-source frameworks and open-weight models that turns digitised newspaper pages into a knowledge graph domain experts can query and trace back to the source, applied to the *Gazzetta dell’Umbria*, an Italian-language newspaper from the second half of the 19th century. Second, we show how the resulting graph supports historical research: combining semantic and topological analysis, we study the debate on the reconstruction of Norcia following the 1859 earthquake, surfacing actors, concepts, and connections, and testing historiographical hypotheses against the source material. This approach to Digital Humanities, and particularly Digital History, views digitisation as an interdisciplinary process

4th Italian Workshop on Artificial Intelligence for Cultural Heritage (IAI4CH 2026, <https://ai4ch.di.unito.it/>), co-located with the 24th International Conference of the Italian Association for Artificial Intelligence (AIIA 2026). 6–9 October 2026, Perugia, Italy

*Corresponding author.

✉ sara.alimenti@unipg.it (S. Alimenti); carla.binucci@unipg.it (C. Binucci); giulio.biondi@unipg.it (G. Biondi); emilio.digiacoia@unipg.it (E. Di Giacomo); fabrizio.grosso@unipg.it (F. Grosso); marco.legittimo@unipg.it (M. Legittimo); giuseppe.liotta@unipg.it (G. Liotta); regina.lupi@unipg.it (R. Lupi); paola.venuti@unipg.it (P. Venuti)

✉ 0009-0003-5407-2958 (S. Alimenti); 0000-0002-5320-9110 (C. Binucci); 0000-0002-1854-2196 (G. Biondi); 0000-0002-9794-1928 (E. Di Giacomo); 0000-0002-5766-4567 (F. Grosso); 0000-0002-4532-7678 (M. Legittimo); 0000-0002-2886-9694 (G. Liotta); 0000-0001-8875-340X (R. Lupi); 0009-0007-5006-9590 (P. Venuti)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

that integrates automated computation into the analysis of large-scale historical data (see, for example, Spina [1]). The first part of the project explores how computational methods—especially knowledge graphs and AI—can open new perspectives on the history of ideas, building on recent work in the field [2].

1.1. The Case Study

On 22 August 1859, an earthquake struck Norcia (Umbria), destroying a large part of the town; the debate that followed – preserve, renovate, or transform its damaged heritage – intersects with the construction of the idea of *nation* in post-unification Italy, a central theme in Risorgimento historiography. Our primary source is the newspaper *Gazzetta dell’Umbria* (Perugia, 1860–1870, 3187 issues, hosted in the Emeroteca Digitale Italiana[3]). The case study is organised around the term *pubblica utilità* (public utility) as an entry point into the reconstruction debate, relating it to *Norcia*, *terremoto* (earthquake), and *ricostruzione* (reconstruction).

1.2. Related Work

Recent work applies AI to CH research along several lines: retrieval over cultural artefacts, NER on historical text, and knowledge graph (KG) construction; historians also apply network and semantic analysis to structured historical data, independently of extraction. On retrieval, [4] use multimodal RAG for artefact provenance, while [5] ground an LLM on a hand-crafted ontology of a few known artworks. On NER, [6] show ChatGPT underperforms BERT-based models on 19th–20th century French newspapers, especially on domain-specific entities and OCR-induced errors, related to our disambiguation problem (Sect. 2.2.1). On KG construction, we build on KGGen [7] (Sect. 2) for its schema-free extraction; most CH efforts instead rely on ontology engineering and manual annotation [8, 9, 10]. Closest to us, [11] combine LLM extraction with a target ontology (SEBI); unlike them, we avoid a predefined schema to preserve information an ontology might exclude. Network and semantic analysis are likewise established historical-research tools [12, 13], e.g. to trace the concept of *nation* over time, paralleling our study of *pubblica utilità* (Sect. 4.1).

2. Framework Architecture

2.1. OCR

Rather than proprietary solutions such as Transkribus, in line with our approach we rely on open-weight transformer-based Visual Language Models (VLMs), which process document images holistically and use attention to capture long-range layout dependencies – an advantage for the column segmentation, irregular fonts, and print degradation common in 19th-century newspapers. We use Dots-MOCR [14] as our transcription engine: it extracts embedded images and tables jointly with text and proves robust to layout irregularities, correctly disregarding ink spillovers and bleed-through as non-textual and avoiding hallucinated transcriptions. About 10% of issues required manual preprocessing (rotation, skew, print artefacts); page-level transcriptions are merged into per-issue documents before KG extraction.

2.2. Knowledge Graph Extraction

We adopt KGGen [7], a schema-free text-to-KG extractor reported to outperform GraphRAG and OpenIE by up to 18% on the MINE1 benchmark. We avoid restricting extraction to a predefined ontology, retaining as much information as possible at the cost of higher downstream complexity [15], addressed via embedding-based canonicalization (Sect. 2.2.1). We compare candidate setups – 3 LLMs (see Tab. 1) and 2 prompt languages – on a subset of 17 documents containing both “Norcia” and “pubblica utilità” (Sect. 3), and select Llama 3.3 with the Italian prompt for the full corpus, since it offers the best balance between extraction quality and computational efficiency (see Tab. 2). We preserve relations’ traceability for context expansion by annotating each relation with its source document.

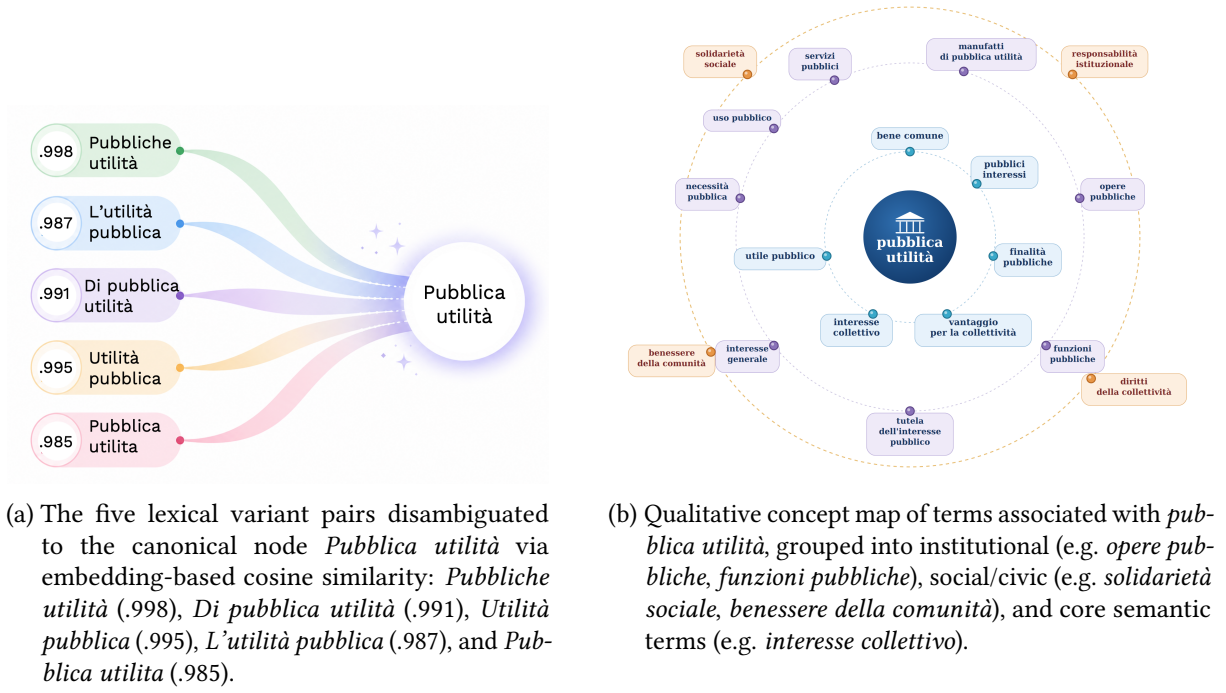


Figure 1: Left: disambiguation of lexical variants to the canonical node. Right: qualitative concept map of semantically associated terms.

2.2.1. Disambiguation

The full extraction yields approximately 0.59M nodes and 1.33M edges, a scale at which lexical variation (e.g. *pubblica utilità*, *utilità pubblica*, *pubblica utilita*) dominates node duplication. We disambiguate using a purely semantic criterion: node labels are embedded with `multilingual-e5-large-instruct` (ME5), and pairs above a cosine-similarity threshold are merged into a single canonical node; observed similarities range from 0.985 to 0.998, as illustrated in Fig. 1a. This step precedes the semantic field *expansion* described in Sect. 4.1: disambiguation collapses redundant surface forms of the same node, whereas expansion grows the node set with related but genuinely distinct concepts.

3. Performance comparison

We perform a preliminary comparison of three models' extraction performance using the [7] MINE1 benchmark methodology: (i) **Facts extraction**: short, self-contained facts are extracted by an LLM from each document as ground truth, manually screened; distinct models are used for fact extraction, KG extraction, and judging (Tab. 1) to avoid self-preference bias [16]. (ii) **Subgraph induction**: facts and KG nodes are embedded with ME5 (Tab. 1); for each fact, the top-k similar nodes and all nodes at distance ≤ 2 induce a subgraph. (iii) **Subgraph analysis**: an LLM-as-a-Judge assesses whether the ground-truth facts can be reconstructed from the subgraph. The comparisons run on two Nvidia H200 GPUs (141GB VRAM) via VLLM; facts are extracted with GPT-OSS and judged with Gemma 4 (English or Italian prompt). Model versions are in Tab. 1.

We select documents containing the keywords "Norcia" and "pubblica utilità" (or its variants), yielding a small set of 17 documents suitable for manual analysis; the two terms are chosen for their relevance and contrasting nature – a proper name vs. an abstract two-word term. Results are in Tab. 2.

4. AI-augmented Historical Analysis

Historical language exhibits both lexical continuities and divergences with modern Italian, particularly for terms such as *pubblica utilità* (public utility), whose surface form varies while its discursive function

Model Short Name	Model full name	Task
ME5	multilingual-e5-large-instruct	Embedding
GPT-OSS	openai/gpt-oss-120b	Facts Extraction
Llama 3.1	meta-llama/Llama-3.1-8B-Instruct	KG Extraction
Qwen 3.6	Qwen/Qwen3.6-35B-A3B-FP8	KG Extraction
Llama 3.3	nvidia/Llama-3.3-70B-Instruct-FP8	KG Extraction
Gemma 4	google/gemma-4-31B-it	LLM-as-a-Judge

Table 1

The models used in the performance comparison. The first four are the compared models. GPT-OSS was used to extract the facts, and Gemma 4 was used as the judge for the MINE1 benchmark.

Model	Prompts Language	Avg. Extraction Time (mm:ss)	Entities	Relations	MINE1 Score (AVG)
Llama3.1	ENG	03:02	7131	5663	0.31
Llama3.1	ITA	03:01	7213	5172	0.38
Qwen3.6	ENG	06:42	10553	11729	0.41
Qwen3.6	ITA	06:49	11256	12417	0.44
Llama3.3	ENG	04:37	8665	10407	0.38
Llama3.3	ITA	04:32	8689	10976	0.42

Table 2

Extraction performance across candidate LLMs and prompt languages on the 17-document evaluation subset: average extraction time, extracted entity/relation counts, and MINE1 benchmark scores [7] computed against Italian and English LLM-as-a-judge prompts.

Term	utile pubblico	servizio pubblico	pubblico servizio	pubblici servizi	pubblici interessi	pubblici servizi	servizio pubblico	servizi pubblici	servizi pubblici	interessi pubblici
Cosine similarity	0.9746	0.9687	0.9682	0.9668	0.9663	0.9658	0.9653	0.9653	0.9648	0.9639

Table 3

A selection of terms closest to *pubblica utilità* in the embedding space (multilingual-e5-large-instruct).

remains stable. We combine semantic and topological analysis to study this and related concepts in the Norcia reconstruction debate.

4.1. Semantic Analysis

KGGen provides two automated entity-aggregation methods – SemHash and a hybrid clustering+LLM-judge approach – but inspection of the merged word clusters shows neither delivers satisfying results on our dataset. We therefore opt for manual thresholding on the four main keywords of our study, i.e. *Norcia*, *Pubblica utilità*, *Terremoto*, *Ricostruzione*: we embed all entities and rank them by cosine similarity to each keyword. Domain-expert supervision, increasingly necessary as content specificity and language variation grow, enables two further actions: (i) **Embedding-based disambiguation**: pairs with cosine similarity in $[0.985 - 1]$ are merged into a single node, with merged labels stored as node attributes. (ii) **Semantic field expansion**: the distance-1 subgraph on the four core terms (984 nodes, 5160 edges) is expanded by embedding similarity to include synonyms and related concepts (e.g. *interesse pubblico*), yielding 3046 nodes and 29710 edges. The disambiguation threshold is chosen by inspecting the top similar entities to a keyword, typically lexical variants; the expansion threshold is relaxed instead, capturing semantically close but distinct terms. For *pubblica utilità*, the terms in Tab. 3 show distinct valences distinguishable by context and frequency; the term appears predominantly in public/institutional discourse (royal decrees, municipal deliberations, administrative acts) and is nearly absent from private/commercial contexts, consistent with its role as a rhetorical-institutional operator in post-unification state discourse. Fig. 1b gives a complementary, qualitative view of Tab. 3.

4.2. Topological Analysis

Alongside semantic analysis, we inspect the graph’s topology, starting from the neighbourhood of the four keywords and their semantically close terms (see Sect. 4.1); we refer to each such entity as a *keyword*. Notably, this neighbourhood is not tree-like: cycles arise from relations between a keyword’s neighbours coming from documents other than those relating the neighbours to the keyword, letting the scholar see how contexts relate across documents. We further study how keywords relate via graph paths, restricting the analysis to the distance-1 induced graph so it remains explorable in Gephi[17], considering paths up to length 3. Path analysis surfaces concrete evidence for historian-hypothesised connections: in issue 2326 of the “La Gazzetta dell’Umbria”, a length-3 path links *Norcia* to “*opera di pubbliche utilità*”, confirming a connection the domain expert had presupposed but not yet verified – turning a hypothesis into a verifiable query at a fraction of the manual time it would require. Separately, a single-term neighborhood restricted to *pubblica utilità* alone shows strong connectivity (160 nodes, 220 edges), a quantitative indicator of the term’s salience that reduces the selection bias inherent to manually reading the full 3187-issue corpus.

5. Conclusions and Future Perspectives

We present a fully open-source, open-weight pipeline for extracting knowledge graphs from digitised historical newspapers, applied to the reconstruction of Norcia after the 1859 earthquake using the *Gazzetta dell’Umbria* as primary source. Combining schema-free LLM-based triple extraction with embedding-based disambiguation and semantic field expansion lets domain experts move from keyword search to structured, traceable queries. This is a preliminary step in an ongoing project; several directions remain open: (i) **Evaluation**. A manually annotated ground truth would enable a more rigorous comparison than the LLM-judged MINE1 benchmark [7] (Sect. 3); we also plan to benchmark transcription against alternative OCR/VLM engines (WER, CER). (ii) **Extraction pipeline**. Comparing KGen against other frameworks, e.g. ATOM [18] and LLM-Graph-Builder [19], would clarify how much of our results is attributable to the extraction system rather than the schema-free design; a more robust page-joining and chunking strategy (Sect. 2) would also help. (iii) **Graph analysis**. Community detection and centrality could surface actors beyond those selected a priori (Sect. 4.2); temporal information [20, 18] would let us trace how the discourse evolves over 1860–1870. (iv) **Generalization**. Our schema-free design is not specific to this source; applying it to other newspapers, document types, or periods would test how well our findings generalise.

Acknowledgments

The project was founded by Ce.Di.Pa. - PNC Programma unitario di interventi per le aree del terremoto del 2009-2016 - Linea di intervento 1 sub-misura B4 - “Centri di ricerca per l’innovazione” CUP J37G22000140001.

Declaration on Generative AI

Beyond its use within the described framework, Generative AI was solely employed to improve the writing quality of selected paragraphs during the preparation of this article.

References

- [1] S. Spina, From digitisation to computation: Digital history, invented archives, and the epistemology of historical knowledge in the age of artificial intelligence, *Journal of History Research* (2026). URL: <https://api.semanticscholar.org/CorpusID:285294682>.

- [2] Explorations in the Digital History of Ideas: New Methods and Computational Approaches, Cambridge University Press, 2023.
- [3] I. C. per il Catalogo Unico, Gazzetta dell'umbria, Emeroteca Digitale Italiana, 2026. URL: <https://www.internetculturale.it/it/913/emeroteca-digitale-italiana/periodic/testata/9886>.
- [4] T. Zhang, Y. Sun, R. Liu, Provenance analysis of archaeological artifacts via multimodal rag systems, 2025. URL: <https://arxiv.org/abs/2509.20769>. arXiv:2509.20769.
- [5] R. Loffredo, M. De Santo, Using ontologies for llm applications in cultural heritage, in: CEUR workshop proceedings, volume 3865, 2024, pp. 37–43.
- [6] C.-E. González-Gallardo, E. Boros, N. Girdhar, A. Hamdi, J. G. Moreno, A. Doucet, Yes but.. can chatgpt identify entities in historical documents?, in: 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2023, pp. 184–189. doi:10.1109/JCDL57899.2023.00034.
- [7] B. Mo, K. Yu, J. Kazdan, P. Mpala, L. Yu, C. Kanatsoulis, S. Koyejo, Kggen: Extracting knowledge graphs from plain text with language models, in: Advances in Neural Information Processing Systems, volume 38, Curran Associates, Inc., 2025, pp. 30092–30115.
- [8] G. Corona, D. Guidotti, L. Pandolfo, Constructing a knowledge graph for italian cinema divas' autobiographies 3865 (2024) 22–29.
- [9] M. Ruskov, M. Dagioglou, M. Kokol, S. Montanelli, G. Petasis, et al., A knowledge graph of values across space and time, in: CEUR Workshop Proceedings, volume 3536, CEUR-WS, 2023, pp. 8–20.
- [10] S. Giacomini, C. Martignano, G. Rubin, A. Bardi, M. Buzzoni, M. Daquino, R. Del Gratta, A. M. Del Grosso, F. Fischer, R. Rosselli Del Turco, et al., Atlas: A knowledge graph to enhance the findability of international scholarly research on the italian digital cultural heritage, Umanistica Digitale 10 (2026) 45–76. doi:10.60923/issn.2532-8816/21765.
- [11] A. Schimmenti, V. Pasqual, F. Vitali, M. van Erp, Knowledge graphs generation from cultural heritage texts: combining llms and ontological engineering for scholarly debates, Journal of Documentation 82 (2026) 206–250. doi:<https://doi.org/10.1108/JD-07-2025-0203>.
- [12] S. de Valeriola, Can historians trust centrality? historical network analysis and centrality metrics robustness, Journal of Historical Network Research 6 (2021) 85–125. doi:10.25517/JHNR.V6I1.105.
- [13] S. Hengchen, R. Ros, J. Marjanen, M. Tolonen, A data-driven approach to studying changing vocabularies in historical newspaper collections, Digital Scholarship in the Humanities 36 (2021) ii109–ii126. doi:10.1093/llc/fqab032.
- [14] H. Zheng, Y. Li, K. Zhang, L. Xin, G. Zhao, H. Liu, J. Chen, J. Lou, Q. Fu, R. Yang, S. Jiang, W. Luo, W. Su, W. Zhang, X. Zhu, Y. Li, Y. ma, Y. Chen, Y. Ji, Z. Yu, G. Yang, C. Zhang, L. Zhang, Y. Liu, X. Bai, Multimodal ocr: Parse anything from documents, 2026. URL: <https://arxiv.org/abs/2603.13032>. arXiv:2603.13032.
- [15] B. Zhang, H. Soh, Extract, define, canonicalize: An LLM-based framework for knowledge graph construction, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 9820–9836. doi:10.18653/v1/2024.emnlp-main.548.
- [16] K. Wataoka, T. Takahashi, R. Ri, Self-preference bias in llm-as-a-judge, arXiv preprint arXiv:2410.21819 (2024).
- [17] M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, in: Proceedings of the international AAAI conference on web and social media, volume 3, 2009, pp. 361–362. doi:10.1609/icwsm.v3i1.13937.
- [18] Y. Lairgi, L. Moncla, K. Benabdeslem, R. Cazabet, P. Cléau, Atom: Adaptive and optimized dynamic temporal knowledge graph construction using llms, arXiv preprint (2025). arXiv:2510.22590.
- [19] Neo4j, Llm knowledge graph builder, 2026. URL: <https://neo4j.com/developer/genai-ecosystem/llm-graph-builder/>.
- [20] C.-E. González-Gallardo, E. Boros, E. Giamphy, A. Hamdi, J. G. Moreno, A. Doucet, Injecting temporal-aware knowledge in historical named entity recognition, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 377–393.