Ministry of Science and Higher Education
Republic of Poland

NATIONAL PROGRAMME FOR THE DEVELOPMENT OF HUMANITIES

Rocznik Filozoficzny Ignatianum
The Ignatianum Philosophical Yearbook
Vol. 32, No. 1 (2026), s. 289–311
PL ISSN 2300–1402
DOI: 10.35765/rfi.2026.3201.17

Tomasz Śmigla

ORCID: 0009-0002-2005-333X
Ignatianum University in Cracow

The Significance of Modern Digital Methods for the Study of Jesuit Missions in China[1]

Znaczenie nowoczesnych metod cyfrowych dla badań nad misjami jezuickimi w Chinach

Abstract

This article addresses the application of modern digital tools in research on the Jesuit missionary heritage in China, using the project of a critical edition of *Historiae Sinarum Imperii* (HSE) by Tomasz Szpot Dunin, SJ, as a case study. Drawing on the theoretical assumptions of digital humanities – understood not as a separate discipline but as a research strategy – the author analyzes the complete "digital workflow" employed in the project: from digitization and HTR (Handwritten Text Recognition) transcription, through text structuring and critical editing in the Classical Text Editor (CTE) environment, to advanced corpus analysis and the experimental use of large language models (LLMs) and RAG technology.

1 This research was funded by the state budget under the programme of the Minister of Education and Science, *National Programme for the Development of the Humanities* (project no. NPRH/U22/SP/0021/2023/12), entitled *Historiae Sinarum Imperii and Collectanea Historiae Sinensis by Tomasz Szpot Dunin SJ – the Polish Contribution to the Study of Chinese Culture and History in the Early Modern Period*. The funding amount was PLN 1,566,263.42, which is also the total value of the project.

The aim of the article is to demonstrate that the integration of technology into the research process does not replace the scholar but rather expands researchers' cognitive capacities, enabling efficient interpretation of sources of considerable scale and complexity. In this context, the new possibilities available to researchers of archival texts using digital humanities methods are as significant as the careful selection of tools and technologies appropriate to the established research objectives.

Keywords: digital humanities, Jesuit missions, HTR, Transkribus, critical edition, LLM, RAG, large language models, Classical Text Editor, *Historiae Sinarum Imperii*, Tomasz Szpot Dunin

Abstrakt

Niniejszy artykuł podejmuje problematykę zastosowania nowoczesnych narzędzi cyfrowych w badaniach nad dziedzictwem jezuickim w zakresie misji w Chinach na przykładzie projektu edycji krytycznej dzieła *Historiae Sinarum Imperii* (HSE) Tomasza Szpota Dunina SJ. Wychodząc od teoretycznych założeń humanistyki cyfrowej, rozumianej nie jako odrębna dyscyplina, lecz strategia badawcza, autor analizuje kompletny „cyfrowy workflow" wykorzystywany w projekcie, od digitalizacji i transkrypcji HTR (Handwritten Text Recognition), przez strukturyzację i krytyczną edycję tekstu w środowisku Classical Text Editor (CTE), aż po zaawansowaną analizę korpusową i eksperymentalne wykorzystanie modeli językowych (LLM) i technologii RAG. Celem pracy jest wykazanie, że integracja technologii w procesie badawczym nie prowadzi do zastąpienia badacza, lecz do poszerzenia jego zasobów poznawczych, co pozwala na wydajną interpretację źródeł o znacznej objętości i złożoności. Nowe możliwości, które otwierają się przed badaczem tekstów archiwalnych wykorzystującym metody z zakresu humanistyki cyfrowej, są w tymże kontekście tak samo istotne jak właściwe dobranie środków i technologii odpowiednich do realizowania ustalonych celów badawczych.

Słowa klucze: humanistyka cyfrowa, misje jezuickie, HTR, Transkribus, edycja krytyczna, LLM, RAG, duże modele językowe, Classical Text Editor, *Historiae Sinarum Imperii*, Tomasz Szpot Dunin.

Historiae Sinarum Imperii: Source Context

The archives of the Society of Jesus are among the most important repositories of knowledge on early modern culture, particularly with regard to Jesuit missions in the Far East. Of special significance within these holdings are documents relating to the China mission, which testify

to a unique encounter between Eastern and Western cultures: between European science and Christianity on the one hand, and the Confucian tradition on the other. They also illuminate the many cultural dilemmas involved in Jesuit accommodation in the Far East.

Among the materials preserved in the Roman Archives of the Society of Jesus (ARSI), the legacy of the Polish Jesuit Tomasz Szpot Dunin, SJ (1644/1645–1713), deserves particular attention. His remarkably extensive works – most notably *Historiae Sinarum Imperii* and *Collectanea Historiae Sinensis* – are not merely chronicle-style accounts of events surrounding the transition from the Ming to the Qing dynasty, but also form a highly complex corpus of geographical, ethnographic, and theological knowledge.

These manuscripts pose a significant research challenge not only because of the richness of their content, but also because of their physical form. They consist of large-scale volumes, together exceeding one thousand folio pages, written in a heterogeneous Latin and interwoven with Chinese terminology cited by the author, often using non-standard transliteration systems. The state of preservation of the material – characterized by localized abrasions, ink bleed-through, and variations in handwriting – combined with marginal annotations written in a very small script, further complicates the process of traditional reading and manual transcription of the original text.

Working with such an extensive and heterogeneous body of source material, in which many folios contain dense marginalia, deletions, authorial corrections, and multilingual terminology, presents the research team with a number of significant challenges. Traditional philological and historical methods, while remaining an indispensable foundation of scholarly practice, often prove insufficiently efficient in terms of time and operations when confronted with the sheer scale of the data[2] (amounting, in a sense, to the "big data" of the sixteenth and seventeenth centuries). Manual transcription, indexing, collation, and categorization of such a vast corpus within a traditional, linear workflow thus becomes an extremely labor-intensive process, potentially delaying substantive analysis and, consequently, the publication of research results together with critical editions and translations of the source texts.

In response to the many challenges posed by the source material, numerous solutions have emerged, not only of a technological nature but also methodological, allowing researchers to plan their work in scholarly projects more efficiently and systematically. One such solution is the

2 Ann M. Blair, *Too Much to Know: Managing Scholarly Information before the Modern Age* (New Haven–London: Yale University Press, 2010), 132–133.

implementation of a specific digital humanities strategy in the processing and editing of archival data. This does not involve merely transferring work to a computer or digitalization understood simply as taking photographs of manuscripts, but rather designing a coherent "digital workflow"[3] that enables more efficient achievement of the research goals. This process naturally encompasses the entire life cycle of the document: from its high-resolution digital capture, through machine-assisted reading using dedicated HTR software, the structuring of data in critical editions with editors such as the Classical Text Editor (CTE), to semantic analysis and exploration of the collected data supported by artificial intelligence. In recent years (2020–2025), new advances in generative AI have increasingly played a role across the tools and methods of digital humanities.[4] This approach allows not only for the optimization of working time, but also for the possibility of formulating new research questions and expanding the set of tools available to the researcher in direct engagement with the source material.[5]

Digital humanities as a research strategy

Before proceeding to a detailed discussion of the specific technological solutions applied in the *Historiae Sinarum Imperii* project, it is necessary to outline the theoretical framework that structures our approach to new technologies supporting historical research. Digital humanities is often, in general discourse, mistakenly reduced to the use of particular software tools, the creation of websites, databases, or digital repositories. Such an instrumental understanding, however, significantly oversimplifies the nature of the ongoing transformations in the methodology of the humanities, brought about both by well-established solutions that have been in use for years and by entirely new technologies that are only now finding their place within the toolkit of the modern humanist. As I have argued in an earlier publication, digital humanities should be understood

3 In the sense of a comprehensive strategy for organizing research work using digital tools, encompassing the preparation of resources appropriate to the research goal, as well as the full sequence of interconnected stages of data processing and source analysis.

4 Jiangfeng Liu, Xinglan Ma, Lanyu Wang, Lei Pei, "How can generative artificial intelligence techniques facilitate intelligent research into ancient books?," *Proceedings of the ACM on Computing and Cultural Heritage* 7/4 (2024): 1–57.

5 Silke Schwandt (ed.), *Digital Methods in the Humanities: Challenges, Ideas, Perspectives* (Bielefeld: transcript Verlag, 2021).

less as a set of concrete technical instructions addressed to the user of a more or less complex information system, and more as a transdisciplinary methodological and epistemological approach.[6] Its core lies in the modeling and representation of humanities data in a way that enables their processing by quantitative methods and, to some extent, by qualitative or hybrid approaches. This, in turn, may lead not only to increased efficiency in scholarly work, but also to new research insights that are often difficult to obtain through intuition alone or through traditional close reading.[7] This approach, as should always be emphasized, does not stand in opposition to humanities research conducted predominantly with traditional methods – such as manual transcription combined with typesetting using simple text editors – but rather constitutes its natural extension and reinforcement, as well as a logical continuation in an era that offers researchers an unprecedented range of powerful tools and methodological possibilities.[8]

In the context of working on the manuscripts of Tomasz Szpot Dunin, the adopted strategy aligns with the author's proprietary 3A classification model, which distinguishes three principal functions of digital tools within the research process, and at the same time defines categories of their usefulness corresponding to successive stages of work on source texts with the use of new technologies: Automation, Aggregation, and Augmentation.

1. **Automation** serves to accelerate routine, repetitive, and time-consuming tasks such as text transcription (HTR), preliminary part-of-speech tagging, the identification of named entities, or the extraction of semantic entities from the source material (NER). Its primary aim is to free up the researcher's cognitive and temporal resources, allowing them to be devoted instead to conceptual, analytical, and interpretative work.
2. **Aggregation** enables the organization and structuring of vast volumes of data into complex, searchable databases or digital critical editions. In a scholarly landscape where the primary challenge

6 Tomasz Śmigla, "Humanistyka cyfrowa jako strategia badawcza: typologie narzędzi, model pracy i przykład edycji tekstu źródłowego," *Rocznik Filozoficzny Ignatianum* 31/3 (2025): 321–344.

7 Danuta Smołucha, *Humanistyka cyfrowa w badaniach kulturowych. Analiza zjawiska na wybranych przykładach* (Kraków: Wydawnictwo Naukowe Akademii Ignatianum, 2021).

8 Magdalena Szpunar, *Humanistyka współczesna, Słowniki społeczne*, vol. 11, ed. Bogusława Bodzioch-Bryła (Kraków: Wydawnictwo Naukowe Uniwersytetu Ignatianum, 2024), 132–133.

is no longer the lack of access to sources but rather their overabundance and fragmentation, aggregation tools make it possible to bring order to informational chaos and to construct coherent narratives. In the context of the *Historiae Sinarum Imperii* (HSE) project, aggregation tools include, among others, software used for the typesetting and critical editing of source texts, as well as tools that support the preparation of translations and scholarly commentaries.

3. **Augmentation**, which is a somewhat more complex process, can be understood in general terms as the extension or enhancement of the researcher's cognitive capacities. This may take place, for example, through data visualization, quantitative analysis (such as stylometry or frequency analysis), or semantic exploration and searching of texts using basic NLP algorithms or large language models. Such approaches can reveal patterns, trends, and anomalies that would likely remain invisible to a scholar working with the sources without the support of an appropriate set of digital tools.

The application of this type of division, which can be used to classify categories of tools suitable for use by modern humanists, makes it possible to design the research process efficiently on the way to achieving the intended goals, regardless of which specific tools or technological solutions are involved. It is important to note that, due to the very rapid development of the technological landscape and the associated hardware and software, the methodological foundations for the use of particular tools and programs in digital humanities remain largely unchanged. The hybrid paradigm outlined here assumes that technology, from the very beginning of research planning, is not a "black box" that produces ready-made answers in a quasi-magical way, but rather a precisely selected instrument whose effectiveness depends directly on the quality of the input data and on the critical judgment and decisions of the expert – whether historian or philologist – regarding its use. In this model, the researcher does not become a "machine operator," but an architect of the research process, who consciously selects tools appropriate to the specificity of the problem while retaining full control over the structure and organization of the research workflow and the interpretation of its results.

Automation of the research process: from scan to text

The starting point for all analytical activities is the transformation of the physical document into digital form. In the case of the *Historia*

Sinarum Imperii project, this process began with a renewed initial digitization of the source material. Previously available digital copies, produced according to older standards, were characterized by low resolution, which caused significant difficulties in implementing semi-automatic text transcription using HTR models. Effective application of handwriting recognition algorithms requires a clear separation of ink lines from the paper background; therefore, the quality of the input material – most importantly the high resolution of the prepared images – is of crucial importance. The use of professional manuscript-scanning equipment, commissioned privately, made it possible to obtain high-quality image files covering the first two volumes of HSE. These scans, produced at high DPI resolution and with a high level of detail preserved, became an indispensable foundation for the first stage of applying HTR algorithms.

Adaptation of HTR models in the Transkribus environment

HTR technology stands, in a sense, in contrast to classical OCR (Optical Character Recognition), which dominates work with printed texts and operates on entirely different principles. Rather than recognizing individual, segmented letters through pattern matching, HTR most often relies on convolutional neural networks (CNNs) to analyze entire lines of text sequentially.[9] The system learns the specific characteristics of an individual author's handwriting, the contextual occurrence of characters in relation to one another, and typical ligatures. Such solutions also make it possible to easily identify and tag side texts and marginalia, separating them from the main body of the document. The platform selected for this task was Transkribus, developed by the READ-COOP consortium, which has become a de facto standard in many archival and editorial projects.[10] Other solutions of this type include, for example, eScriptorium.[11] At present, there exists a wide range of base and hybrid

9 READ-COOP, *OCR vs. HTR or "What is AI, actually?*, Transkribus blog, 9 May 2021, https://blog.transkribus.org/en/insights/ocr-vs-htr.

10 Joe Nockels, Paul Gooding, Sarah Ames, Melissa Terras, "Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of Transkribus in published research," *Archival Science* 22/3 (2022): 367–392.

11 eScriptorium Documentation, *eScriptorium Documentation – About this documentation*, Read the Docs, https://escriptorium.readthedocs.io/en/latest/ (accessed on: 18.12.2025).

methods for handwritten text recognition, differing in performance, intended use cases, and levels of accuracy.[12]

The experience gained from implementing HTR in the *Historiae Sinarum Imperii* project clearly illustrates the difference between generic and dedicated approaches, as well as the importance of selecting an appropriate base model. In the preliminary phase, a comparative experiment was conducted in which several ready-made, publicly available HTR models were tested. One of them was *The German Giant I*, a model trained on an extensive corpus of nearly three million lines of text and fifteen million words, encompassing German and Latin writings from the sixteenth to the twenty-first centuries. Despite its impressive training base, the model proved largely unreliable: in initial tests it achieved a character error rate (CER) of 8.30%. In practice, this meant that a significant number of incorrectly recognized words required manual correction, rendering the process relatively inefficient in terms of time. In parallel, models dedicated specifically to Latin were tested, including *Pylaia_NeoLatin_Ravenstein* and *Italian Administrative Hands 1550–1700*. Although these performed somewhat better with respect to overall letter forms, none of them provided the level of precision required for efficient scholarly editing. In particular, they generated numerous errors in the interpretation of complex ligatures, abbreviations, and numerals.

A breakthrough came with the strategic decision to fine-tune a custom model using the infrastructure provided by Transkribus. This process – fine-tuning – required the research team to prepare so-called ground truth data,[13] which involved the careful manual transcription of a representative sample of the source material. For the training of the first model, pages of varying degrees of difficulty were selected, together amounting to 21,310 words (approximately 2,800 lines of text, corresponding to about fifty manuscript pages). This material served as training data for the base model, enabling it to adapt to the accurate classification of the distinctive features of Szpot Dunin's handwriting – figuratively speaking, his way of forming the letter *s*, joining prepositions with words, employing specific ligature patterns, and, more broadly, the characteristic "hand" evident in his manuscript writing.

12 Husam Ahmad AlHamad, Mohammad Shehab, Moh'd Khaled Y. Shambour, Muhannad A. Abu-Hashem, Ala Abuthawabeh, Hussain Al-Aqrabi, Mohammad Sh. Daoud, Fatima B. Shannaq, "Handwritten Recognition Techniques: A Comprehensive Review," *Symmetry* 16/6 (2024): 681.

13 Fine-tuning is, in machine-learning terminology, the process of further training an existing model on reference data (in the case of the Transkribus environment, ground truth) in order to adapt it to the specifics of a particular task.

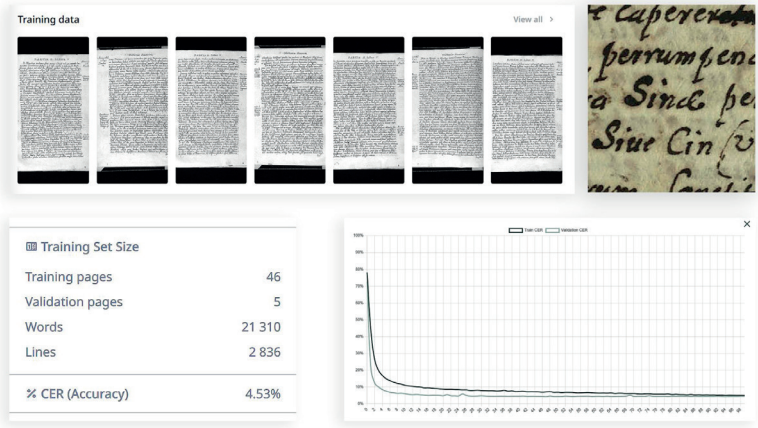


Fig. 1. Fine-tuning process of the HTR model in the Transkribus environment: training dataset (Ground Truth) consisting of manually transcribed manuscript pages (top), statistics of the first training set (bottom left), and learning curves illustrating the decrease in Character Error Rate (CER) on the training and validation sets (bottom right). Source: Resources of the *Historiae Sinarum Imperii* project.

The results of training the custom model were significant. The error rate on the validation set (the portion of data excluded from training for verification purposes) dropped dramatically from the initial >8% CER observed when using default experimental models to 4.53% CER in the first version of the fine-tuned model. In subsequent training cycles, after preparing a larger amount of Ground Truth pages, it was possible to reduce the CER to 3–4% across successive test trials. At this level of accuracy (when the system makes an error approximately once every few dozen characters), the role of the human shifts from manually transcribing the authorial text and painstakingly verifying their own work to acting as a "proofreader" of an already highly efficient HTR system, which significantly reduces the time required to process a single page of text. Analysis of the learning curves indicated that the error rate stabilizes relatively quickly when consistent, high-quality training data are provided, confirming the thesis that the time invested in preparing the initial Ground Truth dataset is one of the most cost-effective elements of the automation strategy, yielding multiple returns when processing hundreds of manuscript pages.

Segmentation challenges and new horizons (Vision-Language Models)

Contemporary digital humanities continue to evolve rapidly, and the boundaries of technological possibilities are shifting dynamically. While text recognition itself has reached a relatively high level of accuracy, a significant challenge in Szpot's volumes – and more broadly in many early modern manuscripts – remains layout analysis. These manuscripts feature highly complex topography: the main text rarely exists in isolation and is surrounded by dense networks of marginal notes, headings, corrections, and authorial annotations. Traditional HTR engines, which rely on simple baseline detection algorithms, often fail in such environments, frequently merging marginalia with the main column or losing the correct reading order. This necessitates painstaking manual correction of text regions by the researcher before running text recognition.

In response to these challenges, new opportunities arise for leveraging the latest advancements at the intersection of computer vision and natural language processing, namely the so-called Vision-Language Models (VLMs). Recent publications, such as *Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks*,[14] *An HTR-LLM Workflow for High-Accuracy Transcription and Analysis of Abbreviated Latin Court Hand*,[15] and *Unlocking the Archives: Using Large Language Models to Transcribe Handwritten Historical Documents*,[16] point to the enormous potential of this technology. Models such as Florence-2 (developed by Microsoft) and multimodal versions of Mistral models (e.g., Pixtral) represent a new approach to semi-automatic reading of manuscript folios. They do not "read" text line by line in a mechanical fashion; instead, they analyze the image holistically, semantically "understanding" the visual structure of the document. They can identify a block of text as a marginal note not only based on its position but also through visual context (e.g., smaller letter size, angle of writing, narrative style, etc.) as well as linguistic context. This, with a high degree of certainty, represents a direction

14 Seorin Kim, Julien Baudru, Wouter Ryckbosch, Hugues Bersini, Vincent Ginis, *Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical records*, arXiv, 20 January 2025, arXiv:2501.11623.

15 Joshua D. Isom, *An HTR-LLM workflow for high-accuracy transcription and analysis of abbreviated Latin court hand*, arXiv, 5 July 2025, arXiv:2507.04132.

16 Mark Humphries, Lianne C. Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray & Elizabeth Spence, *Unlocking the Archives: Using Large Language Models to Transcribe Handwritten Historical Documents*, arXiv, 2 November 2024, arXiv:2411.03340.

that could positively impact the preliminary document processing stage in the near future, eliminating the need for simplified algorithmic detection of text regions and enabling fully automated page structuring in combination with transcription and translation processes. In many cases, VLM-based solutions prove surprisingly efficient compared to CNN-based dedicated applications such as Transkribus.[17] Similarly, multimodal LLMs are already being applied to simpler tasks such as extracting text from printed documents or those containing elements of handwritten text.[18]

During work on the *Historiae Sinarum Imperii* material, a method of double HTR→VLM confirmation was also employed. In this mode of visual source analysis, the first stage of transcription is carried out in the Transkribus environment using a fine-tuned HTR model; the second stage is handled by a multimodal large language model with image-processing capabilities (in this specific case, GPT-5); and the third stage belongs to the researcher, who performs the final correction in consultation with the original source folios. At the first stage, the researcher receives an initial version of the automatic transcription of a folio generated by Transkribus. At the second stage, the scan of that folio, together with the preliminary transcription output, is passed to the VLM, along with a prompt adapted to the specific model, instructing it to correct the transcription on the basis of the attached image of the source text within the inference environment. At the final stage, through comparative analysis, the researcher performs the definitive correction of the transcript, working simultaneously with the original source folio, the initial HTR transcription, and the reference working correction produced by the VLM. This process significantly facilitates and accelerates work with the text, while also minimizing the risk of transcription errors by multiplying points of reference – effectively resembling a consultation between the transcriber and multiple researchers working in parallel on the same text.

17 Giorgia Crosilla, Lukas Klic & Giovanni Colavizza, Benchmarking Large Language Models for Handwritten Text Recognition, arXiv, 19 March 2025, arXiv:2503.15195.

18 For example: Wei Luo, *Multimodal-LLM as A Reliable Tool for Information Extraction from Historical Documents: A Digital Humanities Approach to Swedish Patent Cards (1945–1975), master's thesis*, Uppsala University, 2025, 30 HE, 70 pp., Theses within Digital Humanities 56.

Aggregation and structure: challenges of critical editing

Raw text, even when obtained with a high level of accuracy on the order of 96–97%, naturally constitutes only the foundation of a scholarly or critical edition of a source. For such material to become useful to the academic community, it must be structured, equipped with a critical apparatus, verified, and prepared for publication. Within the 3A model, this process belongs to the category of aggregation. In the case of *Historiae Sinarum Imperii*, the choice of tools for this stage was dictated by specific philological requirements which – due to the need for a smooth workflow in the creation and editing of footnotes, indices, marginalia, and, more broadly, the critical apparatus – ruled out the use of standard office software. Popular word processors such as Microsoft Word, and even advanced DTP layout programs like Adobe InDesign or Scribus (an open-source solution), often offer limited capabilities for the dynamic management of multiple layers of annotations and textual variants, or are ill-suited to flexible work on complex, densely annotated material – an issue of crucial importance in projects of this kind. The problem becomes even more pronounced when it is necessary to combine the Latin alphabet with Chinese ideographs in Unicode and to handle sophisticated forms of critical formatting.

The solution adopted in the project – which forms the core of the aggregation stage – is a relatively niche tool: Classical Text Editor (CTE), developed by Stefan Hagel at the Austrian Academy of Sciences.[19] It is specialist software designed from the ground up for humanists and editors of classical sources.[20] What distinguishes CTE is that it treats the text not as a linear sequence of characters, but as a relational database. Each word in the main text is assigned a unique identifier to which multiple layers of information can be "attached." This makes it possible to create an edition in which independent critical apparatuses can coexist in parallel:

- *Apparatus criticus* – recording variant readings, deletions, and authorial corrections;
- *Apparatus fontium* – identifying cited sources;
- *Apparatus commentarii* – containing historical and philological explanations.

19 Classical Text Editor, *Classical Text Editor*, Austrian Academy of Sciences – CTE, https://cte.oeaw.ac.at/ (accessed on: 18.12.2025).

20 Stephan Hagel, "The Classical Text Editor: An attempt to provide for both printed and digital editions," in *Digital Philology and Medieval Texts*, eds. Arianna Ciula, Francesco Stella (University of Michigan: Pacini, 2007), 77–84.

Crucially, from the standpoint of work efficiency, CTE enables work in a WYSIWYG (What You See Is What You Get) mode, offering the researcher a constant preview of the final page layout with correctly formatted, multi-level footnotes and line numbering. This eliminates a common problem in the humanities – the discrepancy between the working version of an edited text and the typeset version – and makes work on annotations simpler and more efficient. In *Historiae Sinarum Imperii*, Chinese names occur extremely frequently, introduced by the author in phonetic notation using seventeenth-century transcription systems that are difficult to identify today and require classification and clarification at the stage of compiling indexes of names, places, and persons appearing in the text. In a critical edition, it is necessary to provide modern equivalents in the Pinyin system as well as in the original Hanzi characters. In this context, CTE ensures stable handling of mixed scripts. The tool also allows the export of the finished layout to PDF (print-ready) or TEI XML format, opening the way to further digital publication without the need for an additional typesetting stage. The implementation of CTE within the project thus fulfills the postulate of aggregation in its fullest sense: it integrates the results of automatic HTR transcription with the researcher's unique expert knowledge into a single, coherent, and stable publishing whole, enabling the professional presentation of the results of source-text analysis.

Augmentation – new analytical perspectives

The third and final category of the usefulness of digital tools in humanities research marks an exceptionally interesting stage of work in our workflow. This stage is augmentation. While automation contributes to optimizing the time devoted to a given stage of research, and aggregation introduces a degree of order that enables the collection, categorization, and classification of source data, augmentation aims to contribute to the generation of new knowledge on their basis. It consists in using a digital corpus to formulate new research questions. Once the full text of just the first two volumes of *Historiae Sinarum Imperii* has been obtained (with a volume of nearly 250 recto verso leaves per volume, corresponding to approximately 500 pages of text), it becomes possible to apply digital tools for macro-scale analysis (distant reading), complementing the traditional expert-led analysis (close reading).[21]

21 Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, Gerik Scheuermann, "Visual Text Analysis in Digital Humanities," *Computer Graphics Forum* 36/6 (2017): 226–250.

Corpus tools – AntConc and Voyant tools

One of the more popular free tools for so-called text mining – encompassing, among other things, corpus, semantic, and linguistic analysis across large-scale textual datasets – are AntConc[22] and Voyant Tools.[23] AntConc, for example, allows for rapid searching of large language corpora in order to identify linguistic and thematic patterns. The software offers, among other features, a Keyword in Context (KWIC) function, which makes it possible to track and analyze how an author constructs a narrative and how (that is, in what immediate context) key figures or objects are described. Another important function of AntConc (as well as of virtually any corpus analysis program) is the N-gram search engine, which identifies recurring sequences of words occurring in the text. This type of functionality is useful, for instance, in identifying and standardizing specific geographical terminology, but also in more detailed exploration of a source with regard to the co-occurrence of particular lexical constructions or phrases built around a single core concept. A further crucial component is the collocation analysis module. It reveals words that most frequently co-occur with a given search term, providing the user with frequency coefficients for joint occurrence as well as the probability of their appearance in immediate proximity. In addition, the software offers capabilities such as cluster analysis and straightforward visualization of the most frequently occurring terms in the form of tag clouds. Text mining, along with many of the methods it encompasses, is widely used across numerous humanities disciplines that engage with large textual corpora, including historical, anthropological, and cultural studies, as well as research in the field of scholarly text analysis.[24]

22 Laurence Anthony, *AntConc: A Freeware Corpus Analysis Toolkit for Concordancing and Text Analysis*, *LaurenceAnthony.net*, https://www.laurenceanthony.net/software/antconc/ (accessed on: 10.12.2025); AntConc Documentation, *Introduction – AntConc Manual*, *Read the Docs*, https://antconc-manual.readthedocs.io/en/latest/intro.html (accessed on: 10.01.2025).

23 Lexos/Voyant Tools Team, *Voyant Tools, https://voyant-tools.org/* (accessed on: 10.12.2025); A. Miller, Text Mining Digital Humanities Projects: Assessing Content Analysis Capabilities of Voyant Tools, *Journal of Web Librarianship* 12/3 (2018): 169–197.

24 Mark Etheridge, Alex M. Boulton, "Excavating Archaeological Texts: Applying Digital Humanities to the Study of Archaeological Thought and Banal Nationalism," *Internet Archaeology* 53 (2020): 1–30.

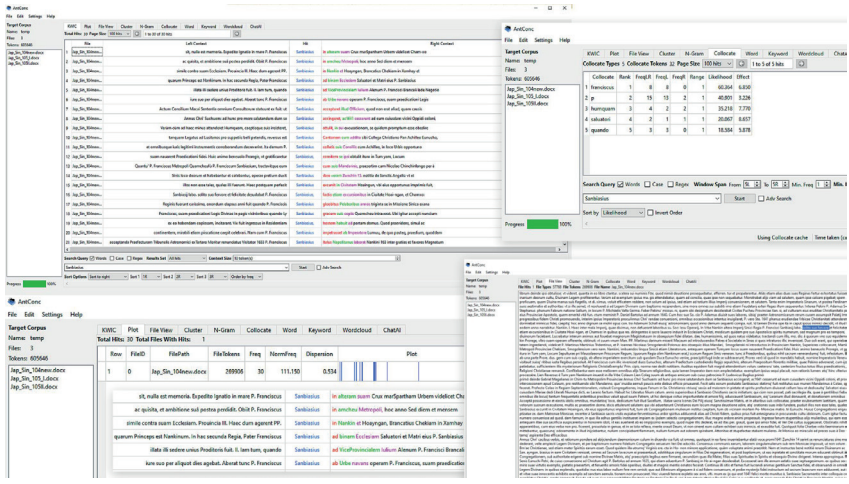


Fig. 2. Corpus analysis in the AntConc environment: KWIC search results view for the entity "Sanbiasius" (left), collocation analysis with a ranking of co-occurring lexemes (upper right), and a visualization of the distribution of the term's occurrences within the corpus (lower left). Source: Resources of the *Historiae Sinarum Imperii* project.

To provide a simple and illustrative example of the use of some of the most elementary tools of semantic analysis applied to a textual corpus, one may refer to a text-mining case that takes a person-entity as its point of departure. The ability to retrieve all occurrences of the surname "Sanbiasius" across the imported datasets made it possible, in the example illustrated above, to generate simple distribution charts that visually indicate which parts of the work (and which corpus documents) describe this figure and where such descriptions occur with the greatest density. Collocation analysis also yields particularly interesting results in this context. It quickly and visually reveals several important points of reference within the text. Franciscus Sanbiasius frequently appears in the vicinity of words such as Nankin (one of the places of his activity), *fervor* (zeal), and terms related to pastoral effectiveness (*felicitate*, *conversio*, *fructus*). The use of such basic analytical tools to examine a selected persona thus allows for the establishment of foundations for a kind of preliminary reconstruction of the figure's "rhetorical profile" as perceived by the author. A tool similar to AntConc, but published in an open-source model, is the already mentioned Voyant Tools, available as a web application and also usable locally. Voyant Tools additionally offers significantly expanded possibilities for the visualization of textual data without the need to install software on the end device, making it

a highly attractive entry point into semantic and structural linguistic analysis for researchers who have not previously had the opportunity to incorporate such tools into their work.[25]

The use of LLM and RAG technologies

At present, the most avant-garde area of augmentation is the use of Large Language Models (LLMs) for source analysis through Retrieval-Augmented Generation (RAG) technology.[26] This approach constitutes a direct response to a fundamental problem of generative models such as GPT-5, Claude, or Gemini: the risk of so-called "hallucinations." Trained on vast quantities of data, these models tend to generate content that sounds highly credible and fluent, yet may be factually incorrect. In the work of a historian – where fidelity to source texts is, by its very nature, a paramount value – this phenomenon disqualifies the straightforward use of most publicly available inference environments with their default settings.

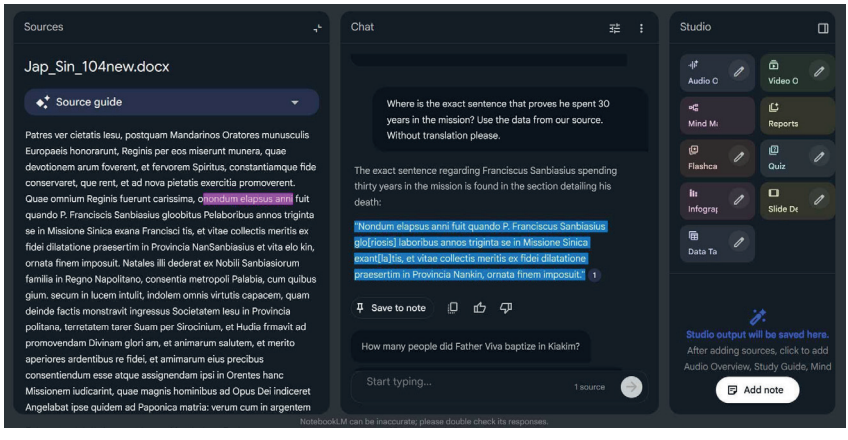


Fig. 3. Application of the Retrieval-Augmented Generation (RAG) architecture in Google NotebookLM: a natural-language query and the model's response grounded in source material through dynamic citations. Source: Resources of the *Historiae Sinarum Imperii* project.

25 Ella Alhudithi, "Review of Voyant tools: See through your text," *Language Learning & Technology* 25/3 (October 2021): 43–50.

26 Kim Martineau, *What is retrieval-augmented generation?*, IBM Research, 22 August 2023, https://research.ibm.com/blog/retrieval-augmented-generation-RAG.

RAG technology partially mitigates this problem by fundamentally changing the mode of interaction with the model, enabling it to draw on an external, verified knowledge base – in our case, the edited and preliminarily corrected transcription of the first volumes of *Historiae Sinarum Imperii*. An excellent example of the practical application of this technology at the exploratory stage is the use of Google NotebookLM.[27] As illustrated in the example above, based on the imported source file Jap_Sin_104 (a raw transcript of one of the Latin-language volumes), the system enables rapid interaction with the text without the need to manually configure complex RAG pipelines[28] and vector databases. A simple natural-language query concerning the figure of Fr. Franciscus Sanbiasius ("Find all the information you can about the figure…") resulted, within a matter of seconds, in the generation of a preliminary biographical note ready for inspection and verification. Crucially from the standpoint of historical methodology, the model not only correctly identified the date and place of the missionary's death (Canton, 1649) and his affiliation with the Neapolitan province in accordance with the source, but also accompanied each of these pieces of information with an interactive footnote (citation) linking directly to the fragment of the source text that substantiates the claim. This mechanism – visible in the interface as numbered references – fulfills the postulate of "grounding" AI-generated knowledge in evidentiary material, allowing the researcher to immediately verify the correctness of the response firsthand. In this way, the tool functions as a dynamic semantic index, enabling the synthesis of information dispersed across hundreds of manuscript pages and accessing it through natural language queries. This constitutes a classic example of augmentation: the researcher is not relieved of interpretive responsibility, but is instead granted immediate access to a preliminary synthesis of facts whose manual preparation would otherwise require many hours of archival research.

To fully understand the potential and limitations of this solution, it is worth taking a closer look at how the technology operates by explaining the technical aspects of RAG in a brief and accessible manner. First, the source text – in our case, the transcription of Szpot's work – is divided into smaller fragments (so-called chunks), which are then transformed by a specialized model (an embedder) into sequences of numbers known

27 Google, *Learn about NotebookLM*, Google Support, https://support.google.com/notebooklm/answer/16164461 (accessed on: 10.01.2025).

28 A pipeline, in general terms, refers to a data-processing architecture based on the sequential linking of independent components (e.g., an embedder, a vector search engine, and a generative model).

as vectors or embeddings. These vectors represent the semantic meaning of the text; put simply, within this space the words "Nankin" and "province" will be mathematically closer to each other than "Nankin" and "Rome." When a researcher poses a query, it too is converted into a vector. The system then searches the database to find those fragments of the source text whose vectors are "closest" to the query vector (this is typically measured using so-called cosine similarity). Only these selected, most relevant fragments of the manuscript are then passed to the generative model (LLM), together with a system instruction requiring that the response be formulated exclusively on their basis.[29] It is precisely this mechanism that makes RAG so effective in working with historical sources and largely resistant to model "hallucinations," curbing its "creativity" in favor of fidelity to the supplied context. RAG technology and its derivatives are already being used relatively frequently in projects related to archival studies, as well as in the management of large textual databases across various areas of humanities research.[30]

While tools such as NotebookLM and other popular applications implementing RAG, offered via model interfaces from companies like OpenAI, Google, or Anthropic, provide a low entry threshold and impressive exploratory capabilities, when working with archival materials of particular significance, data sovereignty and privacy become crucial and non-negotiable concerns. Using commercial "cloud-based" solutions – i.e., off-site inference relative to the infrastructure of institutions or organizations conducting research on such materials – requires transmitting transcriptions to external servers. In the case of unpublished critical editions or sensitive materials, this can often be problematic or raise concerns regarding data security.[31] For this reason, within the developed workflow, parallel to tests on public platforms, local model-based

29 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, arXiv, 22 May 2020, arXiv:2005.11401.

30 Yaming Fu, Jie Song, Xinran Zhang, Jingyun Bi, "Innovative practice of archival data development workflow in the AGI era: a case study of scientist archives project," *Information Research: an international electronic journal* 30, iConf (2025): 349–360; Ha Dung Nguyen, Thi-Hoang Anh Nguyen, Thanh Binh Nguyen, *A Proposed Large Language Model-Based Smart Search for Archive System*, arXiv, 13 January 2025, arXiv:2501.07024.

31 Sonali Tyagi, Yufeng Gong, Umit Karabiyik, "Forensic analysis and privacy implications of LLM mobile apps: a case study of ChatGPT, Copilot, and Gemini," *Forensic Science International: Digital Investigation* 54 (2025): 301974; Oliver Cartwright, Harriet Dunbar, Theo Radcliffe, *Evaluating Privacy Compliance in Commercial Large Language Models – ChatGPT, Claude, and Gemini,* Research Square, preprint, 26 July 2024.

solutions are being explored. By using software such as LM Studio or Ollama, it is possible to run open-source language models (e.g., from the LLaMA 3 or Mistral families, as well as many other models currently being released by various organizations and companies, practically quarter by quarter[32]) directly on a researcher's workstation without any network connection.[33] Thanks to quantization techniques, which reduce a model's memory requirements with minimal loss of quality, even modern consumer GPUs with relatively modest video memory (>16GB) can handle many useful, though simplified, local RAG implementations.[34] This allows full confidentiality of the examined manuscripts while still taking advantage of the benefits offered by smaller, fully private LLMs or VLMs. Such a working model – local, relatively secure, and fully controlled by the researcher – represents the intended direction for the development of the digital historian's toolkit in the coming years. A remaining challenge, too extensive to address in detail here, is still the significant disparity in size and complexity – and therefore in performance and quality of responses – among the various open-source models currently available. Accessibility also remains an issue: larger, more advanced models require investment in hardware with substantial RAM/VRAM, which can exclude researchers who do not have the infrastructure to engage in these kinds of experiments.[35]

A second aspect of this problem is, naturally, the relatively higher entry threshold associated with local solutions. Interfaces of publicly available tools are generally far more accessible to users who are not well-versed in the technological aspects of LLM operation, whereas locally

32 ApX Machine Learning, *Local LLMs Directory – Latest Local Language Models*, *ApX Machine Learning*, https://apxml.com/models?sort=release_date (accessed on: 10.01.2025). The page contains a catalog of LLMs/VLMs adapted for local use, including information on their number of parameters, maximum context length in tokens, and release date.

33 Teemu Kivimäki, *Usability Evaluation of the Local Large Language Models*, master's thesis, University of Turku, 2025, https://urn.fi/URN:NBN:fife2025061670380; Tom Smigla, *List of Local LLM Software Compatible with NVIDIA and AMD Cards*, TechTactician.com, 7 October 2025, https://techtactician.com/list-of-local-llm-software-compatible-with-nvidia-and-amd-cards/.

34 Elias Frantar, Saleh Ashkboos, Torsten Hoefler & Dan Alistarh, *GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers*, arXiv, 31 October 2022, arXiv:2210.17323v1.

35 Tom Smigla, *LLMs & Their Size in VRAM Explained – Quantizations, Context, KV-Cache*, TechTactician.com, 28 June 2025, https://techtactician.com/llm-gpu-vram-requirements-explained/. An original article explaining how the size of language models translates into GPU (VRAM) / RAM requirements, and how quantization techniques and caching affect memory demand.

hosted tools often require the researcher to possess broader technical knowledge compared to popular web applications such as ChatGPT or Google AI Studio, as well as active engagement in deploying and configuring the environment in a manner appropriate to their research objectives. These are just some of the factors that make publicly available solutions still the most commonly used option for exploring the use of LLMs and VLMs in the context of digital humanities methodologies.

Conclusion

The digital research ecosystem presented in this article – which spans from scanning, through transcription using HTR models, to structuring and editing texts in the Classical Text Editor, followed by exploration with corpus analysis software, and ultimately the use of Large Language Models (LLMs) – represents a practical realization of the principles of digital humanities in terms of consciously planning and constructing a workflow tailored to the researcher's needs. The tools and tasks aligned with the logic of the 3A model are not intended to replace the researcher but to equip them with instruments appropriate to the capabilities offered by continuously evolving technologies. The solutions discussed here, along with dozens of others available on the market or emerging in the future, do not constitute a "magic button" that solves interpretive or analytical problems. Rather, they provide new lenses through which historical sources can be examined at different scales and with greater precision, and, in many cases, significantly relieve the researcher from repetitive and laborious aspects of their work. The future of research on Jesuit missions, and more broadly of early modern history in the context of the ever-growing volume of available sources, lies in the skillful synthesis of traditional erudition with technological possibilities. The workflow developed for *Historiae Sinarum Imperii* demonstrates that, with a consciously designed research process, the machine can become the historian's most valuable ally in the pursuit of deeper understanding, transforming the historian from a reader into an architect of knowledge.

Bibliography

Manuscripts

Archivum Romanum Societatis Iesu (ARSI)
Jap. Sin. 104
Jap. Sin. 105 I
Jap. Sin. 105 II

Books and monographs

Blair Ann M., *Too Much to Know: Managing Scholarly Information before the Modern Age* (New Haven & London: Yale University Press, 2010).

Schwandt Silke (ed.), *Digital Methods in the Humanities: Challenges, Ideas, Perspectives* (Bielefeld: transcript Verlag, 2021).

Smołucha Danuta, *Humanistyka cyfrowa w badaniach kulturowych. Analiza zjawiska na wybranych przykładach* (Kraków: Wydawnictwo Naukowe Akademii Ignatianum, 2021).

Journals

Alhamad Husam Ahmad, Shehab Mohammad, Shambour Moh'd Khaled Y., Abu-Hashem Muhannad A., Abuthawabeh Ala, Al-Aqrabi Hussain, Daoud Mohammad Sh., Shannaq Fatima B., "Handwritten Recognition Techniques: A Comprehensive Review," *Symmetry* 16/6 (2024): 681.

Alhudithi Ella, "Review of Voyant tools: See through your text," *Language Learning & Technology* 25/3 (2021): 43–50.

Fu Yaming, Song Jie, Zhang Xinran, Bi Jingyun, "Innovative practice of archival data development workflow in the AGI era: a case study of scientist archives project," *Information Research: An International Electronic Journal* 30(iConf) (2025): 349–360.

Jänicke Stefan, Franzini Greta, Cheema Muhammad Faisal, Scheuermann Gerik, "Visual Text Analysis in Digital Humanities," *Computer Graphics Forum* 36/6 (2017): 226–250.

Liu Jiangfeng, Ma Xinglan, Wang Lanyu, Pei Lei, "How can generative artificial intelligence techniques facilitate intelligent research into ancient books?," *Proceedings of the ACM on Computing and Cultural Heritage* 7/4 (2024): 1–57.

Miller A., "Text Mining Digital Humanities Projects: Assessing Content Analysis Capabilities of Voyant Tools," *Journal of Web Librarianship* 12/3 (2018): 169–197.

Nockels Joe, Gooding Paul, Ames Sarah, Terras Melissa, "Understanding the application of handwritten text recognition technology in heritage contexts: a systematic review of," *Archival Science* 22/3 (2022): 367–392.

Śmigla Tomasz, "Humanistyka cyfrowa jako strategia badawcza: typologie narzędzi, model pracy i przykład edycji tekstu źródłowego," *Rocznik Filozoficzny Ignatianum* 31/3 (2025): 321–344.

Tyagi Sonali, Gong Yufeng, Karabiyik Umit, „Forensic analysis and privacy implications of LLM mobile apps: a case study of ChatGPT, Copilot, and Gemini," *Forensic Science International: Digital Investigation* 54 (2025): 301974.

Chapters in monographs

Hagel Stephan, "The Classical Text Editor: An attempt to provide for both printed and digital editions," in *Digital Philology and Medieval Texts*, eds. Arianna Ciula, Francesco Stella, (University of Michigan: Pacini, 2007), 77–84.

Szpunar Magdalena, "Humanistyka cyfrowa," in *Humanistyka współczesna, Słowniki społeczne*, vol. 11, ed. Bogusława Bodzioch-Bryła (Kraków: Wydawnictwo Naukowe Uniwersytetu Ignatianum, 2024), 132–133.

Online

AntConc Documentation, *Introduction – AntConc Manual. Read the Docs* (accessed on: 10.12.2025).

Anthony Laurence, *AntConc: A Freeware Corpus Analysis Toolkit for Concordancing and Text Analysis*, LaurenceAnthony.net (accessed on: 10.12.2025).

ApX Machine Learning, Local LLMs Directory – Latest Local Language Models. ApX Machine Learning (accessed on: 10.01.2025).

Cartwright Oliver, Dunbar Harriet, Radcliffe Theo, *Evaluating Privacy Compliance in Commercial Large Language Models – ChatGPT, Claude, and Gemini*, Research Square preprint, 26 July 2024.

Classical Text Editor, *Classical Text Editor. Austrian Academy of Sciences – CTE* (accessed on: 18.12.2025).

Crosilla Giorgia, Klic Lukas & Colavizza Giovanni, *Benchmarking Large Language Models for Handwritten Text Recognition*, arXiv, 19 March 2025, arXiv:2503.15195.

eScriptorium Documentation, *eScriptorium Documentation – About this documentation*, Read the Docs.

Frantar Elias, Ashkboos Saleh, Hoefler Torsten, Alistarh Dan, *GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers*, arXiv, 31 October 2022, arXiv:2210.17323v1.

Google, *Learn about NotebookLM*, Google Support (accessed on: 10.01.2025).

Humphries Mark, Leddy Lianne C., Downton Quinn, Legace Meredith, McConnell John, Murray Isabella & Spence Elizabeth, *Unlocking the Archives: Using Large Language Models to Transcribe Handwritten Historical Documents*, arXiv, 2 November 2024, arXiv:2411.03340.

Isom Joshua D, *An HTR-LLM workflow for high-accuracy transcription and analysis of abbreviated Latin court hand*, arXiv, 5 July 2025, arXiv:2507.04132.

Kim Seorin, Baudru Julien, Ryckbosch Wouter, Bersini Hugues, Ginis Vincent, *Early evidence of how LLMs outperform traditional systems on OCR/HTR tasks for historical records*, arXiv, 20 January 2025, arXiv:2501.11623.

Kivimäki Teemu, *Usability Evaluation of the Local Large Language Models*. Master's thesis, University of Turku, 2025.

Lewis Patrick, Perez Ethan, Piktus Aleksandra, Petroni Fabio, Karpukhin Vladimir, Goyal Naman, Küttler Heinrich, Lewis Mike, Yih Wen-tau, Rocktäschel Tim, Riedel Sebastian, Kiela Douwe, *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, arXiv, 22 May 2020. arXiv:2005.11401.

Lexos/Voyant Tools Team, *Voyant Tools* (accessed on: 10.12.2025).

Luo Wei, *Multimodal-LLM as A Reliable Tool for Information Extraction from Historical Documents: A Digital Humanities Approach to Swedish Patent Cards (1945–1975). Master's thesis*, Uppsala University, 2025. Theses within Digital Humanities 56. 70 pp.

Martineau Kim. What is retrieval-augmented generation?, IBM Research, 22 August 2023.

Nguyen Ha Dung, Nguyen Thi-Hoang Anh & Nguyen Thanh Binh, *A Proposed Large Language Model-Based Smart Search for Archive System*, arXiv, 13 January 2025, arXiv:2501.07024.

READ-COOP, *"OCR vs. HTR" or "What is AI, actually?"*, Transkribus blog post, 9 May 2021.

Smigla Tom, *LLMs & Their Size in VRAM Explained – Quantizations, Context, KV-Cache*, TechTactician.com, 28 June 2025. https://techtactician.com/llm-gpu-vram-requirements-explained.

Smigla, Tom, *List of Local LLM Software Compatible with NVIDIA and AMD Cards*, TechTactician.com, 7 October 2025.

BOOK REVIEWS