Interrogating the Machine:
A Human-AI Workflow for Structured Extraction From Historical Sources

Robin Cherix, University of Applied Sciences and Arts Western Switzerland, Switzerland
Melinda Fleury, University of Geneva, Switzerland
Elena Mugellini, University of Applied Sciences and Arts Western Switzerland, Switzerland

The European Conference on Arts & Humanities 2026
Official Conference Proceedings

Abstract

Text-analysis technology has evolved rapidly with the advent of Large Language Models (LLMs). As these tools reshape how we process language, the humanities face a pressing question: can LLMs meaningfully support—rather than replace—the historian's craft of extracting information from primary sources? Extracting precise data such as events, characters, and locations from historical texts remains a labor-intensive process that resists scaling. While computational approaches such as named entity recognition have long been applied to textual sources, they typically deliver outputs without offering historians the means to interrogate or refine the extraction process itself. We argue that a human-AI collaboration enables historians to process more sources, more systematically, while retaining interpretive control. Our approach automatically extracts events, characters, and locations from historical texts. Historians then review, validate, correct, and enrich the proposed data. They can also interact with the LLM through a chatbot interface to query the reasoning behind each extraction, preserving transparency and scholarly agency. To evaluate this approach, we invited historians to test our system, which implements Mistral Large 2.1, on a 16th-century historical source. Elements extracted by the LLM closely matched those identified manually by historians, demonstrating the tool's reliability. Qualitative feedback further indicated that this human-AI collaboration empowered historians to work at greater scale while maintaining a clear understanding of what was extracted and why, notably through the ability to request explanations from the model.

Keywords: digital humanities, human-machine interaction, human-AI collaboration

Introduction

The rapid evolution of natural language processing and Large Language Models (LLMs) has begun to fundamentally alter computational workflows across numerous academic disciplines. However, their integration into the humanities presents challenges. Primary source analysis remains an intrinsically qualitative, interpretive craft. Extracting structured entities such as historical actors, precise locations, and complex events from primary documents is a highly labor-intensive process that traditionally resists automatic scaling. Existing computational tools like standard Named Entity Recognition (NER) often present opaque, rigid outputs that leave little room for scholarly interrogation, risk stripping context, or fail to handle historical linguistic variations.

To address these challenges, this paper presents a human-AI collaborative framework designed to assist historians in extracting structured data while maintaining their interpretive agency. Developed in collaboration with scholars at the Institut d'Histoire de la Réformation (Institute of Reformation History, IHR), specialists in the early modern period with a particular focus on 16th-century France, this project explores how LLMs can serve as interactive, transparent research partners rather than uncritical automation engines.

We introduce a web-based proof-of-concept application leveraging Mistral Large 2.1 to automatically propose historical entities, which scholars can subsequently review, enrich, or challenge via a chatbot interface. To evaluate the efficacy and scholarly reception of this approach, we conducted a user study with six historians from the IHR. Participants performed comparative extraction tasks on 16th-century primary texts and evaluated the system's interactive features. The results demonstrate high alignment between LLM extractions and expert human annotations for characters and locations. Crucially, qualitative feedback confirmed that the ability to query the model's rationale significantly enhanced trust, transparency, and analytical rigor.

Literature Review

NER enables the identification of people, places, dates, and specific events in a text (Keber et al., 2024). Several different approaches and combinations of NER systems have been used in recent years, with significant progress made. Initially, these systems were mostly based on pattern matching or specific word recognition, e.g., a capitalized word was likely a person or a place, a number might be a date (Keber et al., 2024).

However, the current research in this field is increasingly moving towards deep learning, allowing the detection of mentions of actors or events even when they are not explicitly named. This ability is particularly important with regard to the context of this study, where a complete and nuanced extraction of all relevant entities, including the implicit ones, is mandatory. Several works mention the use and combination of existing NER models such as BiLSTM-CRF (Chen et al., 2020; Li et al., 2023; Yan & Li, 2023), BiGRU-CRF (Zhang et al., 2021), RoBERTa (V. D. Lai et al., 2021), and ERNIE (Zhang et al., 2021). All four show satisfactory results, but they miss on important factors such as implicit references, language evolution and context. For instance, “the king” does not refer to the same person if the text was written in France in 1568 or in England in 1412. In recent years, deep learning approaches have shifted toward LLMs, which have proven highly effective for text-processing tasks while enabling natural language interaction. Naturally, the paradigm changed from human-machine interaction to human-machine collaboration.

The collaboration between humans and AI represents an ongoing endeavor that has gained substantial traction in both research and industrial communities over the last five years (Berretta et al., 2023; Puerta-Beldarrain et al., 2025). The underlying ambition is to combine the scalability and computational throughput of AI systems with the contextual understanding and domain expertise of human practitioners (Puerta-Beldarrain et al., 2025).

Such collaborative paradigms have been investigated across a number of concrete application domains, with the relative weight of the human and the AI varying with the stakes and the type of expertise involved. In healthcare, it has been proposed as a means of mitigating the shortage of qualified workers and the heavy workload of clinicians (Y. Lai et al., 2021).

Recent foundation models such as LEADS have shown measurable gains in recall and time savings when clinicians collaborate with AI to mine medical literature (Wang et al., 2025), and patient-centered tools such as WoundAIssist illustrate the same principle in routine clinical care (Borst et al., 2025). Beyond healthcare, decision-support studies cluster around finance, business, and law (Gomez et al., 2025), while related work has demonstrated the benefit of human-AI collaboration for managing large biological-specimen collections (Stenhouse et al., 2023) and for information extraction from biomedical and legal documents (Mondal et al., 2023). To the best of our knowledge, however, the application of human-AI teaming to the humanities remains largely unexplored: in the systematic review of 108 AI-assisted decision-making studies by Gomez et al. (2025), humanities does not appear among the surveyed domains.

Bringing human-AI collaboration to the humanities offers significant opportunities. Scholars working on historical sources routinely face vast corpora that are simply too large for exhaustive close reading, yet whose interpretation requires a level of expertise that no general-purpose model can be expected to match. Through a human-AI collaboration approach, AI capacities can be leveraged to manage such corpora at scale while scholars retain control over the analytical process and over the interpretation of results (El-Hajj et al., 2023; Natarajan et al., 2025). By addressing humanities, we open human-AI collaboration to a field left unexplored so far. However, this opportunity comes with specific challenges. Historical texts often differ markedly from their contemporary counterparts: Old French, for instance, is sensibly different from modern French, with non-standardized orthography and grammatical rules that are still nascent and fluctuate across time and region. Mainstream language models, trained predominantly on contemporary corpora, may therefore misinterpret such material.

Adapting AI to these specificities is itself an active line of research, with dedicated models such as BERTrade for Old French (Grobol et al., 2022) and bilingual general-purpose models such as CroissantLLM for French and English (Faysse et al., 2025), as well as analogous efforts for other historical languages, including a BERT-CRF model trained on Rabbinic Hebrew that outperforms its modern-Hebrew counterpart on reference extraction in Jewish Responsa literature (Ben-Gigi et al., 2025).

The humanities' community has not waited for the formal vocabulary of human-AI teaming to start building such workflows in practice. A long-standing line of work on text reuse and intertextuality detection has produced a steady succession of tools that scholars use to navigate large historical corpora, from early n-gram fingerprinting approaches (Ganascia et al., 2014), to sequence-alignment systems such as TextPAIR and BLAST-based methods deployed on 18th-century French corpora (Roe, 2024; Roe et al., 2023) and on the cross-lingual reuse of

classical Chinese sources by Matteo Ricci (McManus et al., 2025), and up to semantic level based on word embeddings and Word Mover's Distance (Scheirer et al., 2016; Schubert, 2020).

Building upon these foundational efforts, our work connects automatic entity extraction with a human-AI collaborative loop designed specifically for historical texts. Rather than treating information extraction as a static, fully automated pipeline, we position the language model as an interactive assistant that proposes, explains, and adjusts entity extractions under the direct supervision of historical domain experts.

Methodology

To demonstrate LLMs can be used in a human-AI collaboration paradigm to extract structured data from primary sources, we built a web application embedding a chatbot linked to an LLM. This LLM is in charge of extracting the data from primary sources uploaded to the web application by historians. Once extracted, the data can be visualized in different ways from the web application.

Model Selection

Our objective was to establish an automated proof-of-concept pipeline capable of identifying characters, locations, and events within 16th-century French historical texts. To determine the most suitable architecture for this task, we defined four key evaluation criteria: zero-shot foundation model, no fine-tuning requirements, broad linguistic understanding, especially French, and accessibility. First, the model required strong zero-shot reasoning to operate effectively without initial retraining. Second, to ensure generalizability and minimize technical overhead within humanities workflows, the system needed to function without extensive domain-specific fine-tuning. Third, the model had to exhibit broad linguistic comprehension, seamlessly navigating non-standardized spelling and archaic syntax across both 16th-century and contemporary French. Finally, to support sustainable academic deployment, the underlying model needed to remain cost-effective and accessible.

Based on these prerequisites, we selected Mistral Large 2.1. It powers an automated extraction pipeline that parses ingested historical texts, outputs structured entities (characters, places, and events), and serves as the backend logic for an interactive visualization and annotation platform.

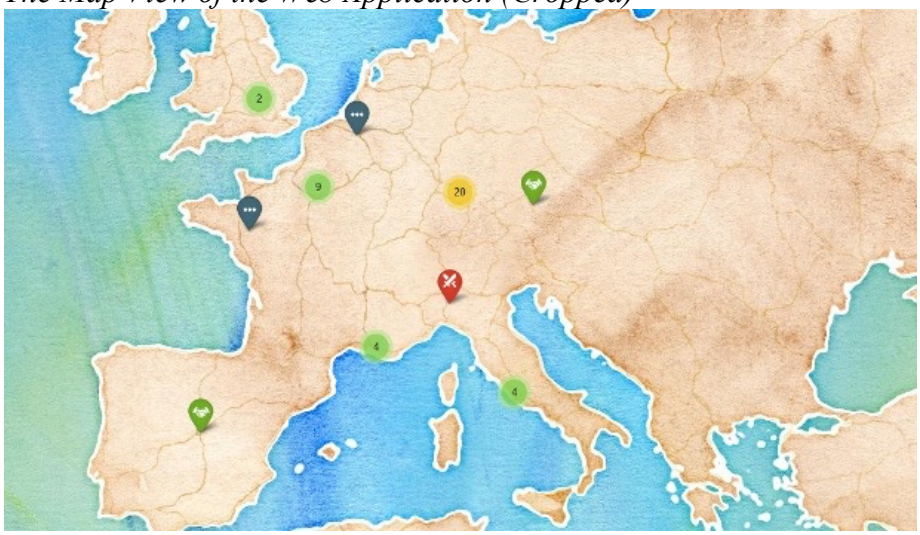
Web Application

To support human-AI teaming, we designed a web-based application featuring an integrated backend processing pipeline and an interactive front-end interface. Primary sources uploaded to the platform trigger the Mistral Large 2.1 extraction pipeline.

Following extraction, users interact with the historical data through three complementary visualization perspectives. In the Map View (cf. Figure 1), events are geographically situated on an interactive cartographic interface that uses icons for event categorizations and features a temporal slider for period-based filtering. The Timeline View complements this spatial representation by ordering events chronologically to accentuate temporal clusters and sequential relationships. Lastly, the Relation Graph View abstracts connections among characters, locations, and events into an interactive network, enabling scholars to filter specific entities and evaluate complex relational structures.

Figure 1

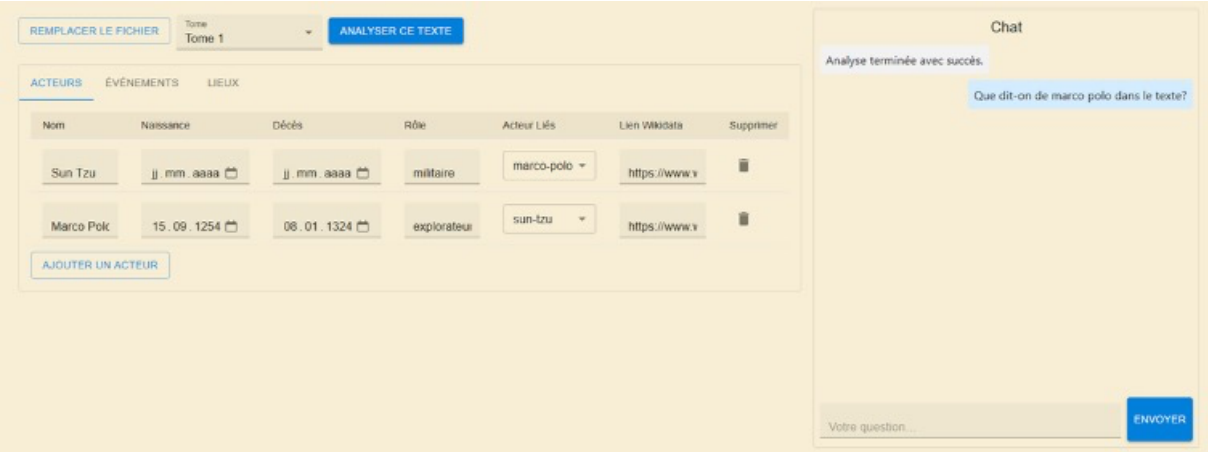
The Map View of the Web Application (Cropped)



To maintain transparency and scholarly agency, the web application features an interactive workspace sidebar containing a chatbot connected to the LLM. Historians can directly query the model about its decisions (e.g., asking "Why was this entity classified as an event?" or "What context supports this location reference?"). Furthermore, scholars retain full editing control: they can directly update or delete model-generated extractions, or manually add new entities missed by the automated pipeline. This page is shown on Figure 2.

Figure 2

The Data Extraction View, With the Chatbot Interface on the Right



User Study

To evaluate the system's effectiveness and usability, we conducted a user study with six professional historians specializing in the early modern period from the Institut d'histoire de la Réformation (IHR). The evaluation sessions were conducted remotely via Zoom using a structured protocol:

1. App Familiarization: Participants were introduced to the web interface and its visualization tools.
2. Manual Extraction Task: Historians read a selected 16th-century primary text and manually identified characters, locations, and events.
3. AI-Assisted Task: Historians performed the same extraction process on a text using the

LLM-assisted workflow, reviewing and refining the proposed entities.

4. Comparative Analysis: The manually extracted entities, model-only extractions, and human-validated extractions were systematically compared.

Throughout the session, participants were encouraged to "think aloud" to capture spontaneous qualitative feedback regarding trust, interface interaction, and analytical workflow. Following the tasks, we conducted semi-structured interviews to gather deeper insights into their experiences with the chatbot interface and human-AI collaboration.

Results

The user study yielded key insights into how LLMs handle complex historical texts compared to expert human analysis. Strong alignment was observed between the LLM and the historians regarding the extracted characters. Six out of the seven characters identified manually by at least two historians were also successfully extracted by the LLM pipeline. The model demonstrated high accuracy with locations entities. For example, a location identified by five out of six historians was independently extracted by the model. Additionally, a location extracted in different ways by two historians ("Étang" and "Marais") was extracted as one combined location: "Étang près d'un marais." Event extraction proved to be the most complex and contested task. Manual human extraction exhibited high variability, ranging from scholars who extracted zero distinct events to those who considered "the whole text is an event." Consequently, strict automated matching for events was less distinct, reflecting fundamental differences in how scholars conceptualize historical occurrences. The extraction comparison between historians and the LLM is shown on Table 1. The interactive dialogue feature directly impacted analytical outcomes: two out of the six participants altered or refined their extracted data after querying the LLM through the chatbot interface to inspect its rationale.

Table 1

Automatic LLM Extraction; in Green the Entities Found by at Least One Historian

Characters	Locations	Events
Le Roi	Thoul sur Mozelle	Conflit avec le Marquis Albert
L'évêque de Bayonne	Étang près d'un marais	Séjour du Marquis Albert près de Thoul sur Mozelle
Le Duc D'aumalle		
Marquis Albert		
Les Germains		
Les Français		

Discussion

The findings from our study illustrate both technical and conceptual aspects of integrating AI into historical workflows. The discrepancy in event extraction highlights a core domain reality: historical "events" lack a universal, rigid definition and is tied to individual interpretation. While computational models rely on boundary definitions, human scholars interpret events through varying historiographical lenses, ranging from discrete granular actions to broad narrative contexts. To a lesser extent, geographical locations present similar interpretative flexibility. Although it can be a source of debate in the community, we argue that the integration of AI in the process of primary source analysis does not impact the interpretative nature of the process, as AI suggests extraction, but historians keeps its full scholarly agency.

Rather than viewing model variances as extraction errors, our results suggest that an LLM functions effectively as a “second point of view.” Much like collaborating with a peer researcher, interacting with the model provides a secondary, probabilistic perspective informed by its training data.

Qualitative feedback from the IHR historians was positive, particularly regarding the system's collaborative balance. Scholars explicitly appreciated that the tool did not impose opaque outputs. The ability to interrogate the LLM via the conversational interface, asking why an extraction was made or where implicit context was derived, transformed the model from an unreliable “black box” into an interactive source of information, successfully preserving scholarly control and enhancing critical engagement.

Conclusion

In this paper, we introduced a human-AI collaborative framework tailored for historical text analysis, evaluated through a proof-of-concept application built with Mistral Large 2.1 and tested alongside domain experts from the Institut d'histoire de la Réformation. Our findings demonstrate that while LLMs achieve high extraction fidelity for concrete entities like characters and locations, their primary value in complex tasks, such as event identification, lies in their role as an interactive partner rather than a fully autonomous reader. By pairing automated extraction with visual analytics and an explainable, dialogue-based query interface, this approach allows historians to scale their analytical reach across large primary source corpora without sacrificing interpretive rigor or scholarly authority.

Acknowledgements

The web app and user study were realized as part of Pierre Masserey’s Master thesis at the University of Fribourg, Switzerland. This work was applied to the project ARCHEO-POL led by Pr. Paul Alexis Mellet (IHR) and funded by the Swiss National Science Foundation

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The author declares that Gemini was used for redaction help, improving phrasing and fixing grammar. The content, ideas, design, procedures, findings, analyses, and discussion are originally written and derived from careful and systematic conduct of the research.

References

Ben-Gigi, N., Zhitomirsky-Geffet, M., Schler, J., & Katzoff, B. (2025). Automatic Construction of the Citation Network from the Medieval Jewish Responsa Literature. *Journal on Computing and Cultural Heritage*, *18*(2), 1–18. https://doi.org/10.1145/3712013

Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., & Kluge, A. (2023). Defining human-AI teaming the human-centered way: A scoping review and network analysis. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1250725

Borst, V., Riedmann, A., Dege, T., Müller, K., Schmieder, A., Lugrin, B., & Kounev, S. (2025). *WoundAIssist: A Patient-Centered Mobile App for AI-Assisted Wound Care With Physicians in the Loop* (arXiv:2506.06104). arXiv. https://doi.org/10.48550/arXiv.2506.06104

Chen, Y., Ding, Z., Zheng, Q., Qin, Y., Huang, R., & Shah, N. (2020). A History and Theory of Textual Event Detection and Recognition. *IEEE Access*, *8*, 201371–201392. https://doi.org/10.1109/ACCESS.2020.3034907

El-Hajj, H., Eberle, O., Merklein, A., Siebold, A., Shlomi, N., Büttner, J., Martinetz, J., Müller, K.-R., Montavon, G., & Valleriani, M. (2023). Explainability and transparency in the realm of digital humanities: Toward a historian XAI. *International Journal of Digital Humanities*, *5*(2), 299–331. https://doi.org/10.1007/s42803-023-00070-1

Faysse, M., Fernandes, P., Guerreiro, N. M., Loison, A., Alves, D. M., Corro, C., Boizard, N., Alves, J., Rei, R., Martins, P. H., Casademunt, A. B., Yvon, F., Martins, A. F. T., Viaud, G., Hudelot, C., & Colombo, P. (2025). *CroissantLLM: A Truly Bilingual French-English Language Model* (arXiv:2402.00786). arXiv. https://doi.org/10.48550/arXiv.2402.00786

Ganascia, J.-G., Glaudes, P., & Del Lungo, A. (2014). Automatic detection of reuses and citations in literary texts. *Literary and Linguistic Computing*, *29*(3), 412–421. https://doi.org/10.1093/llc/fqu020

Gomez, C., Cho, S. M., Ke, S., Huang, C.-M., & Unberath, M. (2025). Human-AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science*, *6*. https://doi.org/10.3389/fcomp.2024.1521066

Grobol, L., Regnault, M., Suárez, P. O., Sagot, B., Romary, L., & Crabbé, B. (2022). BERTrade: Using Contextual Embeddings to Parse Old French. *13th Conference on Language Resources and Evaluation (LREC 2022)*, 1104–1113. https://doi.org/10.5281/ZENODO.6461220

Keber, M., Grubišić, I., Barešić, A., & Jović, A. (2024). A Review on Neuro-symbolic AI Improvements to Natural Language Processing. *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, 66–72. https://doi.org/10.1109/MIPRO60963.2024.10569741

Lai, V. D., Nguyen, M. V., Kaufman, H., & Nguyen, T. H. (2021). Event Extraction from Historical Texts: A New Dataset for Black Rebellions. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 2390–2400). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.211

Lai, Y., Kankanhalli, A., & Ong, D. (2021). *Human-AI Collaboration in Healthcare: A Review and Research Agenda*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2021.046

Li, Y., Geng, C., & Yang, D. (2023). Research on Chinese Event Extraction Method with No-trigger. *2023 6th International Conference on Computer Network, Electronic and Automation (ICCNEA)*, 91–95. https://doi.org/10.1109/ICCNEA60107.2023.00028

McManus, S. M., Li, Y., Tam, L., Qiu, S., Liu, S., Ng, D., Wong, W., & Yu, L. (2025). Measuring accommodation: A quantitative text-reuse analysis of Matteo Ricci's reliance on chinese classics. *International Journal of Digital Humanities*. https://doi.org/10.1007/s42803-025-00104-w

Mondal, I., Yuan, M., N, A., Garimella, A., Ferraro, F., Blair-Stanek, A., Van Durme, B., & Boyd-Graber, J. (2023). *InteractiveIE: Towards Assessing the Strength of Human-AI Collaboration in Improving the Performance of Information Extraction* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2305.14659

Natarajan, S., Mathur, S., Sidheekh, S., Stammer, W., & Kersting, K. (2025). Human-in-the-loop or AI-in-the-loop? Automate or collaborate? *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'25/IAAI'25/EAAI'25*, *39*, 28594–28600. https://doi.org/10.1609/aaai.v39i27.35083

Puerta-Beldarrain, M., Gómez-Carmona, O., Sánchez-Corcuera, R., Casado-Mansilla, D., López-de-Ipiña, D., & Chen, L. (2025). A Multifaceted Vision of the Human-AI Collaboration: A Comprehensive Review. *IEEE Access*, *13*, 29375–29405. https://doi.org/10.1109/ACCESS.2025.3536095

Roe, G. (2024). Text reuse as cultural practice: Intertextuality in the 18th-century digital archive. *Digital Enlightenment Studies*, *2*(1). https://doi.org/10.61147/des.23

Roe, G., Fedchenko, V., & Nicolosi, D. (2023). *Enlightenment Inflluencers: Networks of Text Reuse in 18th-century France*. https://doi.org/10.5281/ZENODO.8107964

Scheirer, W., Forstall, C., & Coffee, N. (2016). The sense of a connection: Automatic tracing of intertextuality by meaning. *Digital Scholarship in the Humanities*, *31*(1), 204–217. https://doi.org/10.1093/llc/fqu058

Schubert, C. (2020). Intertextuality and Digital Humanities. *It - Information Technology*, *62*(2), 53–59. https://doi.org/10.1515/itit-2019-0036

Stenhouse, A., Fisher, N., Lepschi, B., Schmidt-Lebuhn, A., Rodriguez, J., Turco, F., Toms, E., Reeson, A., Paris, C., & Thrall, P. (2023). Improving Biological Collections Data through Human-AI Collaboration. *Biodiversity Information Science and Standards*, *7*. https://doi.org/10.3897/biss.7.112488

Wang, Z., Cao, L., Jin, Q., Chan, J., Wan, N., Afzali, B., Cho, H.-J., Choi, C.-I., Emamverdi, M., Gill, M. K., Kim, S.-H., Li, Y., Liu, Y., Ong, H., Rousseau, J., Sheikh, I., Wei, J. J., Xu, Z., Zallek, C. M., … Sun, J. (2025). *A foundation model for human-AI collaboration in medical literature mining* (arXiv:2501.16255). arXiv. https://doi.org/10.48550/arXiv.2501.16255

Yan, C., & Li, S. (2023). Exploring Multiview Interactive Visualization for Chinese Historical Texts Using Knowledge Aggregation. *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 165–168. https://doi.org/10.1109/JCDL57899.2023.00031

Zhang, J., Jing, X., Zheng, J., & Shi, B. (2021). A Chinese Document-level Event Extraction Method based on ERNIE. *2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT)*, 13–17. https://doi.org/10.1109/CECIT53797.2021.00011

Contact email: robin.cherix@hes-so.ch