Du manuscrit à l'algorithme

L'ESR et les institutions patrimoniales à l'heure de l'*Handwritten Text Recognition*

De Craene Valentin [1]

13/03/2024

1. Ingénieur d'études - humanités numériques, chargé du traitement des données scientifiques, MESHS-CNRS (UAR 3185). @: valentin.de-craene@univ-lille.fr

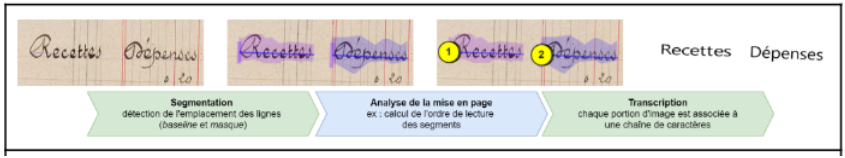


Figure – L'HTR en une image : les trois phases de traitement de la donnée

- ***Handwritten Text Recognition (HTR)*** ou reconnaissance automatique des écritures manuscrites
- ***Deep-learning*** ou apprentissage profond

Enjeux principaux :

Un changement d'échelle dans le traitement et la gouvernance des données ?
Un transfert de compétence entre ingénierie et recherche ?

Une nouvelle technologie en effervescence ? Un bref historique de l'HTR

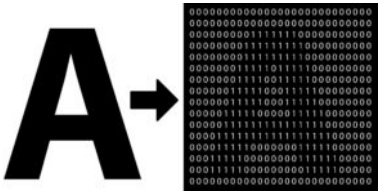


Figure – Processus d'OCR par binarisation

Figure – Processus d'HTR par segmentation

OCR	HTR
Performance : Taux d'erreur sur les caractères inférieur à 2 %, fonctionne uniquement sur les documents imprimés	Performance : Taux d'erreur sur les caractères entre 5 et 10 %, fonctionne sur les documents manuscrits
Outils : Abby (adobe, commercial); Tesseract 4 (gratuit, code ouvert); OCR4all; OCRD	Outils : Transkribus (commercial) ou Kraken (gratuit, code ouvert)
Fonctionnement : Modèles génériques par langue préexistants et s'appuie sur des fontes de caractères	Fonctionnement : nécessite la constitution d'un corpus d'entraînement pour entraîner un modèle

Figure – *Benchmark* des performances de l'OCR et de l'HTR par A. Pinche

Le deep-learning au service du paléographe : fonctionnement de l'HTR

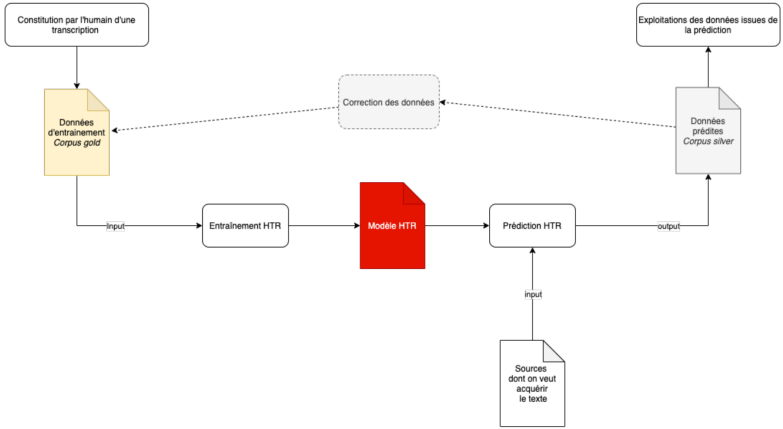
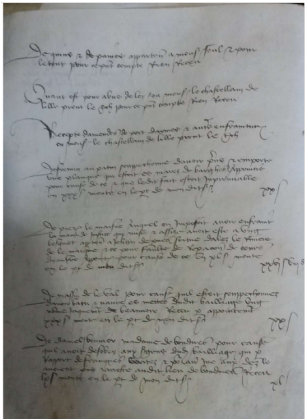
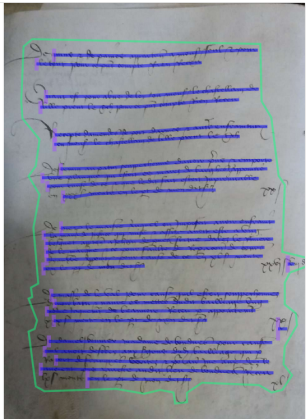


Figure – Structuration d'un *workflow* d'HTR

Infrastructures et outils de l'HTR





1 De quins et de paniés appartenant a monseigneur seul et pour
2 le tout pour ce present compte rien receu
3 Quant est pour abus de loy ou monseigneur le chastellain de
4 Lille prent le tierche pour ce present compte rien receu
5 Recepte d'amendes de port d'armes et autres enfraintures
6 ou monseigneur le chastellain de Lille prent le tierche
7 De Fremin au Pati souppechonné d'avoir pris et emporté
8 une planque qui estoit es mettes de Barghes apportié
9 pour cause de ce et que le dit fait estoit improuvable
10 en XXX sous monté en part de mon seigneur XX sous
11 De Pierre le Maistre auquel on imposoit avoir enfrainct
12 la main de justice qui mise et assize avoit estéa ung
13 bosquet appartenant a Jehan de Pernes situé ?alez le riviere
14 de la marque et ce pour faulte de reparacion de cours
15 Appointé pour cause de ce en XL sous monté
16 en le part de mon dit seigneur
17 De Massin de le Val pour cause qu'il estoit souppechonnez
18 d'avoir batu et manié es mettes dudit bailliage ung
19 nommé Jacquement de Beaunove receu par appointement
20 XXX sous monté en le part de mon dit seigneur
21
22
23 De Daniel Bennier madame de Bondues pour cause
24 qu'il avoit desobey aus sergens dudit bailliage qui par
25 rapport de franques veritez et pour l'an jue aux dez le
26 avois pris et arresté au dit lieu de Bondues receu
27 LX sous monté en le part de mon dit seigneur

Figure – Interface de transcription HTR d'e-Scriptorium

Mettre en oeuvre l'HTR : un bref état de l'art

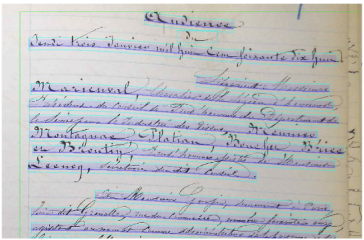


Figure – ANR TimeUS : segmentation via Transkribus

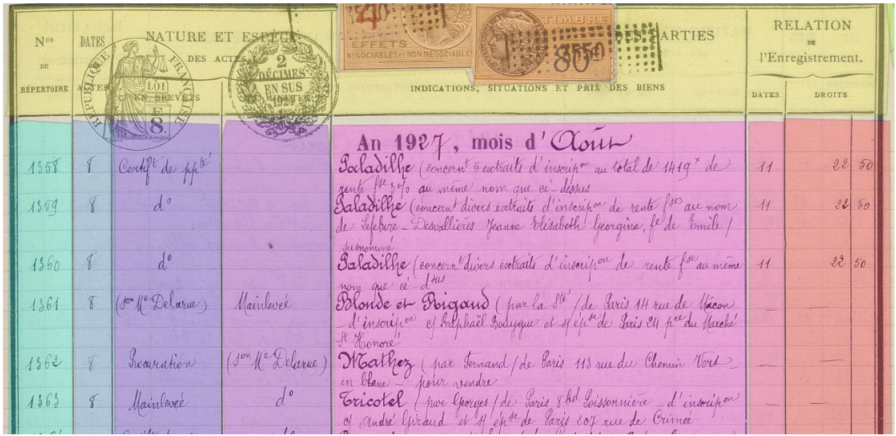


Figure – Projet LecTauRep : segmentation via E-Scriptorium

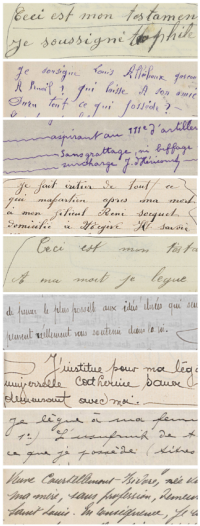
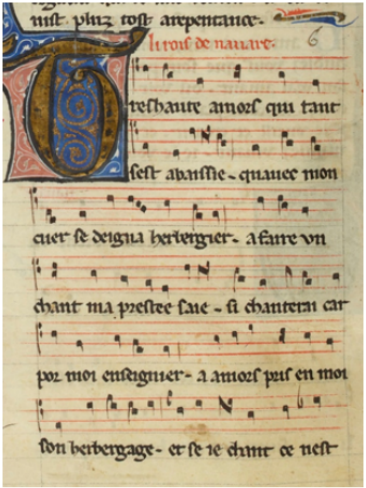


Figure – Diversité des documents transcrits au cours du projet Testaments de Poilus



Chansonnier M (Paris, BnF, fr. 844, f. 12vd)

Figure – ANR Maritem : *Manuscrit du Roi*

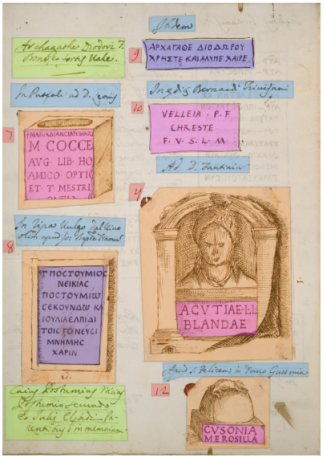


Figure 1. Venice, Marciana Library, Marc. Lat. XIV, 200 (4336), f. 1v; regions of interest coloured by type.

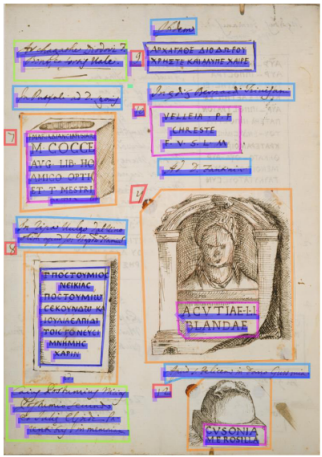


Figure 2. Venice, Marciana Library, Marc. Lat. XIV, 200 (4336), f. 1v; baseline recognition.

Figure – Projet Episearch : segmentation et reconnaissance des lignes

Figure – Binarisation des numérisations du projet ANR ChamDoc

Mutualiser et partager les vérités de terrain : le cas HTR-United

Figure – Logo d'HTR-United

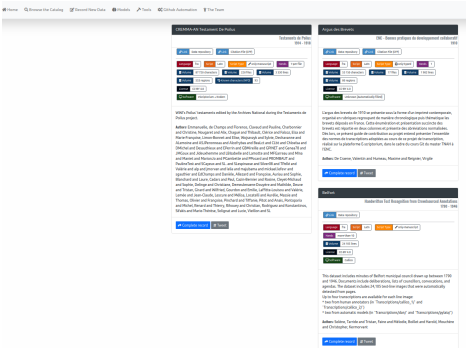


Figure – Interface web du catalogue HTR-United

Les enjeux et évolutions contemporaines de l'HTR

- Vers des modèles génériques (ex : CREMMA et HTRomance)
- Améliorer la segmentation des modèles (besoin de corrections manuelles pour le moment)
- Pérenniser les données et insérer l'HTR dans les institutions patrimoniales

L'HTR en bibliothèque et en archives : position stratégique et gouvernance de la donnée

- Au sein de la BnF :
  1. Feuille de route 2022-2026 « Intelligence Artificielle »
  2. **Cellule IA** d'expertise et mise en production

- Archives Nationales :
  1. **Stratégie 2021-2025** dont point 16 sur l'intelligence artificielle
  2. Accueil de projets (ex : LecTauRep) et structuration d'un « pôle d'excellence »