{BnF | Bibliothèque nationale de France

Article

La transcription automatique d'écritures manuscrites : premiers pas à la BnF

Automatic transcription of handwriting : first steps at the BnF

Sébastien Cretin[a]

[a] : Expert OCR et formats éditoriaux, département de la Conservation, BnF

Mots-clés: reconnaissance des écritures manuscrites (HTR), Reconnaissance optique de caractères (OCR), modèle, Intelligence artificielle (IA), Recherche &Développement (R&D)

Keywords: handwritten text recognition (HTR), Optical Character Recognition (OCR), model, Artificial Intelligence (AI), Research &Development (R&D)

Depuis une décennie, l'HTR (*Handwritten Text Recognition*, ou reconnaissance d'écritures manuscrites) représente la promesse d'accélérer radicalement l'accès aux contenus des documents manuscrits numérisés. Le formidable essor de l'intelligence artificielle (IA) -en particulier de l'apprentissage machine (*deep learning*)- a permis aux technologies d'HTR d'atteindre un niveau de développement permettant la mise en œuvre de programmes d'expérimentation préparant de futures phases de production, notamment à la BnF et dans d'autres institutions patrimoniales détentrices de fonds manuscrits.

Comment ça marche, et pour quels résultats ?

L'HTR, c'est une sorte de super OCR (*Optical Character Recognition*, ou reconnaissance de caractères imprimés). Les deux technologies ont un principe commun : elles visent à identifier, sur des images numériques, les signes (lettres, diacritiques, chiffres, ponctuation) formés par des ensembles de pixels foncés adjacents sur un fond plus clair. Ces technologies requièrent souvent un traitement préalable de l'image : la binarisation, c'est-à-dire la conversion de la trame trichromique[1] de l'image en une trame composée de pixels uniquement noirs ou blancs. L'HTR et l'OCR sont donc fondamentalement des processus de *computer vision*[2].

Ce qui les distingue, c'est la difficulté de l'exercice. Là où l'OCR traite de glyphes[3] séparés les uns des autres et distribués dans des polices qui sont majoritairement consignées dans les bases de référence des moteurs[4], l'HTR, lui, doit souvent composer avec des écritures cursives[5], qui plus est très différentes les unes des autres car il y a autant d'écritures que de scripteurs.

Mais cette difficulté n'empêche pas des résultats probants, spécifiquement quand les modèles[6] d'IA sont entraînés sur des écritures similaires à celle à transcrire, ou mieux, de la même main. Les textes produits par l'exploitation de ces modèles présentent des taux d'erreur au caractère suffisamment bas pour accélérer considérablement les travaux - de transcription ou de recherche - des archivistes, des bibliothécaires ou des paléographes.

Une planification pluriannuelle

Sa faisabilité technique étant établie, et ses promesses d'amélioration sûres, l'HTR s'impose comme un cas d'usage d'IA évident et essentiel pour les institutions patrimoniales détentrices de fonds manuscrits.

Les Archives nationales, déjà impliquées dans trois projets d'HTR (Himanis, LectAuRep et Simara[7]), se sont assigné l'objectif, dans leur "Stratégie 2021-2025", de "devenir pôle d'excellence" en la matière. Elles entendent "dépasser le stade de l'expérimentation" par un "programme ambitieux capable de lever des verrous en matière d'accès à certains fonds et d'offrir une pénétration rapide de leurs contenus "[8].

La BnF a elle aussi adopté une stratégie de planification pluriannuelle, via une "Feuille de route IA pour la période 2022-2026", dont les projets structurants, coordonnés par une "Cellule IA", intègrent l'HTR [9].

Impliqué dans les travaux exploratoires en cours, le "BnF DataLab[10]"[11] accueille les équipes de recherche lauréates de ses appels à projets. Financés et accompagnés - scientifiquement et techniquement - les chercheurs ont un accès facilité aux collections numériques de l'établissement, y compris celles sous droits. Cette année deux projets d'HTR sont hébergés :

- **HTRomance** : porté par Thibault Clérice (Centre Jean-Mabillon, École nationale des Chartes, Inria[12]) et Alix Chagué (Inria, Université de Montréal), sa haute ambition est de produire un modèle de transcription résistant aux changements de mains, voire de langues, afin de traiter des manuscrits littéraires en latin et langues romanes, du XI^{e} au XIX^{e} siècle.
- **Valorisation numérique du fonds Dulaurier** de la **BnF** : porté par Bernard Coulie et Bastien Kindt (Institut orientaliste de Louvain, UCLouvain) et Chahan Vidal-Gorène (Calfa[13]), il traite des manuscrits arméniens copiés ou par Édouard Dulaurier (1807-1881), ou qu'il avait fait copier. Ce projet, dont la qualité des premières restitutions est à souligner, se propose d'être une preuve de concept d'un *modus operandi* pour le traitement de langues à graphie non latine avec peu de données d'apprentissage.

L'année dernière était hébergé le projet **Gallic(orpor)a** : porté par Benoît Sagot (Inria), Simon Gabay (Université de Genève) et Jean-Baptiste Camps (École nationale des Chartes), il a produit une chaîne de traitement des documents anciens de Gallica, des premiers manuscrits aux imprimés révolutionnaires. Cette chaîne exporte non seulement le texte des documents mais aussi des informations sémantiques comme les entités nommées[14].

Un pipeline de tests

L'hébergement de ces projets comporte de précieux avantages pour la BnF. Le dialogue avec les chercheurs permet de juger finement de l'état de l'art, de suivre son évolution rapide, et facilite l'utilisabilité des fichiers informatiques, outils et modèles produits dans le cadre de ces projets. Cette dernière intention a abouti à la mise en place d'un programme interne de recherche et développement (R&D) sur l'HTR. Supervisé par un groupe de travail qui fait le lien entre la "Cellule IA" et le "DataLab", il consiste à expérimenter les modèles HTR existants pour fixer l'horizon de la phase d'industrialisation et le début de versements réguliers dans Gallica de textes produits par HTR.

Deux plateformes existent et pouvaient constituer des options techniques pour cette R&D : Transkribus[15] et eScriptorium[16]. Une journée d'étude leur a d'ailleurs été consacrée à la BnF en mai 2022[17] Pour diverses raisons, elles ont été écartées du cœur de la stratégie technique de la R&D, au profit de l'exploitation directe de Kraken, un moteur HTR *open source*[18] déjà massivement utilisé par la communauté des Humanités numériques[19]. Le pipeline de traitement construit dans un environnement Linux[20] permet de :

- consulter les métadonnées descriptives des modèles HTR disponibles. Elles sont de première importance pour déterminer quels documents sont traitables par quel modèle : ampleur des données d'entraînement, langue(s) et caractères couverts, gestion des abréviations ou des ratures, mesure de performance sur les données test, etc. ;
- télécharger les modèles et les appliquer. Le suivi de la segmentation des écritures et de leur transcription se fait dans un terminal ;
- évaluer rapidement l'efficience des modèles : les données quantitatives sont synthétisées dans une interface (*Fig. 1*) qui donne accès à tous les fichiers informatiques produits lors du processus. L'utilisateur peut voir le résultat de la binarisation des images, visualiser la segmentation des lignes, et lire le texte transcrit (*Fig. 2*) ;
- soumettre des "vérités terrain"[21] pour obtenir, par comparaison avec les prédictions du modèle, une mesure objective de sa précision.

In fine, ce pipeline produit évidemment des fichiers XML[22] ALTO ; format qui porte aujourd'hui les OCR consultables dans Gallica.

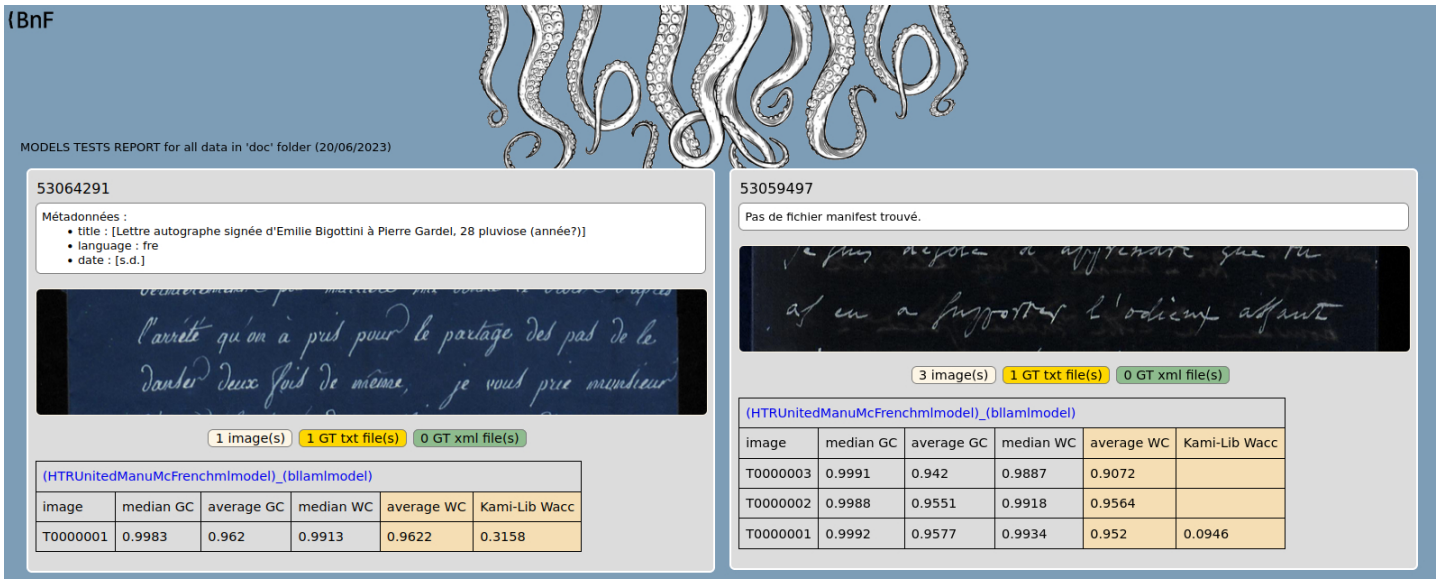

Figure 1 : Capture d'écran de l'interface de consultation des résultats des tests. Chaque document traité est identifiable par des métadonnées techniques ou bibliographiques. Sous la vignette donnant un aperçu de l'écriture, sont affichés plusieurs indicateurs de précision de la transcription. ©BnF. Sébastien Cretin

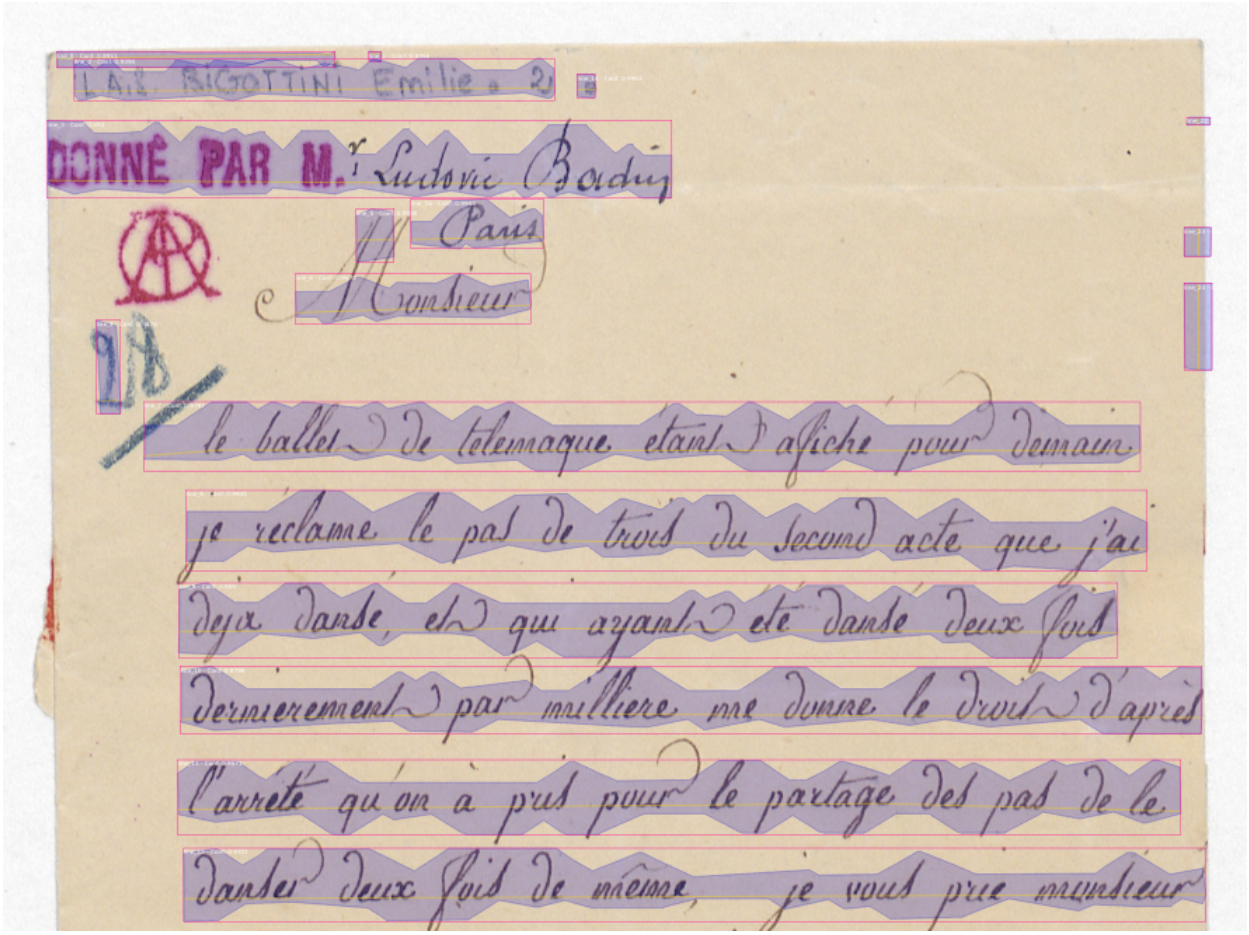

Figure 2 : Capture d'écran d'un des exports des outils de test ; il sert à visualiser la segmentation des lignes. Sur la page du manuscrit, sont tracés des polygones représentant les zones d'écriture identifiées lors du traitement ©BnF. Sébastien Cretin

La tour (de Babel) infernale

La R&D HTR à la BnF ne compte que quelques mois d'existence, et la difficulté de l'HTR s'y impose continuellement comme une évidence. Cela ne contredit pas ce que nous avons écrit plus haut : les résultats sont probants quand le modèle utilisé a été entraîné sur des écritures similaires à celle à transcrire. En dehors de cette situation, l'efficience des modèles plonge.

Les collections physiques de la BnF comptent plus de 400 000 manuscrits - dont 182 000 sont numérisés et disponibles dans Gallica -, écrits dans une grande variété de langues, par un nombre incalculable de scripteurs, depuis l'Antiquité

jusqu'à nos jours. Vouloir les déchiffrer tous avec un ordinateur c'est chercher son chemin dans une sorte de tour de Babel infernale. Le fil d'Ariane que l'on tire aujourd'hui pour en sortir, c'est l'utilisation de modèles dits "génériques", c'est-à-dire entraînés sur des matériaux volumineux et très hétérogènes, qui peuvent dans de rares cas fournir des résultats acceptables (autour de 50% de fidélité au mot). Il en existe quelques-uns exploitables avec Kraken ; on peut citer :

- CREMMA pour les manuscrits en latin et en ancien français du VIII^{e} au XV^{e} siècle ;
- LECTAUREP pour le français cursif contemporain ; entraîné dans le cadre d'un projet dirigé par les Archives nationales, sur des documents administratifs (1742 à 1928) ;
- Manu McFrench pour le français cursif moderne et contemporain (1600 à 2000) ;
- HTRomance, cité plus haut, qui devrait rejoindre cette liste dans les prochains mois.

One model to rule them all ?*

La fin de 2023 et l'année 2024 seront encore, à n'en pas douter, des phases exploratoires, pendant lesquelles le dialogue avec les chercheurs créateurs de modèles aura une importance stratégique majeure. Les technologies de l'IA ont connu ces dernières années, et même ces derniers mois, des évolutions impressionnantes. L'HTR est pris dans cette effervescence. De là à croire en l'émergence prochaine de modèles tellement robustes qu'ils pourraient déchiffrer toutes les mains d'écriture, de toutes les époques ? Non. Mais croire en d'énormes progrès à venir, et avoir l'ambition de les suivre, sans nul doute !

* Référence humoristique au Seigneur des Anneaux ("And one ring to rule them all..."). On pourrait traduire par "Un modèle pour maîtriser toutes [les écritures] ?

Notes

1. **Trichromie** : Les pixels des images numériques sont colorés selon un dosage de 3 couleurs : rouge, vert, bleu.

2. ***Computer vision*** : La vision par ordinateur est un domaine de l'IA qui permet aux ordinateurs d'interpréter ce qui est représenté sur les trames en pixels des images numériques ou des vidéos. L'IA permet aux ordinateurs de "penser" ; la vision par ordinateur leur permet de "voir". La détection d'objets, la segmentation d'images, la classification d'images, la reconnaissance de formes ou de mouvements, sont les quatre grandes catégories d'algorithmes utilisés dans ce domaine.

3. **Glyphe** : représentation graphique d'un signe typographique. Change avec la police de composition.

4. **Moteur** : synonyme de logiciel, ici.

5. **Écriture cursive** : style d'écriture manuscrite dans lequel les lettres sont liées lors du tracé.

6. **Modèle** : En IA, un modèle est la construction algorithmique générant une déduction ou une prédiction à partir de données d'entrée. Le modèle est "entraîné" avec des données constituées ou vérifiées par des humains.

7. **Projets HTR des Archives nationales** :

Himanis : https://www.irht.cnrs.fr/fr/recherche/les-programmes-de-recherche/himanis

LectAuRep : https://lectaurep.hypotheses.org/

Simara : https://speakerdeck.com/etalabia/20220203-datadrink-simara

8. Archives nationales (2021) Stratégies 2021-2025 / Bruno Ricard, dir. Paris : Archives nationales, 29 p. [pdf]

https://francearchives.gouv.fr/file/139ee2d26405dd90a644f09865a0fe96e53cea60/AN_Strat%C3%A9gie_2021-2025.pdf

9. *La BnF et l'intelligence artificielle* https://www.bnf.fr/fr/feuille-de-route-ia#bnf-feuille-de-routeBNF%20feuille%20de%20route%20ia

10. **BnF DataLab** : Ouvert en octobre 2021, c'est un service à destination des chercheurs qui souhaitent travailler sur les collections numériques de la BnF. Une convention de partenariat a été signée avec la très grande infrastructure de recherche (TGIR) Huma-Num. Il propose aux chercheurs des outils, des environnements, des accompagnements adaptés aux différentes étapes de leurs projets numériques (constitution de corpus, traitements et analyses, valorisation, préservation des données…). Le BnF DataLab est implanté en salle X, en bibliothèque de recherche.

11. **Bnf. Datalab**. Les projets de recherche à la Bnf. https://www.bnf.fr/fr/les-projets-de-recherche-bnf-datalab

12. **Inria** : Institut national de recherche en informatique et en automatique.

13**. Calfa** : Structure associative qui conçoit et développe des outils et des ressources pour l'étude de l'arménien (dictionnaires multilingues, étymologiques, de synonymes, moteur HTR, analyseur de structures et de texte). https://calfa.fr/

14. **Entité nommée** : Dans un texte, mot (ou un groupe de mots, ou nombre) qui identifie une personne, une organisation, un lieu, un évènement, une date, une quantité, une distance, etc. La reconnaissance d'entités nommées est une tâche basique et primordiale de l'extraction d'information dans un document écrit.

15. **Transkribus** : Plateforme de transcription et d'annotation de documents manuscrits numérisés, elle permet aussi d'entraîner ses propres modèles. Développée par l'université d'Innsbruck, elle est payante. https://readcoop.eu/transkribus/

16. **eScriptorium** : Plateforme de transcription et d'annotation de documents manuscrits numérisés, elle permet aussi d'entraîner ses propres modèles. Développée par l'université PSL (Paris Sciences & Lettres), elle est gratuite. https://ephenum.hypotheses.org/1412

17. "*Point HTR 2022 : Transcrire, annoter et éditer numériquement des documents d'archives*" ; Dassonneville Gautier et al. (2022) « Compte-rendu de la journée d'étude *Point HTR 2022 » Transkribus / eScriptorium : Transcrire, annoter et éditer numériquement des documents d'archives* : 9 mai 2022 Bibliothèque nationale de France – Site François-Mitterrand, [Rapport de recherche] CAPHES - UMS 3610 CNRS/ENS; AOROC

https://hal.science/hal-03692413

18. https://kraken.re/main/index.html

19. **Humanités numériques** : Elles se proposent de redéfinir les méthodes et les sujets de recherche dans le champ des arts, lettres, sciences humaines et sciences sociales, grâce aux outils informatiques (fouille de données, traitement automatique du langage…).

20. **Linux** : Contrairement à Windows ou MacOS, Linux est un système d'exploitation *open source* développé de façon collaborative. Il est indispensable au moteur Kraken utilisé dans le programme de R&D HTR de la BnF.

21. **Vérité terrain** : Transcription d'un document réalisée et vérifiée par un humain, à laquelle est comparée la transcription automatique réalisée par un ordinateur. Cette comparaison permet de mesurer précisément l'efficience de l'ordinateur.

22. **XML**: *eXtensible Markup Language*. Langage de structuration de données textuelles basé sur l'emploi de balises qui signalent la nature sémantique des données. Il sert à partager et archiver l'information.

Contact :
CRETIN Sébastien : sebastien.cretin@bnf.fr

Résumé :

Cet article décrit le contexte de mise en œuvre du programme de recherche et développement de la BnF dans le domaine de la transcription automatique des écritures manuscrites.

Abstract :

This article describes the background of the implementation of BnF's research and development program in the field of automatic handwriting transcription