ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Mohammed K. I. Mechentel
licencié ès lettres

Reconnaissance automatique de l'écriture manuscrite au service d'une grande institution muséale.

Mémoire pour le diplôme de master
« Technologies numériques appliquées à l'histoire »

2024

Résumé

Ce mémoire a été réalisé dans le cadre d'un stage au Musée du Louvre. Il présente un état des lieux de la reconnaissance automatique de l'écriture manuscrite (HTR) en retraçant son histoire et son évolution. Le document explore ensuite en détail les principes fondamentaux des projets HTR, ainsi que les diverses contraintes techniques et méthodologiques associées à leur mise en œuvre.Dans le cadre de ce travail, un projet HTR a été mis en place sous la forme d'un projet pilote, élaboré par le service de l'ingénierie documentaire, des images et de la traduction du Louvre. Ce projet pilote porte spécifiquement sur les sources de documentation du Département des Antiquités grecques, étrusques et romaines (DAGER) et du Service de l'histoire du Louvre (SHL). Le mémoire analyse ainsi les différentes étapes de ce projet, de sa conception théorique à son exécution pratique, en mettant en lumière les défis techniques et organisationnels rencontrés, ainsi que les résultats obtenus et les perspectives pour l'avenir de la reconnaissance automatique des écritures manuscrites au Louvre.

This master thesis was produced as part of an internship at the Musée du Louvre. It presents an overview of handwritten text recognition (HTR), tracing its history and evolution. The paper then explores in detail the fundamental principles of HTR projects, as well as the various technical and methodological constraints associated with their implementation.As part of this work, an HTR project was set up in the form of a pilot project, developed by the Louvre's docuemntary, image and translation engineering department. This pilot project focuses specifically on the documentation sources of the *Département des Antiquités grecques, étrusques et romaines* (DAGER) and the *Service de l'histoire du Louvre* (SHL). The dissertation analyzes the various stages of this project, from its theoretical conception to its practical implementation, highlighting the technical and organizational challenges encountered, as well as the results obtained and prospects for the future of handwritten text recognition at the Louvre.

Mots-clés : HTR ; Musée du Louvre ; registres d'inventaires ; cahier d'inventaires de fouille ;

Informations bibliographiques : Mohammed K. I. Mechentel, *Reconnaissance automatique de l'écriture manuscrite au service d'une grande institution muséale*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Emmanuelle Bermès, Sybille Clochet, École nationale des chartes, 2024.

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude envers mes deux responsables de stage, Sybille Clochet et Emmanuelle Bermès, qui ont permis à cette expérience de se dérouler dans les meilleures conditions possibles. Mes remerciements s'adressent également à l'ensemble de l'équipe du service de l'ingénierie documentaire, des images et de la traduction du Musée. Je garde en mémoire les liens enrichissants que nous avons tissés au cours de ce stage et les nombreux enseignements tirés de nos échanges. Je souhaite aussi remercier chaleureusement tous les membres des services de documentation du Musée, en particulier Noémie Latte et Laura Favreau, qui m'ont accueilli avec une grande bienveillance et enthousiasme. Enfin, mes pensées reconnaissantes se tournent vers ma famille — ma mère, mon père, mes frères et mes sœurs — pour leur soutien inconditionnel et précieux.

Bibliographie

[1] BARRERE (Killian), *Architectures de Transformer légères pour la reconnaissance de textes manuscrits anciens*, Thèse de doctorat, INSA de Rennes, [s.l.], 20 décembre 2023. URL : `https://hal.science/tel-04385383`. Consulté le 6 août 2024.

[2] CAMPS (Jean-Baptiste), CLÉRICE (Thibault) et PINCHE (Ariane), « Noisy medieval data, from digitized manuscript to stylometric analysis : Evaluating Paul Meyer's hagiographic hypothesis », *Digital Scholarship in the Humanities*, vol. 36, Supplement_2 (novembre 2021), p. ii49-ii71.

[3] CAPURRO (Carlotta), PROVATOROVA (Vera) et KANOULAS (Evangelos), « Experimenting with Training a Neural Network in Transkribus to Recognise Text in a Multilingual and Multi-Authored Manuscript Collection », *Heritage*, vol. 6, n 12 (novembre 2023), p. 7482-7494.

[4] CHAUDHURI (Arindam), MANDAVIYA (Krupa), BADELIA (Pratixa) et al., *Optical character recognition systems for different languages with soft computing*, Cham, 2017 (Studies in fuzziness and soft computing, n Volume 352). ISBN : 978-3-319-50252-6.

[5] DIMAURO (G.), IMPEDOVO (S.), PIRLO (G.) et al., « Automatic Bankcheck Processing : A New Engineered System », *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 11, n 04 (juin 1997), p. 467-504.

[6] DIMOND (T. L.), « Devices for reading handwritten characters », dans *Papers and discussions presented at the December 9-13, 1957, eastern joint computer conference : Computers with deadlines to meet on XX - IRE-ACM-AIEE '57 (Eastern)*, présenté à Papers and discussions presented at the December 9-13, 1957, eastern joint computer conference : Computers with deadlines to meet, Washington, D.C., ACM Press, 1958. URL : `http://portal.acm.org/citation.cfm?doid=1457720.1457765`. Consulté le 16 août 2024, p. 232-237.

[7] ESTILL (Laura) et LEVY (Michelle), « Chapter 12Evaluating digital remediations of women's manuscripts », *Digital Studies / Le champ numérique*, vol. 6, n 6 (juillet 2016). URL : `https://www.digitalstudies.org/article/id/7296/`. Consulté le 17 août 2024.

[8] GILLOUX (Michel), « Research into the new generation of character and mailing address recognition systems at the French post office research center », *Pattern Recognition Letters*, vol. 14, n 4 (avril 1993), p. 267-276.

[9] GORSKI (Nikolai), ANISIMOV (Valery), AUGUSTIN (Emmanuel) et al., « Industrial bank check processing : the A2iA CheckReaderTM », *International Journal on Document Analysis and Recognition*, vol. 3, n 4 (mai 2001), p. 196-206.

[10] GOVINDAN (V.K) et SHIVAPRASAD (A.P), « Character recognition — A review », *Pattern Recognition*, vol. 23, n 7 (janvier 1990), p. 671-683.

[11] CAMPS (Jean-Baptiste), « Homemade manuscript OCR (1) : OCRopy » (février 2017). URL : `https://graal.hypotheses.org/786`. Consulté le 15 août 2024.

[12] LECUN (Yann), BOTTOU (Léon), BENGIO (Yoshua) et al., « Gradient-based learning applied to document recognition », *Proceedings of the IEEE*, vol. 86, n 11 (novembre 1998), p. 2278-2324.

[13] LEEDHAM (C.), « Historical perspectives of handwriting recognition systems », dans , [s.l.], [s.n.], 11 mars 1994. URL : `https://www.semanticscholar.org/paper/Historical-perspectives-of-handwriting-recognition-Leedham/687ba32285fd386af22837ef3884bda6b165b97c`. Consulté le 17 août 2024.

[14] LIT (L. W. C. van), *Among digitized manuscripts : philology, codicology, paleography in a digital world*, Leiden ; Boston, 2020 (Handbook of oriental studies = Handbuch der Orientalistik. Section one, The Near and Middle East, n volume 137). Z105. ISBN : 978-90-04-40035-1.

[15] MUEHLBERGER (Guenter), SEAWARD (Louise), TERRAS (Melissa) et al., « Transforming scholarship in the archives through handwritten text recognition : Transkribus as a case study », *Journal of Documentation*, vol. 75, n 5 (septembre 2019), p. 954-976.

[16] PINCHE (Ariane), *Guide de transcription pour les manuscrits du Xe au XVe siècle*, [s.l.], juin 2022. URL : `https://hal.science/hal-03697382`. Consulté le 7 août 2024.

[17] PINCHE (Ariane), CLÉRICE (Thibault), CHAGUÉ (Alix) et al., « CATMuS-Medieval : Consistent Approaches to ing ManuScripts », dans , présenté à DH2024, [s.l.], [s.n.], 5 août 2024. URL : `https://inria.hal.science/hal-04346939`. Consulté le 7 août 2024.

[18] PLAMONDON (R.) et SRIHARI (S.N.), « Online and off-line handwriting recognition : a comprehensive survey », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n 1 (janvier 2000), p. 63-84.

[19] POMART (Julien), « L'apport des outils numériques dans une stratégie de valorisation des archives », Billet, *Archives de la FMSH*, 11 novembre 2014. URL : `https://archivesfmsh.hypotheses.org/1297`. Consulté le 18 août 2024.

[20] RABINER (L.R.), « A tutorial on hidden Markov models and selected applications in speech recognition », *Proceedings of the IEEE*, vol. 77, n 2 (février 1989), p. 257-286.

[21] RAKESH (S.), KUSHAL REDDY (P.), PRASHANTH (V.) et al., « Handwritten text recognition using deep learning techniques : A survey », *MATEC Web of Conferences*, éd. K. Satyanarayana, P.B. Bobba, A. Perveen, et al., vol. 392 (2024), p. 01126.

[22] SCHANTZ (H. F.), *The History of OCR, optical character recognition*, Manchester Center, Vt, 1982. ISBN : 978-0-943072-01-2.

[23] SCHEITHAUER (Hugo), CHAGUÉ (Alix), ROSTAING (Aurélia) et al., « Production d'un modèle affiné de reconnaissance d'écriture manuscrite avec eScriptorium et évaluation de ses performances », dans , présenté à Les Futurs Fantastiques - 3e Conférence Internationale sur l'Intelligence Artificielle appliquée aux Bibliothèques, Archives et Musées, AI4LAM, [s.l.], [s.n.], 8 décembre 2021. URL : `https://inria.hal.science/hal-03538195`. Consulté le 6 août 2024.

[24] STUTZMANN (Dominique), KERMORVANT (Christopher), VIDAL (Enrique) et al., « Handwritten Text Recognition, Keyword Indexing, and Plain Text Search in Medieval Manuscripts », dans , présenté à ADHO / EHD 2018 - Mexico City, [s.l.], [s.n.], 2018. URL : `https://dh-abstracts.library.cmu.edu/works/6324`. Consulté le 7 août 2024.

[25] TERRAS (Melissa), « The rise of digitisation : An overview », dans *Digital Perspectives*, 2010. URL : `https://www.research.ed.ac.uk/en/publications/the-rise-of-digitisation-an-overview`. Consulté le 17 août 2024, p. 3-20.

[26] TERRIEL (Lucas), « Atelier : Production d'un modèle affiné de reconnaissance d'écriture manuscrite avec eScriptorium et évaluation de ses performances. Évaluer son modèle HTR/OCR avec KaMI (Kraken as Model Inspector) », dans *Les Futurs Fantastiques - 3e Conférence Internationale sur l'Intelligence Artificielle appliquée aux Bibliothèques, Archives et Musées*, Paris, France, AI4LAM and Bibliothèque nationale de France, décembre 2021. URL : `https://hal.science/hal-03495762`. Consulté le 6 août 2024.

[27] TUFFÉRY (Christophe), « Retour d'expériences sur l'utilisation comparée de plusieurs dispositifs de transcription numérique d'archives de fouilles archéologiques », dans *Documents Anciens et HTR*, Paris, École nationale des chartes, 2022.

[28] UL-HASAN (Adnan), BUKHARI (Syed Saqib) et DENGEL (Andreas), « OCRoRACT : A Sequence Learning OCR System Trained on Isolated Characters », dans *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, présenté à *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, San-

torini, Greece, IEEE, avril 2016. URL : `http://ieeexplore.ieee.org/document/7490113/`. Consulté le 17 août 2024, p. 174-179. ISBN : 978-1-5090-1792-8.

[29] *Recent trends in image processing and pattern recognition : Third International Conference, RTIP2R 2020, Aurangabad, India, January 3–4, 2020 : revised selected papers. Part 2*, éd. K.C. Santosh, B. Gawali, présenté à *RTIP2R*, Singapore, Springer, 2021 (Communications in computer and information science, n 1381). ISBN : 9789811604935.

[30] *Système de Reconnaissance d'Ecriture Manuscrite*, 1. Aufl., Saarbrücken Éditions universitaires européennes 2015, 2015. ISBN : 978-3-8417-4433-3.

[31] *Handbook of document image processing and recognition : with 98 tables*, éd. D.S. Doermann, London Heidelberg, Springer, 2014 (Springer reference). ISBN : 978-0-85729-858-4.

Introduction

Créé à l'aube de la Révolution, le musée du Louvre a connu au cours de son histoire de nombreux changements de statut. Depuis 1992, il est reconnu comme un établissement public, ce qui lui confère des missions générales définies par l'État. À ce titre, le musée du Louvre a pour mission de conserver, protéger et restaurer pour le compte de l'État les œuvres qui font partie des collections inscrites sur ses inventaires, tout en les présentant au public. Il doit également assurer l'accueil des visiteurs, encourager la fréquentation du musée et promouvoir la connaissance de ses collections par tous les moyens appropriés. Le musée est également chargé de contribuer à l'éducation, la formation et la recherche dans les domaines de l'histoire de l'art, de l'archéologie et de la muséographie, ainsi que de gérer un auditorium et d'élaborer sa programmation. Enfin, il a pour mission de conduire des études scientifiques sur ses collections et de préserver, gérer et mettre en valeur les immeubles dont il est doté.

Pour documenter et avoir une connaissance de ses collections, le Louvre dispose de ressources documentaires uniques et très rares. Elles sont primordiales pour la production de recherche autour de tout ce que le musée peut conserver.

En 2020, le musée a lancé un programme de numérisation de ces ressources pour en accroître l'accessibilité et permettre l'extraction et l'exploitation des données qu'elles contiennent. Les documents manuscrits, aux côtés des fonds photographiques, forment une part importante des documents qui enrichissent les collections. Faciliter l'accès aux données qu'ils renferment est un enjeu majeur pour la documentation du musée.

Cet objectif peut désormais être appréhendé dans un contexte de développement des technologies de reconnaissance automatique d'écriture manuscrite (HTR) [1]. Dans le but de développer une stratégie efficace pour les prochaines années, prenant en compte les contraintes, les spécificités et les exigences de la technique, le musée du Louvre a décidé de lancer un projet pilote. Parmi l'ensemble de ses ressources documentaires, le choix s'est porté sur une sélection de registres d'inventaire du Département des Antiquités

1. « HTR » est l'abréviation la plus couramment employée pour référer à la reconnaissance automatique d'écriture manuscrite. Elle provient du terme originel anglais *handwritten recognition*.

grecques, étrusques et romaines (), ainsi que des cahiers d'inventaire des fouilles de la Cour Napoléon [2] du Service de l'histoire du Louvre (SHL), deux sources numérisées entre 2022 et 2023.

Ce projet pilote s'inscrit dans le cadre d'un stage de 4 mois, comprenant une partie théorique visant à établir un état des lieux autour du HTR et une partie pratique mettant en œuvre la technique. Le présent mémoire de stage restituera les recherches effectuées à cet égard tout en documentant la mise en œuvre du projet pilote.

Mais qu'est-ce que l'HTR ? Tout comme la reconnaissance de caractère optique (OCR) dont elle est voisine, on pourrait la définir comme la « prédiction d'un contenu textuel à partir d'une image de la source, par une intelligence artificielle entraînée par des données qui peut s'intégrer dans un processus alternant phases d'intervention humaines et phases de calcul. ». La différence entre ces deux techniques réside simplement dans le type d'écriture auquel elles sont confrontées : l'écriture imprimée pour l'OCR, l'écriture manuscrite pour l'HTR. En d'autres termes, la reconnaissance d'écriture manuscrite est une technologie qui utilise l'intelligence artificielle pour repérer les éléments manuscrits d'une image et les convertir en texte en devinant à quel type de caractères correspondent les formes des lettres.

Le processus de reconnaissance d'écriture manuscrite repose sur le machine learning (apprentissage automatique), un domaine de l'intelligence artificielle où une machine apprend et s'améliore à partir de données. En analysant une grande quantité d'exemples, les algorithmes de machine learning identifient des motifs qui améliorent leurs prédictions. Le résultat de l'application d'un algorithme à un ensemble de données est appelé un "modèle". C'est l'objet final, prêt à être utilisé, capable de faire des prédictions sur de nouvelles données. Néanmoins, cela ne signifie pas que le modèle ne peut pas être amélioré par la suite : il est possible de l'entraîner avec de nouvelles données. Ainsi, il est en constante évolution en fonction des nouvelles données dont il dispose.

Dans le cadre du HTR, les modèles sont entraînés à l'aide de ce que l'on nomme la « vérité terrain ». Il s'agit d'images contenant une écriture manuscrite accompagnées de leur transcription exacte. Ces transcriptions, réalisées par des humains, servent de « réponses correctes » que le modèle doit apprendre à reproduire lorsque lui soumettra d'autres images contenant des écritures manuscrites. Par ailleurs, plus ces données sont nombreuses et proviennent d'une même écriture ou d'écritures voisines, plus les prédictions du modèle sont susceptibles d'être précises.

2. Ces fouilles ont été réalisées entre 1984 et 1986.

Longtemps chasse gardée de la recherche en informatique et en intelligence artificielle, le HTR n'a été introduit que récemment dans le monde des humanités numériques et des institutions patrimoniales. Contrairement à l'OCR, qui est utilisé depuis le début des années 2000, les premiers grands projets mobilisant le HTR remontent au milieu des années 2010. Devant cet état de fait, il nous paraît pertinent de nous demander quels sont les principaux défis et perspectives auxquels une grande institution muséal comme le Louvre peut être confrontée lorsqu'il souhaite utiliser la reconnaissance d'écriture manuscrite. Pour apporter des éléments de réponse, nous commencerons par examiner l'histoire de cette technique afin de comprendre ses récentes évolutions et la dynamique qui les accompagne. Ensuite, nous identifierons les principes fondamentaux d'un projet qui l'utilise, tout en analysant les contraintes techniques rencontrées lors de sa mise en œuvre, pour déterminer la solution la plus adéquate. Enfin, nous évaluerons les besoins du musée dans ce domaine et les premiers résultats de son application, afin de mieux cerner ses promesses et son potentiel.

Chapitre 1

Comprendre l'environnement et l'esprit autour de la reconnaissance d'écriture manuscrite : une histoire synthétique de la technique.

1.1 1950 - 2016 : un domaine de recherche de l'informatique et de l'intelligence artificielle.

La reconnaissance d'écriture, qu'il s'agisse de caractère manuscrit ou imprimé, a constitué, très tôt, un domaine important de la recherche en informatique (computer science) dès le milieu du XX^{e} et un des premiers problèmes qui a été posé à l'intelligence artificielle [1]. Une succession d'évolutions techniques dans le domaine ont permis l'amélioration de la reconnaissance d'écriture au cours des décennies. L'emploi des statistiques avancées dans les années 1980 notamment de type modèles de Markov cachés, de la reconnaissance de forme dans les années 1990, puis le développement des réseaux de neurones entre 2000 et 2010 ont permis aux algorithmes de reconnaissance d'écriture d'être de plus en plus performants [2]. Cette amélioration générale des performances des algorithmes est également due à l'évolution matérielle de l'informatique, notamment à la démocratisation croissante de processeurs de plus en plus puissants.

1. Voir à titre d'exemple le travail précurseur de DIMOND (T. L.), « Devices for reading handwritten characters », dans *Papers and discussions presented at the December 9-13, 1957, eastern joint computer conference : Computers with deadlines to meet on XX - IRE-ACM-AIEE '57 (Eastern)*, présenté à Papers and discussions presented at the December 9-13, 1957, eastern joint computer conference : Computers with deadlines to meet, Washington, D.C., ACM Press, 1958. URL : `http://portal.acm.org/citation.cfm?doid=1457720.1457765`. Consulté le 16 août 2024, p. 232-237.

2. RABINER (L.R.), « A tutorial on hidden Markov models and selected applications in speech recognition », *Proceedings of the IEEE*, vol. 77, n 2 (février 1989), p. 257-286.

Si l'on s'intéresse à la production académique sur le développement des techniques de reconnaissance d'écriture, on s'aperçoit que la reconnaissance optique de caractère (OCR) et la reconnaissance d'écriture manuscrite (HTR) sont étroitement liées, pour ne pas dire souvent associées. En 2016 encore, des articles académiques d'importances traitaient encore les deux à la même enseigne [3]. En d'autres termes, l'écriture manuscrite et imprimée étaient donc étudiées sous un même paradigme technique. Ceci n'était évidemment pas sans poser de véritables difficultés aux chercheurs en informatique. Les textes imprimés constituaient pour l'apprentissage machine des données définies, standardisées, prévisibles et régulières. Les textes manuscrits, à l'exception de formes très standardisées de calligraphies, étaient aussi variables que le nombre de mains qui les produisaient [4]. Ainsi, jusqu'à l'avènement des réseaux de neurones qui a permis de la démocratiser, la reconnaissance des écritures manuscrites a longtemps constitué de grands défis pour les chercheurs en informatique [5].

Si notre stage a pour but d'appliquer cette technique sur registres d'inventaire et des cahiers de fouilles, il est importer de noter que les premières sources sur lesquelles l'application de la technique s'est appliquée étaient issues principalement du secteur financier et commercial (notamment pour la vérification des écritures sur les chèques) [6] et le secteur postal [7]. Sa mise en œuvre sur des sources patrimoniales ne s'est démocratisée que dans les années 2010. Ce n'est qu'après les programmes de numérisation de masse par des institutions (principalement des bibliothèques et des archives) qui avaient comme visée de permettre un plus grand accès à leur collection que la reconnaissance d'écriture sur ce type de document est devenue plus fréquente [8]. Elle répond en effet à une demande de l'usager ou du chercheur d'effectuer des recherches textuelles dans les textes numérisées en ligne. Cette dynamique a ainsi permis de faire progresser les techniques de reconnaissance

3. UL-HASAN (Adnan), BUKHARI (Syed Saqib) et DENGEL (Andreas), « OCRoRACT : A Sequence Learning OCR System Trained on Isolated Characters », dans *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, présenté à *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, Santorini, Greece, IEEE, avril 2016. URL : `http://ieeexplore.ieee.org/document/7490113/`. Consulté le 17 août 2024, p. 174-179.

4. PLAMONDON (R.) et SRIHARI (S.N.), « Online and off-line handwriting recognition : a comprehensive survey », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n 1 (janvier 2000), p. 63-84.

5. CAPURRO (Carlotta), PROVATOROVA (Vera) et KANOULAS (Evangelos), « Experimenting with Training a Neural Network in Transkribus to Recognise Text in a Multilingual and Multi-Authored Manuscript Collection », *Heritage*, vol. 6, n 12 (novembre 2023), p. 7482-7494.

6. GORSKI (Nikolai), ANISIMOV (Valery), AUGUSTIN (Emmanuel) et al., « Industrial bank check processing : the A2iA CheckReaderTM », *International Journal on Document Analysis and Recognition*, vol. 3, n 4 (mai 2001), p. 196-206.

7. GILLOUX (Michel), « Research into the new generation of character and mailing address recognition systems at the French post office research center », *Pattern Recognition Letters*, vol. 14, n 4 (avril 1993), p. 267-276.

8. TERRAS (Melissa), « The rise of digitisation : An overview », dans *Digital Perspectives*, 2010. URL : `https://www.research.ed.ac.uk/en/publications/the-rise-of-digitisation-an-overview`. Consulté le 17 août 2024, p. 3-20.

d'écritures manuscrites dans leur globalité mais également d'insuffler une émulation dans le champ des sciences humaines, plus particulièrement dans le champ des humanités numériques où ont émanés de plus en plus de projets requérant la reconnaissance d'écriture optique ou manuscrite [9].

1.2 2016 - 2019 : l'effervescence de la technique dans les humanités numériques.

Comme nous venons de le voir, la numérisation massive de collections patrimoniales par leurs institutions couplée à leur catalogage et au développement d'instruments de recherche en ligne a permis au public et aux chercheurs un accès aisé à une grande salve de documents historiques. La reconnaissance d'écriture imprimée et manuscrite a été mobilisée pour faciliter un peu plus cet accès en permettant à l'usager de repérer plus facilement des sujets, des mots, des lieux, des personnes, ou encore des événements qu'ils souhaitaient trouver au sein des documents numérisés. De manière générale, le rapport aux textes et aux questions de recherches en lien avec ceux-ci s'est transformé en permettant des possibilités de recherche plus accrues, plus larges ou mieux ciblées. Laura Estill docteure en humanités numériques et Michelle Levy docteure en littérature perçoivent cette évolution comme potentiellement révolutionnaire dans un chapitre d'ouvrage consacré à l'évaluation des remédiations numériques de manuscrits féminins en 2016 :

> « Les numérisations de manuscrits les plus fonctionnelles s'adressent à des utilisateurs de différents niveaux ayant des compétences variées et des questions de recherche diverses. Elles permettront à la fois des explorations fortuites et des études plus ciblées. Avec d'autres projets de numérisation de manuscrits en cours, de nouvelles technologies telles que les images 3D de pages de manuscrits (Manuscrits de la cathédrale de Lichfield 2015), et des projets qui s'attaquent à la reconnaissance optique de caractères (OCR) pour l'écriture manuscrite, la prochaine génération de manuscrits numérisés promet d'étendre et de révolutionner une fois de plus l'étude des documents manuscrits historiques [10] »

À compter de cette période, une certaine effervescence dans le domaine de la reconnaissance d'écriture devient palpable. L'avènement des processeurs graphiques grand

9. MUEHLBERGER (Guenter), SEAWARD (Louise), TERRAS (Melissa) et al., « Transforming scholarship in the archives through handwritten text recognition : Transkribus as a case study », *Journal of Documentation*, vol. 75, n 5 (septembre 2019), p. 954-976.

10. ESTILL (Laura) et LEVY (Michelle), « Chapter 12Evaluating digital remediations of women's manuscripts », *Digital Studies / Le champ numérique*, vol. 6, n 6 (juillet 2016). URL : `https://www.digitalstudies.org/article/id/7296/`. Consulté le 17 août 2024.

public permettant un meilleur entraînement des modèles et l'accès de plus en plus démocratisé à des bibliothèques de programmation (*libraries*) telle que *PytTorch* mènent les techniques de reconnaissance d'écriture à connaître de profondes transformations. Ainsi, plusieurs projets de plateformes et de logiciels conçus pour permettre aux chercheurs ou au grand public de recourir aux techniques de reconnaissances d'écriture se développent.

1.2.1 Une initiative à l'échelle européenne : le projet *Transkribus*.

Pour comprendre les origines de , il est nécessaire de comprendre le contexte institutionnel qui a permis sa naissance. En 2016, L'*European Association for Digital Humanities*[11] lance le projet *The Recognition and Enrichment of Archival Documents (READ)*. Il s'inscrit dans la continuité des recherches et avancées faites dans le cadre du projet *tranScriptorium*[12]. Ayant pour visée générale de « rendre les documents d'archives les plus accessibles possible grâce à l'utilisation de technologies de pointe »[13], son objectif principal était de fournir une plateforme permettant la reconnaissance automatique d'écriture manuscrite de documents afin de « révolutionner l'accès aux archives écrites »[14]. Cette initiative se traduira non seulement par la création de mais aussi par un nombre important d'articles académiques portant sur les aspects computationnels du HTR ainsi que des jeux de données (*data sets*).

C'est surtout par sa philosophie que la plateforme se démarquait. Les membres du projet READ considéraient que devait reposer sur une « approche coopérative » (*cooperative approach*)[15] en la rendant gratuite et accessible autant aux institutions patrimoniales (bibliothèques et archives) et de recherche qu'au grand public. Cette approche était également pensée pour améliorer les performances de la plateforme grâce aux principes de l'apprentissage machine. Entraînée par une communauté aux demandes nombreuses et variée, serait en mesure de traiter une typologie étendue de sources manuscrites.

En ce sens, elle est à différencier des solutions commerciales existantes sur le marché

11. Fondée en 1973 sous le nom d'*Association for Literacy and Linguistic Computing*, cette association avait d'abord pour but de soutenir l'utilisation de l'informatique dans les études de langue et littérature, se focalisant en particulier sur le traitement informatique de corpus linguistiques et de textes littéraires. Avec le développement sans cesse croissant de nouvelles technologies numériques, l'éventail des prérogatives de l'association s'est étendu à l'histoire, l'histoire de l'art, la musique, l'étude des manuscrits, le traitement de l'image ou encore aux éditions électroniques. Son nom actuel, adopté en 2012, reflète cette évolution. Voir la site officiel de l'association : `https://eadh.org/`.

12. Lancé en 2013 et achevé en 2015, il était issu d'un partenariat entre treize universités et institutions européennes patrimoniales et de recherche. Composé d'une large équipe d'informaticiens, de développeurs et de chercheurs en sciences humaines, il étudiait et cherchait à développer les techniques de HTR et OCR. Voir la documentation officielle du projet : `https://eadh.org/projects/transcriptorium`.

13. Voir la documentation officielle du projet : `https://eadh.org/projects/transcriptorium`.

14. *Ibidem.*

15. Muehlberger et al., 2018

de l'époque qui ne permettaient pas à l'usager de participer au développement et au processus d'entraînement de la machine. On citera à titre d'exemple la société britannique *Adam Matthew Digital* ou encore l'éditeur français *A2iA* qui participa au projet *HIstorical MANuscript Indexing for user-controlled Search* (HIMANIS), sujet de la prochaine section.

1.2.2 HIMANIS et Horae : des projets de recherche français pionniers.

Durant la période, en parallèle du développement de , l'Institut de recherche et d'histoire des textes (IRHT), et plus particulièrement le responsable de sa section de paléographie latine, Dominique Stutzmann, ont été à l'origine de deux projets précurseurs dans l'utilisation de la reconnaissance automatique de l'écriture manuscrite sur des sources historiques médiévales : HIMANIS et HORAE.

1.2.2.1 *HIstorical MANuscript Indexing for user-controlled Search* (HIMANIS)

Le projet européen HIMANIS a été lancé en 2015 grâce au financement du programme européen *JPI Cultural Heritage and Global Change*. Il est issu d'un partenariat entre trois institutions académiques européennes : l'IRHT, l'Université de Groningen et l'Université polytechnique de Valence et une société française, *A2iA*, à l'époque l'un des leaders mondiaux en matière d'intelligence artificielle et d'analyse d'images. Ce partenariat a permis une collaboration entre institutions patrimoniales, chercheurs en sciences humaines et chercheurs en informatique et intelligence artificielle. [16]

Les membres d'HIMANIS sont partis d'un constat : malgré les efforts de numérisation des institutions patrimoniales européennes – phénomène que nous avons décrit précédemment – les informations contenues dans les images numérisées demeuraient encore trop inaccessibles. En effet, les technologies numériques ne parvenaient pas encore à répondre à la demande des chercheurs et du public, qui souhaitaient interroger avec plus de précision les sources historiques consultées en ligne. Avec comme corpus d'essai la collection de registres produits par la chancellerie royale française entre les XIV^{e} et XV^{e} siècles, numérisée par les Archives nationales, la collaboration avait pour ambition de développer des technologies de recherche et d'indexation sur les sources historiques, jusqu'à transformer, à terme, le paradigme de la recherche historique sur les archives. [17]

Les résultats du projet se sont avérés extrêmement probants. Les membres d'HIMANIS sont parvenus non seulement à convertir et structurer automatiquement douze

16. Voir la documentation officielle du projet : `https://www.irht.cnrs.fr/fr/recherche/les-programmes-de-recherche/himanis`

17. Consortium Oriflams, dans « Himanis », `https://himanis.hypotheses.org/1`

inventaires de la chancellerie royale, mais également à indexer le texte de ces registres médiévaux. Pour ce faire, il a évidemment fallu recourir aux technologies de reconnaissance automatique de l'écriture manuscrite.

La première transcription du projet a été produite en février 2017 et portait sur les douze inventaires de la chancellerie royale. Demandée à A2iA, le processus aurait duré quatre semaines. Plus de 1 400 pages ont pu être transcrites automatiquement, avec une vérité-terrain correspondant uniquement aux quatre premières pages de chaque inventaire, issues d'éditions anciennes. Deux éléments expliquent la réussite de cette transcription automatique : les images numérisées par les Archives nationales étaient d'excellente résolution, et le corpus de 1 400 pages était rédigé par une seule et même main. À terme, l'intégralité des registres médiévaux de la chancellerie royale française a pu être automatiquement transcrite. [18].

1.2.2.2 *Hours - Recognition, Analysis, Editions* (HORAE)

Le projet HORAE s'inscrit dans la continuité du projet HIMANIS. Bien qu'il se concentre sur un autre type de source médiévale – les livres d'heures numérisés par la médiathèque de Poitiers – et qu'il soit financé par l'Agence nationale de la recherche, ce qui en fait avant tout un projet français, on y retrouve à nouveau l'IRHT et Dominique Stutzmann [19]. Cette continuité se reflète également dans le choix des partenaires techniques du projet : en plus du Laboratoire des sciences du numérique de Nantes, l'entreprise française *Teklia*, spécialisée dans le traitement des documents par l'intelligence artificielle, a été retenue. Il est à noter que Christopher Kermorvant, fondateur et dirigeant de *Teklia*, avait joué un rôle central dans le projet HIMANIS en tant que directeur de l'équipe de recherche d'*A2iA*.

Parmi les objectifs du projet HORAE figuraient le développement de nouveaux logiciels en accès libre de reconnaissance d'écriture manuscrite ainsi que des outils de segmentation et de détection de plagiat adaptés aux manuscrits médiévaux [20]. *Teklia* a ainsi pu concevoir un système de reconnaissance automatique spécifiquement adapté aux livres d'heures, jetant les bases de ce qui deviendra la plateforme de traitement de documents [21]. En septembre 2019, le projet parvient à produire une transcription automatique de

18. Nous n'avons pas pu obtenir d'informations détaillées sur les processus de transcription qui ont suivi celle de février 2017.

19. Dans le carnet *Hypothèses* d'HIMANIS, Dominique Stutzmann fait d'ailleurs un clin d'œil aux liens entre HIMANIS et HORAE dans un billet intitulé « Disant ses heures : from Himanis to Horae ». Il y souligne l'importance des livres d'heures dans les pratiques de lecture médiévales en s'appuyant sur un acte transcrit et indexé par le projet HIMANIS. Voir : `https://himanis.hypotheses.org/189`

20. Voir la documentation officiel du projet : `https://www.irht.cnrs.fr/fr/recherche/les-programmes-de-recherche/horae`

21. Nous parlerons de cette plateforme plus en détails dans les prochaines parties.

400 manuscrits de livres d'heures.

Les avancées réalisées par l'IRHT ont ainsi marqué une voie importante dans le développement des techniques de reconnaissance automatique d'écriture manuscrite, tant en France qu'à l'échelle européenne. Contrairement à l'approche coopérative adoptée par , l'IRHT a choisi de s'appuyer sur des partenaires issus du secteur privé pour la mise en œuvre technique de ses projets. Cependant, privé ne signifie pas forcément fermé. En effet, en partenariat avec le consortium READ – à l'origine de – un jeu de données issu du projet HIMANIS a été rendu disponible en libre accès [22]. Quant au projet HORAE, il a permis, comme l'avons dit, le développement de la plateforme en libre accès *Arkindex*.

1.2.3 SCRIPTA-PSL et le développement de *Kraken* et d'eScriptorium.

SCRIPTA-PSL est une initiative de recherche interdisciplinaire et stratégique (IRIS) lancée en 2017 par l'École Pratique des hautes études et l'École française d'Extrême-Orient, respectivement membre et partenaire de l'Université Paris Sciences et Lettres (PSL). Ce programme de recherche a pour objectif de réunir les sciences de l'écrit (principalement la paléographie et la codicologie), les sciences humaines et les sciences informatiques afin d'étudier l'écrit ancien, principalement non latin, en s'appuyant sur l'évolution des technologies numériques appliquées à l'écriture [23]. .

Dans le cadre de ce projet, l'aspect numérique est coordonné par l'équipe eScripta [24]. Entre 2017 et 2019, cette équipe a développé deux outils dans le domaine de la reconnaissance automatique de l'écriture : *Kraken* et eScriptorium. Le premier, développé par Benjamin Kiessling, membre d'eScripta, est le fruit de ses recherches universitaires débutées en 2014 sur la reconnaissance optique de caractères (OCR), principalement appliquée aux langues anciennes et non latines (arabe, persan, syriaque, grec ancien, ou encore hébreu). Cette expertise a contribué au développement de *Kraken*, rendu public en 2017. Il s'agit d'un outil d'analyse de mise en page et de reconnaissance d'écriture manuscrite fondé sur l'apprentissage profond (*deep learning*), mais accessible uniquement en ligne de commande.

Par la suite, l'équipe eScripta, et principalement Benjamin Kiessling, ont travaillé sur une adaptation de *Kraken* et ont développé eScriptorium. Conçu comme un logiciel

22. Il s'agit du jeu de données « HIMANIS Guérin » qui fournit une vérité terrain de plus de 30 000 lignes (1200 images transcrites environ) pour l'entraînement HTR. Voir `:https://zenodo.org/records/553530`.

23. Voir la documentation officielle du projet : `https://psl.eu/actualites/scripta-histoire-et-pratique-de-lecrit`

24. Voir le carnet Hypotheses de l'équipe : `https://escripta.hypotheses.org/`

libre avec une interface utilisateur web (contrairement à *Kraken*), le logiciel se voulant plus accessible permet de segmenter un document, de détecter les lignes, de les transcrire, et d'entraîner un modèle pour l'appliquer à ses propres sources en se branchant directement sur le moteur de *Kraken*.

Ainsi, dans la seconde moitié des années 2010, trois grands axes se distinguent dans le développement des technologies de reconnaissance d'écriture manuscrite dans les humanités numériques en France et en Europe. Sous l'égide de l'European Association for Digital Humanities et du projet READ, la plateforme Transkribus, en libre accès, a été développée selon une approche coopérative. Parallèlement, l'Institut de recherche et d'histoire des textes (IRHT) a fait appel à des solutions privées pour ses deux grands projets de recherche, HIMANIS et HORAE, en collaborant d'abord avec *A2iA* puis avec *Teklia*. Enfin, l'équipe eScripta a développé *Kraken* et son adaptation eScriptorium.

Il est important de souligner que ces trois voies ne sont pas hermétiques et que de nombreuses collaborations ont émergé entre ces projets. Rappelons, par exemple, qu'un partenariat entre le consortium READ et HIMANIS a permis la mise en libre accès d'un jeu de données issu des transcriptions du projet. De plus, Teklia a également collaboré avec eScripta pour l'amélioration d'eScriptorium. Ce dynamisme collaboratif autour de la communauté HTR (Handwritten Text Recognition) perceptible vers la fin de la décennie devient un phénomène de plus en plus prégnant jusqu'à nos jours, ce qui sera exploré dans la section suivante.

1.3 2019 à nos jours : les derniers développements marqués sous le signe de la collaboration.

1.3.1 L'équipe-projet ALMAnaCH.

En 2017, une équipe-projet nommée *Automatic Language Modelling and Analysis and Computational Humanities* (ALMAnaCH) est fondée au sein de l'Institut national de recherche en sciences et technologies du numérique (INRIA). Cette équipe pluridisciplinaire en intelligence artificielle est dédiée aux domaines du traitement automatique des langues et des humanités numériques. Si elle se concentre principalement sur les modèles de langue neuronaux, la traduction automatique, la modélisation du dialogue et l'intelligence artificielle interactive, elle entame en 2018 ses premiers travaux dans le domaine de la reconnaissance d'écriture manuscrite dans le cadre du projet *LECTure Automatique de REPertoires* (LectAuRep), en collaboration avec le Minutier des notaires de Paris des Archives nationales.

1.3.1.1 Le projet LectAuRep et le développement d'eScriptorium (2018 - 2022).

En partenariat avec le Ministère de la Culture, le projet visait à repenser l'utilisation des registres des actes notariés des Archives nationales — la source historique la plus consultée par les chercheurs et le public des archives notariales — en explorant le traitement de ces sources par la reconnaissance automatique d'écriture. À terme, le projet prévoyait de réaliser une lecture approfondie et sémantique des actes afin de mener des études quantitatives. Contrairement aux projets HIMANIS et HORAE, qui utilisaient pour l'entraînement de l'intelligence artificielle une vérité de terrain produite par des spécialistes des sources médiévales, LectAuRep s'est principalement appuyé sur des transcriptions collaboratives.

Il est important de rappeler qu'au début du projet, les solutions de reconnaissance d'écriture manuscrite étaient encore assez limitées. *Kraken* et Transkribus venaient tout juste d'être développés, et le recours à des entreprises privées comme *Teklia* n'a pas été envisagé. De manière générale, les membres du projet considéraient le traitement numérique des sources comme étant de nature « expérimentale et applicative » [25].

Durant la première phase du projet, de 2018 à 2019, Transkribus avait été privilégié pour la segmentation des pages et des colonnes des tableaux issus des registres [26]. Cependant, en 2019, la plateforme a changé de modèle économique et est devenue payante. Ce changement s'inscrit dans la stratégie du consortium *READ*, qui a choisi de créer une société coopérative, *READ-COOP*, afin de maintenir et développer les services de la plateforme. En réponse à cette évolution, les membres du projet LectAuRep ont décidé de se tourner vers la plateforme du projet SCRIPA, eScriptorium, la solution *open source* et gratuite la plus développée sur le marché.

Au cours des deuxième et troisième phases, entre 2019 et 2022, l'évolution d'eScriptorium est étroitement liée aux progrès du projet LectAuRep. Des réunions hebdomadaires se tiennent entre les membres du projet, notamment Alix Chagué ingénieure en chef, et l'équipe eScripta, afin de travailler sur le développement du logiciel. Ainsi, bien qu'eScriptorium soit à l'origine une création du projet SCRIPTA-PSL, l'équipe-projet ALMAnaCH de l'INRIA a largement contribué à son amélioration dans plusieurs domaines à compter de cette date.

25. La plupart des informations concernant le projet LectAuRep ont été obtenues lors d'un entretien avec Hugo Scheithauer réalisé le 5 juin 2024. Aujourd'hui ingénieur en recherche et développement à l'INRIA, il était un membre du projet.

26. La segmentation est une étape fondamentale pour toute reconnaissance automatique de caractère optique ou d'écriture manuscrite. Cette notion sera expliquée en détaille dans la prochaine partie.

Tout d'abord, ALMAnaCH a successivement mis en place plusieurs serveurs dimensionnés pour le déploiement d'eScriptorium, permettant aux utilisateurs de profiter des traitements offerts par la plateforme de manière bien plus rapide que sur ses propres infrastructures. Entre 2018 et 2020, un serveur nommé « LectAuRep », mis en place sous forme de machine virtuelle dans le cadre de l'offre de services informatiques de l'INRIA, a été dédié au projet. Cependant, ce serveur ne disposait pas de processeur graphique (GPU).

En 2020, pour l'équipe ALMAnaCH a acquis pour le projet LectAuRep un nouveau serveur nommé « Traces6 », financé par le Dispositif de soutien à l'archivistique et aux humanités numériques (DAHN) [27]. Équipé quant à lui d'un processeur graphique, ce serveur a permis un entraînement en apprentissage automatique beaucoup plus rapide qu'avec LectAuRep, passant de plusieurs heures à quelques minutes. En plus de cette augmentation de la puissance de calcul, Traces6 a pu accueillir de nouveaux utilisateurs d'eScriptorium, permettant à une dizaine de projets de recherche de tester la plateforme [28].

Enfin, en 2021, l'architecture de Traces6 a évolué grâce au financement du projet d'équipement CREMMA - qui sera l'objet de la prochaine section - par la region Île de France. Ce projet a pour but de créer un environnement équilibré et robuste pour le déploiement de la plateforme eScriptorium et sa mise à disposition de la communauté des humanités numériques. Dans ce cadre, Traces6 fait partie d'un des serveurs clés sur lesquels repose l'architecture de l'environnement.

Au-delà de la mise à disposition de ses serveurs, l'équipe ALMAnaCH a également joué un rôle majeur dans l'amélioration du code source d'eScriptorium en mettant à contribution des cas d'usage. Tout d'abord, des réunions hebdomadaires entre les développeurs d'eScriptorium et les membres du projet LectAuRep ont permis d'identifier les axes d'amélioration de la plateforme. Ensuite, avec le déploiement du serveur Traces6, le serveur LectAuRep n'a pas été abandonné ; il a été réaffecté aux tests et à l'amélioration du code source. Par ailleurs, l'élargissement de l'accès à eScriptorium à d'autres projets que nous avons mentionné a permis de recueillir des retours d'utilisation, contribuant également à l'amélioration de ce code. Enfin, l'équipe ALMAnaCH est aussi à l'origine d'une documentation ouverte et collaborative pour aider l'utilisateur à se servir de la plateforme [29].

27. Il s'agit d'un dispositif mis en place par le Ministère de l'Enseignement supérieur pour mettre en place une chaîne d'édition scientifique numérique en TEI. Ce dispositif est en partenariat avec l'École des hautes études en sciences sociales (EHESS) ET l'Université du Mans.

28. Nous citerons à titre d'exemple et sans être exhaustif des projets tels que e-NDP, Artl@s, *Paris Time Machine* - Annuaire de Paris ou encore *TIME US*.

29. Consultable à l'adresse suivante : https ://escriptorium.readthedocs.io.

1.3.1.2 Développement d'outils autour du HTR (2020 - 2024).

En plus du développement d'eScriptorium, l'équipe ALMAnaCH a conçu plusieurs outils liés à la reconnaissance d'écriture manuscrite. Le premier axe concerne les données d'entraînement en accès libre. Dans le cadre du projet CREMMA, l'équipe a créé en 2022 *WikiCremma*, un corpus de données pour l'entraînement des machines sur le français contemporain extrait de sources provenant de *Wikipedia*. Un autre corpus, *HTRomance*, propose des vérités de terrain pour le français du XII^{e} au XIX^{e} siècle. Dans le cadre de l'initiative CATMuS (*Consistent Approaches to Transcribing ManuScripts*), un modèle HTR et un jeu de données ont été élaborés et publié en 2024. Ce modèle, fruit d'une collaboration entre les projets CREMMA, GalliCorpora, HTRomance et DEEDS, est entraîné sur l'ancien et le moyen français, le latin, l'espagnol et l'italien. Il suit des directives précises que nous détaillerons dans la section suivante. Le jeu de données couvre plus de 200 manuscrits et incunables en dix langues différentes, allant du VIII^{e} au XVI^{e} siècle.

Mais l'équipe est surtout à l'origine du catalogue collaboratif *HTR United*, un écosystème conçu pour fournir des corpus de référence afin de former rapidement des modèles sur des collections plus petites. Ce catalogue regroupe les hébergements et les descriptions de jeux de données d'entraînement d'une multitude de projets pour le HTR et l'OCR, couvrant des périodes de l'Antiquité à nos jours, principalement en français, mais aussi en arabe, grec et hébreu. Enfin, l'équipe a développé KaMI-Lib, une bibliothèque Python permettant d'évaluer un modèle de transcription HTR et OCR, qui sera également expliquée plus précisément dans la partie suivante.

1.3.2 Le projet CREMMA et CREMMALAB : réflexions théoriques et infrastructure.

En 2017, la région Île-de-France annonce la création d'un *Domaine de recherche et d'innovation majeur axé sur les matériaux anciens et patrimoniaux* (DIM-MAP). Ce réseau de recherche s'intéresse aux matériaux historiques en favorisant une collaboration interdisciplinaire entre plusieurs champs : sciences humaines, sciences physico-chimiques, tout en intégrant la participation des acteurs économiques et sociaux du secteur patrimonial. Dans le cadre de cette initiative, divers projets peuvent solliciter un financement. Le *projet Consortium reconnaissance d'écriture manuscrite des matériaux anciens (CREMMA* fait partie des initiatives sélectionnées.

Ce projet se présente comme un consortium qui vise à créer un service de mise à disposition de ressources serveurs pour faciliter l'accès aux outils de reconnaissance d'écriture manuscrite via la plateforme eScriptorium. Il regroupe l'équipe ALMAnaCH

de l'INRIA, le centre Jean Mabillon, l'École nationale des chartes, l'Institut de recherche et d'histoire des textes (IRHT), le Laboratoire de médiévistique occidentale de Paris (LAMOP), la section des sciences historiques de l'École pratique des hautes études, et l'École française d'Extrême-Orient. Le but est de proposer ce service en premier lieu aux laboratoires et projets partenaires, avec une ouverture future aux initiatives externes après approbation du consortium.

Par ailleurs, les laboratoires partenaires prévoient de rendre disponibles des jeux de données issus de divers matériaux, tels que des manuscrits médiévaux et modernes ou encore des supports muraux. Ces ensembles de données visent à fournir des modèles aux nouveaux projets. Le projet prévoit aussi une documentation pour accueillir et former les utilisateurs de la plateforme eScriptorium.

En 2021, un nouveau volet du projet voit le jour : CREMMALAB. Soutenu également par la région Île-de-France dans le cadre du DIM-MAP, ce volet accompagne l'ouverture du service de transcription en développant une documentation sous forme de manuels mais également des ateliers, des formations, et de retours d'expérience, à partir de l'analyse d'un corpus médiéval. À cet effet, Ariane Pinche, post-doctorante, est recrutée pour diriger cette initiative.

Comme présenté dans les sections précédentes, plusieurs de ces objectifs ont été atteints. En 2021, en s'appuyant sur l'infrastructure du serveur du projet LectAuRep, Traces6, une nouvelle architecture plus performante est déployée par l'INRIA pour offrir eScriptorium dans des conditions optimales. En 2022, les jeux de données *WikiCremma* et *HTRRomance* sont créés. En décembre de la même année, le consortium CREMMA propose finalement aux projets sélectionnés un accès gratuit à eScriptorium, déployé sur ses serveurs.

Outre les aspects techniques, CREMMALAB a organisé des discussions scientifiques collaboratives autour de la reconnaissance de texte manuscrit. Cinq séminaires ont été tenus, abordant des sujets tels que les modèles HTR, la segmentation et la transcription de signes médiévaux spécifiques. Ces échanges ont abouti à un colloque intitulé « Documents anciens et reconnaissance automatique des écritures manuscrites », qui a eu lieu les 23 et 24 juin 2022 à l'École nationale des chartes. Il s'agissait d'une succession de présentations de projets HTR. Les enjeux, les contraintes techniques et matérielles autour de chaque projet ont fait l'objet de discussions. Cet événement a été moteur pour appréhender le futur du HTR puisqu'il a permis de dégager des perspectives d'amélioration de la technique appliquée aux documents historiques.

Chapitre 2

Principes généraux de la mise en oeuvre d'un projet HTR.

Pour mener à bien un projet de reconnaissance d'écriture, il ne suffit pas de comprendre l'environnement institutionnel dans lequel il s'est développé ; il est également essentiel d'en appréhender les différentes étapes ainsi que les contraintes techniques potentielles qui peuvent survenir. Il est tout aussi important de connaître en détail le fonctionnement, les avantages et les inconvénients des solutions existantes pour la mise en œuvre de ce type de projet. Ce chapitre présentera donc d'abord les phases fondamentales de tout traitement HTR, en soulignant les obstacles susceptibles d'en perturber le déroulement. Il se concentrera ensuite sur les solutions les plus appropriées, en les comparant afin de mettre en lumière leurs forces et leurs faiblesses.

2.1 Du fichier image à un texte structuré.

Dans cette section, nous présentons ce que nous considérons comme les étapes fondamentales et complémentaires de tout traitement HTR. Ces connaissances ne proviennent pas de recommandations définies par une quelconque institution, mais résultent de notre observation empirique et de notre analyse de la manière dont les projets de reconnaissance d'écriture ont été réalisés.

2.1.1 L'importation des images.

Bien que cette étape puisse sembler assez rudimentaire à première vue, il est essentiel de prêter attention aux formats des images que l'on souhaite importer. En effet, une mauvaise qualité d'image pourrait considérablement affecter le processus de reconnaissance d'écriture manuscrite. Les formats d'images que l'on peut envisager d'importer dépendent de la politique des outils que l'on utilise pour le traitement HTR. Nous verrons au sein du troisième chapitre comment la méconnaissance des restrictions d'une plate-

forme peut grandement entraver un processus de reconnaissance d'écriture. Néanmoins, en règle générale, les formats classique de fichiers images sont acceptés (`JPEG`, `PNG`, `BMP`, `TIFF`).

Notons que de plus en plus de plateformes et d'outils de reconnaissance d'écriture manuscrite permettent l'importation d'images via leur manifeste IIIF [1]. Ce moyen est de plus en plus privilégié pour plusieurs raisons. Tout d'abord, il garantit une qualité d'image optimale. Ensuite, dans le cas où l'on souhaite constituer un corpus d'images provenant de différentes institutions, il permet d'accéder directement aux sources sans avoir à les télécharger ni à les réorganiser.

2.1.2 La segmentation.

Dans un processus de reconnaissance d'écriture manuscrite, la méthode utilisée consiste à reconnaître le texte une ligne à la fois contrairement à la reconnaissance optique de caractère qui traite les caractères optiques sur la page entière. Ainsi, avant de procéder à une transcription, chaque zone de texte sur l'image doit être identifiée et au sein de ces zones, chaque ligne de base (baslines), c'est-à-dire, la ligne qui traverse le bas des caractères alignés. En résumé, cette étape permet de repérer les zones où le texte est contenu, d'identifier et numéroter les lignes qui composent ces zones dans l'ordre de lecture. Sans cette segmentation, la transcription produite ne suivra aucune cohérence puisqu'on ne connaîtra ni le sens de lecture, ni l'ordre des caractères.

Au-delà de préparer la transcription automatique, le processus de segmentation permet également à terme d'analyser la mise en page de la source que l'on traite. Plus précisément, il est possible dans cette étape de préparer l'annotation et la description des différentes parties qui composent une page. À ce titre, la communauté scientifique a pu proposer des initiatives de vocabulaire harmonisé et contrôlé. Citons à titre d'illustration le projet SegmOnto : A Controlled Vocaabulary to Describe the Layout of Pages [2]. Les recommandations du projet (*guidelines*) dispose d'une ontologie générale permettant de décrire la mise en page d'une variété de sources dans leur aspect matériel.

La segmentation est l'une des étapes les plus difficiles du traitement de la reconnaissance d'écriture manuscrite. Il est possible de segmenter manuellement une page,

1. Le protocole IIIF est un ensemble de spécifications qui a pour but de standardiser l'accès et l'interopérabilité des images numériques à haute résolution sur le web, provenant généralement des collections numérisées en ligne de grandes institutions patrimoniales. Ces institutions rendent ainsi via ce protocole leur collection d'image accessible et interopérable. Voir le site officiel : `https://iiif.io/`.

2. Voir : Simon Gabay, Ariane Pinche, Kelly Christensen, Jean-Baptiste Camps, Nicola Carboni, SegmOnto, *A Controlled Vocabulary to Describe the Layout of Pages*, version 0.9, Genève/Lyon/Paris, 2023, `https://segmonto.github.io/`.

mais cela représente une tâche extrêmement longue et fastidieuse, voire impossible dans le cadre de la plupart des projets HTR de grande envergure. Cette étape est néanmoins automatisable : il est possible de recourir à des modèles de segmentation automatiques et de les entraîner. Cependant, la plupart de ces modèles s'avèrent peu performants et leur entraînement est souvent assez long.

Il existe au sein de la communauté scientifique une prise de conscience de la difficulté du processus de segmentation. Des recherches récentes ont, par exemple, proposé de renouveler les méthodes utilisées. Thibault Clérice suggère ainsi de passer de la polygonisation fondée sur la classification des pixels [3] à la détection d'objets [4]. Dans le cadre de ses travaux [5] YALTAi, un modèle de segmentation fondés sur la détection d'objet a pu développé. Il s'agit d'une adaptation pour la reconnaissance d'écriture manuscrite de modèles de détection d'objets préexistant : les modèles YOLO.

L'un des derniers défis du processus de segmentation qui mérite d'être relevé est la reconnaissance de formes complexes telles que les tableaux. À cet égard, certaines plateformes telles que Arkindex et Transkribus se targuent de proposer des modèles de segmentation automatiques performants pour des sources incluant ce type de contenu. Divers projets ont également mis en lumière l'enjeu que représente la segmentation des tableaux. On peut citer, à titre d'exemple, le Projet d'océrisation des recensements parisiens (POPP) [6], qui portait sur des documents de recensement de population comportant des tableaux, ou encore le projet LectAuRep, que nous avons abordé en détail dans le premier chapitre.

2.1.3 La transcription.

La transcription consiste à identifier chaque caractère d'un texte manuscrit à l'aide d'un modèle, après que le texte a été segmenté en zones et en lignes. Ce modèle doit être entraîné avec ce que l'on appelle une « vérité-terrain », c'est-à-dire une transcription correcte et validée du texte manuscrit, qui sert de référence pour améliorer la précision

3. Méthode de traitement d'image où chaque pixel est d'abord classifié en fonction de ses caractéristiques (par exemple, couleur ou intensité) pour déterminer à quel zone il appartient. Les pixels classifiés sont ensuite regroupés pour former des polygones qui délimitent précisément les différentes zones de l'image. Cette méthode permet ainsi de regrouper les écriture dans une même zone.

4. Technique qui permet à une machine d'identifier et de localiser automatiquement des objets spécifiques dans une image. Cette méthode utilise des modèles d'intelligence artificielle pour repérer et encadrer les objets d'intérêt présents dans le contenu visuel. Dans le cadre d'une adaptation pour le reconnaissance d'écriture manuscrite, l'objet reconnu est la zone de l'image contenant l'écriture.

5. Clérice Thibault. *You Actually Look Twice At it (YALTAi) : using an object detection approach instead of region segmentation within the Kraken engine*, 2023.

6. Constum (Thomas), « Reconnaissance et extraction d'informations dans des tableaux manuscrits historiques : vers une compréhension des recensements de Paris de l'entre-deux guerre », dans *Documents Anciens et HTR*, Paris, École nationale des chartes, 2022.

de la transcription automatique. Cette transcription peut être effectuée manuellement par une personne capable de lire et de retranscrire le texte sans erreur, ou elle peut déjà exister, par exemple sous forme d'édition en ligne.

2.1.3.1 Recourir à des modèles.

Il est également possible de recourir à des modèles déjà entraînés sur des sources ayant une écriture similaire à celle que l'on souhaite traiter. En utilisant la vérité-terrain, ces modèles peuvent être ajustés grâce à de nouveaux entraînements, devenant ainsi plus efficaces sur ces sources. Par exemple, dans le cadre de notre stage, nous avons appris que le projet e-NDP s'est appuyé principalement sur les jeux de données du projet HIMANIS, abordé dans le premier chapitre, pour mener ses processus de reconnaissance automatique des écritures des registres du cloître de Notre-Dame [7].

Un enjeu courant se pose lorsque la source à traiter présente des écritures de plusieurs auteurs. Il s'agit alors de décider s'il faut entraîner un modèle « mixte », c'est-à-dire un modèle unique pour toutes les pages, ou un modèle distinct pour chaque type d'écriture. Cette décision dépend largement des caractéristiques des sources et des modèles préalablement disponibles. Par exemple, le projet LectAuRep, mentionné au premier chapitre, a obtenu d'excellents résultats en utilisant un modèle mixte.

Il est important de noter que les modèles de transcription sont généralement accessibles, la communauté scientifique adoptant majoritairement une approche de science ouverte, comme nous l'avons montré dans la première partie. Ainsi, des modèles éprouvés dans le cadre de certains projets peuvent être réutilisés pour des sources similaires.

Bien que certains projets HTR ne partagent pas toujours leurs modèles, la grande majorité met à disposition la vérité de terrain qui a permis de les développer. Le catalogue HTR-United, que nous avons précédemment cité, offre des vérités de terrain adaptées à différents types de sources, qu'elles soient de périodes ou de langues diverses.

Notons qu'à mesure que de nouveaux projets de reconnaissance d'écriture manuscrite voient le jour, il devient de plus en plus probable de trouver des modèles préalablement entraînés sur des écritures similaires à celles sur lesquelles on souhaite travailler.

2.1.3.2 Évaluer son modèle.

Enfin, il est important de savoir que l'évaluation des performances d'un modèle est une étape importante, tant lors de son entraînement qu'après avoir réussi à obtenir

7. Informations recueillies lors d'un entretien réalisé avec Sergio Torres Aguilar au sujet du projet e-NDP, le 20 juin 2024.

une transcription satisfaisante. Elle également essentielle lorsqu'on souhaite partager un modèle avec la communauté scientifique afin de renseigner les futurs utilisateurs sur son taux de précision. Mais comment évaluer ces performances ? L'évaluation se fait en comparant une chaîne de référence, issue d'une transcription correcte, avec une chaîne générée automatiquement. On identifie alors les écarts entre les deux chaînes en fonction de trois opérations distinctes :

— Suppressions : caractères absents dans la transcription générée.
— Substitutions : caractères remplacés par d'autres.
— Insertions : caractères ajoutés par rapport à la transcription correcte.

En utilisant la distance de Levenshtein, un procédé mathématique qui mesure l'écart entre deux chaînes de caractères, on peut estimer leur similarité syntaxique. Deux principales métriques de performance sont utilisées pour évaluer un modèle :

— Taux d'erreur par caractères (CER)
— Taux d'erreur par mots (WER)

Des outils comme KaMI, que nous avons mentionné dans le premier chapitre, disponibles sous forme d'application ou de bibliothèque Python, permettent de calculer ces métriques [8].

2.1.4 Post-correction.

L'étape de post-correction, qui consiste à corriger la transcription, plus ou moins longuement selon la qualité du texte produit par le modèle, a été envisagée sous un nouvel angle par des travaux récents. Alors qu'elle se faisait largement à la main, on envisage désormais d'utiliser de grands modèles de langage (LLM) pour améliorer les transcriptions générées [9]. Étant donné leur introduction récente, il manque encore de recul sur leur efficacité [10]. Cependant, certaines études montrent qu'ils peuvent être très performants

8. Le dépôt Github de la bibliothèque Python est disponible au lien suivant `https://github.com/KaMI-tools-project/KaMi-lib`.

9. Un LLM (Large Language Model) est un modèle d'intelligence artificielle conçu pour comprendre et générer du texte en langage naturel. Entraîné sur de vastes ensembles de données textuelles, il apprend les structures linguistiques, les règles grammaticales ou encore le vocabulaire. Appliqué à des transcriptions, il pourrait résoudre les suppressions, les insertions ou les substitutions de caractères.

10. Les LLM ont commencé à devenir accessibles au grand public à partir de la fin des années 2010 et du début des années 2020. Le plus poulaire d'entre eux, GPT avait été introduit par OpenAI en 2019.

pour corriger des transcriptions [11]. Ils pourraient donc représenter une option intéressante à considérer pour tout projet de reconnaissance d'écriture manuscrite.

2.2 Solutions de reconnaissance d'écriture manuscrite : analyse comparative.

Cette section vise à recenser et analyser les principales solutions de reconnaissance d'écriture manuscrite disponibles. Nous nous sommes concentrés sur celles disposant d'une interface graphique. Nous n'avons pas jugé pertinent d'appliquer la même démarche aux outils HTR en ligne de commande déployables sur des frameworks comme PyTorch ou TensorFlow, ou sur des environnements de programmation, en raison de leur abondance et de leur faible intuitivité pour un utilisateur non familier avec le développement informatique. Il est à noter que, parmi ces outils, Kraken, en dehors de son adaptation sur eScriptorium, est la solution la plus utilisée par la communauté des humanités numériques. Depuis l'intégration de la technologie Transformers, Kraken a considérablement amélioré ses performances, mais reste très coûteux en termes d'infrastructure. Par ailleurs, nous ne mentionnerons pas non plus les solutions de reconnaissance d'écriture manuscrite sur-mesure qui peuvent être demandées lors de prestations, que ce soit dans le cadre d'une numérisation ou non.

2.2.1 Plateformes *open source* avec interface : eScriptorium et Arkindex.

À l'heure actuelle, il existe deux grandes plateformes open source de traitement de documents utilisant la reconnaissance d'écriture manuscrite et disposant d'une interface : Arkindex, développée par Teklia, et eScriptorium, créé initialement dans le cadre du projet SCRIPTA-PSL.

Rappelons qu'une plateforme *open source* permet à ses utilisateurs d'inspecter et de modifier son code en fonction de leurs besoins. Le code source d'eScriptorium, issu du projet SCRIPTA, est disponible dans sa version la plus récente sur le Gitlab de l'INRIA [12]. Le code source d'Arkindex est également accessible sur le Gitlab de Teklia [13].

11. Boros Emanuela et al. « Post-Correction of Historical Text Transcripts with Large Language Models : An Exploratory Study. » In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage*, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), pages 133–159, St. Julians, Malta. Association for Computational Linguistics.

12. À l'adresse suivante : `https://gitlab.inria.fr/scripta/escriptorium`

13. À l'adresse suivante : `https://gitlab.teklia.com/arkindex`

Disposer d'une interface utilisateur graphique (GUI) présente certains avantages par rapport à l'utilisation en ligne de commande. D'abord, la navigation et la compréhension des fonctionnalités de la plateforme sont facilitées par des menus, des boutons et des icônes clairs et intuitifs. En outre, la manipulation des documents contenant de l'écriture manuscrite est plus directe : les sections spécifiques à transcrire peuvent être sélectionnées et agrandies.

Au-delà de ces caractéristiques, les deux plateformes permettent également l'auto-hébergement. En d'autres termes, il est possible de déployer Arkindex et eScriptorium sur ses propres serveurs. Cela peut être utile, tant pour l'utilisateur individuel que pour une institution, afin d'éviter les contraintes imposées par les services qui proposent d'héberger l'application. Par exemple, Teklia facture ce service pour Arkindex, tandis que le CREMMA, demande, pour eScriptorium, de partager sous licence ouverte les données d'entraînement et les modèles de transcription créés sur ses serveurs. Héberger ces plateformes sur ses propres infrastructures permet donc, dans le premier cas, de réaliser des économies, et dans le second, de protéger ses données personnelles.

Les deux plateformes offrent des fonctionnalités de segmentation et de transcription, manuelles ou automatiques. Pour les opérations manuelles, l'interface permet de sélectionner précisément les zones à segmenter ou à transcrire. Pour les opérations automatiques, il est possible d'importer des modèles de segmentation et de transcription, ainsi que d'exporter les modèles créés ou améliorés par l'utilisateur. Des modèles par défaut sont également proposés sur la version d'eScriptorium déployée sur les serveurs du CREMMA.

Enfin, l'équipe-projet ALMAnaCH et celle de Teklia ont rédigé une documentation pour aider les nouveaux utilisateurs à maîtriser ces plateformes.

Ainsi, Arkindex et eScriptorium sont toutes deux open source et possèdent des fonctionnalités similaires pour la reconnaissance d'écriture manuscrite, avec des performances équivalentes. Cependant, il est important de noter que, contrairement à eScriptorium, Arkindex a été conçu pour un traitement plus complet des documents, incluant la reconnaissance de formes, l'identification d'entités nommées, la classification de pages ou encore l'extraction de valeurs clés.

2.2.2 Plateformes payante avec interface : Transkribus.

Comme mentionné dans la première partie de ce travail, Transkribus, initialement gratuit, a adopté un modèle économique basé sur des forfaits fonctionnant avec des crédits. Quatre forfaits sont proposés : « *individual* », « *scholar* », « *team* » et « *organisation* ». Les

trois premiers varient de zéro à 59,90 € par mois, tandis que pour le forfait « *organisation* », un devis personnalisé peut être établi après avoir rempli un formulaire en ligne sur le site de Transkribus.

Ces forfaits donnent accès aux services de segmentation et de transcription automatiques, qui fonctionnent sur un système de crédits. Chaque mois, 100 crédits sont alloués, quel que soit le forfait, et chaque opération coûte un crédit. Si les crédits mensuels ne suffisent pas, des recharges payantes sont disponibles.

Bien que le nombre de tâches pouvant être effectuées soit identique pour chaque forfait, l'espace de stockage, ainsi que le nombre de sessions d'entraînement de modèles par mois, varient. Le forfait "organisation" offre un espace de stockage illimité, un nombre illimité de sessions d'entraînement, une assistance client prioritaire et la possibilité de créer des modèles sur mesure.

Dans le cadre des opérations de transcription et de segmentation automatiques, tous les forfaits permettent d'accéder aux modèles du catalogue de Transkribus, issus de projets coopératifs précédents. Parmi ces modèles, Transkribus met en avant ses « *Supermodels* », des modèles de reconnaissance d'écriture manuscrite entraînés sur une grande variété de documents provenant de différentes époques, langues et écritures. Par exemple, le supermodèle « Text Titan 1 » reconnaît des écritures manuscrites en six langues (allemand, néerlandais, français, finnois, suédois et anglais) couvrant une période allant du XVI^{e} au XXI^{e} siècle.

Comme eScriptorium et Arkindex, Transkribus dispose d'une interface qui permet la transcription et la segmentation manuelles directement sur l'image source, et propose une documentation pour aider l'utilisateur. La plateforme permet également de créer des modèles à partir de ses propres documents manuscrits et données d'entraînement.

Cependant, contrairement aux deux plateformes open source, Transkribus ne permet pas de peaufiner des modèles autres que ceux qui ont été réalisé avec la bibliothèque PyLaia ou issus de son propre catalogue. De plus, bien que les utilisateurs puissent rendre leurs modèles accessibles à tous sur la plateforme, il n'est pas possible de les exporter hors de cet écosystème. Enfin, le code source de Transkribus n'est pas ouvert, et la plateforme ne peut être déployée que de manière payante, exclusivement sur les serveurs de READ-COOP, l'entreprise propriétaire de Transkribus.

2.2.3 Tableau récapitulatif

	Arkindex	**eScriptorium**	**Transkribus**
Objectif principal	Traitement complet des documents	OCR / HTR	OCR / HTR
Open source	Oui	Oui	Non
Auto-hébergement	Possible	Possible	Impossible
Segmentation manuelle	Possible	Possible	Possible
Transcription manuelle	Possible	Possible	Possible
Segmentation automatique	Possible	Possible	Possible
Transcription automatique	Possible	Possible	Possible
Importation de modèles	Libre	Libre	Limitée
Exportation de modèles	Libre	Libre	Limitée
Documentation	Oui	Oui	Oui

Chapitre 3

La reconnaissance d'écriture manuscrite dans le contexte d'une grande institution muséale : mise en oeuvre sur les registres d'inventaire du DAGER et du SHL.

3.1 Les besoins du Musée du Louvre en matière de reconnaissance d'écriture manuscrite.

En complément de la mission qui nous a été confiée dans le cadre de ce stage, il était nécessaire d'élargir notre réflexion et de comprendre les besoins de l'ensemble des services de documentation du musée, en tenant compte de la diversité et des spécificités des départements auxquels ils sont rattachés. Chaque département ayant ses propres intérêts et une histoire particulière, les documentations varient considérablement et ne rencontrent pas les mêmes enjeux. Par exemple, il peut s'agir d'archives de conservateurs difficiles à déchiffrer ou de la nécessité de retrouver la documentation d'un objet dans un catalogue de vente d'objets d'art. Par ailleurs, la réflexion ne s'est pas seulement porté sur la documentation, mais également sur l'application de la reconnaissance d'écriture manuscrite sur les objets des collections eux-mêmes (manuscrits, fragments de textes, papyrii).

Ce recueil de besoins est fondé sur des entretiens menés avec les services de documentations de la quasi-totalité des départements du Musée, à l'exception du Département des sculptures.

3.1.1 Quelques expériences préliminaires de reconnaissance d'écriture manuscrite.

Alors que la majorité des services de documentation du musée n'ont aucune expérience avec la reconnaissance de l'écriture manuscrite et, pour la plupart, méconnaissent complètement cette technique, le Département des Arts de Byzance et des Chrétientés orientales (DABCO) ainsi que le Musée national Eugène-Delacroix font figure d'exception.

Au sein du DABCO, le HTR avait été utilisé dans le cadre d'une recherche de provenance portant sur des documents d'inventaires anciens issus de collections privées. Plus précisément, le département disposait d'une copie d'une page d'inventaire rédigée en allemand à la fin du XIX^{e} siècle. En raison de l'extrême difficulté à déchiffrer l'écriture, le recours à la reconnaissance d'écriture manuscrite s'était avéré précieux.

Le Musée national Eugène-Delacroix, quant à lui, se consacre à l'heure où nous écrivons ces lignes sur l'exploration de son histoire ainsi que de celle de ses bâtiments. Il a entrepris de transcrire les archives relatives aux travaux de construction et de rénovation de l'appartement et de l'atelier occupés autrefois par Delacroix, espaces qui abritent aujourd'hui le musée. Ce fonds d'archives, conservé à l'Institut national de l'histoire de l'art, a été numérisé et est désormais accessible via sa bibliothèque numérique. La décision de procéder à la transcription de ces archives s'inscrit dans le cadre d'une étude scientifique visant à mieux comprendre les aménagements intérieurs des lieux à l'époque de Delacroix.

Pour mener à bien ce projet, une vacation a été proposée entre la fin de l'année 2023 et le début de l'année 2024. Pauline Charrier, étudiante en deuxième année de master « Technologies numériques appliquées à l'histoire » à l'École nationale des chartes, a été employée pour réaliser cette tâche de transcription. L'utilisation d'une plateforme de reconnaissance d'écriture manuscrite avait été envisagée, et le choix s'était d'abord porté sur Transkribus. Cependant, en raison de résultats jugés insuffisants, le recours à cette plateforme a été abandonné au profit d'une transcription manuelle.

Précisons que, dans tous les départements, l'OCR est mis en œuvre régulièrement depuis 2019 sur la documentation envoyée en numérisation chez des prestataires. C'est notamment le cas du service de documentation du Département des Arts de l'Islam (DAI), qui prévoit actuellement d'océriser des catalogues de vente dans le cadre de recherches de provenance.

3.1.2 Aujourd'hui : des besoins, du matériau, et des idées.

Bien que le HTR reste largement méconnu de la plupart des services de documentation du musée, la présentation de cette technique et les échanges réalisés au cours du stage ont permis de faire émerger de nouvelles perspectives d'utilisation. Dans tous les services, l'exploitation de ces documents servent généralement deux grands prérogatives majeures : le récolement et la recherche de provenance.

3.1.2.1 Le récolement et la recherche de provenance : deux prérogatives essentielles des services de documentation.

Le récolement.

Le récolement, dit « décennal », est une opération réglementaire obligatoire pour tous les musées de France depuis 2002, encadrée par le Ministère de la Culture, comme stipulé dans l'article L 451-2 du Code du patrimoine et l'article 12 de la loi du 4 janvier 2002 relative aux musées de France. Cette loi repose sur le principe que les collections des musées français, étant inaliénables et imprescriptibles, doivent faire l'objet d'une vérification tous les dix ans pour s'assurer que leur propriétaire est en mesure de prouver leur appartenance et leur présence. Concrètement, le récolement consiste à « vérifier, sur pièce et sur place, à partir d'un bien ou de son numéro d'inventaire, la présence de ce bien dans les collections du musée, sa localisation, son état, son marquage, et la conformité de l'inscription à l'inventaire avec l'objet lui-même, ainsi que, le cas échéant, avec les différentes sources documentaires, archives, dossiers d'œuvre, catalogues » [1].

Dans des institutions muséales aussi riches que le Louvre, le récolement décennal représente un véritable défi pour les départements qui disposent de plusieurs centaines de milliers d'objets. C'est le cas de certains des départements avec lesquels nous nous sommes entretenus comme le DAGER.

La recherche de provenance.

La recherche de provenance est une démarche d'investigation visant à retracer l'historique de la possession d'une œuvre d'art depuis sa création jusqu'à son acquisition, ou son projet d'acquisition, par un musée. Cette démarche permet de mieux comprendre l'œuvre, son statut, son influence et son parcours historique.

1. Extrait de l'arrêté du 25 mai 2004 fixant les normes techniques relatives à la tenue de l'inventaire, du registre des biens déposés dans un musée de France et au récolement.

Elle répond avant tout à des impératifs éthiques et juridiques, notamment ceux des Principes de Washington, qui découlent de la Conférence de Washington sur les œuvres d'art volées par les nazis [2]. Chaque musée a ainsi la prérogative de s'assurer que les œuvres qu'il possède n'ont pas été illégalement confisquées à leurs propriétaires légitimes durant la période allant de l'arrivée au pouvoir du régime nazi en Allemagne à la fin de la Seconde Guerre mondiale (1933-1945).

Au-delà de ces impératifs, la documentation générale des collections à des fins scientifiques est également envisageable. Certains départements se sont déclarés ouverts à des projets de reconnaissance de l'écriture manuscrite dans le cadre de partenariats scientifiques avec des institutions de recherche, par exemple en histoire de l'art. Le projet HTR, un temps envisagé par le Musée national Eugène-Delacroix sur le fonds d'archives relatives aux travaux de construction et de rénovation de l'appartement de Delacroix, numérisé par l'Institut national d'histoire de l'art, en est une illustration.

3.1.2.2 Une typologie documentaire manuscrite variée.

En prenant en compte ces enjeux, chaque service de documentation avait identifié lors de nos entretiens des ressources manuscrites spécifiques à exploiter. Nous avons relevé principalement trois types de sources : les inventaires, les archives privées de conservateurs et les cahiers de fouilles.

Les inventaires.

L'inventaire désigne, de manière générale, une ressource documentaire qui recense et décrit en détail l'ensemble des objets possédés par le département d'un musée. En pratique, il peut se décliner sous différents formats, généralement sous forme de "livres d'entrées" ou de "registres d'entrées." Ces documents contiennent des informations plus ou moins détaillées sur chaque objet ou œuvre, telles que le numéro d'inventaire, le titre, l'artiste ou le créateur, la date d'entrée, la provenance, l'état, le mode d'acquisition (achat, don, ou legs.), ainsi que d'autres détails jugés importants par le conservateur.

Certains départements, comme le Département de peinture, disposent d'un autre type de support pour leur inventaire : le "livre de mouvements." Ces livres sont principalement conçus pour suivre tous les déplacements des objets ou œuvres de la collection sur une période définie, que ce soit à l'intérieur du musée (entre les salles d'exposition, les

2. La Conférence de Washington sur les œuvres d'art volées par les nazis, tenue en décembre 1998, a réuni 44 pays pour discuter de la restitution des biens culturels pillés par les nazis pendant la Seconde Guerre mondiale. Les « Principes de Washington » ont été adoptés par les participants à l'issue de la conférence.Il s'agit d'un ensemble de recommandations visant à identifier les œuvres volées, à encourager la transparence des archives, et à faciliter la restitution des biens aux héritiers légitimes.

réserves ou les ateliers de restauration) ou à l'extérieur (prêts pour expositions, restaurations externes, etc.). Ils contiennent, en plus de ces informations, l'essentiel de ce que l'on trouve dans les registres d'inventaires classiques.

La gestion des inventaires peut parfois être complexe. Plusieurs départements disposent d'inventaires fragmentés répartis sur différents registres, ce qui est souvent le cas des départements récemment créés. Certains étaient auparavant des collections thématiques au sein de départements plus vastes : par exemple, le Département des Arts de l'Islam faisait autrefois partie du Département des Objets d'Art. Le DABCO, créé en 2022, a quant à lui intégré une partie des collections en lien avec sa thématique, provenant du Département des Antiquités égyptiennes et du Département des Antiquités grecques, étrusques et romaines. Sa documentation dépend donc de l'inventaire de ces deux départements.

Centraux dans la vie d'un musée puisqu'ils sont indispensables au récolement, les inventaires font partie des documents à prioriser dans le cadre d'un traitement HTR selon la majorité des services de documentation. Généralement, les registres d'inventaires font partie des documents les mieux tenus du musée, en particulier les plus anciens. Tenus par des conservateurs, l'écriture est généralement soignée et provient sur de nombreuses pages d'une même main. Plus le registre est ancien, plus l'écriture est standardisée. Au vu de ces caractéristiques, nous pouvons considérer qu'il s'agit des ressources les mieux adaptées pour une reconnaissance d'écriture manuscrite. Nous livrerons plus de détails techniques autour des registres d'inventaire du DAGER dans la prochaine section.

Les archives privées de conservateurs.

Le deuxième type de document qui intéresse le plus les services de documentation sont les archives privées de conservateurs. Qu'ils aient travaillé au musée ou non, ces archives peuvent représenter une mine d'informations riches et précieuses pour un service cherchant à approfondir l'étude des collections de son département.

Cependant, ces archives peuvent parfois être très difficiles à lire, car il s'agit souvent de documents personnels (carnets, correspondances, fiches manuscrites) rédigés par le conservateur. L'écriture peut donc être moins soignée que dans le cadre d'un inventaire. Les compétences d'un documentaliste ayant étudié ces archives jusqu'à s'en familiariser peuvent s'avérer indispensables pour les déchiffrer. Au moment du stage, un projet de reconnaissance de l'écriture manuscrite était en cours sur ce type de documents au sein du musée. Sous la direction de Françoise Barbe, conservatrice au département des Objets d'art, et en collaboration avec le service d'ingénierie documentaire, un stage se déroulait en ce sens. Il avait pour objectif la reconnaissance de l'écriture des fiches manuscrites

consacrées aux émaux peints de Limoges, créées par l'ancien conservateur du département, Jean-Joseph Marquet de Vasselot, entre 1902 et 1933.

De manière générale, un projet HTR sur les archives privées d'un conservateur pourrait s'avérer extrêmement pertinent et rentable. En effet, ces archives contiennent un grand nombre de fragments manuscrits rédigés par une même main, ce qui permettrait aux modèles de transcription automatique de s'adapter plus facilement aux particularités de l'écriture, à condition de bénéficier d'un entraînement rigoureux. La transcription automatique serait bien plus performante sur ce type d'archive que sur des sources comportant des écritures variées. De plus, dans le cadre de la production d'une vérité terrain, l'aide de conservateurs ayant étudié ces archives est souvent envisageable. Des difficultés pourraient toutefois surgir si l'écriture du conservateur venait à manquer de soin, ce qui est fréquemment le cas.

Les cahiers de fouilles.

Le dernier type de document souvent mentionné au cours de nos entretiens est le cahier de fouilles. Ce support de travail est utilisé par les archéologues pour consigner de façon détaillée l'ensemble des observations et des données recueillies durant une fouille archéologique. Il regroupe des descriptions et des dessins des objets découverts. Ces cahiers peuvent fournir un complément documentaire précieux pour renseigner les pièces détenues par le département issues de fouilles.

L'efficacité d'un projet HTR sur ce type de source dépend largement du nombre de mains et de la qualité de l'écriture des archéologues qui l'ont rempli. Plus de précisions seront données lorsque nous aborderons plus en détail les cahiers de fouilles conservés par Service de l'histoire du Louvre dans la prochaine section.

L'absence de numérisation : une contrainte surmontable ?.

De manière générale, si, durant nos entretiens, un certain enthousiasme a pu être manifesté quant aux perspectives offertes par la reconnaissance d'écriture manuscrite sur les sources que nous venons de recenser, il est important de noter que beaucoup d'entre elles – surtout les archives privées – ne sont pas entièrement numérisées. Cet état de fait entrave grandement les possibilités d'un traitement HTR sur ces sources et doit être pris en compte.

Une piste pouvant surmonter l'absence de numérisation a été proposée lors de nos entretiens pour : la recherche collaborative. En effet, plusieurs projets de recherche ayant

mobilisé le HTR ont pu être financés et ont consacré une part non négligeable de leur budget à la numérisation du corpus étudié [3]. À cette fin, plusieurs départements s'étaient déclarés enthousiastes à l'idée de collaborer avec des institutions de recherche dans le cadre de projets HTR.

3.1.2.3 HTR et base de données numérique : le cas particulier du Département des Arts de Byzance et des chrétientés orientales.

Parmi les départements récents, le DABCO se distingue par sa base de données documentaire unique et ses projets novateurs. Un entretien avec son chef, Maximilien Durand, nous a permis de mieux comprendre les initiatives envisagées par le département. Dans ces projets, le HTR est envisagé sous plusieurs aspects. La base de données du DABCO est constituée d'une arborescence de fichiers accessible via l'explorateur Windows. Cette arborescence comprend des "dossiers d'œuvres" contenant leur documentation. Qu'il s'agisse d'articles scientifiques ou de catalogues de vente, sous forme d'images numérisées ou de captures d'écran, tout est au format `JPEG`. Ce choix a été fait dans une perspective de pérennité, notamment pour s'adapter aux futurs développements de l'intelligence artificielle, par exemple dans la détection d'images et la reconnaissance de texte. Le fait que cette base de données soit construite avec l'explorateur Windows permet qu'elle puisse être alimentée à l'avenir par des non-spécialistes de l'informatique disposant de notions bureautiques rudimentaires.

Par ailleurs, une océrisation ou un processus de HTR sur les images de cette base sont envisagés. Une fois obtenue, la transcription des images serait ajoutée aux métadonnées des fichiers, ce qui permettrait d'indexer chaque œuvre avec des mots-clés. La navigation dans l'explorateur Windows deviendrait ainsi plus rapide et efficace, car les dossiers d'œuvres contenant ces termes apparaîtraient lors de la recherche par mots-clés. Le fait que la communauté des humanités numériques s'intéresse de plus en plus à la reconnaissance d'écritures non latines est une perspective réjouissante pour le DABCO, dont certaines ressources documentant les œuvres peuvent être en serbe, bulgare, copte ou encore en arabe.

3. Notre expérience personnelle au sein du projet TariMa a été relatée lors de nos entretiens. Il s'agit d'un partenariat entre l'Institut de recherches et d'études sur les mondes arabes et musulmans (IREMAM), Aix-Marseille Université - Maison Méditerranéenne des Sciences de l'Homme (MMSH) et la Bibliothèque universitaire des langues et civilisations (BULAC), portant sur la numérisation et le traitement HTR d'une dizaine de manuscrits maghrébins conservés à la BULAC. Les financements alloués au projet ont grandement contribué à la numérisation de ces sources.

3.2 Mise en oeuvre sur les sources du DAGER et du SHL.

3.2.1 Enjeux et contraintes.

3.2.1.1 Réflexions autour du choix du corpus.

Rappelons que, dans le cadre du stage, il était prévu d'initier un projet test de mise en œuvre de reconnaissance d'écriture manuscrite sur une sélection des registres d'inventaires manuscrits du DAGER et des cahiers d'inventaire du mobilier de fouilles archéologiques de la Cour Napoléon du SHL.

La question de la sélection précise du corpus a été discutée à plusieurs reprises avec les deux référentes scientifiques du projet : Laura Favreau, cheffe du service d'études et documentation du DAGER, et Noémie Latte, documentaliste-scientifique du SHL. À l'issue de ces discussions, nous avons décidé de traiter un corpus de 200 fichiers images provenant de chaque source. Ces fichiers images issus de numérisations récentes étaient stockés dans les serveurs du Louvre au format `TIFF`.

Pour le bon déroulement de la mise en œuvre de la reconnaissance d'écriture manuscrite, il était nécessaire que les fichiers d'images issus des numérisations soient d'une qualité suffisante pour être reconnus par les modèles de segmentation et de transcription. Si toutes les numérisations des cahiers d'inventaires du mobilier de fouilles présentaient une qualité adéquate, ce n'était pas le cas des différents registres d'inventaires. C'est pourquoi la sélection du corpus des sources du DAGER devait se concentrer sur les livres d'entrée MN et N, tenus durant le Second Empire. Le corpus de 200 images a été rapidement constitué : il s'agissait de la totalité des 124 fichiers images de numérisation du livre d'entrée MN et des 76 premiers fichiers images de numérisation du livre d'entrée N.

Si l'écriture des livres d'entrées MN et N provenait de la même main, ce n'était pas le cas des cahiers d'inventaires du SHL. Sur un seul cahier de 80 pages, plus d'une dizaine de mains différentes pouvaient être identifiées. Notre première idée était de sélectionner un corpus comprenant le plus petit nombre possible de mains pour obtenir de meilleurs résultats à l'issue du traitement HTR. Cette idée a été rapidement abandonnée pour deux raisons. D'abord, les mains alternant de manière extrêmement irrégulière, leur identification et leur comptage s'avéraient extrêmement fastidieux. Ensuite, les résultats obtenus sur un tel corpus ne seraient pas représentatifs de la diversité réelle des écritures présentes au sein des cahiers, ce qui les rendrait tronqués et trompeurs.

Il a donc été décidé de sélectionner un corpus de 200 images choisi au hasard parmi le millier d'images disponibles, afin de représenter la variété des mains présentes dans les sources. Ainsi, le corpus de 200 images des cahiers d'inventaires du SHL comprenait l'intégralité des cahiers VI et VII (160 pages à eux deux) ainsi que les 40 premières pages du cahier VIII.

	Nombre d'images
Livre d'entrées MN	124
Livre d'entrées N	76
Total	200

TABLE 3.1 – Corpus du DAGER

	Nombre d'images
Cahier VI	80
Cahier VII	80
Cahier VIII	40
Total	200

TABLE 3.2 – Corpus du SHL

3.2.1.2 Caractéristiques du corpus final : difficultés attendues et réponses envisagées.

Une fois le corpus défini, une réflexion approfondie sur ses caractéristiques et les difficultés qu'elles pouvaient susciter lors d'un traitement HTR a pu être menée. Le premier grand défi serait probablement la segmentation, car la quasi-totalité des images des deux corpus contenait des inventaires sous forme de tableaux, à l'exception de 9 encarts présents dans le registre MN. Pour les livres d'entrées du DAGER, la forme des tableaux était parfaitement régulière et semblait être pré-imprimée. Ce n'était pas le cas dans les cahiers d'inventaires des fouilles du SHL : les tableaux étaient faits à la main et n'avaient pas toujours la même forme. L'entraînement de modèles sur ce corpus promettait donc d'être plus long et fastidieux.

Les écritures constituaient le second grand défi du processus HTR. L'écriture des livres d'entrées du DAGER était standardisée et régulière puisqu'elle provenait de la même main. Le recours à un seul et même modèle pré-entraîné sur des écritures du XIX^{e} siècle s'imposait donc comme une évidence. Parmi tous les modèles dont nous avions connaissance, celui produit dans le cadre du projet LectAuRep nous paraissait le plus pertinent, ayant été éprouvé lui aussi sur des registres. Il est à noter que, bien que l'écriture date du XIX^{e} siècle, sa lecture n'était pas toujours aisée. Dans le cadre de la production de vérité terrain, cette difficulté devait être prise en compte.

Provenant de mains diverses, l'écriture des cahiers d'inventaires du SHL était irrégulière mais plus lisible. La question qui se posait légitimement dans cette situation était de savoir s'il fallait recourir à un modèle mixte, c'est-à-dire un seul et même modèle pour tout le corpus, ou utiliser autant de modèles qu'il y avait de mains différentes. La deuxième option était envisageable dans le cas où une même main apparaissait sur plusieurs pages avant d'être remplacée par une autre. Dans le cas des cahiers d'inventaires, ce n'était presque jamais le cas et il pouvait y avoir jusqu'à 4 ou 5 mains sur une seule et même page. Le recours à un modèle mixte, si possible pré-entraîné sur une variété d'écritures du XX^{e} siècle, semblait donc l'option la plus pertinente. Cependant, dans ce domaine, nous n'avons pas pu trouver de modèles existants. Un jeu de données nommé Tapus Corpus, produit par Alix Chagué et disponible sur HTR-United, avait néanmoins attiré notre attention : il s'agissait d'une vérité terrain basée sur une variété de documents dactylographiés français du XX^{e} siècle, contenant des extraits de pièces de théâtre, de poèmes, de lettres et de rapports administratifs. Ces données pouvaient être utiles dans le cadre de l'entraînement de notre modèle mixte.

Si, de manière générale, la qualité des pages des livres d'entrées du DAGER était impeccable, ce n'était pas toujours le cas des cahiers d'inventaire de fouilles du SHL. Plusieurs taches d'encre et ratures pouvaient être disséminées au fil des pages. Nous ne percevions cependant pas cet état de fait comme un problème suffisamment important pour entraver le processus HTR.

Enfin, dans le cadre de la transcription, des questions pouvaient légitimement se poser quant aux abréviations présentes dans les deux corpus. Il ne nous a pas paru nécessaire de les développer, étant donné qu'elles étaient très compréhensibles pour la plupart. De plus, une liste des abréviations avec leur signification était incluse dans les cahiers d'inventaire de fouille du SHL.

3.2.1.3 Un environnement prédéfini : de Nakala à eScriptorium.

Dans le cadre de la mise en œuvre de la reconnaissance d'écriture manuscrite, une grande partie des solutions techniques ont été imposées par les conditions du stage.

Premièrement, le corpus d'images devait nécessairement être préalablement entreposé dans Nakala. Il s'agit d'un entrepôt de données géré par Huma-Num, une infrastructure de recherche dédiée aux disciplines des lettres, des sciences humaines et sociales, ainsi qu'aux humanités numériques, financée et mise en œuvre par le Ministère de l'Enseignement supérieur et de la Recherche. Cette migration des fichiers images des serveurs du Louvre vers cet entrepôt de données avait déjà été décidée en amont du projet HTR.

Elle permet non seulement de bénéficier d'un hébergement gratuit et pérenne des données, mais également d'assurer leur interopérabilité. En ce sens, un projet de bibliothèque numérique basé sur Omeka S et utilisant les images stockées dans Nakala est envisagé pour le futur. Enfin, la plateforme permet de disposer, grâce à son API, d'un manifeste garantissant une qualité d'image optimale et une interopérabilité des données.

Deuxièmement, eScriptorium a été choisie comme la plateforme à employer pour la mise en œuvre du HTR sur notre corpus. Plusieurs raisons expliquent ce choix. D'abord, l'accès aux serveurs du CREMMA était facilité puisque le consortium est porté par l'École nationale des chartes et l'INRIA. Cet accès se faisait via le remplissage, au préalable d'un formulaire renseignant le projet, ses sources ainsi que les fonctionnalités de la plateforme auxquelles nous souhaitions avoir accès.

Le recours à ces serveurs est précieux. D'abord, parce que le Louvre ne dispose pas d'une telle infrastructure. Ensuite, parce que, comme cela a été démontré dans les chapitres précédents, il n'est pas possible de trouver une infrastructure serveur gratuite offrant une performance similaire. La seule contrepartie est le partage, sous licence ouverte et selon une méthode documentée, de la vérité de terrain produite, ainsi que des éventuels modèles de transcription automatique. Les registres d'inventaires et les cahiers d'inventaire des fouilles étant des archives publiques, les données peuvent tout à fait être partagées.

3.2.2 Établissement d'un protocole

Une fois les enjeux et les contraintes liés aux sources cernés, une méthodologie a été établie pour la mise en œuvre, en respectant les principes fondamentaux de tout projet HTR. Chaque étape du processus devait être chronométrée et évaluée afin de donner une idée précise du rapport entre le temps investi et les résultats escomptés.

3.2.2.1 Téléchargement au préalables des modèles.

La première étape était le téléchargement des modèles de transcription dont nous avions besoin. Le modèle LectAuRep que nous souhaitions utilisé était disponible sur le *Github* du projet [4]. Après son téléchargement, il suffisait de l'importer dans la page « *My models* ».

3.2.2.2 Importation et segmentation.

L'importation des fichiers images dans eScriptorium se ferait en utilisant le manifeste Nakala.Pour la segmentation, nous comptions évaluer, pour les deux corpus, la

4. À l'adresse suivante : `https://github.com/lectaurep/lectaurep_base_model`.

pertinence de la segmentation automatique avec les modèles intégrés dans l'instance eScriptorium du CREMMA. Si le résultat n'était pas concluant, nous envisagions, pour les cahiers d'entrées du DAGER, de recourir à un entraînement de modèle de segmentation pré-entraînés sur des tableaux. Pour les cahiers d'inventaire des fouilles, nous procéderions à un entraînement avec une segmentation manuelle.

3.2.2.3 Transcription

Pour la transcription des cahiers d'entrées du DAGER, nous envisagions de fournir 5 types de transcriptions différentes, fonctionnant par paliers :

1. Une transcription automatique fondée uniquement sur le modèle LectAuRep, sans aucune vérité terrain ni entraînement supplémentaire.
2. Une transcription issue d'un entraînement du modèle LectAuRep avec une vérité terrain équivalente à 5 % du corpus, soit la transcription manuelle de 10 pages.
3. Une transcription issue d'un entraînement du modèle LectAuRep avec une vérité terrain équivalente à 10 % du corpus, soit la transcription manuelle de 20 pages.
4. Une transcription issue d'un entraînement du modèle LectAuRep avec une vérité terrain équivalente à 15 % du corpus, soit la transcription manuelle de 30 pages.
5. Une transcription issue d'un entraînement du modèle LectAuRep avec une vérité terrain équivalente à 20 % du corpus, soit la transcription manuelle de 40 pages.

Pour la transcription, nous comptions pour les cahiers d'entrées du SHL fournir également 5 types de transcription différentes fonctionnant également par pallier :

1. Une transcription issue d'un entraînement d'un modèle créé avec une vérité terrain équivalente à 5 % du corpus, soit la transcription manuelle de 10 pages.
2. Une transcription issue d'un entraînement d'un modèle créé avec une vérité terrain équivalente à 10 % du corpus, soit la transcription manuelle de 20 pages.
3. Une transcription issue d'un entraînement d'un modèle créé avec une vérité terrain équivalente à 15 % du corpus, soit la transcription manuelle de 30 pages.
4. Une transcription issue d'un entraînement d'un modèle créé avec une vérité terrain équivalente à 20 % du corpus, soit la transcription manuelle de 40 pages.
5. Une transcription issue du modèle créé et entraîné sur les 5 % du corpus, mais aussi sur l'intégralité des jeux de données du corpus Tapus Corpus.

Ces différents paliers permettent d'estimer les résultats de l'entraînement de modèles dans une multitude de cas, afin de déterminer si un investissement minimum, moyen ou important est nécessaire pour obtenir des résultats convaincants. Pour parvenir à cette estimation, une évaluation de chaque modèle sera réalisée en utilisant l'outil KaMI.

3.2.2.4 Exploitation de la transcription : le format de sortie.

3.2.3 Une mise en œuvre pleine de désagrément

Cette section vise à présenter la mise en œuvre du protocole que nous avons établi, ainsi qu'à mettre en lumière tant les réussites que nous avons obtenues que les difficultés auxquelles nous avons été confrontés.

3.2.3.1 Entreposage des données dans Nakala.

L'entreposage des données dans Nakala devait se faire en deux étapes. Tout d'abord, il fallait créer un fichier .csv répertoriant tous les fichiers image de notre corpus, en y ajoutant des métadonnées au format Dublin Core, à savoir : le nom du fichier (*file*), le statut de la donnée (*status*), le type de donnée (type), le titre de la donnée (*title*), l'auteur (*author*), la date de création de l'image (*date*), la licence associée (*licence*), la description (*description*) et une liste de mots-clés (*keywords*). Ce fichier .csv permet l'import par lot dans Nakala via un script Python, évitant ainsi l'importation et l'ajout manuel de chaque fichier.

Le script Python a été abandonné au profit d'un outil plus intuitif et rapide : l'application web *Mynkl*. Cette application permet également l'import par lot d'une grande quantité de fichiers. Il suffit de déposer dans l'application les fichiers image et le fichier .csv associé que nous venons de décrire. L'entreposage de 200 images s'effectue ainsi en moins de cinq minutes.

3.2.3.2 Importation : réajustements.

Lors de la mise en œuvre de notre protocole, nous avons rapidement rencontré une première difficulté majeure : l'importation sur eScriptorium d'un manifeste Nakala généré via son API n'était pas possible.

Cette situation a nécessité une recherche de solutions alternatives pour avancer dans notre projet. Pour ce faire, nous avons initié un échange avec Hugo Scheithauer, un des membres de l'équipe ALMAnaCH, faisant partie des responsables de la maintenance d'eScriptorium sur les serveurs du CREMMA. Cet échange a permis d'identifier une première piste de solution : nous a été recommandé d'avoir recours à une bibliothèque

Python capable de générer un manifeste IIIF à partir des fichiers images d'une donnée Nakala. Cette méthode avait précédemment fait ses preuves pour résoudre un problème d'affichage d'images dans des visionneuses lorsqu'un manifeste généré par l'API de était utilisé. Le recours à cette bibliothèque s'est avéré inefficace : le manifeste généré ne pouvait permettre l'importation de l'image.

Nous avons alors envisagé une autre approche en testant une importation directe des fichiers image dans leur format originel `TIFF`. Dans l'attente de retours complémentaires de l'équipe ALMAnaCH, cette solution temporaire a nous a permis de d'initier l'exécution du protocole.

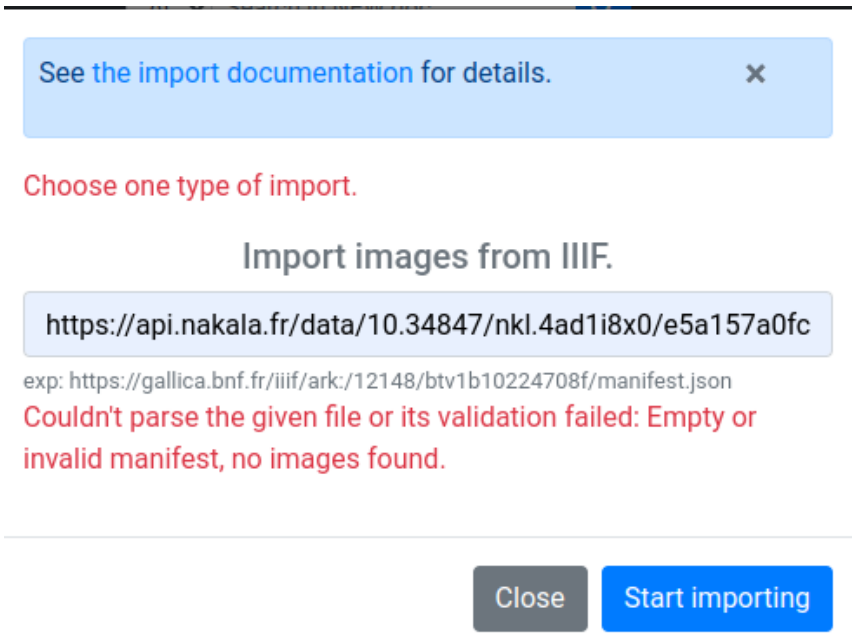

FIGURE 3.1 – Message d'erreur de l'application eScriptorium après l'importation d'un manifeste IIIF Nakala.

Rapidement, nous avons constaté que la segmentation automatique intégrée à eScriptorium était extrêmement efficace. En un temps relativement rapide, quelques secondes seulement, le modèle intégré à l'application a pu détecter les zones d'écriture et les lignes de base des images, à l'exception de certaines notes marginales.

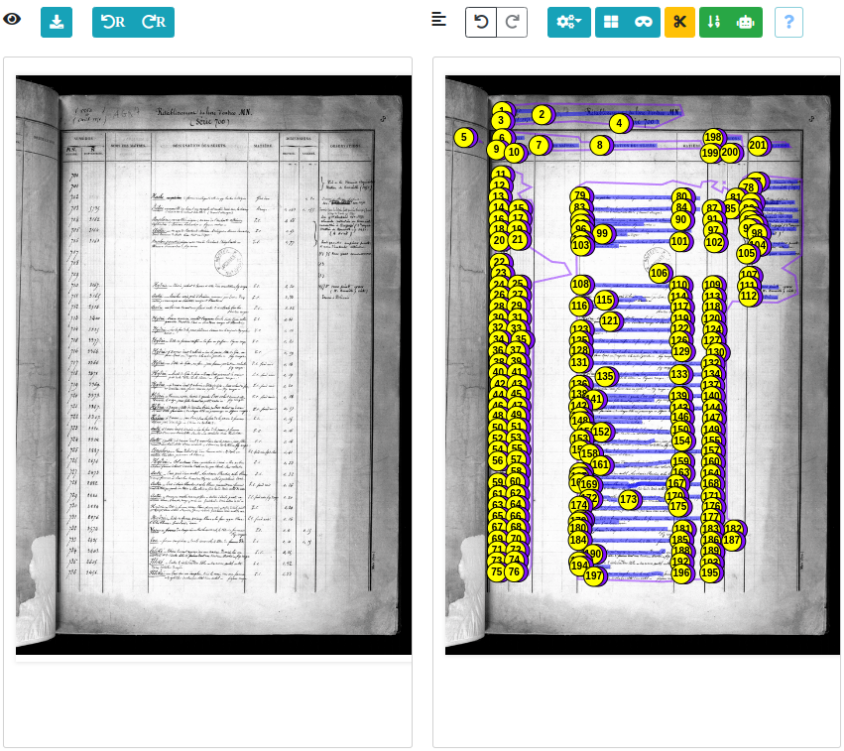

FIGURE 3.2 – Résultats issus de la transcription automatique par le modèle intégré.

Toutefois, la première tentative de transcription automatique avec le modèle Lectaurep, sans entraînement préalable grâce à une vérité-terrain, a donné des résultats désastreux : chaque ligne de texte n'a été transcrite que par un seul caractère n'ayant aucun lien avec le caractère d'origine.

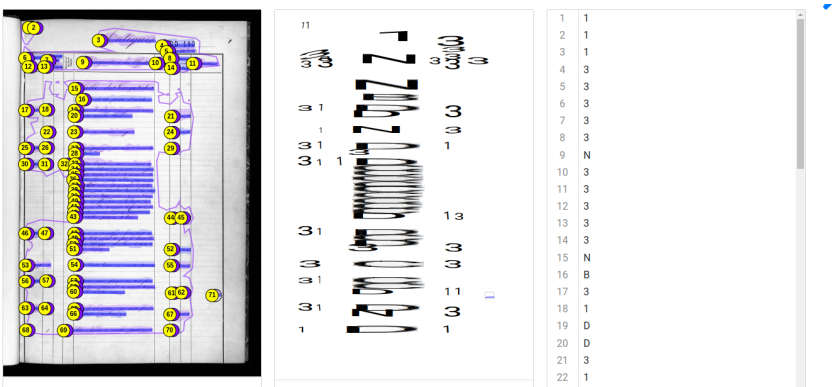

FIGURE 3.3 – Application du de la transcription automatique via le modèle Lectaurep sur un fichier `TIFF`.

Nous avons alors émis l'hypothèse que ce dysfonctionnement était lié à une mauvaise gestion des fichiers `TIFF` par eScriptorium, ce qui était particulièrement visible

lorsque l'on zoomait sur les images dans la visionneuse, celles-ci manquant cruellement de netteté.

Face à cette impasse, nous avons exploré une autre possibilité mentionnée dans la documentation d'eScriptorium : l'importation au format `PDF`. Nous avons donc converti toutes les images du corpus en format `PDF` à l'aide d'Adobe Acrobat, puis procédé à leur importation et à la remise en œuvre du protocole. La segmentation a alors montré une précision encore supérieure à celle obtenue avec les fichiers `TIFF`.

Cependant, la transcription automatique effectuée avec le modèle LectAuRep, toujours sans "vérité terrain", est restée extrêmement médiocre. Seuls les nombres que contenaient la page étaient parfaitement transcrits. Il est ainsi devenu évident que l'entraînement du modèle avec une vérité terrain était nécessaire pour obtenir des résultats satisfaisants. Malheureusement, ces contretemps ont entraîné un retard considérable dans le déroulement prévu du stage. Au moment de cette nouvelle tentative sur les fichiers PDF, il ne restait que quelques jours avant la fin de celui-ci, ce qui réduisait considérablement le temps disponible pour l'entraînement du modèle.

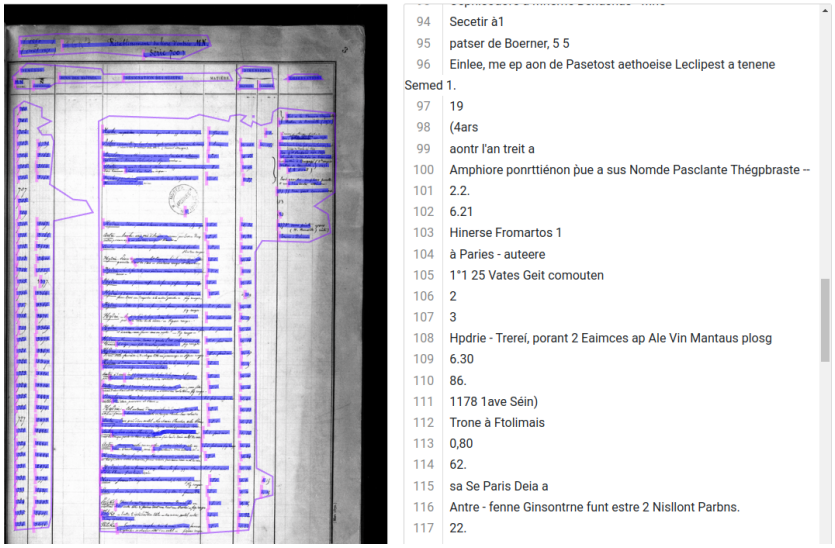

FIGURE 3.4 – Application du de la transcription automatique via le modèle Lectaurep sur un fichier `PDF`.

3.2.3.3 Un entraînement impossible.

Dans les derniers jours du stage, nous avons constaté que, bien qu'ayant explicitement indiqué dans le formulaire du CREMMA notre intention d'entraîner notre modèle HTR, cette fonctionnalité n'apparaissait pas sur l'application.

Nous avons donc de nouveau sollicité Hugo Scheithauer ainsi que Thibault Clérice, mais le temps qu'une réponse soit apportée, le stage était déjà terminé. Par conséquent,

seule une infime partie du protocole a pu être réalisée en raison des problèmes d'importation et de l'impossibilité d'entraîner les modèles.

3.2.3.4 Bilan du protocole.

La première difficulté que nous avons rencontré concernant l'importation découle notamment du choix initial de recourir à un manifeste Nakala, imposé dans le cadre du stage. Une connaissance insuffisante de la compatibilité des formats pouvant être importés dans eScriptorium, couplée à une communication imparfaite avec l'équipe chargée de sa maintenance, a considérablement entravé notre progression. Dans un contexte autre que celui d'un stage limité dans le temps, ces étapes auraient été tout à fait surmontables. Il est important de noter que la phase de segmentation automatique avait montré un potentiel extrêmement prometteur.

En conclusion, eScriptorium demeure à nos yeux un excellent choix pour ce type de projet, notamment pour l'accès à une infrastructure gratuite et puissante. Cette plateforme permet également de jouir des avantages de la communauté autour du HTR, très orienté vers l'open source, en permettant l'importation de modèles extérieurs. Malgré les échecs rencontrés, le protocole mis en place nous paraissait pertinent et respectait les principes fondamentaux qui doivent guider tout projet de reconnaissance automatique de texte manuscrit

Conclusion.

Le premier chapitre nous a montré que la reconnaissance d'écriture manuscrite ne s'est décloisonnée que très récemment du champ de l'informatique et de l'intelligence artificielle. Elle a pénétré le domaine des humanités numériques au milieu des années 2010. Cette arrivée est étroitement liée aux grands programmes de numérisation de masse entrepris par les institutions patrimoniales. Depuis la deuxième moitié des années 2010 et jusqu'à nos jours, le développement de la reconnaissance d'écriture manuscrite a été marqué par une grande effervescence et une collaboration initiées et portées par une multitude d'acteurs provenant des institutions de recherche en sciences humaines et en informatique, mais aussi du monde patrimonial. Une institution muséale, n'étant pas en lien direct avec cet univers, peut vite se retrouver désorientée face à une telle quantité d'acteurs, de projets et d'outils, et pourrait ne pas saisir l'esprit qui anime leur développement.

Connaître l'histoire de la reconnaissance d'écriture manuscrite revient également à comprendre les dynamiques qui entourent cet outil. Les cinq dernières années ont été marquées par une dynamique collaborative qui a permis le développement d'une multitude d'outils accompagnant la HTR, mais qui a également posé les fondements de nombreuses réflexions autour de la technique. Afin de s'inscrire, en tant qu'établissement public, dans ce domaine et de tirer parti de ces réalisations en faisant des choix pertinents, une connaissance de l'histoire de la technique et des projets d'envergure récents réalisés en la matière nous paraît primordiale.

Le second chapitre nous a montré que le HTR comporte également des règles et des contraintes techniques qui lui sont propres. Les connaître est essentiel avant d'amorcer n'importe quel projet, d'autant plus que ces règles et ces contraintes sont susceptibles d'évoluer avec les développements de l'intelligence artificielle, qui participe au renouvellement ou à l'amélioration des principes fondamentaux de la méthode. Nous avons ainsi pu observer comment la segmentation est susceptible d'évoluer, ou comment les phases de correction commencent à utiliser les grands modèles de langage. À partir de cette connaissance, en sachant ce qui est possible ou non de faire dans le cadre de son institution, il devient plus facile d'envisager et d'évaluer la pertinence des différentes solutions techniques sur le marché, disponibles pour la réalisation de nos projets. Des solutions open

source gratuites comme Arkindex ou eScriptorium existent et rivalisent, voire offrent plus de possibilités techniques que des plateformes payantes comme Transkribus.

Le troisième chapitre a montré que, maîtrisée, la HTR peut représenter pour un musée une technique prometteuse pour faciliter des tâches fondamentales d'un service de documentation, telles que le récolement et la recherche de provenance. On pouvait se demander si les sources d'un musée pouvaient être traitées par des plateformes habituellement confrontées à des sources provenant plutôt de bibliothèques ou d'archives. Les manipulations effectuées au cours de ce stage ont montré quelques promesses : des modèles préentraînés sur des sources similaires à celles des musées existent. Mais elles ont également révélé que les difficultés peuvent surgir, non pas d'un manque de maîtrise technique, mais de problèmes de communication entre institutions.

Ainsi, en résumé, nous pensons qu'une institution muséale comme le Louvre peut nourrir de grands espoirs dans le HTR à l'avenir pour réaliser ses ambitions. Mais pour ce faire, il faut non seulement connaître l'environnement autour de cette technique, qu'il s'agisse de son histoire, de ses acteurs ou de ses principes techniques, mais également participer activement au renforcement du dialogue entre les institutions muséales et le monde des humanités numériques.

Index

Table des matières