Digital Heritage (2025)
S. Campana, D. Ferdani, H. Graf, G. Guidi, Z. Hegarty, S. Pescarin, and F. Remondino (Editors)

Enhancing the study of historical figures through AI-powered interactive data visualizations

Giovanni Profeta[1], Joseph Cornelius[2] and Fabio Rinaldi[2]

[1]University of Applied Sciences and Arts of Southern Switzerland SUPSI
[2]Dalle Molle Institute for Artificial Intelligence IDSIA USI-SUPSI

Abstract

One of the most important tasks for a historian is to identify key historical figures across multiple cultural archives and analyze their impact on history. The extensive effort of cultural institutions in digitizing historical archival materials and distributing them through online digital archives have significantly enhanced the study of historical figures. However, current historical digital archives, which rely on keyword-based search methods, often return numerous but imprecise results making it challenging for historians to understand chronological and contextual events surrounding a historical figure. We would like to present the result of the Mini-Muse project. It was a preliminary research project aimed at identifying data visualization models and user-friendly interface features to help historians visually explore historical figures and their actions. The project leverages Natural Language Processing (NLP) algorithms to extract metadata from unstructured text and generate structured data about key figures. It also applies data visualization techniques to support the visual analysis of each figure's timeline. The project adopts a user-centered design approach to ensure that the user interface features meet the needs of historians. It involves a pool of Swiss and Italian historians to gather insights on their research practices and validate a working prototype. The findings of the preliminary research project suggests that the introduction of an "action flow view", an interactive timeline displaying the historical figure's actions extracted automatically, can significantly improve the identification and study of historical figures.

CCS Concepts

• Information systems → Web interfaces; Users and interactive retrieval; Digital libraries and archives; • Human-centered computing → Human computer interaction (HCI); Graphical user interfaces; Interaction design; Visualization systems and tools; • Computing methodologies → Natural language processing;

1. Introduction

Identifying key historical figures and evaluating their impact on history is one of the most relevant tasks for a historian. It is extremely important for several societal needs including providing support for political decisions based on historical precedents [Ber18], constructing and refining historical narratives [ZA20], and their long-term impacts on subsequent generations [RNNA22]. Research on historical figures can be based on either primary or secondary sources. When focusing on primary sources, namely the original materials from the period or event being studied, the research methodology involves searching for, analyzing, and interpreting these sources to produce a written work, also called secondary source [SKKM22]. When focusing on secondary sources, namely textual interpretations and analyses of historical events based on primary sources, the research methodology adopts a critical analysis that includes contextual evaluation, cross-source comparison and synthesis of the findings [Thi03] [SK93]. In the study of historical figures through secondary sources, cultural digital archives play a key role by providing online access to digitized documents. However, these archives often suffer from usability issues [VKT10] [DCS*15] and frequently fail to meet the research needs of historians. We introduce the Mini-Muse research project, which aims to gather preliminary knowledge on how the combination of research coming from the data science and information design domains can overcome the current digital archives issues and enhance historical research. It integrates recent Natural Language Processing (NLP) methods with information visualization methodologies to produce experimental interactive user interfaces to foster the study of historical figures and their actions.

2. Methodology

The research project employed a user-centered methodology involving a pool of historians and heavy users of cultural digital archives to ensure the gathering of real needs and expectations. The project ran from October 2023 to September 2024 and consisted of three work packages: user research, design and implementation of a working prototype of AI-powered interactive data visualizations and user test. The user research allowed us to gather histori-

DOI: 10.2312/dh.20253150

ans' needs regarding the use of digital archives for conducting research on historical figures based on secondary sources. This work package essentially included one-to-one semi-structured interviews with a group of 14 people (expert historians and history students, archivists and documentalists) and the collection of features of interest for those people (see Table 1).

Feature	Request*
Getting extrinsic elements (about the publication, i.e. author and publication date)	Very high
Getting intrinsic elements (about the content of the publication, i.e. historical figure and locations mentioned, summary)	High
Getting the reference to the point where the information comes from	High
Getting results in the language selected by the user (translated if it is needed)	High
Advanced search (or filter) options (i.e. by date, and timespan)	Medium
Need for algorithm transparency (information about the methodology and the technology adopted to elaborate a filter/prompt and return an output)	Medium
Search or filter content in a specific language	Low
Getting the action flow: what happens during a period of time	Low
Compare different articles about the same topic	Low
Getting information about the rights of use of the document	Low
Getting links to other resources for expanding/improving the research (LOD principles)	Low
See articles with similar topic	Low
Index (or visual overview), with titles and authors, of the whole collection	Low
See the N-gram visualization	Very low
Getting other intrinsic elements (i.e. keywords, and mentions to primary source)	Very low

Table 1: *Features of interest for the 14 people we involved in the study.*
**The request column refers to how many interviewed users expressed that request (very high: 80%-100%, high: 60%-79%, medium: 40%-59%, low: 20%-39%, very low: less than 20%).*

The analysis of the features which emerged from the user interviews allowed us to define a set of two relevant use cases for the implementation of the working prototype:

A. Providing users with "action flows", namely all the actions a historical figure undertook as mentioned in a set of secondary sources
B. Providing users with custom summaries about a selected document.

The design and implementation of the working prototype allowed us to test the technical feasibility of the selected use cases and to evaluate the efficacy of the user interface in providing meaningful and relevant information. The work package included the implementation of a set of NLP algorithms, to automatically extract information from a set of articles from the Swiss Historical Journal (published quarterly by the Swiss History Society since 1951 and made available online by a service by the ETH Library called E-Periodica), and the design of a user interface that visually displays data extracted from a set of NLP algorithms. The user tests of the working prototype consisted of a set of online meetings with the experts involved in the user research and a follow-up anonymous survey. Through the user tests it was possible to gather feedback about the working prototype in terms of usability, efficacy in supporting the historians' main research tasks, and suggestions for further studies on how the integration of NLP algorithm and data visualization can enhance historical research.

3. NLP algorithms

We developed a pipeline approach incorporating a suite of NLP methods to build the working prototype. This pipeline consists of three main stages, each targeting specific analytical objectives aligned with our two use cases. First, we focused on extracting relevant information from the combined content of plain-text articles and their associated metadata files. This stage involved entity-level information extraction through the following methods:

- **Named Entity Recognition (NER)**: This method integrates several techniques, including rule-based date extraction using regular expressions, off-the-shelf pretrained language models (we use spaCy de_core_news_md v3.7.0) for identifying entities such as locations, dates, and persons, and dependency parsing combined with dictionary-based approaches to extract actions [FCR19].
- **Entity Linking**: Historical figures and other entities are linked using a hybrid method that combines rule-based strategies with vector-based similarity heuristics (Genea and Hofman, 2017).

Second, we aimed to identify higher-level, contextual information by analyzing the relationships between extracted entities—particularly focusing on actions, their associated actors, and the corresponding spatial and temporal contexts. This was achieved by using:

- **Action Flow Extraction**: To capture and structure sequences of actions within context, we applied NER techniques from the first stage alongside structured prompting strategies using large language models, more specifically GPT-4-0125-preview via the OpenAI API [AAA*25].

Third, we implemented methods for both summarization and article-specific question answering to support interpretability and user interaction:

- **Text Summarization**: This component includes both extractive summarization using Latent Semantic Analysis (LSA) [GL01] [OCA10] and abstractive summarization with large language models [BS23] [LSH*24], applied individually to each article.
- **Article-Based Closed Question Answering**: To enable users to pose detailed questions about individual articles, we employed retrieval-augmented LLMs [LPP*21] capable of generating precise, content-grounded answers.

This systematic integration of NLP algorithms facilitated detailed text analysis and enhanced the ability to extract meaningful insights from the texts. All annotations, extracted content, and summarized documents are efficiently stored within the Neo4j graph database to facilitate data retrieval and relationship mapping. Neo4j was selected because our main tasks consisted of the extraction of entities and their relationships from text, which are well-suited to graph-based representation. It also enables efficient querying of relational patterns and supports graph visualization for easier exploration. The system ensured seamless accessibility and integration of various annotations and services through a secure web API. The experiments were performed on a system equipped with an NVIDIA Tesla V100 (32 GB, PCIe) GPU and a dual-socket 40-core x86_64 CPU with 760 GB of RAM.

4. Interactive data visualizations

We developed a frontend consisting of two web pages designed to support the selected use cases: the action flow view and the article inquiry view.

- **Action flow view**, based on the use case A) providing users with "action flows": It allows users to see the action flows (all the actions a historical figure undertook), through an interactive timeline. In this view, the user can filter the action flows according to the type of historical figure (person or institution) and sort them according to a list of parameters, including the number of actions per figure and the completeness of the actions' metadata (see Figure 1).
- **Article inquiry view**, based on the use case B) providing users with custom summaries of a selected document: It allows users to ask questions about an article through a chatbot connected, via API, with the NLP algorithms. This view lists the whole set of analyzed articles with their related action flows, historical figures and locations, and an automatically generated short summary. The view allows users to sort items by publication date and author surname (see Figure 2).

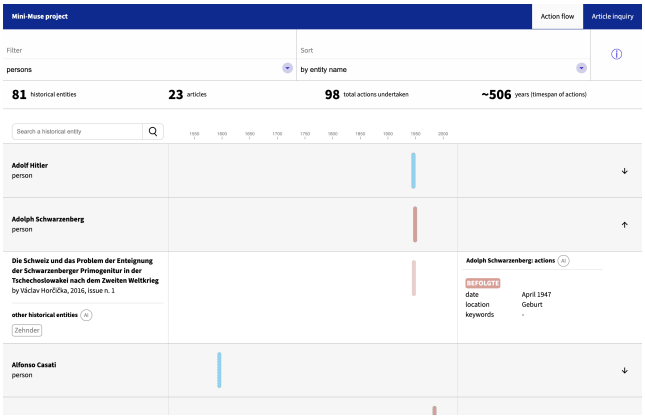

Figure 1: *The action flow view shows the lists of historical figures mentioned in the articles.*

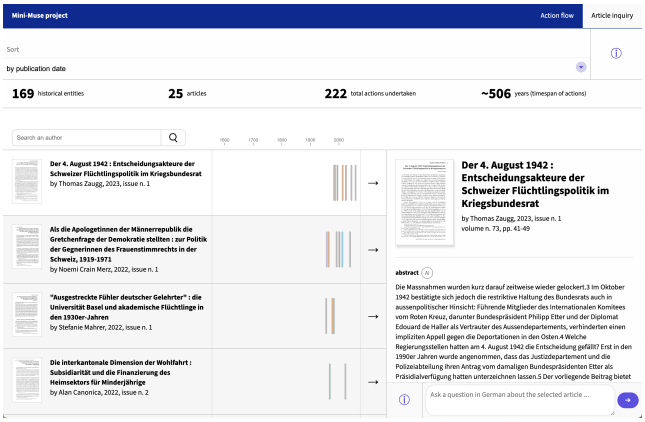

Figure 2: *The article inquiry view shows the full list of articles and allows users to ask questions about a selected one through a chatbot.*

5. Conclusions and future works

The Mini-Muse project allowed us to gather an overall understanding of possible research directions regarding the implementation of AI-assisted search tools to support historical research. According to the pool of experts, the most promising features of the working prototype were the list of articles mentioning a key historical figure, the chatbot to ask questions about articles, and the list of historical entities for every article. The most innovative aspects for historical research that require further studies were the color code to visually distinguish different types of actions (i.e. actions related to figures' movements and statements), the sort by completeness of the results/actions' metadata, and the AI-generated abstract if the one written by the author is not available. These results showed that AI extraction of action flows, and related graphical visualizations, can have a great impact on historical research activities in terms of both speeding up and enhancing the quality of the work. The action flow allows historians to easily find correlations among historical figures and relevant documents mentioning them and so fostering the comparison of different authors' perspectives on the same event/figure. We have also gathered evidence that the action flow, combined with a chatbot that allows users to ask questions about a selected article, can lead to the introduction, within cultural digital archives, of specific AI-powered interactive data visualizations to support historians' activities. In future works, we plan to investigate how to scale our NLP algorithms and data visualization models for larger collections of digitized publications and how to integrate the analysis of other historical entities, such as processes (i.e. urbanization, colonialism and decolonization) and flows of people and goods across territories (i.e. migration, slave trade and spice trade) for more advanced historical research.

Acknowledgments

This work was supported by Hasler Stiftung and was developed in partnership with ETH-Zurich Library.

References

[AAA*25] ACHIAM J., ADLER S., AGARWAL S., AHMAD L., AKKAYA I., ALEMAN FLORENCIA L., ..., ZOPH B.: Gpt-4 technical report. *arXiv preprint* (2025). doi:10.48550/arXiv.2303.08774. 2

[Ber18] BERRIDGE V.: Why policy needs history (and historians). *Health Economics, Policy and Law 13*, 3–4 (2018), 369–381. doi:10.1017/S1744133117000433. 1

[BS23] BASYAL L., SANGHVI M.: Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models, 2023. URL: https://arxiv.org/abs/2310.10449, arXiv:2310.10449. 2

[DCS*15] DANI A., CHATZOPOULOU C., SIATRI R., MYSTAKOPOULOS F., ANTONOPOULOU S., KATRINAKI E., GAROUFALLOU E.: Digital libraries evaluation: Measuring europeana's usability. In *Proceedings of the Research Conference on Metadata and Semantics Research* (09 2015), pp. 225–236. doi:10.1007/978-3-319-24129-6_20. 1

[FCR19] FURRER LENZ A. J., COLIC N., RINALDI F.: Oger++: hybrid multi-type entity recognition. *Journal of Cheminformatics 11*, 1 (2019), 7. URL: https://doi.org/10.1186/s13321-018-0326-3, doi:10.1186/s13321-018-0326-3. 2

[GL01] GONG Y., LIU X.: Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2001), SIGIR '01, Association for Computing Machinery, p. 19–25. URL: https://doi.org/10.1145/383952.383955, doi:10.1145/383952.383955. 2

[LPP*21] LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., TAU YIH W., ROCKTÄSCHEL T., RIEDEL S., KIELA D.: Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL: https://arxiv.org/abs/2005.11401, arXiv:2005.11401. 2

[LSH*24] LIU Y., SHI K., HE K. S., YE L., FABBRI A. R., LIU P., RADEV D., COHAN A.: On learning to summarize with large language models as references, 2024. URL: https://arxiv.org/abs/2305.14239, arXiv:2305.14239. 2

[OCA10] OZSOY M. G., CICEKLI I., ALPASLAN F. N.: Text summarization of turkish texts using latent semantic analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics* (USA, 2010), COLING '10, Association for Computational Linguistics, p. 869–876. 2

[RNNA22] RIVERO P., NAVARRO-NERI I., ASO B.: Who are the protagonists of history? exploratory study on historical relevance after completing compulsory secondary education in spain. *Social Sciences 11*, 4 (2022). URL: https://www.mdpi.com/2076-0760/11/4/175, doi:10.3390/socsci11040175. 1

[SK93] STEWART D. W., KAMINS M. A.: *Secondary research : information sources and methods*, 2nd ed ed. Applied social research methods series v. 4. SAGE, Newbury Park, Calif, 1993. 1

[SKKM22] SULISTYO W. D., KHAKIM M. N. L., KURNIAWAN B., MASROIR M. I.: Utilization of historical sites as a learning source based on outdoor learning. *KnE Social Sciences 2022*, 3rd International Conference on Geography and Education (ICGE) (2022), 289–300. URL: https://doi.org/10.18502/kss.v7i16.12174, doi:10.18502/kss.v7i16.12174. 1

[Thi03] THIES C. G.: A pragmatic guide to qualitative historical analysis in the study of international relations. *International Studies Perspectives 3*, 4 (02 2003), 351–372. URL: https://doi.org/10.1111/1528-3577.t01-1-00099, arXiv:https://academic.oup.com/isp/article-pdf/3/4/351/5133710/3-4-351.pdf, doi:10.1111/1528-3577.t01-1-00099. 1

[VKT10] VORA P., KOMURA N., TEAM S. U.: The n00b wikipedia editing experience. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration* (New York, NY, USA, 2010), WikiSym '10, Association for Computing Machinery. URL: https://doi.org/10.1145/1832772.1841393, doi:10.1145/1832772.1841393. 1

[ZA20] ZUL A., AISIAH A.: Historical empathy learning model for strengthening character education 2013 curriculum. In *Proceedings of the International Conference On Social Studies, Globalisation And Technology (ICSSGT 2019)* (2020), Atlantis Press, pp. 561–571. URL: https://doi.org/10.2991/assehr.k.200803.070, doi:10.2991/assehr.k.200803.070. 1