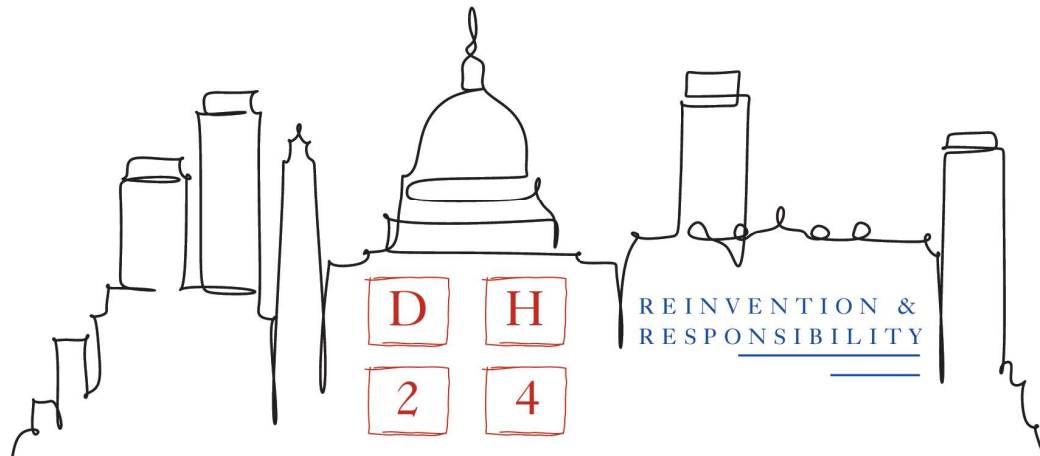
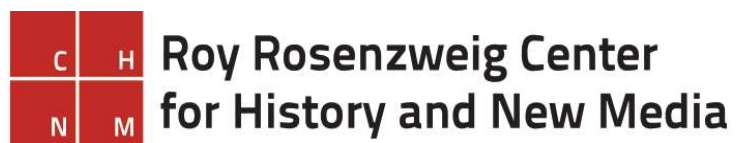




Book of Abstracts

August 6-9, 2024

Arlington, VA, U.S.A.



AI Models and Services for the Leopardi Digital Library

XML

1. Introduction

2. Giacomo Leopardi is one of the most important authors of Italian literature. Since 2017, the National Center for Leopardian Studies (Recanati), the National Library of Naples and the University of Macerata are leading a project aimed at the digitalization of the most comprehensive corpus of archival documents authored by or related to the author born in Recanati (Melosi & Marozzi, 2021). Currently, a large part of the material is digitized using open standards. An important challenge which still remains untackled is the design and application of Artificial Intelligence (AI) techniques from Natural Language Processing (NLP) and Computer Vision (CV) which could support both librarians and users. Thus, the main goal of this contribution is to provide a preliminary answer on which services and methodologies are needed to enhance the Leopardi Digital Library with AI and which challenges have to be faced in order to apply them responsibly and successfully.

3. Models and Services

4. One of the most fundamental tasks in the digitalization process of a manuscript is Handwritten Text Recognition (HTR). Currently, services such as Transkribus (Nockels et al. 2022) publicly provide HTR models trained on large corpora from the cultural heritage domain which are capable of recognizing texts from images. However, HTR is an important task since errors in this step can influence the performances of other NLP techniques applied on the extracted text. For this reason, Spell Checking systems have been used in order to enhance the quality of HTR-ed text, even though the effectiveness of these algorithms is still unclear (Huynh et al., 2020).

5. Another crucial task is Entity Linking (EL), which consists in linking references to entities and concepts in the text to a Knowledge Graph, such as Wikidata. Currently, few pre-trained models are available for multilingual texts (Ayoola et al. 2023); however, EL on historical and literary texts has shown to be more difficult, due to the different linguistic varieties and lack of contextual information. In order to improve EL, domain-specific Knowledge Bases using Wikipedia and Wikidata have been designed (Labusch & Neudecker, 2020).

6. In addition, another important task for which LMs are of help is Semantic Search. Usually, a Semantic Search system is based on a LM, e.g., a Sentence Transformer (Reimers & Gurevych 2019), which encodes each sentence of a document into a vector representation and a vector storage for similarity search. One of the advantages of using semantic search is that by leveraging the distributional semantics encoded by LMs, the search system of a library can return to the user relevant items based on a query without being affected by syntactic or spelling variations.

7. Design of an AI-based Architecture

8. *Figure 1* presents the design of an AI-based architecture for the Leopardi Digital Library integrating the models and services discussed. While the Web visualization presents the manuscript image along with its automatic transcription and metadata in XML, the search engine allows users to explore the Leopardian collection by looking for specific metadata or by querying specific sentences. Other essential features are public APIs to ingest both new documents and their manually curated transcriptions and linked open data to share digital objects and related metadata in the Semantic Web. An example of integration of TEI/XML editions with LOD standards was presented in (Giovannetti & Tomasi, 2022).

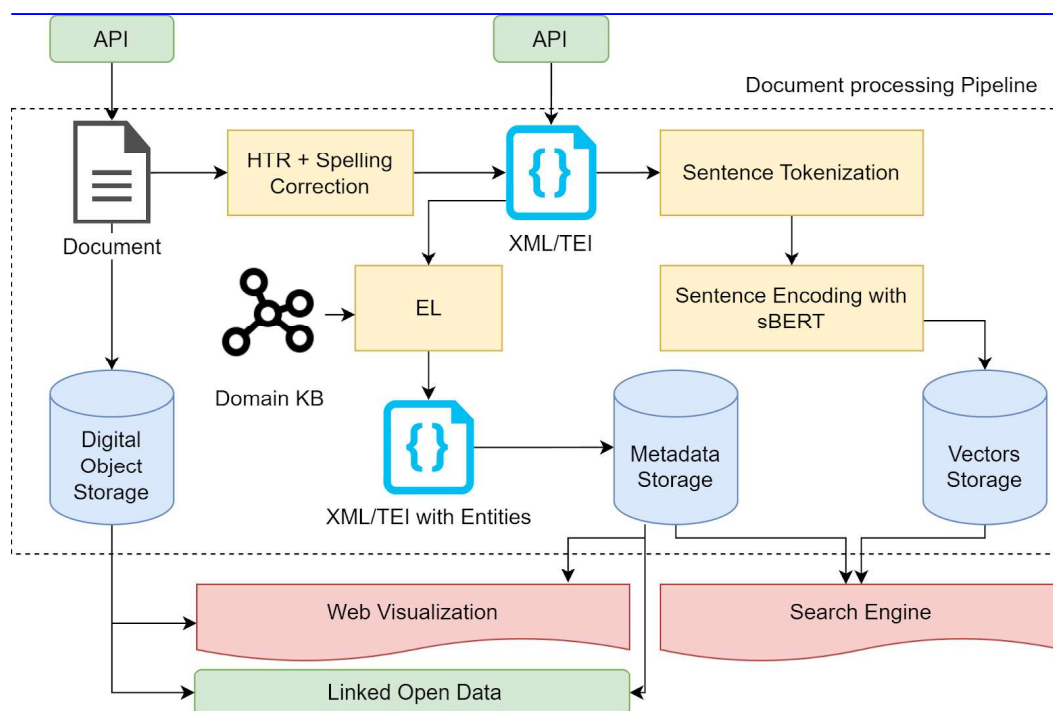


Figure 1. Figure 1: Integration of AI Models and Services in the Leopardi Digital Library.

Figure 1.

Conclusion

In conclusion, the integration of AI techniques into the Leopardi Digital Library project represents a crucial step towards the comprehensive digitization and enrichment of Giacomo Leopardi's literary legacy. The current focus on digitization using open standards has laid a solid foundation, but the potential for leveraging AI technologies opens new avenues for efficiency and accessibility in exploring this vast cultural heritage. The discussion has highlighted key AI-based services and methodologies that can significantly contribute to the enhancement of the digital library as well as how to integrate them together in order to provide user-friendly visualizations and powerful information retrieval.

As the Leopardi Digital Library project evolves, the challenges identified in implementing AI technologies must be met with a responsible approach. By addressing these challenges, the Leopardi Digital Library is poised to become a pioneering example of how AI can be effectively harnessed to unlock the full potential of cultural heritage, making it accessible to a global audience and preserving it for future generations.

Appendix A

Bibliography

- References
- Giovannetti, Francesca / Tomasi, Francesca** (2022): "Linked data from TEI (LIFT): A Teaching Tool for TEI to Linked Data Transformation", in: Digital Humanities Quarterly 016 (2).
- Huynh, Vinh-Nam / Hamdi, Ahmed / Doucet, Antoine** (2020): "When to Use OCR Post-correction for Named Entity Recognition?", in: 22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020. 33–42. DOI: 10.1007/978-3-030-64452-9_3.
- Johnson, Jeff / Douze, Matthijs / Jégou, Hervé** (2017): "Billion-scale similarity search with GPUs", arXiv, <<http://arxiv.org/abs/1702.08734>> [30.05.2024].
- Labusch, Kai / Neudecker, Clemens** (2020): "Named Entity Disambiguation and Linking Historic Newspaper OCR with BERT", in: Cappellato, Linda / Eickhoff, Carsten / Ferro, Nicola / Névél, Aurélie (eds.): Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum (= CEUR Workshop Proceedings). Thessaloniki, Greece: CEUR. (= CEUR Workshop Proceedings), <https://ceur-ws.org/Vol-2696/#paper_163> [30.05.2024].
- Limkonchotiwat, Peerat / Cheng, Weiwei / Christodouloupoulos, Christos / Saffari, Amir / Lehmann, Jens** (2023): "mReFinED: An Efficient End-to-End Multilingual Entity Linking System", in: Bouamor, Houda / Pino, Juan / Bali, Kalika (eds.): Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 15080–15089, <<https://aclanthology.org/2023.findings-emnlp.1007>> [30/05/2024].
- Melosi, Laura / Marozzi, Gioele** (2021): "Il progetto Biblioteca Digitale Leopardiana: per una catalogazione e digitalizzazione dei manoscritti autografi di Giacomo Leopardi", in: DigItalia 16 (1): 65–81. DOI: 10.36181/digitalia-00026.
- Nockels, Joe / Gooding, Paul / Ames, Sarah / Terras, Melissa** (2022): "Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research", in: Archival Science 22 (3): 367–392. DOI: 10.1007/s10502-022-09397-0.
- Reimers, Nils / Gurevych, Iryna** (2019): "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", arXiv, <<http://arxiv.org/abs/1908.10084>> [30/05/2024].

