Connecting Metropolitan and Colonial Societies through Notarial Archives: Digital Challenges for Information Extraction in Socio-Historical Research

Caroline Koudoro-Parfait

Abstract
Brief summary (4,000 characters max) of the research proposal

This project aims to develop new methods for the large-scale digital processing of handwritten archival sources from metropolitan France and overseas colonial archives. At its core, the project seeks to advance the integration of Natural Language Processing (NLP) tools—especially Handwritten Text Recognition (HTR) and Named Entity Recognition (NER)—in historical research by enabling the systematic processing of rich corpora of manuscripts which currently resist automatic information extraction. By working with collections preserved at the Archives nationales (France) and the Archives nationales d'outre-mer (ANOM), the project addresses both methodological challenges and historiographical opportunities in the digital humanities. From a technical standpoint, the project will leverage recent advances in NLP to design and evaluate robust models for HTR, NER, keyword extraction, and network analysis. The first phase will focus on metropolitan notarial archives, including minutes that have already been processed by HTR and the methodical inventory of the insinuations of the Châtelet de Paris. The latter, a particularly complex and voluminous source, recently digitized from microfilms, presents a significant challenge for automatic processing and model training. Its transcription would mark a substantial breakthrough for both historical NLP and digital archival scholarship. The second phase expands the geographical and thematic scope by incorporating colonial archives, particularly materials related to Guadeloupe, Martinique, and French Guiana from the 17th to the early 19th century. These include civil records, naval archives, and administrative correspondence. This corpus will be used to analyze long-term patterns of mobility—including slave trade routes, military transfers, and the movements of merchants and clergy—with particular attention to the social and political transformations following the French Revolution. A complementary research strand will map circulations across colonies and between European powers in the Caribbean, contributing to a more integrated understanding of colonial and imperial networks. The two phases are closely interlinked, methodologically and thematically. The first provides a testbed for developing and refining models that will be deployed in the second, while both share a focus on recovering social structures through prosopographic and relational analyses. The insinuations register offers exceptional potential for network reconstruction among French elites, while the colonial corpus allows for the reconstruction of social landscapes and mobility flows in the Caribbean context. Together, the two components enable a connected history of metropolitan and colonial societies. Beyond historical insights, the project addresses broader issues of interoperability, reproducibility, and open science. One of its key objectives is to create reproducible workflows for processing noisy handwritten historical data—protocols that could be reused and adapted by other researchers working with similar sources. Emphasis will be placed on performance evaluation, validation of extracted data, and the creation of high-quality training datasets. In doing so, the project contributes to a more inclusive and transparent digital infrastructure for historical research and promotes the long-term accessibility and valorization of heterogeneous archival corpora. Ultimately, this project stands at the intersection of technological innovation and historical inquiry, aiming not only to deepen our understanding of early modern social dynamics but also to transform the way scholars interact with vast and complex archival materials.

Keywords: Digital Humanities, Handwritten Text Recognition, Named Entity Recognition, Archival Data, Noisy data

1 14b. Brief, convincing presentation of your innovative research idea: What is new about your hypothesis / approach / method / theory? (500 characters max)

This project combines handwritten text recognition and named entity extraction with historical research to process large-scale notarial archives from both metropolitan France and the colonial territories. It develops tools to train

robust models on diverse, degraded manuscripts, opening new prosopographical and socio-historical perspectives. The aim is to build reproducible, interoperable tools for digital archival analysis.

2 14c. What do you think will be the impact of your research on the further development of your own academic profile? (2,000 characters max)

This project will significantly contribute to the further development of my academic profile by allowing me to pursue in-depth the methodological challenges I began addressing during my PhD, particularly regarding the variability and complexity of input data in corpora of ancient texts. In close collaboration with historian Damien Tricoire—an expert in early modern history with a strong background in colonial studies and ongoing research on Parisian notarial archives, notably the "Insinuations du châtelet" —as well as the team from the Department of Modern History, I will address technical issues such as optical character recognition (OCR) and Handwritten Text Recognition (HTR) noise. Together, we will also tackle the challenges of processing heterogeneous historical sources, including digitized printed material and microfilmed manuscripts. These formats remain largely underexplored in automated text analysis, and our interdisciplinary collaboration aims to help close this gap.

The project will also enhance my expertise in exploring notarial archives and designing tools tailored to socio-historical research. By engaging with various archival collections and user communities, I aim to better align historians' needs with NLP capabilities, ensuring that the systems we develop are both technically solid and epistemologically grounded. Through this collaboration, I will also deepen my understanding of the sources and the historical questions they raise, while gaining valuable experience in corpus acquisition

Working in an international and interdisciplinary environment will allow me to engage with new research structures and academic cultures, expand my professional networks, and create opportunities for future collaborations and scientific events. Finally, the project will strengthen my knowledge of large language models and their application to historical data, positioning me at the intersection of digital humanities, historical inquiry, and computational linguistics.

References