From Manuscript to Data: An Integrated Pipeline for Handwriting Recognition, Editing, and Indexing

Vincent Jolivet, Lucas Terriel, Olivier Canteaut

From Manuscript to Data: An Integrated Pipeline for Handwriting Recognition, Editing, and Indexing

Introduction

Notre-Dame de Paris is not only the cathedral of stone, wood, and stained glass that has drawn intense scientific attention since the 2019 fire in efforts to guide its reconstruction. It is also a place that has been animated and maintained by men and women all along the centuries since the Middle Ages. These individuals are the focus of the *e-NDP* project (*Notre-Dame de Paris et son cloître : les lieux, les gens, la vie, 1326-1504*), which aims to reconstruct and analyze the social system that defined Notre-Dame over the last two centuries of the medieval period[1].

The Notre-Dame cathedral chapter forms the core of this institutional and social system. It was composed of fifty one canons who, alongside the bishop, were responsible for liturgical celebrations in the cathedral throughout the year. They are also charged with the upkeep of the cathedral itself, of the churches under its jurisdiction, and of the personnel involved in religious services: minor canons, chaplains, choirboys, and others. In addition, the canons managed a space located in the south of the cathedral and placed under their jurisdiction: the cloister. This area included around twenty houses, which served as residences for some of the canons, as well as common buildings—most notably the chapter house, where the chapter convened regularly. However, the cloister buildings did not represent the entierty of the chapter's properties. Through the accumulation of pious donations, the chapter also became a major landowner and feudal lord, holding various rights and lands across Paris, throughout the Île-de-France region, and even beyond.

During its frequent meetings —three times a week— the chapter addressed a wide range of matters. Internal governance occupied a central place in its deliberations, particularly the election or nomination of new canons, as well as the appointment and supervision of the personnel serving in the cathedral and in the churches under its jurisdiction. The chapter also concerned itself with managing its assets and its justicial rights, and settling the many disputes in which it was involved. In doing so, it ruled over a small world, which emerge through its decisions. These deliberations began to be systematically recorded from 1326: for each meeting, the names of the attending canons were noted, followed by a statement of conclusions, written in Latin. In the early years, these records were brief, and many entries still remain obscure today, due to their cryptic and allusive phrasing. Over time, the proceedings became increasingly detailed, sometimes even including copies of additional documents. In total, several thousand pages were accumulated by the chapter over the years: the capitular registers were continuously kept until they came to a temporary halt at the time of the French Revolution.

The majority of the registers thus compiled have survived and are now preserved at the National Archives of France, where they were selected in the 19th century to be included in the *Monuments ecclésiastiques* series, under the signature LL. Twenty-six registers cover the period from 1326 to 1504, with the exception of two significant gaps between 1330 and 1346, and between 1371 and 1392[2]. Although well known to historians of Notre-Dame de Paris, they have been surprisingly underused. Their sheer volume—14,000 pages—has often discouraged researchers, who rather relied on the excerpts compiled in the 18th century by Canon Claude

1 The e-NDP project brings together teams from Université Paris-Cité (laboratoire ICT-Les Europes, UR 337), École nationale des chartes-PSL (Centre Jean-Mabillon, EA 3624), Université Paris 1-Panthéon-Sorbonne (Laboratoire de médiévistique occidentale, UMR 8589), Archives nationales, Bibliothèque nationale de France and Institut de France (Mazarine library). It was funded by the French National Research Agency (project ANR-20-CE27-0012).

2 Detailed numerical inventory of the medieval registers of Notre-Dame de Paris (LL 105 to LL 127-128) by Hugo REGAZZI, under the direction of Sébastien NADIRAS, available in the virtual reading room of the National Archives: https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?irId=FRAN_IR_059635,

Sarasin, which are methodically organized[3]. Yet the complete registers are essential for understanding the social, economic and liturgical dynamics of the chapter community. The e-NDP project therefore sets itself the goal of producing and sharing a full-text, annotated version of this extensive documentation, drawing on recent advances in handwritten text recognition (HTR).

In this article, we present the processing pipeline developed to automate the reading and linguistic annotation of this extensive medieval corpus. With the aim of facilitating navigation through heterogeneous data collections, we have built an application designed as a central access point to all the resources generated over the course of the project. Finally, particular attention was given to documenting the processing pipeline, so that researchers can understand the potential biases in the shared data, reproduce the processing steps for validation purposes, or contribute to improve the available resources.

1. A Processing Pipeline

In recent years, advances in machine learning have significantly expanded the possibilities for acquiring and processing large corpora of unpublished historical texts. The use of HTR, which enables automatic transcription and annotation of digitized manuscript collections, is among the most promising developments. The e-NDP research program was designed to explore and evaluate these emerging technological capabilities, particularly in the context of large-scale corpora composed of handwritten materials. The processing pipeline developed for the automatic transcription and annotation of the text faced several challenges, notably the variability of handwriting styles among the many scribes who contributed to the registers, as well as the evolution of script forms over the course of two centuries. Additional issues included the high degree of orthographic variation in medieval Latin and the highly specialized vocabulary characteristic of capitular conclusions. These factors made the task of producing a linguistically annotated text of 4.5 million words particularly complex.

The objective was not to produce a scholarly edition conforming the traditional philological rules. While current HTR systems still produce a considerable number of transcription errors, particularly when dealing with noisy or heterogeneous input, these inaccuracies do not significantly hinder the use of such outputs for computational purposes. In fact Natural Language Processing (NLP) tools are often robust enough to extract meaningful patterns and linguistic features even from imperfect transcriptions. When effective, HTR can transform otherwise inaccessible manuscript archives into searchable, indexed textual resources, thus opening up new opportunities for historical and linguistic research. However several critical challenges remain to be addressed. Among these are the limited availability of high-quality training data, the significant variability in individual handwriting and writing systems, and the complexity of historical page layouts, which often include marginalia, interlinear additions, and non-linear reading orders.

To address the challenges inherent in training an effective HTR model—particularly those related to handwriting variability and data scarcity—we collaborated with a team of expert palaeographers and transcribers. Using the eScriptorium platform[4], they produced approximately 500 high-quality pages of ground truth data, specifically tailored for model training and evaluation. The transcription guidelines followed a semi-diplomatic approach: all abbreviations were systematically expanded, punctuation was normalized, and allographic variations (i.e., different graphical representations of the same letter) were regularized to a canonical form. This approach strikes a balance between preserving linguistic fidelity and optimizing data consistency for machine learning purposes. These 500 pages were integrated into an already substantial body of transcribed material, notably including data from previous initiatives such as the HOME-

3 National Archives of France, LL 233 to 354.

4 https://escriptorium.readthedocs.io/en/latest/

Alcar project. Altogether, the aggregated ground truth corpus comprises nearly one million tokens, providing a solid foundation for the development of robust HTR models[5].

On the basis of this enriched dataset, we conducted a series of experimental studies aimed at evaluating model performance under different configurations and training conditions. The preliminary findings of this work were informally documented in a paper which outlines the key technical results and methodological insights derived from our experiments[6]. In our efforts to enable automatic transcription of Latin and Old French medieval manuscripts, we developed two dedicated HTR models, each corresponding to a distinct script type, as well as a third, cross-script model designed to handle both simultaneously. The two targeted scripts are: *Textualis*, a formal book hand widely used from the late 11th to the 13th century and *Cursiva*, a more rapidly executed script that emerged in the 13th century and remained in use through the 15th century.

While a detailed technical presentation of these models exceeds the scope of this article—particularly given the rapid and ongoing advances in the field of medieval manuscript transcription—it is worth noting that the performance of our initial models, especially the cross-script model, was already promising. This cross-script model achieved character-level accuracy rates of 94%, a noteworthy result given the complexity and variability of medieval handwriting[7]. Subsequent developments have built upon this foundational work. In particular, S. Torres Aguilar has since undertaken substantial efforts to refine and improve the early models. An updated version, TRIDIS, incorporating improved training data and architectural adjustments, has been released and is available for broader use[8].

One of the core objectives of our project was to develop generic HTR models that could be easily fine-tuned to new corpora with minimal manual effort. To that end, we conducted a series of experiments aimed at determining the minimum amount of ground truth data required to achieve significant performance improvements through fine-tuning. The results were encouraging: a fine-tuning procedure involving as few as ten pages of annotated transcriptions led to a reduction in Character Error Rate (CER) ranging from 2% to 7%, depending on the script and model configuration. When expressed in terms of word-level accuracy, this improvement corresponds to approximately 20% fewer incorrectly transcribed words, which is particularly significant when considering the scale of medieval corpora. More unexpectedly—and perhaps even more importantly—our experiments on multilingual and multiscript training revealed that cross-training does not degrade model accuracy. In other words, combining Latin and vernacular texts as well as different scripts within the same training corpus did not result in any observable performance loss compared to monolingual or single-script models.

This finding is of particular relevance to the real-world complexity of late medieval documentary sources, which frequently contain multiple languages and writing systems. For instance, administrative and legal manuscripts often alternate between Latin and vernacular languages, sometimes even within the same document or page. Legal phraseology in Latin, in particular, remained in widespread use well into the late Middle Ages, even in regions where the vernacular dominated. Thus, the development of robust, multilingual and multiscript HTR models appears to be both feasible and aligned with the intrinsic characteristics of medieval textual production. These findings support the creation of flexible models capable of handling heterogeneous inputs—a crucial step for advancing large-scale automatic transcription in medieval studies.

The processing pipeline does not stop at the automatic transcription of manuscript pages. In order to make the resulting texts truly usable for scholarly research, a substantial post-processing phase was undertaken. This involved several key steps. First, an automated correction of the transcribed text was carried out, as far

5 https://zenodo.org/records/7401833

6 TORRES AGUILAR, Sergio/ JOLIVET, Vincent, “Handwritten Text Recognition for Documentary Medieval Manuscripts”, in *Journal of Data Mining and Digital Humanities*, 2023. https://doi.org/10.46298/jdmdh.10484

7 https://zenodo.org/records/7547438

8 https://zenodo.org/records/13862096

as the available resources allowed, to mitigate the most frequent errors introduced during the HTR process. Second, the texts underwent lemmatization, enabling a more consistent and linguistically meaningful representation of lexical items across morphological variants. Finally, the processed texts were exported in formats such as Vertical file optimized for computational analysis, particularly with the aim of generating searchable and structured indexes[9]. These indexes play a crucial role in enabling researchers—whether historians, philologists, linguists, or digital humanists—to explore and interrogate this previously unpublished and largely inaccessible corpus. By transforming raw transcriptions into linguistically enriched and queryable data, this workflow significantly enhances the value of the corpus for cross-disciplinary investigations into medieval textual production.

2. The e-NDP Application

To facilitate exploration of the corpus, we developed the e-NDP application[10]. We designed a dedicated interface to enable researchers to fully benefit from several interconnected representations of the corpus: high-definition images, a text layer that can be overlaid to assist readers, and annotated text indexes that support targeted queries.

An Interactive Facsimile. To facilitate reading, the images of the registers and their automatic transcriptions are synchronized within the application's viewer through the IIIF framework. This alignment of text and image in an interactive facsimile enables assisted and contextualized reading of the texts, with the original images linked to their corresponding archival metadata. We enriched the digital inventory of the collection—provided by the National Archives of France and accessible via the virtual reading room[11]—by developing a pseudo-chronological navigation table. This tool allows direct access to the content of each register according to a specified month. For citation purposes, we provided persistent links pointing directly to the consulted register image as well as to the corresponding facsimile page. Researchers can thus cite any of the 14,000 pages in the corpus and return to them easily for further consultation.

Exploring the Corpus. The e-NDP application integrates the NoSketch Engine indexing tool, which supports the Corpus Query Language (CQL)[12]. This query language enables researchers to fully exploit the linguistic annotations of the corpus—such as lemmatization and named entity recognition—as well as its structural metadata, including layout analysis. The combination of indexing and multilayered annotation makes possible, for instance, to retrieve all occurrences of the person named *Barre* specifically within chapter conclusions. Lemmatization, in particular, helps overcome the challenges addressed by Latin declension and orthographic variation.

The immediate access to this uncorrected, noisy corpus represents a significant shift from the traditional model of carefully curated philological editions. It opens up new research avenues—sometimes in unanticipated directions—particularly for medievalists. For example, it enables to extract data related to material and architectural heritage: for instance conclusions about the various chapels in the cathedral can be easily located and analyzed, offering insights into the historical management and evolution of the building. From the search interface, users can seamlessly access the digitized image of the register page containing a given result and, if needed, return to the interactive facsimile for broader contextual reading.

Documenting the Processing Pipeline. Throughout the project, we developed automatic reading models, multiple versions of the text, and a series of scripts to support their correction and annotation. In alignment with Open Science principles, each dataset was deposited in a repository (Zenodo or Nakala) following the FAIR data principles. As previously mentioned, the application brings together different representations of

9 https://www.sketchengine.eu/glossary/vertical-file/
10 https://endp.chartes.psl.eu/
11 https://www.siv.archives-nationales.culture.gouv.fr/siv/IR/FRAN_IR_059635
12 https://www.sketchengine.eu/nosketch-engine/

the register corpus—raw images, ALTO files, IIIF manifests, and indexes—for purposes of both reading and data mining. But beyond simply organizing these elements, we aimed to document the entire processing pipeline. This is a matter of traceability. In this sense, one could speak of a new expression of the editorial practices that have long ensured philological quality—now adapted to the age of machine learning. Researchers must be able to understand how the corpus was constructed, in order to identify potential biases and to replicate the process if needed.

In response to this transparency requirement, the e-NDP application offers a visualization that clarifies the role of each dataset within the processing pipeline[13]. It allows users to quickly grasp how the data were produced and it provides direct access to the relevant repositories, as well as to the source code used for their generation. The relationships between datasets and scripts—often distributed across multiple platforms—are made explicit, offering a synoptic view of a documentation landscape fragmented by current data-sharing practices. This is no trivial issue: while the need to document data and source code is widely acknowledged, we argue that further reflection is required on how to effectively document entire data-processing pipelines.

Conclusion

The application is structured around three core objectives. First, it aims to disseminate historical sources to a broad and diverse audience, making a previously underused corpus accessible beyond a community of specialists. Second, it seeks to address the specific needs of historians, notably by supporting precise citation practices and enabling advanced queries tailored to scholarly research—such as prosopographical investigations, institutional history, or liturgical studies. Third, the application is designed to promote data sharing within an open, interoperable infrastructure, facilitating the reuse of resources across disciplines, with particular relevance to digital humanities research.

Rather than serving merely as a consultation tool, the platform operates as a data-driven interface offering agnostic access to its resources. Data can be retrieved via standardized APIs and is systematically stored in open, persistent repositories that ensure long-term visibility and traceability. Extensive documentation accompanies both the datasets and the processing pipelines, in line with best practices in Open Science. This approach not only ensures reproductibility and transparency but also encourages the integration of these materials into new analytical workflows or comparative corpora. However the aim is not merely to share data. In keeping with the principles of Open Science, the project also seeks to provide access to—and thoroughly document—the entire processing pipeline, which relies on machine learning methods and tools.

The project provided a valuable opportunity to advance the state of the art in the automatic transcription of medieval pragmatic documents, which palaeographic and linguistic variability presents considerable challenges. Beyond the technical achievements, it also highlighted the importance of documenting each stage of a complex data processing chain—from image preprocessing and model training to annotation, indexing, and interface integration. Such documentation is essential not only for ensuring reproducibility, but also for allowing external researchers to assess the quality and potential biases of the resulting datasets, to reuse or adapt the methods for other corpora, and to contribute collaborative improvements. This approach positions the project within a broader effort to define transparency and traceability standards in the application of machine learning to historical sources.

Vincent JOLIVET, Lucas TERRIEL and Olivier CANTEAUT
École nationale des chartes – PSL / CJM

13 https://endp.chartes.psl.eu/ressources