Les registres du Trésor des chartes : des regestes et index publiés à l'indexation par intelligence artificielle

(The registers of the *Trésor des Chartes*: from published regests and indexes to indexing by artificial intelligence)

Dominique Stutzmann and Virgile Reignier

At the core of the French royal *memoria* since the Middle Ages, the Trésor des chartes are a perfect case study to identify the potentials and challenges both of and for using traditional regestes and artificial intelligence. Our paper will highlight our work with the manuscript registers belonging to the Trésor des Chartes and the analytical inventory thereof. Since its inception in 1958 under the supervision of Robert Fawtier, it has gained recognition for its exceptional quality and remarkable precision in handling the contents of these volumes (Paris, National Archives, JJ series). Despite these accomplishments, the inventory remains largely incomplete, as the printed inventories solely encompasses registers JJ 37 to JJ 79B, followed by unpublished, manuscript inventories until JJ98. Additionally, only the first and third printed volumes feature an index facilitating the exploration of subjects, place and person names. To address these limitations, researchers typically resort to supplementary materials such as archival cards, selective inventories and editions, and occasionally resort to perusing the registers themselves extensively.

During the period from 2015 to 2017, the Himanis project made significant progress by offering a word-based full text search engine for the registers of the Trésor des Chartes. Then, in the HOME project from 2018 to 2022 published annotated datasets and developed AI models for transcribing the texts and identifying the names of individuals and locations mentioned within them. In this paper, we aim to present the prospects for advancing this research, with the ultimate goal of achieving automatic indexing of the Trésor des Chartes volumes.

Despite their incompleteness, the available finding aids offer a substantial amount of descriptive information regarding the content of the registers. Moreover, these aids embody a remarkable reservoir of invaluable expertise, not only by extracting essential components of legal dispositions but also by identifying and disambiguating individuals, locations, and organizations, wherein the regests and index mutually complement each other. Providing a complete transcription does not render this expertise and access points obsolete; instead, it enhances its usefulness in navigating the vast wealth of textual information. Therefore, transferring this metadata to a digital information system is required. We first digitized the available material and converted it into XML-TEI documents, ensuring a unified format and employing acts as the fundamental level of granularity. Additionally, we standardized key metadata elements, such as date, language, and unique numbering.

A subsequent challenge emerged: aligning the diverse elements to establish correlations between individual acts and their respective images and metadata. Extensive efforts were invested in meticulously identifying and segmenting textual regions within the images, and matching them with with the inventories' contents for each act. This comprehensive dataset is now accessible to

researchers, enabling exploration of the entire corpus through linear act numbering or metadata searches, particularly employing keywords present in both the inventories and transcriptions.

Despite being at the end of the printed volumes and sometimes remaining unpublished, indexes are, or should be, an integral part of compiling the regests. They offer a complementary means of exploring the register content, particularly within diplomatic, prosopographical, or geographical contexts, and they provide the identification and preferred version for people and place names. The indexed named entities exhibit a concise style, extensively relying on implicit references in their data composition. Each entry is accompanied by a series of descriptive elements with precise meaning and order. Certain sets of entries, connected either by shared themes or associations with individuals or common locations, are condensed into a single paragraph and eliminate repetitive instances. The clarification of these entries was accompanied by enriching the geographical entities, aligning them with online reference systems such as DicoTopo and Geonames.

Furthermore, the entries encompass a multitude of implicit or explicit cross-references, enabling the association of unique entities appearing in various forms. Making all relationships explicit and restoring hierarchical links within nested entries required meticulous processing, resulting in the reestablishment of the index's intrinsic relational database structure. Used as navigational aid for engaging with the entities present within the acts and creating new perspectives such as density maps of places mentioned in the corpus or networks of people and places, this instrument's essence now aligns even more closely with that of a reference data base rather than a descriptive compilation of the acts themselves.

Following the development of the HTR (Handwritten Text Recognition) and REN (Named Entity Recognition) models, which have significantly advanced through several projects, the automatic reading of Trésor des Chartes registers is now embarking on a new objective: entity recognition at large scale based on the noisy HTR output and entity linking. The primary challenge lies in utilizing the existing indexes as ground truth to train the machine to verify the link between each recognized named entity and the entities contained in the index, which are collected within a knowledge base. This task entails two main difficulties: enabling the model to disambiguate homonymous entities based on contextual cues and proposing the addition of new entities when no matches are found during verification.