

E-Rhetoric. A Handwritten Text Recognition infrastructure for Medieval Greek

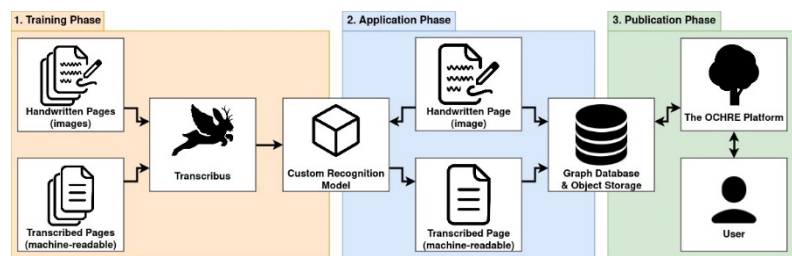
Handwritten Text Recognition (HTR) – i.e. the conversion of handwritten text into digital text through machine learning and pattern recognition techniques – performed on historical texts has improved, accelerated and even revolutionized our ways to understand of the past^{1 2 3}. While successfully applied to Latin characters for at least a decade⁴, this methodology is still in its infancy – albeit rapidly booming⁵ – for texts written in other alphabets, including Greek. And yet, this is where it can be most beneficial, due to a) the amount of un- or poorly edited pre-modern texts and b) the potential to facilitate engagement with palaeographically challenging documents.

E-Rhetoric aims to harness the potential of HTR on Greek Byzantine manuscripts to create a corpus **in support of the Semper Ardens Accelerate project “A Rhetoric for the Empire”** (www.byzherm.com). By targeting informal scholarly Byzantine Greek handwriting from the 13th and the 14th centuries, the infrastructure will lead to: **1.** a digital corpus based on a new HTR model and allowing quantitative analysis; **2.** an open digital edition of previously non-accessible sources.

Source material The infrastructure will first target 1400 handwritten pages from two manuscripts – Vienna, Philologicus Graecus 130 and Vat. gr. 2228 (both late 13th c.) – transmitting theoretical reflections on rhetoric both anonymous and authorial based on a Graeco-Roman handbook ascribed to Hermogenes of Tarsus (2nd-3rd c.). High resolution reproductions have been acquired previously. The sources are ideal for model-training: **1.** they contain large portions of texts written by the same informal hands; **2.** they offer sections of material previously edited; **3.** an unedited portion from Phil. Gr. 130, transcribed during year 1 of “A Rhetoric for the Empire” by post-doc B. MacDougall, provides a ready-made validation dataset to measure the trained HTR model’s accuracy and reliability. HTR will dramatically boost the workflow of “A Rhetoric for the Empire” beyond the goals of WP2 (see letters of support) by making available a new corpus. Once secured digitally, the corpus will also allow for quantitative analysis to test hypotheses of authorship, potentially enriching Greek medieval literary history. Finally, informal handwriting is more rarely targeted through HTR, due to the time and technical skills required. E-rhetoric will transcend these limitations, providing data for further expansions of the model.

Methodology The workflow involves 3 phases: **Training**, **Application**, and **Publication** (see Fig. 1). Based on the work of Elipida Perdiki^{6 7}, **Training** relies on the Transkribus platform, designed for historical manuscript transcription and HTR⁸. We will leverage the PyLaia model within Transkribus⁹ and train it on our specific dataset, optimizing the model for consistent and precise recognition. This step involves segmenting the handwritten text into lines which are then paired with their transcriptions. Higher accuracy is expected due to the focus on a single handwriting style, providing granular recognition of nuances and specificities. The performance of the model on the source material will be rigorously evaluated in the **Application Phase** before being implemented on

additional texts. Unlike alternatives (TensorFlow library, OCR engine Kraken, OCR engine Tesseract) Transkribus offers a powerful pre-built model (PyLaia)¹⁰; performs best in terms of CER (Character Error Rate); does not require coding expertise due to its GUI (Graphical User Interface) and Web-based application. Training requires 15000 words of diplomatic transcription: the known texts copied in the Vatican manuscript exceed such limit, as does the transcribed section of the Viennese manuscript (81000 words). Quantitative analysis to trace authorship will be performed leveraging the framework of the MA program at SDU Department of Mathematics and Computer Science's (IMADA, see below **HUMAN RESOURCES**). The project will move from the [Application Phase](#) to the [Publication Phase](#) using the OCHRE Database Platform and the Online Publication Service (OPS) of the Forum for Digital Culture of the University of Chicago.¹¹ OCHRE implements a semi-structured graph database approach to research data rather than a relational database or document model, which will allow the project to analyze texts on a letter-by-letter basis.¹² Digital publication through OPS provides a dynamic and powerful interface for exploring the manuscripts. OPS, supported by the University of Chicago Library, is committed to quality, durable, and citable digital publications meeting FAIR open data standards. OCHRE/OPS publications benefit from data management and archiving services for the life of the University of Chicago Library.



Timeline The project's life is designed for two years in the best-case scenario but factors in a third year as a contingency plan if data needs further finetuning or more manuscript sources are added: **February 2025-July 2025 Training**; **August 2025-January 2026 Application**; **February 2026-January 2027 (contingency January 2028) Publication**.

Infrastructure and resources HUMAN RESOURCES: A research assistant (RA), Nicklas Sindlev Andersen (PhD holder: <https://tinyurl.com/y9supbwr>) will be hired to train the algorithm and set up the corpus. Based at SDU IMADA he will be work closely with the applicant and **Tariq Yousef**, Assistant Professor in Computer Science and expert in Natural Language Processing and Optical Character Recognition during **Training**. RA will collaborate closely with "A Rhetoric of the Empire" post-doc **Byron MacDougall**, during **Application** contributing with his language expertise and knowledge of the target texts. **A team of international collaborators** will join periodical workshops either in person or remotely (see letters of supports) to assist in developing the model and finetune/expand it as well as solve possible issues. **Publication** will be led by **MacDougall** in liaison with OCHRE and with PI. **SUBSCRIPTIONS:** - Transkribus; - Open AI; - OCHRE (see quotes).

Fair data SDU UCloud¹³ and OCHRE¹⁴ comply with FAIR data. OCHRE records all data throughout the research lifecycle and will access images stored in UCloud through a stable URI link.

References

- ¹ Muehlberger G. et alii, Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study, *Journal of Documentation* 75 (2019), 954-976 <https://doi.org/10.1108/JD-07-2018-0114>.
- ² Spina S., Handwritten Text Recognition as a digital perspective of Archival Science, *AIDAInformazioni* 1-2 (2023) <https://www.aidainformazioni.it/index.php/aidainformazioni/article/view/249/44>
- ³ <https://www.carlsbergfondet.dk/det-har-vi-stoettet/cf22-1426>
- ⁴ Wahlberg F., Dahllöf M., Mårtensson L. and Brun A., Spotting words in Medieval manuscripts, *Studia Neophilologica* 86 (suppl) (2014), 171–186.
- ⁵ <https://www.oeaw.ac.at/imafo/veranstaltungen/detail/introduction-into-handwritten-text-recognition>.
- ⁶ Perdiki E., Review of Transkribus: Reviewing HTR training on (Greek) manuscripts, *RIDE* 15 (2022), <https://ride.i-d-e.de/issues/issue-15/transkribus/>
- ⁷ Perdiki E., Preparing Big Manuscript Data for Hierarchical Clustering with Minimal HTR Training, *Journal of Data Mining and Digital Humanities* (2023), <https://doi.org/10.46298/jdmdh.10419>.
- ⁸ <https://www.transkribus.org/>
- ⁹ <https://help.transkribus.org/data-preparation>
- ¹⁰ <https://docs.cloud.sdu.dk/dev/index.html>
- ¹¹ <https://onlinepublications.uchicago.edu/>
- ¹² Yardney S., Schloen S. R., and Prosser M., New digital tools for a new critical edition of the Hebrew Bible, *Open Theology* 5.1 (2019), 80–94 (82), <https://doi-org.proxy1-bib.sdu.dk/10.1515/opth-2019-0006>.
- ¹³ <https://escience.sdu.dk/index.php/data-management-2/>
- ¹⁴ Schloen S.R. and Prosser M.C., Where Data Comes to Life!, in: *Database Computing for Scholarly Research. Quantitative Methods in the Humanities and Social Sciences*. New York: Springer 2023, 267-310.