

A Multi-Agent Methodological Framework Enabling Knowledge-Based Digital Research Platforms for Cultural Heritage

Michele Colombino¹, Matteo Miceli¹, Laurentiu Jr Marius Zaharia¹,
Francesca Davida Pizzigoni² and Fabio Ciravegna¹

¹Dipartimento di Informatica, Università di Torino, Torino, Italia

²Dipartimento di Filosofia e Scienze dell'Educazione, Università di Torino, Torino, Italia

Abstract

Historical research in Cultural Heritage relies heavily on digitized primary sources, yet their heterogeneous structures, complex visual layouts, and evolving terminology make systematic organization, extraction, and comparison a time-consuming challenge. This complexity is particularly evident in the History of Education and the study of school materiality, a field that analyzes educational supplies as primary sources to reconstruct pedagogical practices over time. Within this discipline, historical commercial catalogues of school materials provide valuable evidence, documenting objects alongside their pedagogical, industrial, and commercial contexts. However, fully leveraging these collections requires methodological approaches capable of systematically organizing and extracting knowledge from unstructured documents. This paper proposes a methodological framework for transforming historical documentary collections into structured, traceable, and queryable knowledge bases. The framework combines methodological design principles with a multi-agent Generative AI (GenAI) architecture that preserves documentary semantics, source traceability, provenance, and supports the continuous evolution of the resulting knowledge base. We present Cata-LOH (Catalogues for Researching Learning Objects and their History), a real-world digital research platform developed using the proposed framework, illustrating how the framework can be applied to the analysis of historical educational-supply catalogues.

Keywords

Generative AI, Multi-Agent Systems, Cultural Heritage, Information Extraction, Knowledge Bases

1. Introduction

Historical research is fundamentally grounded in the analysis of primary sources. Regardless of the specific documentary corpus under investigation, rigorous historical inquiry relies on a set of well-established core principles: Primary Sources, Historical Contextualisation, the Central Role of the Historian, Critical Interpretation, and the Cross-analysis of Multiple Sources [1, 2, 3]. Together, these principles define the epistemological foundations of historical research, ensuring that historical knowledge emerges not simply from the collection of documentary evidence but from its critical interpretation within broader historical contexts [4, 5].

However, applying these principles to Cultural Heritage (CH) documentary corpora introduces challenges that increasingly exceed what traditional research methods can address at scale. Historical evidence is often fragmented across documentary collections, digitized documents remain unstructured and difficult to query, and preliminary source analysis is highly time-consuming [6, 7].

These challenges are particularly evident in the History of Education and, more specifically, in the study of school materiality — a research field that analyzes educational objects as primary sources to reconstruct the evolution of pedagogical practices over time. Within this discipline, historical commercial catalogues of school materials — documenting educational objects alongside their pedagogical,

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

✉ michele.colombino@edu.unito.it (M. Colombino); m.miceli@unito.it (M. Miceli); laurentiujrmarius.zaharia@unito.it (L. J. M. Zaharia); francescadavida.pizzigoni@unito.it (F. D. Pizzigoni); fabio.ciravegna@unito.it (F. Ciravegna)

🆔 0009-0007-3248-1661 (M. Colombino); 0009-0008-6121-8368 (M. Miceli); 0009-0002-3559-8367 (L. J. M. Zaharia); 0000-0002-9117-4027 (F. D. Pizzigoni); 0000-0001-5817-4810 (F. Ciravegna)



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

industrial, and commercial contexts — represent a growing area of scholarly interest [8] yet lack computational methodological frameworks capable of processing their heterogeneous structures, complex layouts, and evolving terminology at scale.

Addressing these challenges requires enabling the exploration, extraction, and analysis of the knowledge contained in historical documentary corpora by recognising the intrinsic structure of documentary sources, identifying historical objects, actors, concepts, and temporal references, and making explicit the relationships among them. Such a representation enables historians to formulate complex queries across documents, connect and compare historical evidence, and support the construction of new historical interpretations.

In this paper, we present a methodological framework for transforming historical documentary collections into structured, traceable, and queryable knowledge bases. The framework preserves primary-source verification, reduces the effort required for source analysis, and supports comparative investigations across distributed documentary collections.

We used the proposed framework to develop a real-world application, Cata-LOH (Catalogues for Researching Learning Objects and their History), a digital research platform supporting the study of historical educational-supply catalogues. Specifically, Cata-LOH targets research in school materiality—a subfield of the History of Education focused on analyzing physical objects (e.g., desks, maps, scientific instruments) as primary evidence of evolving pedagogical practices. The current corpus comprises Italian, Spanish, and French catalogues published between the nineteenth and twentieth centuries, which constitute a primary source for research on the material culture of schools within the History of Education, documenting educational materials together with their pedagogical, industrial, and commercial contexts.

2. Related Work

Historical research has traditionally been a process of reconstructing the past through the critical examination of primary sources. Historians build interpretations by collecting, comparing, and contextualizing evidence from multiple sources [1, 2, 3]. This activity requires navigating archival collections, identifying relevant materials, and reconstructing relationships across dispersed and heterogeneous documents before drawing historically grounded conclusions. Consequently, historical research is inherently iterative and time-intensive, particularly when relevant evidence must be identified within large and geographically dispersed documentary collections [7, 9].

The expansion of digitization has substantially improved access to historical sources, but it has not reduced the cognitive work required to transform documents into historical knowledge [6]. Instead, digitization has shifted the main challenge from accessing sources to efficiently exploring, analysing, and interpreting large knowledge bases. This transformation has contributed to the emergence of Digital Humanities (DH). Computational approaches are integrated into humanities research to support activities such as transcription, metadata creation, information retrieval, and corpus analysis [10, 11].

Recent advances in Generative AI (GenAI), have extended computational support from document organization and retrieval toward the analysis and interpretation of historical sources. AI-based approaches can assist historians in extracting structured information, identifying entities and relationships, generating semantic representations, and interacting with large documentary collections [12, 13, 14]. Moreover, recent studies demonstrate the feasibility of applying these methods to historical corpora, showing that computational approaches can reduce the effort required for large-scale document analysis [15, 16].

Recent research has started to investigate methodological approaches for integrating AI into historical research workflows, emphasizing the importance of maintaining connections between computational methods, documentary evidence, and historians’ interpretative practices [17, 18, 19].

However, applying AI to historical documents remains challenging due to the specific characteristics of these archival sources. Historical documents often contain degraded text, irregular layouts, complex page structures, and OCR errors that affect downstream analysis tasks such as information extraction

and entity recognition [20, 21, 22].

In addition, historical sources frequently include linguistic variations, obsolete terminology, antiquated layout organisations, and ambiguous references that complicate the identification and linking of information within and across documents [23]. Recent GenAI approaches have shown potential for extracting structured information from historical sources, but their reliability depends on document quality, extraction constraints, and mechanisms for controlling incorrect or unsupported results [24].

Taken together, these works highlight a shared limitation: existing AI-assisted approaches to historical document analysis tend to address specific technical challenges in isolation — OCR quality [20], information extraction [25], entity linking [22] — without providing a unified methodological foundation for transforming heterogeneous documentary collections into structured, traceable, and continuously evolving knowledge bases. Bridging this gap requires frameworks that go beyond technical solutions, integrating the epistemological foundations of historical inquiry into the design of AI-assisted research systems.

3. Design Principles

The proposed design principles originate from a close observation of how historians conduct research on documentary sources. Historical inquiry is driven by research questions whose answers are distributed across heterogeneous documentary collections, requiring historians to continuously interpret, compare, connect, and verify evidence. More specifically, the principles were not derived from a formal or systematic user study, but emerged iteratively from discussions, observations, and practical work with the research group of the Department of Philosophy and History of Education at the University of Turin. The collaboration involved researchers with complementary expertise in Computer Science, Digital Humanities, and History of Education, including domain experts trained in the philological and historical analysis of commercial catalogues. While methodologically rigorous, this process was time-consuming and impossible to scale across large collections. These limitations motivated the formulation of five methodological principles to guide the development of a GenAI-assisted system capable of supporting the manual workflow at scale and retaining the domain expertise and standards underlying it. The principles therefore define the requirements for a historical research system capable of producing trustworthy, traceable, interoperable, and continuously evolving historical knowledge bases.

Preserve Documentary Semantics. The meaning of historical documents extends beyond their conceptual content. Editorial features such as layout, sections, tables, columns, hierarchies, graphical elements, and publisher references encode contextual information that shapes how historical evidence should be interpreted. For example, historical commercial catalogues combine multiple textual and visual components within the same page. Titles, descriptive paragraphs, tables, illustrations, and captions contribute to the organisation and interpretation of this type of documents (see Figure 1). Ignoring these structural characteristics by reducing documents to plain text inevitably removes part of the document context, leading to ambiguous interpretations, incorrect associations, and the loss of relevant information. This principle requires document processing pipelines to preserve the structural organization of the original source, treating document structure as semantic information throughout the extraction process. The resulting structured representation must remain faithful to the original source, enabling reliable interpretation and reuse of historical knowledge.

Preserve Documentary Traceability. Primary sources constitute the foundation of historical knowledge. Knowledge extracted from these sources should preserve an explicit, persistent, and verifiable connection to the original document. Traceability extends beyond maintaining a direct link to a page or image; Relate extracted knowledge to its original context concerns historians to conduct in deep analysis strengthened with a continuous fact checking process. Preserving documentary traceability enables historians also to validate or refine AI-generated results. Therefore, the system

must produce verifiable historical evidence rather than isolated information and knowledge, so as to allow transparency, reproducibility, and trustworthiness.

Organize Knowledge Around Historical Research Questions. Historical research is guided by questions whose answers are rarely contained within a single document. Instead, historians must connect, compare, and interpret evidence distributed across heterogeneous documentary sources. Therefore, extracted information and knowledge should be organized to support, discovery, linking, comparison, aggregation, and reorganization across collections. This principle suggests transforming collections of digitized documents into interrogable research knowledge bases, enabling historians to identify recurring patterns, reconstruct historical developments, and generate new interpretations that would be difficult to obtain through document-by-document analysis.

Enable Semantic Interoperability. Historical knowledge should be organized into a shared and interoperable knowledge base that transforms fragmented documentary collections into a unified research resource. Such a knowledge base organizes documentary evidence, represents the relationships identified across the corpus, and enables historians to explore them through different navigation paths. Moreover, it can be made publicly accessible as LOD and, therefore integrated into other such knowledge bases. This enables systematic querying, comparison, and interpretation of historical evidence across collections, institutions, and research domains while preserving the historian’s control over navigation, validation, and interpretation.

Support Knowledge Base Evolution. Knowledge base should be treated as an evolving research asset rather than as a static collection of information. New documentary sources, revised interpretations, and updates to representation models may progressively modify the understanding of historical evidence. Therefore, AI-assisted CH systems should support controlled knowledge base evolution, preserving versioning, provenance, and traceability over time. This principle requires mechanisms for incrementally incorporating new extractions, expert revisions, ontology updates, and semantic re-linking operations without compromising the integrity of the extracted information. By maintaining versioning and provenance, knowledge bases can continuously grow and improve as new evidence and refined interpretations become available.

4. The Framework

We designed a multi-agent framework that orchestrates specialized Generative AI (GenAI) agents to transform digitized historical catalogues into an interlinked and queryable knowledge base (Figure 2). The resulting knowledge base provides an environment for exploration, analysis, and source verification, supporting historians throughout the research process. The framework consists of six specialized agents: **(i)** a coordinating agent, **(ii)** a document pre-processing agent, **(iii)** a document segmentation agent, **(iv)** an object extraction agent, **(v)** an information verification agent, and **(vi)** a knowledge base manager agent. Together, these agents implement the methodological principles presented in the previous section.

The coordinating agent acts as the entry point of the framework, receiving analysis requests and orchestrating the execution of the specialized agents. It plans and coordinates the workflow, monitors the execution state and information flow, manages failures and retries, and maintains the overall status of the analysis process. The coordinating agent first invokes the document pre-processing agent, which transforms historical documents into a structured representation, implementing the methodological principles of *Preserve Documentary Semantics* and *Preserve Documentary Traceability*. It performs Optical Character Recognition (OCR) to extract textual content, detects and segments document regions, and classifies them into semantic elements such as titles, sections, paragraphs, footnotes, captions, tables, and illustrations. The resulting structured document preserves the reading order together with



Figure 1: Two consecutive pages from a Spanish historical commercial catalogue. The pages illustrate the heterogeneous structure of the source. Textual and visual elements are combined to describe educational objects. The coloured bounding boxes highlight the different documentary components present on the pages: blue: titles and subtitles; green: list paragraphs; red: paragraphs; light green: images; light blue: image captions; yellow: tables.

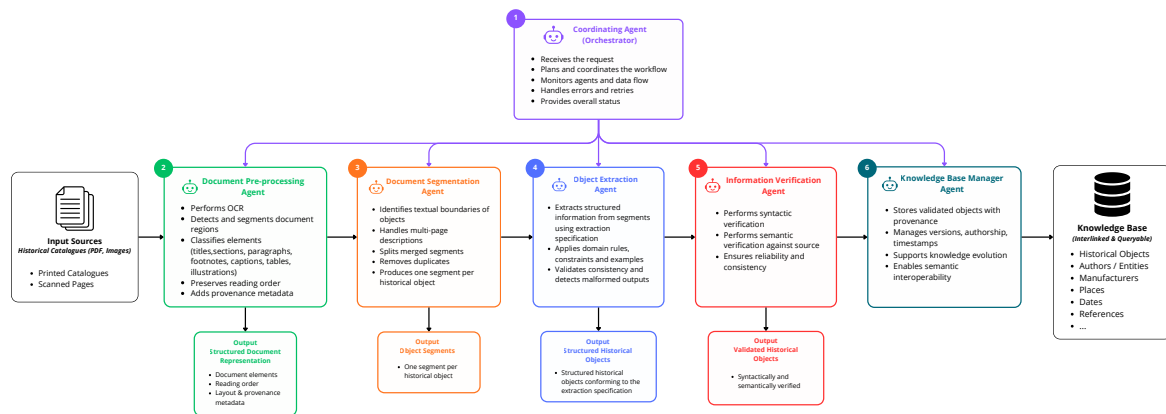


Figure 2: Architecture of the proposed multi-agent framework. Six specialized GenAI agents progressively transform digitized historical catalogues into validated historical objects and organize them into an interlinked, traceable, and queryable knowledge base while preserving documentary semantics, provenance, and traceability.

the metadata required to maintain an explicit and verifiable link to the original documentary source, providing the input for the subsequent agents.

The document segmentation agent receives the structured document representation and identifies the textual boundaries of the historical objects contained in the document. The agent ensures that each object is correctly delimited, even when its description spans multiple pages. The resulting segments undergo a refinement step that splits merged segments and removes duplicates. The final output is a set of object segments, each representing the textual boundaries of a single historical object and providing a contextualized and traceable input for the object extraction agent.

The object extraction agent implements the principle of *Organize Knowledge Around Historical Re-*

search Questions. It receives the refined object segments and extracts structured historical information according to an extraction specification co-designed with domain experts. The specification translates historical research questions into extraction objectives, structured fields, domain-specific rules, constraints, and representative examples that guide the extraction process. The agent also validates the consistency of the extracted objects, detecting malformed outputs.

The information verification agent implements the principle of *Preserve Documentary Traceability* by validating the reliability and consistency of the extracted historical objects. It performs both syntactic and semantic verification. Syntactic verification ensures that the extracted objects conform to the extraction specification, while semantic verification assesses whether the extracted information faithfully and consistently reflects the original documentary evidence.

The knowledge base manager agent implements the principles of *Support Knowledge Base Evolution* and *Enable Semantic Interoperability*. It receives the validated historical objects and incorporates them into an interlinked and queryable knowledge base together with their provenance and versioning metadata. Each extraction is stored as a new version, while subsequent revisions preserve the complete history of modifications, including authorship and timestamps. This enables the knowledge base to evolve incrementally while maintaining full provenance and traceability over time. The resulting structured knowledge base provides a shared semantic representation that enables interoperability with the broader Linked Open Data (LOD) ecosystem, including ontologies, knowledge graphs, and other CH knowledge bases.

4.1. Framework Implementation

The proposed framework is implemented using LangGraph ¹, which orchestrates the execution of the specialized agents through a graph-based workflow operating on a shared memory state. Each agent transition is governed by a directed graph in which the shared state carries the structured document representation and all intermediate outputs across nodes, ensuring a controlled and traceable information flow throughout the pipeline. Each agent receives a well-defined input and produces a structured output that becomes the input for the next stage. The document pre-processing agent relies on MinerU ², a document parsing framework that combines OCR, layout analysis, and the Vision-Language Model (VLM) MinerU2.5-Pro-2605-1.2B to generate the structured document representation. MinerU produces an ordered list of typed blocks per page — titles, paragraphs, list items, tables, captions, and illustrations — each carrying its page number and bounding-box coordinates, which propagate through the entire pipeline as provenance metadata. The document segmentation agent, object extraction agent, and information verification agent are implemented using the NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4 ³ Large Language Model (LLM). All three agents use zero-shot prompting with guided decoding: each call is constrained to produce a valid JSON object conforming to a structured schema defined for that stage, eliminating structural hallucinations at the output level. The segmentation agent processes a sliding window of two consecutive pages, using the previous page as context and the current page as the unit of analysis; multi-block spans flagged as potentially merged undergo a recheck step in which the model is allowed an extended reasoning budget to resolve ambiguous object boundaries. The object extraction agent is guided by an extraction specification co-designed with domain experts, encoding domain-specific rules, field definitions, and constraints that translate historical research questions into structured extraction objectives. The information verification agent performs semantic verification of each candidate extraction against the original source text, checking that every extracted field is directly supported by the documentary evidence; a dedicated flag tracks whether any field was modified during verification, providing a built-in audit trail of the self-correction process. A separate retry mechanism handles technical failures — malformed responses and transient request errors — independently of this semantic verification step. Finally, the knowledge base manager agent relies on MongoDB, where validated historical objects are stored in versioned collections together with provenance and revision

¹LangGraph documentation: <https://docs.langchain.com/>

²MinerU documentation: <https://opendatalab.github.io/MinerU/>

³<https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-NVFP4>

metadata, supporting the incremental evolution of the knowledge base.

5. Cata-LOH

Cata-LOH (Catalogues for Researching Learning Objects and their History) is a digital research platform developed using the proposed methodological framework to support the analysis of historical educational-supply catalogues. The design of Cata-LOH is guided by the methodological principles of research on school materiality within the History of Education, where historical commercial catalogues are treated as philologically grounded primary sources whose original terminology, disciplinary organization, and documentary structure must be faithfully preserved.

These catalogues constitute a rare and valuable source, documenting not only historical teaching objects but also the pedagogical, technological, industrial, and commercial contexts in which they were produced and used. However, the absence of a comprehensive methodological framework, together with the heterogeneity of catalogue structures, disciplinary terminology, and editorial conventions, makes their systematic organization and analysis particularly challenging for researchers.

Cata-LOH provides a queryable knowledge base designed to support historians throughout the research process (see figure 3). The system allows researchers to formulate queries and automatically analyses the complete content of the documentary corpus. Each response is linked to the corresponding documentary evidence, including the original catalogue, source text, page reference, and scanned page image, allowing historians to verify the extracted information directly against the primary source. The platform is structured around three main functionalities: **exploring historical catalogues** to investigate their documentary structure and organization, **searching and analysing individual educational objects** across the corpus through multiple descriptors, and **exploring disciplines and catalogue sections** to study the relationships between educational materials, knowledge domains, and their historical development.

The current implementation is based on a corpus of 98 historical catalogues published between the nineteenth and twentieth centuries, from which approximately 55,000 historical teaching objects have been extracted. Cata-LOH enables researchers to perform cross-document analyses and explore the multidisciplinary information associated with these objects, including pedagogical methods, industrial development, authors, manufacturers, and other contextual information.

6. Conclusion and Future Work

Historical research increasingly requires AI systems that support historians while preserving the methodological foundations of historical inquiry. This work contributes a transferable methodological framework for designing AI-assisted historical research environments. The framework defines a set of methodological principles and realizes them through a multi-agent architecture that transforms documentary collections into structured, traceable, and queryable knowledge bases.

Cata-LOH demonstrates the applicability of the proposed framework within the History of Education, showing how historical commercial catalogues can be transformed into a research environment that supports exploration, verification, comparison, and the continuous evolution of historical knowledge. More broadly, this work shows that AI can contribute to historical research not by replacing historians, but by providing trustworthy and transparent research environments that transform complex documentary collections into continuously evolving knowledge bases, enabling new forms of historical investigation across Cultural Heritage domains.

Several limitations should be acknowledged. The extraction pipeline has been validated on a single documentary corpus — historical educational-supply catalogues in Italian, Spanish, and French — and its transferability to other CH documentary genres has not yet been empirically demonstrated. The framework's design principles are grounded in domain expert observation and co-design, but have not been subject to a formal user study measuring their impact on historians' research workflows. Furthermore, while the multi-agent architecture includes semantic verification to mitigate hallucination

risk, the accuracy of the resulting knowledge base has not yet been systematically evaluated against a ground-truth annotation; such an evaluation, including precision, recall, and expert validation on a representative sample, is planned as part of the next development phase.

Future work will focus on evaluating the resulting knowledge base together with domain experts and extending the framework to additional Cultural Heritage documentary collections. Historians' feedback will be incorporated into the continuous refinement of both the extraction workflow and the evolving knowledge base, further enhancing the proposed framework as a transparent and trustworthy support for historical research.

Acknowledgments

The authors gratefully acknowledge the support of the Dipartimento di Filosofia e Scienze dell'Educazione, Università di Torino, and the History of Education research team for their valuable contribution to the definition of the research requirements and the development of the Cata-LOH Digital Research Platform.

Availability of Data and Materials

The Cata-LOH Digital Research Platform developed as part of this work is available at <https://www.cataloh.it/>. Access to the platform and additional materials associated with the project can be requested by contacting the corresponding author.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: improve language quality, grammar and spelling, and assist with manuscript refinement. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] M. Bloch, *The Historian's Craft*, Manchester University Press, 1953.
- [2] M. C. Howell, W. Prevenier, *From reliable sources: An introduction to historical methods*, Cornell University Press, 2001.
- [3] J. Tosh, *The Pursuit of History: Theory and Methods*, Routledge, 2015.
- [4] J. Van Drie, C. Van Boxtel, Historical reasoning: Towards a framework for analyzing students' reasoning about the past, *Educational Psychology Review* 20 (2008) 87–110.
- [5] M. Kipping, R. D. Wadhvani, M. Bucheli, Analyzing and interpreting historical sources: A basic methodology, *Organizations in time: History, theory, methods* (2014) 305–329.
- [6] D. J. Cohen, *Digital history: A guide to gathering, preserving, and presenting the past on the web*, 2005. URL: <https://api.semanticscholar.org/CorpusID:106625334>.
- [7] I. Milligan, *The transformation of historical research in the digital age*, Cambridge University Press, 2022.
- [8] F. D. Pizzigoni, et al., Nuove piste di ricerca sul catalogo commerciale di oggetti didattici, *Cabás* (2024) 6–24.
- [9] S. Kumpulainen, H. Keskustalo, B. Zhang, K. Stefanidis, Historical reasoning in authentic research tasks: Mapping cognitive and document spaces, *Journal of the Association for Information Science and Technology* 71 (2020) 230–241. URL: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24216>. doi:<https://doi.org/10.1002/asi.24216>. arXiv:<https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24216>.

- [10] D. M. Berry, Introduction: Understanding the digital humanities, in: *Understanding digital humanities*, Springer, 2012, pp. 1–20.
- [11] S. Graham, I. Milligan, S. B. Weingart, K. Martin, *Exploring big historical data: the historian’s macroscope*, World Scientific, 2016.
- [12] O. Eberle, J. Büttner, H. el Hajj, G. Montavon, K.-R. Müller, M. Valleriani, Historical insights at scale: A corpus-wide machine learning analysis of early modern astronomic tables, *Science Advances* 10 (2024) eadj1719. URL: <https://www.science.org/doi/abs/10.1126/sciadv.adj1719>. doi:10.1126/sciadv.adj1719. arXiv:<https://www.science.org/doi/pdf/10.1126/sciadv.adj1719>.
- [13] Y. Xie, M. La Mela, F. Tell, Multimodal llm-assisted information extraction from historical documents: The case of swedish patent cards (1945-1975) and chatgpt, in: *The 9th Digital Humanities in the Nordic and Baltic Countries Conference (DHNb 2025)*, March, 2025.
- [14] S. Galka, G. Vogeler, Annotating, projecting, and interpreting named entities in digital scholarly editions with llms, *International Journal of Digital Humanities* 7 (2025) 483–509.
- [15] M. Terras, J. Baker, J. Hetherington, D. Beavan, M. Zaltz Austwick, A. Welsh, H. O’Neill, W. Finley, O. Duke-Williams, A. Farquhar, Enabling complex analysis of large-scale digital collections: humanities research, high-performance computing, and transforming access to british library digital collections, *Digital Scholarship in the Humanities* 33 (2018) 456–466. URL: <https://doi.org/10.1093/llc/fqx020>. doi:10.1093/llc/fqx020. arXiv:<https://academic.oup.com/dsh/article-pdf/33/2/456/24812007/fqx020.pdf>.
- [16] J. A. Motilla-Chávez, D. Espitia, D. Perez-Martinez, E. Galán-Vásquez, E. Ugalde, M. Lomelí-Jasso, E. Perez-Martinez, H. G. Perez-Gonzalez, F. Carlín-Loza, V. Martínez-Ramírez, et al., A methodological framework for the use of ai tools in automated workflows for generating validated and structured historical datasets, *International Journal of Digital Humanities* (2026) 1–28.
- [17] R. D. Meadows, J. Sternfeld, Artificial intelligence and the practice of history: a forum, *The American Historical Review* 128 (2023) 1345–1349.
- [18] C. Henriot, The ai-augmented research process: A historian’s perspective, arXiv preprint arXiv:2508.01779 (2025).
- [19] L. Jaillant, L. Zhao, Introduction: When data turns into archives: making digital records more accessible with ai, *AI & SOCIETY* 40 (2025) 5787–5791.
- [20] F. Lombardi, S. Marinai, Deep learning for historical document analysis and recognition—a survey, *Journal of Imaging* 6 (2020). URL: <https://www.mdpi.com/2313-433X/6/10/110>. doi:10.3390/jimaging6100110.
- [21] V. Rezanezhad, K. Baierer, M. Gerber, K. Labusch, C. Neudecker, Document layout analysis with deep learning and heuristics, in: *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, 2023, pp. 73–78.
- [22] Z. Abedi, R. M. van Dijk, G. Wijnholds, T. Verhoef, Enriching historical records: An ocr and ai-driven approach for database integration, arXiv preprint arXiv:2512.23710 (2025).
- [23] E. Linhares Pontes, L. A. Cabrera-Diego, J. G. Moreno, E. Boros, A. Hamdi, N. Sidère, M. Coustaty, A. Doucet, Entity linking for historical documents: Challenges and solutions, in: *International Conference on Asian Digital Libraries*, Springer, 2020, pp. 215–231.
- [24] S. D. Stewart, S. Sinha, Retrieving information from unstructured historical sources using large language models, *Computational Humanities Research* 1 (2025) e17.
- [25] S. D. Stewart, S. Sinha, Retrieving information from unstructured historical sources using large language models, *Computational Humanities Research* 1 (2025) e17. doi:10.1017/chr.2025.10019.

A. Cata-LOH Digital Research Platform

This appendix presents screenshots of the Cata-LOH digital research platform, illustrating its main functionalities: the exploration of historical catalogues, the search and analysis of educational objects, and the exploration of disciplines and catalogue sections.

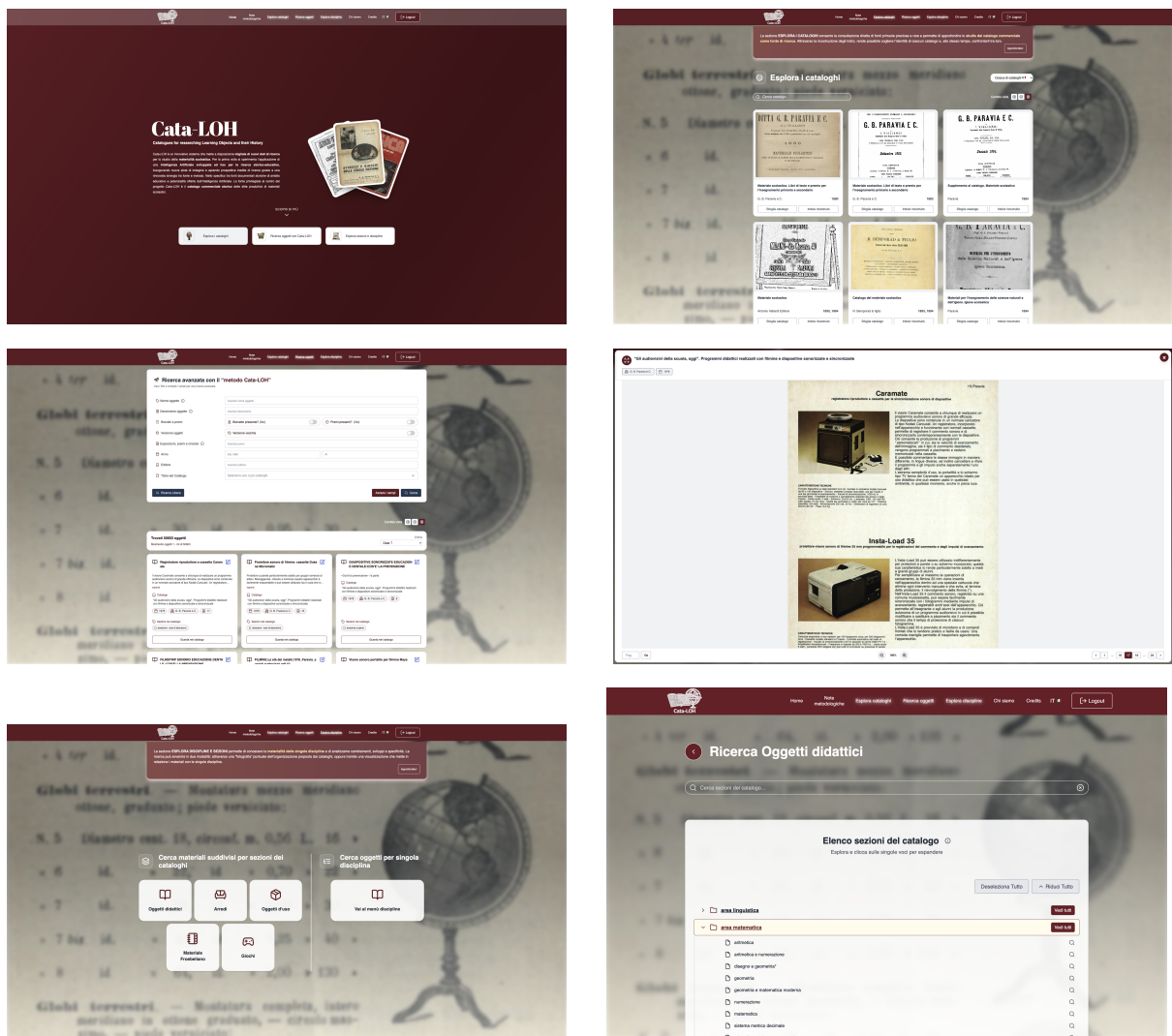


Figure 3: Cata-LOH web application interfaces. The figure shows: (a) homepage; (b) Explore Catalogues page, enabling historians to search catalogues and access digitized sources; (c) Search Objects page: it supports advanced queries through free-text search and structured filters for retrieving educational-supply entities; (d) catalogue visualization interface, providing access to digitized catalogue content; (e) explore disciplines and catalogue sections page. it shows extracted educational-supply entities according to classification areas (e.g., furniture, educational objects, games, and disciplines); (f) disciplinary browsing interface. This page allow historians to explore educational-supply entities by subject categories (e.g., mathematics, history, and geography).