Artificial intelligence at the service of legal history: towards a new way of exploiting sources

Gwenaëlle Callemein

Artificial intelligence at the service of legal history: towards a new way of exploiting sources

GWENAËLLE CALLEMEIN
Associate professor of legal history,
Université Côte d'Azur – ERMES

Abstract: Artificial intelligence has undeniably made its way into the humanities and social sciences. Digital humanities thus provide new perspectives. They, for instance, facilitate access to sources, broaden the dissemination of knowledge and, to a certain extent, contribute to the preservation of written heritage. While they disrupt our traditional work practices, the advancement of artificial intelligence also has the potential to revolutionize them. Robotics streamlines information retrieval, enables the mass processing of data, and, as a result, allows for a different approach to utilizing sources. The processes of Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) exemplify this: artificial intelligence not only facilitates researchers' work but also offers a new reading of ancient texts. It is thus valuable to reflect on the impact of the development of these technologies for researchers and to consider the possibilities they currently offer, as well as those they may soon present in the field of the history of law.

Keywords: Artificial intelligence, Digital humanities, Legal history, historical sources, OCR, HTR, Mass data processing

Artificial intelligence is undeniably an important and widely debated topic today that has sparked a real revolution. The field of computer science is constantly evolving and is experiencing unprecedented growth. Its objective is to develop systems capable of simulating human cognitive processes in order to perform tasks that typically require human intelligence, such as speech recognition, natural language understanding, complex problem-solving, planning, and even creativity. But artificial intelligence is more than just a technological phenomenon. It has a wide impact on our daily lives and is changing our professional practices, raising ethical and legal concerns and, in our particular case, disrupting scientific research. It presents a real challenge for all disciplines, including for legal history.

Nevertheless, the term artificial intelligence is not new. The first theoretical ideas about artificial intelligence date back to the 1950s, notably exemplified by the work of Alan Turing and the publication of his seminal paper "Computing Machinery and Intelligence"[1]. During the same period, John von Neumann and other mathematicians began conceptualizing machines capable of performing complex computations[2]. The digital age was about to begin. In 1956, artificial intelligence was recognized as a specific field of research. It was also in this year that the term was used officially for the first time, during the Dartmouth Conference[3]. There was at that time a real enthusiasm surrounding artificial intelligence, the dream fantasy since Antiquity[4]. The following years were

[1] In this seminal paper, published in 1950, Alan Turing laid the groundwork for modern artificial intelligence by proposing the Turing Test, a criterion designed to evaluate whether a machine could exhibit human-like intelligence. In this article, he asks a central question: "Can machines think?": Alan TURING, "Computing machinery and intelligence", *Mind*, Oxford University Press, vol. 59, n°236, 1950, p. 433-460. Refer to: Benoît Le Blanc, « Alan Turing : les machines à calculer et l'intelligence », Hermès, La Revue, n° 68, 2014/1, p. 123-126 [en ligne].

[2] John von Neumann laid out the principles of the architecture for future computers. The term "ordinateur" appeared in 1955 and was created by philology professor Jacques Perret in response to a request from IBM France. Its executives felt that the word "computer" (or *calculateur* in French) was too restrictive given the capabilities of these machines. *Ordinateur* is thus a specifically French term; elsewhere, the term computer is predominantly used.

[3] The event was organized by mathematician John McCarthy.

[4] It is often argued that the dream of artificial intelligence traces its origins to Antiquity, as humans were already captivated by the notion of creating artificial entities capable of thought or autonomous action. In Greek mythology, Talos, a bronze giant who protected the island of Crete, can be seen as a kind of ancient robot, created by the blacksmith god Hephaestus.

marked by the emergence of new programs, such as ELIZA in 1966, capable of simulating a conversation in natural language[5].

However, the development of artificial intelligence was quickly limited, due to a lack of both computing power and data at that time. Funding declined, and research ultimately stagnated between the 1974s and 1980s. This period came to be known as the "AI winter"[6]. It was not until expert systems[7] and artificial neural networks[8] were developed that research on artificial intelligence started to grow again. However, it was in the 2000s, with the massive increase in computing power and the availability of large quantities of data (Big Data), that it truly experienced a resurgence. Notably, deep learning[9], based on multilayer artificial neural networks[10], enabled spectacular advances in natural language processing and image recognition, as evidenced by AlexNet's victory in the ImageNet 2012 challenge. Artificial intelligence thus gradually became ubiquitous, with practical applications such as virtual assistants, online

---

Another example is the golem, in Jewish culture, a creature made of clay or earth, animated by magical means to serve its creator. These examples illustrate the long history of the idea of creating an artificial life form to accomplish specific tasks.

[5] Created by Joseph Weizenbaum, this program generates the questions of a psychotherapist by reformulating the patient's statements. For example, a statement "A" might lead to the question "Why do you say A?".

[6] This was the first AI winter. A second occurred between 1987 and 1993.

[7] That is, capable of making decisions based on defined rules.

[8] Notably with the work of David RUMELHART, Geoffrey HINTON and Ronald J. WILLIAMS. See their article: "Learning representations by back-propagating errors", *Nature*, vol. 323, 1986, p. 533–536.

[9] Deep learning differs from machine learning in the way it processes data. Deep learning relies on deep neural networks, characterized by a greater number of layers of neurons between the input and output, allowing them to automatically extract complex features. Deep learning excels with large amounts of data, while traditional machine learning approaches are sufficient with smaller datasets.

[10] This progress is attributed to techniques such as backpropagation, a learning method employed to train artificial neural networks. Integral to supervised learning, backpropagation involves iteratively adjusting network parameters to minimize the discrepancy between predicted outputs and ground truth values. The algorithm's significance lies in its capacity to enable efficient learning by refining synaptic weights to reduce training data errors. Backpropagation has been instrumental in advancing deep learning, facilitating the training of highly complex, multi-layered neural architectures.

recommendation systems, and semi-autonomous cars, so much so that today, we all use it.

However, surprisingly, when it is integrated into our daily practices, we no longer really consider it to be artificial intelligence[11]. Artificial intelligence should, in some way, evoke a sense of wonder that reflects its innovative impact. In this sense, the following phrase is attributed to the mathematician John McCarthy: "As soon as it works, no one calls it AI anymore"[12]. This is probably because artificial intelligence can be divided into two broad categories[13]. The first is known as weak artificial intelligence (narrow AI), and is designed to accomplish a specific task (for example, a chess-playing program). This type of artificial intelligence is called weak because it is limited in its capabilities and cannot perform tasks outside its specific area. The second is called strong artificial intelligence (general AI), and aims to reproduce human intelligence in a general sense and be capable of performing any cognitive task that a human could do. This type of artificial intelligence, which both frightens and fascinates, is still hypothetical to this day. However, it is at the heart of many discussions, with constant progress in the field of deep learning, such as language models like GPT-4o, continually pushing the boundaries of machine understanding and text generation.

These advances mean that artificial intelligence is making inroads in every discipline, and the law is no exception[14], as can already be seen with LegalTech[15]

---

[11] For instance, when a GPS system algorithmically determines the most efficient route by processing real-time traffic data.

[12] In the same way, Larry Tesler's theorem (May 1982): "AI is whatever hasn't been done yet".

[13] See: David BRENTE, *L'intelligence artificielle expliquée. Des concepts de base aux applications avancées de l'IA*, St Herblain, Editions ENI, 2024, p. 29-30.

[14] Artificial intelligence is indeed present in all fields, whether it be healthcare, finance, law, defense, marketing, and even archaeology.

[15] The term LegalTech refers to start-ups that use technology, including artificial intelligence, to provide legal services or facilitate the practice of law. It covers a wide range of solutions to automate, simplify, and make legal work more accessible, such as automating the creation of legal documents (contracts, statutes, etc.) or simplifying legal research (finding legal precedents or relevant laws more quickly and accurately). To achieve this, machine learning and natural language processing are used to analyze massive volumes of legal data: Alain BENSOUSSAN and Jérémy BENSOUSSAN, *IA, robots et droit*, Bruxelles, Editions Bruylant, 2019, p. 341-346.

and predictive justice[16]. However, these advancements raise ethical and practical questions that must be considered. It is undeniable that technological progress carries both positive and negative sides. We must be aware of this and remain cautious about the outcomes that arise from it. Artificial intelligence feeds on the data we provide, and this data often contains biases that are difficult to overcome. Additionally, there are questions related to data storage, the risk of data loss, and growing environmental concerns due to the high energy consumption required to train and run artificial intelligence models, particularly large-scale models like those used in deep learning[17]. These aspects are of real importance, however here we focus on the benefits that artificial

---

The Digital Republic Act of October 7, 2016, also aims towards this (a movement in favor of open access) with the provision of judicial decisions outlined in Article 33 of the 2018-2022 programming law. See: Marie-Daphné PERRIN, « La mise à disposition des décisions de justice et son incidence sur la mission juridictionnelle du juge », *in* S. LACROIX-DE SOUSA, P. LARIEU et J. MESTRE (dir.), *Cerveau(x) et Droit. Neurodroit, algorithmes, intelligence artificielle, objets connectés, centre de décision*, Paris, LGDJ, p. 223-237 ; Jean-Michel SOMMER, « Le digital à la Cour de cassation. De la dématérialisation des procédures à l'open data des décisions », *in* B. BEVIERE-BOYER and D. DIBIE (dir.), *Numérique, droit et société*, Paris, Dalloz, 2022, p. 11-18 ; Estelle JOND-NECAND, « L'open data des décisions de justice par la Cour de cassation et ses incidences », *in* B. BEVIERE-BOYER et D. DIBIE (dir.), *Numérique... op. cit.,* p. 261-270 ; Gérard HAAS, « L'open data juridique, processus de démocratie judiciaire », *in* B. BEVIERE-BOYER et D. DIBIE (dir.), *Numérique... op. cit.,* p. 289-299.

[16] Predictive justice refers to the use of artificial intelligence to analyze large numbers of judicial decisions in order to predict the likely outcome of a trial. These tools aim to provide information to lawyers, judges, and parties in conflict (and even legislators) to anticipate possible decisions and better prepare their strategies. See in particular: *La justice prédictive*, Archives de philosophie du droit, tome 60, Paris, Dalloz, 2018 ; J.-B. DUCLARCQ, « Les algorithmes en procès », *RFDA*, 2018 ; Marie-Cécile LASSERRE, « L'intelligence artificielle au service du droit : la justice prédictive, la justice du futur ? », *LPA*, 30 Juin. 2017, n°127v0, p. 6 ; Y. Meneceur, « Quel avenir pour la justice prédictive ? Enjeux et limites des algorithmes d'anticipation des décisions de justice », *JCP*, 2018 doctr. 190 ; Laurence PECAUT-RIVOLIER and Stéphane ROBIN, « Justice et Intelligence Artificielle », *Statistique et société*, n°11, 2, 2023. DOI: https://doi.org/10.4000/statsoc.856.

The DataJust project, led by the French Ministry of Justice, aimed to build a database of judicial decisions in civil matters, rendered on appeal, and related to the compensation of damages resulting from personal injury in the years 2017, 2018, and 2019. However, this project was abandoned in 2022.

[17] A well-known comparison on the environmental impact of artificial intelligence models indicates that training a single model can generate as much carbon emissions as five cars over their entire lifetime. However, researchers are aware of this difficulty, and solutions are already beginning to emerge.

intelligence can offer, specifically in a field related to the history of law: the exploitation of ancient texts.

We already know that the tools of digital humanities provide greater access to sources (through digitization and the creation of digital libraries), they disseminate knowledge more widely (by means of virtual exhibitions and open data) and, to a certain extent, preserve existing corpora. Introducing artificial intelligence into our practices constitutes a further step in transforming our way of approaching sources. While technology is disrupting our work habits, the development of artificial intelligence could revolutionize them. Robotization simplifies information retrieval and enables mass processing of data, thus making it possible to use sources in a new way. There is no doubt that the approaches developed in the field of artificial intelligence represent real scientific challenges and stakes for researchers in the humanities and social sciences. It is therefore worthwhile to examine the impact of these new technologies on research and consider the possibilities they offer and could therefore offer to our discipline. To do this, it is necessary to first explore how artificial intelligence facilitates access to sources (I), and second, how it transforms their analysis (II).

***

I. Artificial intelligence for better access to sources

While it may be surprising to think of artificial intelligence as a tool for reading sources, an essential part of our profession, it can bring to light information that would be very difficult, if not impossible, to discover on a human scale. Indeed, it can reveal and link certain sources together (A) and can even facilitate the reading of ancient documents, thanks to character recognition (B).

A. Revealing new sources and linking them together

The use of artificial intelligence makes it easier to identify and explore existing relevant sources (1) and also to uncover new ones (2).

1. Facilitating access to digitized sources

Sources are often scattered throughout the country and, depending on the field of study, are sometimes even located abroad. This obviously complicates the researcher's work. However, mass digitization tends to eliminate this difficulty. Access has thus been significantly improved in recent years thanks to various digital libraries[18]. In France, the National Library of France has notably developed its digital library, named Gallica[19], which contains several million digitized documents, but there are many others[20] that focus on either a specific field or a particular collection of documents[21].

However, one of the challenges for researchers is to identify the documents they need in the vast amount of potentially relevant results[22]. They often face

[18] Heritage institutions were the first to undertake the digitization of their collections, and then some funding was established to encourage this development, such as the Collex-Persée program, which enabled the promotion of several lesser-known collections. However, the state of digital collections remains, to this day, mostly fragmented.

[19] https://gallica.bnf.fr/accueil/fr/html/accueil-fr See: Bruno BLASSELLE and Gennaro TOSCANO, *Histoire de la Bibliothèque nationale de France*, Paris, éd. BNF, 2022, p. 460-471.

[20] There are also digital libraries at the European level, such as Europeana: https://www.europeana.eu/fr

[21] Without aiming to be exhaustive, there are several digital libraries that collect sources specific to the history of law, such as the David Hoüard digital library dedicated to Norman law: https://droit-normand.nakala.fr/; the 'Corpus Droit' digital library (formerly called Yvette), focusing on medieval legal doctrine: https://numaclay.universite-paris-saclay.fr/s/numaclay/item-set/87843; the Fontes Historiae Iuris digital library, specializing in ancient works related to the history of law and justice: https://fontes-historiae-iuris.univ-lille.fr/bibliotheque-numerique; a digital library dedicated to the Archives of the Critical Legal Movement: https://data-cercrid.inist.fr/s/proces/page/accueil; a digital library that compiles printed materials (pamphlets, legal texts, etc.) from the 16th and 17th centuries in the former Spanish Netherlands: https://dial.uclouvain.be/digitization/fr/search/repository/sm_collection%3A%22Imprim%4oLex%22; and a digital library on colonial law journals: https://revcoleurop.cnrs.fr/

[22] Researchers' habits are gradually changing. Databases are also changing, and searches are increasingly becoming keyword based. However, some terms can be ambiguous, particularly in the legal field. For example, performing a full-text search with the term "prescription" will return results for both the legal sense of the term and the medical one: Sabrine MAS, Michelle CUMYN, David LESIEUR, Cécile GAIFFE and Charles TREMBLAY-POTVIN, « Apports d'une indexation à facettes pour la représentation et le repérage des décisions de justice », *in* J. Michel DOYON (dir.), *L'information et la documentation juridiques au Québec, du manuscrit à l'intelligence artificielle*, Montréal, Editions Yvon Blais, 2021, p. 214-215.

information overload[23]. Artificial intelligence can therefore assist them and improve access to sources by creating interconnected databases[24], where texts are classified and linked together intelligently, thus facilitating research and navigation[25]. Algorithms can be used to group documents with similarities, establish connections between different collections, suggest common themes between disciplines, or organize archives based on their relevance, thereby reducing researchers' search time[26].

Artificial intelligence has the ability to perform searches in large databases in a very short time, thus gathering scattered information and sometimes uncovering texts that might have remained buried under the mass of digital data. Researchers already employ digital humanities tools to facilitate their work, for instance using Boolean operators[27], but the development of artificial intelligence could improve such searches and find documents they might never have consulted without its help[28]. Furthermore, its use is even more impressive when it allows a previously unknown source to be revealed.

---

[23] This is both a challenge for the researcher and a necessity for the development of artificial intelligence, which feeds on data. As Mathieu Courtecuisse reminds us: « l'IA n'a aujourd'hui aucune valeur sans elle [la donnée]. L'IA ne progresse pas sans données, sa performance lui est consubstantielle » (trad.: "AI today has no value without it [data]. AI does not progress without data, its performance is consubstantial with it") : Mathieu COURTECUISSE, *Le saut cognitif. Comment l'intelligence artificielle change le monde*, Paris, First éditions, 2019, p. 39.

[24] As is the case with the Catalogue collectif de France (CCFR), which allows users to find sources held in multiple heritage institutions (5 100 libraries and documentation centers are listed): https://ccfr.bnf.fr/portailccfr/jsp/public/index.jsp. Similarly, Gallica performs a harvesting of collections from certain digital libraries to direct users to these resources.

[25] For example, Predictice allows contemporary judicial decisions to be sorted using specialized filters. See more on this subject: Laura DELMAS et Eloïse HADDAD MIMOUN, « La recherche en science des données chez Predictice », *in* S. LACROIX-DE SOUSA, P. LARIEU and J. MESTRE (dir.), *Cerveau(x) et Droit..., op. cit.,* p. 239-256.

[26] And, at the same time, offering a more comprehensive view of the existing sources.

[27] Boolean operators are tools used in search engines to sort information more effectively (with the operators "and", "or" and "except").

[28] Indeed, unlike traditional keyword searches, artificial intelligence enables semantic searches. This means that, in the future, even if a researcher does not use the exact words contained in a document, artificial intelligence could "understand" the meaning of the query and find relevant documents based on similar concepts.

2. Discovering new historical sources

The exponential progress of artificial intelligence has led to remarkable developments, including the discovery of previously unknown sources, as demonstrated by the recent reading of the carbonized scrolls of Herculaneum[29]. Artificial intelligence played a decisive role in reading the scrolls, a task that had remained impossible for centuries because the documents are extremely fragile[30]. The scrolls were first scanned using advanced imaging techniques, such as high-resolution X-ray tomography, to capture three-dimensional images of the interior of the scrolls without unrolling them physically. However, due to the ink used, the writings were nearly undetectable. Deep learning algorithms were then trained to identify subtle variations in density caused by the ink, even when it was very similar to the parchment's structure[31]. Artificial intelligence thus allowed for the virtual reading of the text through the different layers of the parchment, reconstructing its letters and words. This process required significant computational power, as it involved processing large amounts of image data to reconstruct texts that were invisible to the naked eye, even in the scanned images. Researchers were then able to begin deciphering one of the scrolls without damaging it[32]. This technological feat offered the possibility of discovering works lost for millennia, enriching our understanding of the history of Antiquity[33]. The success of this operation shows the full potential of artificial intelligence and suggests that previously inaccessible historical documents could

[29] A Roman city buried during the eruption of Mount Vesuvius in 79. See: Antonio RICCIARDETTO, Mario CAPASSO, *Les papyrus latins d'Herculanum. Découverte, consistance, contenu*, Liège, Les Éditions de l'Université de Liège, 2011.

[30] It is impossible to unroll them without destroying them completely.

[31] See *infra*.

[32] The first word to be deciphered in October 2023 was "purple". To date, more than a hundred words have been read as part of the Vesuvius Challenge.

[33] In this sense, artificial intelligence is also making its way into the field of archaeology. See: Thomas SAGORY, « L'intelligence artificielle au service des patrimoines et de l'archéologie : une mutation en marche », *Culture & Recherche*, n°147, 2024, p. 95-997; T. BRUGHMANS and M. A. PEEPLLES, « Trends in Archaeological Network Research: a Bibliometric Analysis », *Journal of Historical Network Research*, vol. 1, 2017, p. 1-24.

be read without being physically manipulated (palimpsests maybe[34]), especially now that handwriting recognition processes have been considerably improved.

B. Automatic recognition of printed and handwritten texts

Among the writing recognition techniques, two processes must be distinguished: Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) (1). Both facilitate the study of texts by enabling quick searches for specific information that would otherwise remain difficult to access (2).

1. OCR and HTR techniques

The processes of OCR and HTR demonstrate how artificial intelligence both facilitates the work of researchers and offers a new way of reading sources. These two technologies come into play after the digitization of textual sources and allow for the recognition of character strings within an image file, transforming it into a fully usable digital text[35]. Two techniques must then be differentiated depending on whether the writing is printed or handwritten: OCR is used to convert printed documents into digital text, while HTR transcribes handwritten texts.

The OCR of historical documents is a research field that has significantly developed in recent years, though not without challenges[36]. Old printed texts

[34] In 2023, a conference titled "A deep learning experiment for semantic segmentation of overlapping characters in palimpsests" was held in Italy by Michela Perino, Michele Ginolfi, Anna Candida Felici, and Michela Rosellini.

[35] Using software such as Transkribus, eScriptorium, Tesseract, Calamari or OCR4all. See: M. BUI, C. BRISSON, A. CHAGUE, F. CONSTANT and al., « eScriptorium et l'IA pour la transcription automatique », *Culture & Recherche*, n°147, 2024, p. 55-57.

[36] See : David FLEISCHHACKER, Roman KERN and Wolfgang GÖDERLE, "Improving OCR Quality in 19th Century Historical Documents Using a Combined Machine Learning Based Approach", 2024 [en ligne] ; Ariane PINCHE, Peter Anthony STOKES, "Historical Documents and automatic text recognition", *Journal of Data Mining and Digital Humanities*, 2024 ; Jean-

sometimes show variations in the typographies used, and the general condition of the document can limit the performance of OCR systems. However, the regularity of the characters simplifies the task, and machines trained on this type of writing now provide good results. The difficulties persist, however, with HTR. Handwritten texts are written in a variety of styles that are sometimes difficult to read, making it more challenging to establish a model[37]. There is necessarily a preliminary training of the software to achieve conclusive results[38]. This technology relies on deep learning, which improves from experience without being explicitly programmed for each task. As a result, the outcome becomes more reliable over time, depending on the corrections made by humans[39]. Ultimately, these software programs can recognize the different

---

Baptiste TANGUY, *Océriser pour accéder aux données ? Vers une évaluation non supervisée du bruit dans les données textuelles issues d'OCR de documents du XVII*ème *siècle,* thesis in linguistics, Sorbonne University, 2022 ; Sergio TORRES AGUILAR, "Handwritten Text Recognition for Historical Documents using Visual Language Models and GANs", *HALSHS : archive ouverte en Sciences de l'Homme et de la Société*, 2024 ; Thi Tuyet Hai NGUYEN, *Facilitating access to historical documents by improving digitisation results*, Thesis in computer science, La Rochelle University, 2020.

[37] Even though we have increasingly precise models, for example, the MNIST database, which contains handwritten digits and is often used as a reference for training HTR models. It allows testing the performance of recognition algorithms and guides their adjustment. The use of this technology also enables its enrichment and improves the learning of specific handwriting styles. For instance, both the Edit_Dunhuang and Numerica Sinologica projects are engaged in developing tools aimed at the automatic transcription of historical Chinese documents. Another example is the Manicule project, which makes available training data and HTR models for Medieval texts.

[38] There are many projects focused on manuscript archives. The National Archives have been pioneers in this field with several projects such as HIMANIS, SIMARA, and LECTAUREP. For more on the SIMARA project, see Jean-François MOUFFLET, « L'intelligence artificielle au service du traitement des archives : l'exemple du projet SIMARA », *Culture & Recherche*, n°147, 2024, p. 58-60.
Currently, a project titled PARL-IA-MENT(S) is being led by legal historians (Olivier Descamps and Isabelle Brancourt) for the processing of the archives of the Parliament of Paris using artificial intelligence.

[39] However, the development of generative artificial intelligence tends to reduce the time required for correction. It can be used to detect character recognition errors in the output of an initial OCR process (as OCRonos, for instance, is designed to do). Nevertheless, human verification remains necessary.

characters and forms present in the digitized document so they can be searched and analyzed.

2. The advantages of OCR and HTR

The conversion of images into digital text undoubtedly facilitates the research process[40]. Researchers can, thanks to these technologies, exploit historical sources in new ways. With these methods, it is possible to create online editions of documents and establish searchable digital databases[41]. In practice, this allows for full-text searches, the possibility to reorganize the text, or even to create an index. These techniques not only speed up the research process but also provide a fresh perspective on large-scale corpora. With these tools, researchers can, for example, find occurrences of a word within the same corpus, identify its frequency of use over a given period, or observe linguistic changes.

Beyond these advancements, HTR also allows for the exploitation of documents that are particularly difficult to read. Indeed, some scripts remain illegible to the untrained eye, requiring palaeographic skills to decipher their content. The use of HTR overcomes this obstacle by automating part of the transcription process. When an AI model is trained on a set of representative samples of a given handwriting, it becomes capable of autonomously recognizing and transcribing the remainder of the document or other texts written by the same author[42].

---

[40] Researchers in the humanities and social sciences also use visual sources. Deep learning, particularly through convolutional neural networks, is extremely powerful in image recognition and can be used to analyze collections of images or videos.

[41] Several projects, such as those on customary law, have been developed: the Rin Condé project on Norman customary texts: https://mrsh.unicaen.fr/coutumiers/conde/accueil.html, and also the BIDDC project, which aims to create an Base Internationale de Données sur les Droits Coutumiers : https://www.legiscompare.fr/web/?Base-Internationale-de-Donnees-sur-les-Droits-Coutumiers. Similarly, an ANR project led to the digitization of the Baudouin Collection, which comprises a set of decrees and laws passed during the French Revolution, while also creating a database to enable research within this corpus: http://archives-web.univ-paris1.fr/collection-baudouin/

[42] This therefore necessarily involves prior work and is only of interest when applied to a corpus of a certain size.

Finally, since these processes necessarily involve the digitization of the document beforehand, they contribute, to some extent, to the improved accessibility and visibility of sources, as well as their preservation. In any case, the development of artificial intelligence is undeniably leading to a rethinking of our relationship with sources, as they are increasingly being digitized. This phenomenon is not limited to their access or reading; it also entails the possibility of adopting new approaches to their analysis, thus opening the way to unprecedented perspectives for research.

***

II. Artificial intelligence for analyzing sources

The automation behind artificial intelligence can be used to carry out new analyses, especially in large corpora. In this sense, machine learning[43] offers innovative solutions for solving complex tasks, and in our case, they involve deep learning. This branch of artificial intelligence offers different possibilities in terms of data processing and analysis. It has grown exponentially in recent years, and its application in the humanities and social sciences is significantly transforming research. Its applications are varied, offering unprecedented opportunities to explore, study, and draw value from historical sources. Among these many possibilities, and although they are not the only ones, two aspects prove particularly interesting for the analysis of ancient texts: the ability to automatically process vast corpora (A) and to conduct innovative linguistic studies (B).

A. Processing large corpora

Artificial intelligence makes it possible to analyze a considerable number of texts (1) and, based on these data, to model networks (2).

[43] Machine learning is a field of artificial intelligence that enables a computer system to learn from data and improve with experience, without being explicitly programmed to perform a specific task. Instead of following predefined instructions, a machine learning model is trained on data and can make predictions or decisions based on the examples it has observed. See: David BRENTE, *L'intelligence artificielle expliquée... op. cit.* p. 34-57.

1. Automatically analyzing ancient documents

Researchers in the humanities and social sciences often work with very large corpora, such as literature, letters, parliamentary debates, doctrines, and judicial[44], normative and historical sources[45]. Deep learning makes it possible to detect complex and nonlinear relationships within the mass of data collected, which would not be evident with traditional methods of analysis. It provides a different perspective on a corpus and allows the identification of relationships that would otherwise be imperceptible to the naked eye. Quantitative data analysis[46], for instance, enables the identification of trends in the evolution of laws, judicial decisions, or legal doctrines. Researchers can therefore use this technology either to generate new hypotheses or to verify a specific theory. By automating data extraction and analysis, artificial intelligence frees researchers from the most repetitive tasks, allowing them to focus on interpretation and contextualization of results, while paving the way for more in-depth research.

This ability to process and examine large amounts of data provides researchers with tools to establish connections between texts scattered across time and space, offering a clearer overview of the evolution of legal systems. In legal history, this can be used to identify citations, references, or influences between various normative, doctrinal, or case law sources[47]. The more texts are digitized

---

[44] For example, the E-juris project focuses on the automated analysis of many judicial decisions: https://www.msh-lse.fr/projets/ejuris/

[45] Some researchers are also working on a corpus derived from social media networks by collecting information published on platforms such as X, Instagram, or Facebook, in order to analyze online behaviors or the spread of ideas.

[46] See: Claire LEMERCIER et Claire ZALC, *Méthodes quantitatives pour l'historien*, Paris, La Découverte, 2008, DOI: 10.3917/dec.lemer.2008.01.

[47] By means of named entity recognition, a method used to identify and classify proper names and other specific terms within a text. See: Emanuela Boros and Maud Ehrmann, "Investigating OCR-Sensitive Neurons to Improve Entity Recognition in Historical Documents", *Sustainability and Empowerment in the Context of Digital Libraries* - 26th International Conference on Asia-Pacific Digital Libraries, ICADL 2024, Proceedings ; B. SAGOT and K. GABOR, « Détection et correction automatique d'entités nommées dans des corpus OCRisés », Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2014), Marseille, 2014, p. 437-442 ; Ljudmila PETKOVIC, Motasem ALRAHABI and Glenn ROE, « Impact de la correction automatique de l'OCR/HTR sur la reconnaissance

and OCR-processed, the more artificial intelligence can analyze them and, for example, trace the sources of inspiration of a legal text by searching for its origins in earlier or foreign legislation, detect shifts in legal interpretation, highlight connections between decisions from different jurisdictions, or study the impact of certain legal scholars' works. This approach ultimately enables the mapping of networks of influence and provides insight into how certain concepts have spread over time.

2. Modeling networks and mapping relationships

Artificial intelligence enables the modeling of complex networks by analyzing and structuring information extracted from texts to reveal often hidden relationships between various entities[48]. For instance, machine learning algorithms can extract data related to individuals, places, events, or recurring concepts in documents and establish connections between them. This network modeling allows for the visualization of social, economic, or political dynamics by illustrating how entities interact and evolve over time[49]. By identifying these relationships, it becomes possible to highlight networks of influence and gain a deeper understanding of the dissemination of ideas or beliefs over a given period[50]. The visual representation of these networks, using relational mapping,

---

d'entités nommées dans un corpus bruité », *JIS - Journal of Information Sciences*, 2022, 21 (2), p. 42-57.

[48] Anne BAILLOT, « Visualisation des réseaux : apports, défis et enjeux du travail sur les données historiques. Numérisation de masse et traitement des grands corpus de textes utilisant des méthodes des humanités numériques », Stuttgart, 2015, ⟨halshs-01130425⟩ ; Claire LEMERCIER, « Analyse de réseaux et histoire », *Revue d'histoire moderne et contemporaine*, 52-2, 2005/2, p. 88-112.

For example, an ERC project LostMa (2024-2028), led by Jean-Baptiste Camps at the École nationale des chartes, combines philological expertise and artificial intelligence to study the lost manuscripts of medieval Europe and model the transmission of these texts.

[49] There are several software tools for visualizing and analyzing networks, such as Gephi, which is an open-source software that allows users to explore, represent, and analyze complex graphs.

[50] For example, Nader HAKIM and Annamaria MONTI used a bibliometric method with network analysis to better understand the circulation of legal ideas between France and Italy during the Belle Époque: « Histoire de la pensée juridique et analyse bibliométrique : l'exemple de la circulation des idées entre la France et l'Italie à la Belle Époque », Clio@Themis, 14, 2018, DOI : https://doi.org/10.35562/cliothemis.763

provides researchers with a new way to explore corpora, making it easier to interpret connections between data. For example, the study of legal doctrine could help identify relationships, or even influences, between different schools of thought that have not yet been identified.

It is also possible to undertake text mapping to geographically analyze the origin and identify their owners of certain works, and thus gain a better understanding of the dissemination and, consequently, of knowledge. Establishing these relationships often represents an enormous task on a human scale, whereas artificial intelligence can generate them very quickly once it has the appropriate data. Digital maps allow for better visualization of information and have the advantage of adapting to the researcher's needs by enabling zooming, filtering, or selecting specific data. This interactivity facilitates the exploration and comprehension of complex data. These advancements are made possible, in particular, by the development of another branch of artificial intelligence, which intersects with those previously mentioned: natural language processing.

B. The linguistic study of ancient texts

Automatic natural language processing is an essential branch of artificial intelligence (1). It makes it possible to transcribe texts, sometimes including translations, or to reconstruct missing parts (2).

1. Natural language processing

Natural Language Processing (NLP) focuses on the interaction between computers and human language. Its main purpose is to make it possible for machines to understand, interpret, generate and respond to human language in a natural and useful way[51]. To achieve this, NLP consists of several subfields that work together to ensure effective interaction between humans and machines.

Léo BRUN and Pauline VERDIER also highlight the importance of graphs in the history of law for data exploitation: « Le recours aux humanités numérique en histoire du droit », Chronique Culture, *Revue Crises et Société*, 3, 2024 : https://www.crisesetsociete.com

[51] See: David BRENTE, *L'intelligence artificielle expliquée... op. cit.,* p. 69-74.

These include lexical, syntactic, and semantic analysis[52]. Thanks to this technology, it is possible to dictate a text in Word, seek assistance from a chatbot, use Google Translate, and even develop large language models[53] (LLMs) such as GPT[54] and BERT[55]. So how can NLP be used more specifically in legal history?

NLP enables significant advancements in lexicometry and stylometry through both quantitative and qualitative text analysis[56]. Lexicometry involves the statistical study of word and phrase occurrences within a textual corpus. It helps detect patterns, recurrences, and changes in the use of terms over long periods[57]. By examining the frequency and evolution of specific legal concepts, lexicometry can shed light on how notions such as "victim," "property," or "freedom" have evolved over time[58]. Stylometry, on the other hand, is the study

---

[52] Various techniques are used to process and understand natural language: Bag of Words (BoW), TF-IDF (Term Frequency-Inverse Document Frequency), Recurrent Neural Networks (RNN), Transformers, etc.

[53] The LLM, or Large Language Model, is a type of artificial intelligence model trained on vast amounts of text data to understand and generate human language in a coherent and natural manner.

[54] GPT (Generative Pre-trained Transformer) is a family of language models developed by OpenAI, based on the Transformer architecture, which uses neural networks to autonomously generate text after being pre-trained on large corpora of data.

[55] BERT (Bidirectional Encoder Representations from Transformers) is a language model developed by Google in 2018, also based on the Transformer architecture.

[56] These fields can enrich research in the history of law, but linguists are also interested in syntactic analysis, semantics, lemmatization, and linguistic diachrony. Furthermore, a recent discipline has emerged at the intersection of computer science, linguistics, and history: culturomics. This field refers to the study of cultural phenomena through the analysis of large quantities of textual data. It allows for exploring the evolution of ideas, behaviors, and cultural trends over time. See : Pierre MOUNIER, « Les Humanités numériques, gadget ou progrès ? : Enquête sur une guerre souterraine au sein de la recherche », *Revue du Crieur*, n°7, 2017/2, p. 144-159 [en ligne].

[57] For example, the project *The Proceedings of the Old Bailey* has digitized and indexed criminal trials from the Old Bailey in London, covering more than 200 years of judicial history (1674-1913). Researchers have used NLP tools to analyze word frequency, while categorizing information (victims, defendants, gender, offenses, sentences, etc.) to study social trends, changes in criminal law, and concepts of justice over time.

[58] For instance, software such as Iramuteq or Voyant Tools can be used to automatically extract lexical forms from textual corpora and to generate frequency analyses, word clouds, and co-occurrence graphs.

of writing styles through quantitative analysis of linguistic features such as sentence length, word choice, and punctuation[59]. This approach can be used to attribute anonymous or disputed texts to probable authors by comparing stylistic or lexical characteristics with known works. It can help identify scribes and situate texts within a specific historical period. In legal history, stylometry could also be applied to analyze the formalization of judicial discourse, for example, by examining how a court clerk transcribes testimonies.

In parallel, the linguistic tools being developed enable the translation and reconstruction of ancient texts, thereby facilitating their analysis.

2. Translating and reconstructing ancient texts

Depending on the period and location studied, texts may be written in different languages (sometimes including local dialects) or contain ancient abbreviations. Artificial intelligence tools can thus facilitate the translation and interpretation of these texts by providing contextual analyses that help grasp the nuances of the language used. In the context of an online edition of a text, a researcher may choose to produce a diplomatic edition—one that remains as faithful as possible to the original—or decide to modernize the text to make it more accessible[60]. Artificial intelligence can assist in this process, for example, by translating Latin phrases, which are common in legal writings. However, automatic translation of ancient texts, particularly those in dead languages[61], remains a challenge. While these tools are useful, they do not always account for historical and

[59] It was also used in 2020 in the Gregory case to analyze the anonymous letters sent to the child's parents, by studying the writing style, vocabulary, and sentence structures (a study that is distinct from graphological analysis). This criminal case remains unsolved to this day. It began in 1984, after the body of a 4-year-old boy, Grégory Villemin, was discovered, in a context marked by family rivalries.

[60] In this regard, we highlight the existence of the EMAN platform (Édition de Manuscrits et d'Archives Numériques), which brings together various projects that disseminate and utilize documents and archival collections: https://eman-archives.org/EMAN/

[61] Especially since Latin must, for example, be distinguished from Neo-Latin, as they do not share the same subtleties.

contextual subtleties. Therefore, it is essential to complement their use with specialized expertise to ensure the accuracy and fidelity of the translated text.

Moreover, NLP can help fill gaps in damaged texts by suggesting hypotheses for missing parts. Algorithms can be used to reconstruct missing letters or words[62]. While this technique is already applied to complete writings from ancient Mesopotamia[63], it could also be useful for restoring texts when the paper is damaged – for instance, when ink has bled through the paper or when worms have made holes in it, making certain passages illegible[64].

In conclusion, artificial intelligence could soon become an essential tool for the study of legal history, facilitating access to historical sources and offering new methods of analysis and interpretation. By automating document processing and uncovering previously unseen connections between texts, artificial intelligence enables legal historians to explore research avenues that were previously inaccessible and to deepen our understanding of the evolution of legal systems.

However, this digital revolution also presents a few challenges, particularly in terms of ethics, data biases, accuracy of analyses, and data retention. Researchers

---

[62] NLP can analyze text fragments, understand the context, and predict missing words or phrases.

[63] See for example : E. FETAYA, Y. LIFSHITZ, E. AARON, S. GORDIN, « Restoration of fragmentary Babylonian texts using recurrent neural networks », *Proc. Natl. Acad. Sci. U.S.A.*, 117 (37), 2020, https://doi.org/10.1073/pnas.2003794117 ; Koren LAZAR, Benny SARET, Asaf YEHUDAI, Wayne HOROWITZ, Nathan WASSERMAN and Gabriel STANOVSKY, « Filling the Gaps in Ancient Akkadian Texts: A Masked Language Modelling Approach », in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, p. 4682–4691. A technique that is also used for Greek epigraphy: Yannis ASSAEL, Thea SOMMERSCHIELD and Jonathan PRAG, "Restoring ancient text using deep learning: a case study on Greek epigraphy", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, p. 6368–6375.

[64] One could even consider using artificial intelligence in a manner similar to predictive justice, for instance, when the decision is missing in historical judicial cases. By analyzing similar cases, AI could, based on statistical analyses and predictive models, suggest the most likely ruling that was issued. The potential of artificial intelligence in this field is considerable, though it is essential to approach its application with caution.

will inevitably need to address these issues in order to fully harness the opportunities offered by artificial intelligence. It is essential to have a thoughtful and controlled use of these technologies to fully exploit their potential, while maintaining the rigor and reliability required in scientific research. Projects involving artificial intelligence in the field of legal history are multiplying, but they remain relatively few and, of course, depend on the available funding[65]. To date, digital humanities have been primarily viewed as a means of producing, exchanging, and disseminating legal knowledge[66], rather than as a tool for analyzing sources. Such analyses require technical skills that are sometimes lacking, and often involve quantitative studies, which are not traditionally part of the discipline[67]. Researchers may thus be reluctant. However, artificial intelligence in no way replaces human expertise; it is a complementary tool that primarily serves to enhance researchers' analytical capabilities. This technological contribution, combined with the growing interest it generates, will undoubtedly promote research in this direction[68]. Undeniably, artificial

[65] Legal historians are nonetheless aware of the opportunities brought by the development of artificial intelligence, as demonstrated by the study day "LARTI Project – Law and Artificial Intelligence: Historical Perspective and Contemporary Challenges", held under the direction of Claire Bouglé-Le Roux and Mélanie Clément-Fontaine on February 11, 2025.
Similarly, for the past two years, a series of webinars titled Histoire du Droit 2.0 has been organized by Ninon Maillard and Louise Testot-Ferry.

[66] Most projects focus on providing access to sources and enhancing their visibility, as exemplified by Criminocorpus Lab, a platform dedicated to the history of justice through digital humanities, and ParleFlandre, which facilitates research within the archives of the Parliament of Flanders.
Additionally, virtual exhibitions are being developed, such as those hosted by the Interuniversity Library of Cujas.

[67] Work of this type is quite rare, as Nader Hakim and Annamaria Monti pointed out: art. cit.
Moreover, despite technological advances, researchers still face technical and methodological obstacles, highlighting the importance of an interdisciplinary approach in projects undertaken.

[68] They are also beginning to appear in theses, such as that of Emmanuel RAVESTEIN, *Les hautes juridictions criminelles de l'Ancien Régime à la Révolution : continuité et rupture, de la Provence au département des Bouches-du-Rhône (1781-1795)*, thesis in legal history, University of Aix-Marseille, 2018 or the thesis in progress by Morgane PICA, *L'intertextualité dans la réflexion sur la coutume de Normandie, miroir de l'univers intellectuel des praticiens du droit à l'époque moderne*, under the direction of Patrick Arabeyre, Professor of Legal history at the École nationale des chartes and Tobias Hodel, assistant professor of Digital Humanities at the University of Bern.

intelligence is an important tool that reveals the full potential of data, allowing us to rediscover our past and invent our future.