L'intelligence artificielle au service du **traitement des archives**

Utilisée aux Archives nationales depuis 2015, l'intelligence artificielle permet de traiter les fonds de manière inédite, principalement grâce à la reconnaissance d'écritures manuscrites, au bénéfice des archivistes comme des lecteurs.

ARCHIVES NATIONALES

L'intelligence artificielle, en plus d'être une notion de plus en plus divulguée, est devenue un outil à part entière dans les administrations pour tirer parti des vastes réservoirs de données qu'elles produisent. Le monde de la recherche, de la culture et du patrimoine est lui aussi concerné par ce mouvement de fond : désormais prédomine le sentiment que le stade de l'expérimentation est dépassé pour entrer dans une nouvelle phase des pratiques professionnelles. Aux Archives nationales, c'est à partir de 2015 que l'intelligence artificielle a été employée pour traiter les fonds de manière inédite, principalement en ayant recours à la reconnaissance d'écritures manuscrites, ou *handwritten text recognition* (HTR). Cette technologie s'appuie sur les processus de *machine learning,* qui consistent, à partir de jeux de données, à entraîner une machine à effectuer des actions humaines : ici, en l'occurrence, la transcription de textes manuscrits. Cette technologie répond au rêve longtemps caressé de pouvoir traiter les documents à l'instar de ce qu'était capable d'accomplir l'OCR sur les imprimés. Elle a frayé plusieurs pistes pour le traitement, la diffusion et l'utilisation des archives.

ACCÉDER AU TEXTE DES ARCHIVES : HIMANIS ET ENDP

Lancé fin 2014 par l'IRHT (Institut de recherche et d'histoire des textes) sous l'égide de Dominique Stutzmann, le projet HIMANIS[1] présenta la singularité de choisir comme « terrain de jeu » les registres de la chancellerie royale des XIII[e] et XV[e] siècles conservés dans le Trésor des chartes. Si l'HTR était alors couramment utilisé sur les écritures contemporaines, il n'avait jamais été employé pour transcrire des écritures si anciennes, comportant de nombreuses abréviations. Le corpus à traiter, soit 199 registres pour plus de 83 000 pages, était en outre suffisamment volumineux pour évaluer la maturité de la technologie.

Les premiers résultats ont été obtenus rapidement grâce aux images déjà disponibles au début du projet et aux éditions électroniques d'actes royaux qui fournissaient une « vérité de terrain »[2], c'est-à-dire une transcription d'une partie du corpus assez exacte pour entraîner l'intelligence artificielle à comprendre l'écriture et ses mécanismes d'abréviations. Les données obtenues sont des lemmes alignés sur leur image d'origine et sont interrogeables *via* une interface de recherche[3].

Le chercheur a la possibilité de retrouver directement au cœur du document les termes de son choix et d'effectuer aussi bien des études d'ordre historique que philologique en s'appuyant sur la statistique lexicale. Du point de vue de l'archiviste, ce moteur de recherche pallie l'absence d'inventaire complet de ces registres. On notera néanmoins qu'un tel outil ne peut remplacer un travail d'indexation classique. Face à la matière brute du texte, c'est à l'utilisateur final de retrouver les termes du Moyen Âge qui traduisent un phénomène ou un concept, de connaître les formes lexicales du latin et de l'ancien français pour retrouver toutes les occurrences pertinentes, d'envisager tout ce qu'une indexation matière pouvait lui suggérer.

Dans la continuité d'HIMANIS, il faut aussi évoquer le projet eNDP, débuté en 2020 et qui porte sur les registres des décisions du chapitre de Notre-Dame de Paris. Autre exemple de coopération entre institutions de conservation et structures de recherche, le but de ce projet, qui a actuellement réussi le stade de la transcription par HTR, vise à explorer le contexte social, économique et urbain dans lequel évoluait le chapitre cathédral[4].

ACCOMPAGNER LE TRAITEMENT DES ARCHIVES : LECTAUREP ET SIMARA

Une autre possibilité d'exploiter l'intelligence artificielle consiste à récupérer, à partir des documents d'archives, des informations afin de les réutiliser dans le travail de description documentaire. LectAuRep, projet financé par le ministère de la Culture, mené par le Minutier central des notaires parisiens et l'INRIA, a été conduit de 2018 à 2021[5]. Il porte sur les répertoires de notaires, qui sont les clés d'accès fondamentales à leurs minutes, en donnant la date, l'objet de l'acte et le nom des parties. L'HTR a permis de transcrire ces informations. Elles fournissent un ensemble de métadonnées utiles aussi bien pour décrire les documents que pour se prêter à des explorations statistiques. L'équipe est aussi allée plus loin en ayant recours à la technologie de

[1] L'acronyme signifie : « *HIstorical MANuscripts Indexing for user-controlled Search* ». Carnet de recherche : www.himanis.hypotheses.org

[2] Notamment P. Guérin et L. Celier, *Recueil des documents concernant le Poitou contenus dans les registres de la chancellerie de France,* 14 vol., Poitiers, 1881 ; édition électronique par l'École des chartes : www.corpus.enc.sorbonne.fr/actesroyauxdupoitou

[3] Hébergement par Huma-Num : www.himanis.huma-num.fr/app

[4] Présentation du projet : https://lamop.hypotheses.org

[5] LectAuRep signifie : lecture automatique de répertoires. Carnet de recherche du projet : www.lectaurep.hypotheses.org

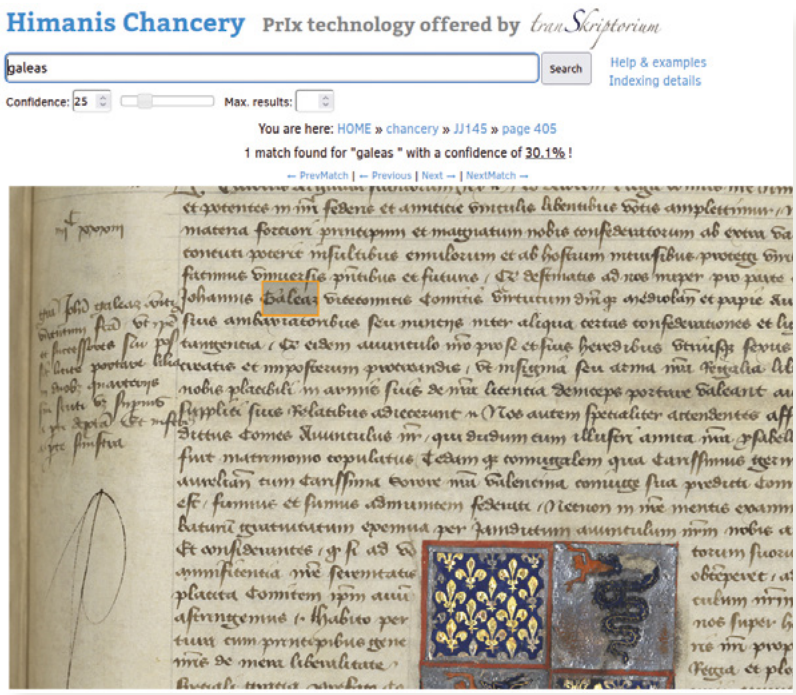
reconnaissance des entités nommées (NER), afin d'accompagner davantage l'archiviste dans le travail de constitution d'index.

Le projet SIMARA[6], soutenu par le plan France Relance, porte quant à lui sur la conversion d'inventaires d'archives anciens en données. Il existe en effet une masse importante d'inventaires manuscrits, allant du XVIIIe au XXe siècle, qui demeurent à ce jour peu accessibles aux lecteurs et qui n'ont fait l'objet d'aucun traitement, hormis une numérisation en mode image. SIMARA vise à traiter environ 800 000 fiches et 100 000 pages. Il s'agit d'une plateforme Web développée par la société Teklia qui l'a adossée à ses infrastructures de traitement HTR. Son but est d'accomplir deux tâches chronophages auparavant réalisées séparément et manuellement par les archivistes : la saisie bureautique de l'inventaire et la structuration des informations en XML EAD. Ces deux opérations sont effectuées conjointement par la plateforme : les outils de segmentation et d'HTR se chargent de transcrire le texte manuscrit, tandis que des outils d'identification des contenus répartissent les informations dans des champs correspondant à des éléments et des attributs XML. L'archiviste peut ainsi se concentrer uniquement sur la relecture des données, en corrigeant les transcriptions, contribuant au passage à améliorer les modèles de reconnaissance d'écritures. Il peut ensuite récupérer le tout au format EAD pour le publier. Ce travail de relecture demeure forcément long et exigeant mais la prise en charge par l'intelligence artificielle des tâches les plus fastidieuses de saisie et d'encodage permettent de réduire significativement le temps de traitement.

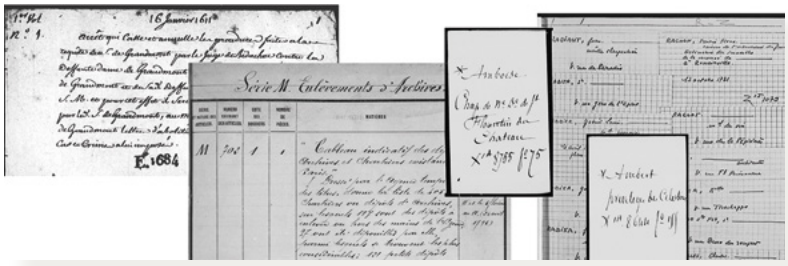
UN PROJET DE PLATEFORME COLLABORATIVE

Face au succès de la solution, une plateforme collaborative pour associer le public à la relecture des transcriptions automatiques sera mise en œuvre (projet GIROPHARES). Les expériences conduites aux Archives nationales ont essentiellement exploité la reconnaissance d'écritures manuscrites, afin d'aller plus loin dans la transformation des fonds d'archives en données. Cette transformation se fait tant au bénéfice des archivistes, pour produire métadonnées et inventaires, que des lecteurs qui réutilisent ces données pour évaluer les phénomènes historiques. De tels projets se multiplient aussi dans le réseau des archives départementales, notamment avec le projet SOCFACE[7] qui, dans la même logique, explore les recensements de population de 1836 à 1936.

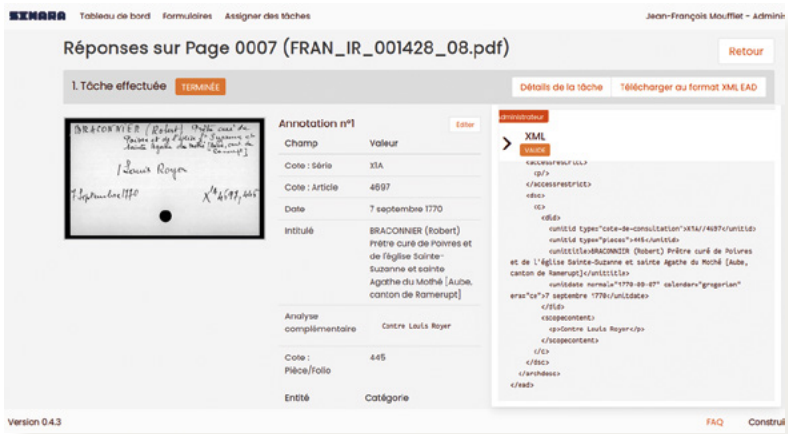
JEAN-FRANÇOIS MOUFFLET
Conservateur en chef du patrimoine, responsable de fonds au département du Moyen Âge et de l'Ancien Régime, Archives nationales
jean-francois.moufflet@culture.gouv.fr



Recherche et reconnaissance d'un même terme au cœur du document.



Les registres des actes notariés sont l'une des sources historiques les plus consultées aux Archives nationales.



Structuration du texte dans un document XML conforme aux spécifications de la *Text Encoding Initiative*.

[6] Acronyme signifiant : Saisie d'Inventaires Manuscrits Assistée par Reconnaissance Automatique. Présentation en ligne pour Etalab : www.speakerdeck.com/etalabia/20220203-datadrink-simara

[7] Présentation du projet : www.socface.site.ined.fr