Paléographie et deep-learning, introduction à l'Handwritten Text Recognition (HTR)

Valentin De Craene

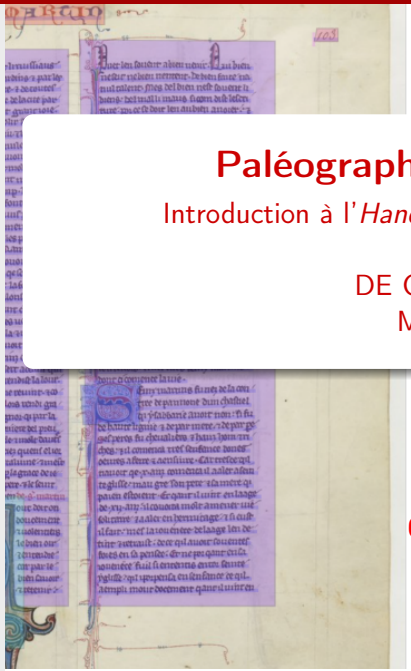
Introduction

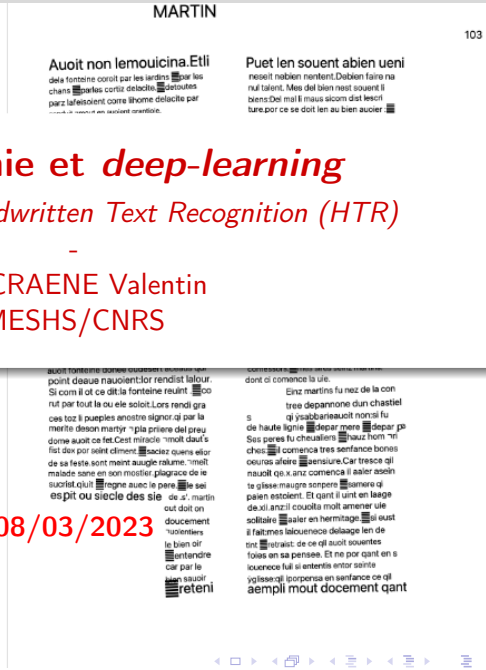
L'HTR, pourquoi et pour qui ? Etat des lieux et principes généraux

Mettre en oeuvre l'HTR : développement d'une chaîne de traitement

Les limites de l'HTR et ses développement à venir

Bibliographie





Paléographie et *deep-learning*

Introduction à l'*Handwritten Text Recognition (HTR)*

-

DE CRAENE Valentin
MESHS/CNRS

08/03/2023

Introduction

- Mai 2022, HTR United : + 220 000 lignes / + 4 millions caractères
- Engouement passager ou maturité d'une (nouvelle) technologie ?

Figure – Logo de l'initiative HTR-United

Introduction
L'HTR, pourquoi et pour qui ? Etat des lieux et principes généraux
Mettre en oeuvre l'HTR : développement d'une chaîne de traitement
Les limites de l'HTR et ses développement à venir
Bibliographie

Handwritten Text Recognition(HTR)

Processus, outils et données produites dans le cadre de la transcription automatisée d'un texte manuscrit à l'aide d'une intelligence artificielle

2 composantes majeures :

- Transcription « automatisée » des documents patrimoniaux = **processus**
- Modéles de transcription (intelligence artificielle) = **outil**

Introduction

L'HTR, pourquoi et pour qui ? Etat des lieux et principes généraux

Mettre en oeuvre l'HTR : développement d'une chaîne de traitement

Les limites de l'HTR et ses développement à venir

Bibliographie

Le développement de l'HTR implique-t-il de reconsidérer la place du paléographe dans l'édition et le traitement des sources manuscrites ?

Paléographie et *deep-learning*

Introduction

L'HTR, pourquoi et pour qui ? Etat des lieux et principes généraux

Mettre en oeuvre l'HTR : développement d'une chaîne de traitement

Les limites de l'HTR et ses développement à venir

Bibliographie

Principes généraux et définitions

Pourquoi et dans quelle mesure recourrir à l'HTR ?

- Échelle du **chercheur** : transcription massive + démarche d'ouverture des données de la recherche.
- Échelle d'un **projet** : « passage à l'échelle » de la phase de production et acquisition de la donnée.
- Échelle de l'**ESR** : mutualiser les corpus + ontologies communes + transparence et reproductibilité des données produites.

Introduction

L'HTR, pourquoi et pour qui ? Etat des lieux et principes généraux

Mettre en oeuvre l'HTR : développement d'une chaîne de traitement

Les limites de l'HTR et ses développement à venir

Bibliographie

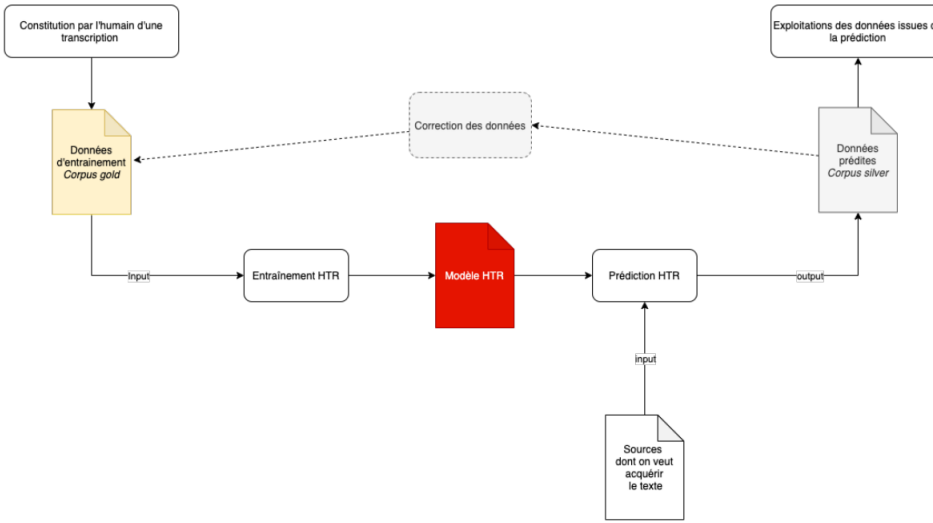


Figure – *Workflow* général d'un processus d'HTR

Paléographie et *deep-learning*

Introduction

L'HTR, pourquoi et pour qui ? Etat des lieux et principes généraux

Mettre en oeuvre l'HTR : développement d'une chaîne de traitement

Les limites de l'HTR et ses développement à venir

Bibliographie

Quelques notions essentielles :

Machine Learning (intelligence artificielle)

Permettre aux ordinateurs d'améliorer leurs performances dans la réalisation d'une tâche répétitive + effets socio-culturels de ses technologies.

Apprentissage supervisé

Méthode d'entraînement par annotation (ou transcription) massive des données en entrée.

Deep learning (apprentissage profond)

Branche du *machine learning* apprentissage par modélisation d'un réseau neuronal + haut niveau d'abstraction.

De l'OCR à l'HTR : historiciser l'intelligence artificielle

Occular Character Recognition (OCR) :

- Reconnaissance caractères par caractères
- Massification dans les années 90. De nos jours : plafond de verre. Outils libres et gratuits très puissants : Tesseract.

Handwritten Text Recognition :

- Changement de paradigme car prédiction (vectorisation) et segmentation par zone (double modèle de prédiction)
- Début années 2010 : apport des GAFAM à l'IA. Spécialisation des technologies et ouverture aux données du patrimoine. 2016 : Transkribus, puis 2019 : e-Scriptorium

Cf. travaux de Th. Poibeau (DR CNRS, LATTICE) :

- 1949 : John McCarthy, distinction fondamentale entre processus d'apprentissage artificiel et « intelligence ».
- 1950-60 : engouement autour de la « cybernétique » et interdisciplinarité.
- 1970-80 : 1e hiver de l'IA.
- 1980' : renouveau avec développement des premières BDD relationnelles massives.
- 1990'-2010 : 2e hiver de l'IA car coût d'entrainement trop élevé.
- depuis 2010 : seconde renaissance avec apports des corpus massifs par les GAFAM + accroissement de la puissance de calcul disponible.

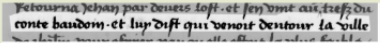
La recherche (française) à l'heure de l'HTR : état des lieux

- Édition de corpus : développement initial autour de la paléographie médiévale.

Test scores out-of-domain

	BnF, ms, fr. 17229, 13th c.	BnF, ms, fr. 185, 14th c.	BnF, NAF 6213, 15th c.	ALL
Cortado	92.71%	92.07%	87.48%	90.95%
1.1.0 Bicerin	91.64%	91.34%	83.40%	89.23%
1.0.1 Bicerin	90.66%	88.45%	79.67%	86.50%

Example, Cortado model on BnF, NAF 6213



conte baudoi. et luy dist qui uenoit dentour la ville

Figure – Modèle saint Martin (BnF fr 412) et son modèle Cremma-Lab, ENC

- Traitement massifs des données textuelles.

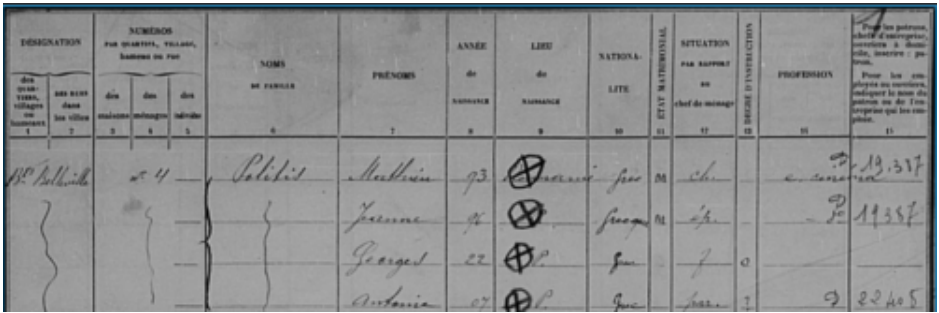


Figure – POPP Projet d'Océrisation des Recensements de la Population Parisienne (LARHRA - GED Condorcet)

- Segmentation des documents anciens et épigraphie.

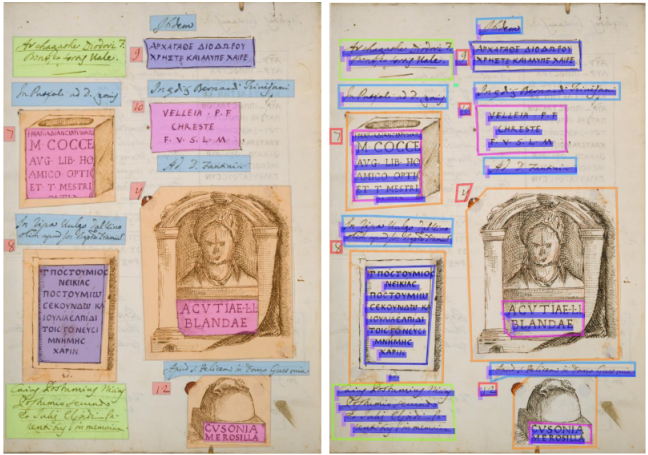


Figure 1. Venice, Marciana Library, Marc. Lat. XIV, 200 (4336), f. 1v; regions of interest coloured by type.

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

Figure – Projet EpiSearch, Université de Venise

Introduction

L'HTR, pourquoi et pour qui ? Etat des lieux et principes généraux

Mettre en oeuvre l'HTR : développement d'une chaîne de traitement

Les limites de l'HTR et ses développement à venir

Bibliographie

- Préservation et étude d'écritures rares ou anciennes.

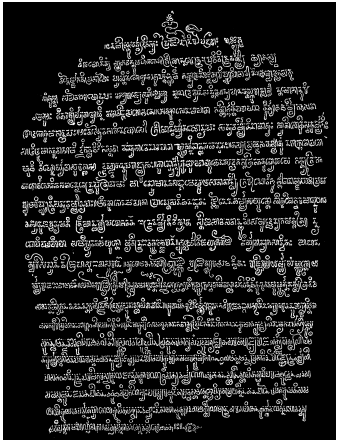


Figure – Projet CHAMDOC (préservation du cham ancien, Vietnam, VIe au XVIIe siècle

Paléographie et *deep-learning*

Introduction

L'HTR, pourquoi et pour qui ? Etat des lieux et principes généraux

Mettre en oeuvre l'HTR : développement d'une chaîne de traitement

Les limites de l'HTR et ses développement à venir

Bibliographie

De la segmentation à la reconnaissance des écritures : le cas d'e-Scriptorium

- Utilisation du logiciel e-Scriptorium sur les serveurs de l'INRIA.
- e-Scriptorium (2019) : développement au sein de l'ESR à partir du modèle d'algorithme Kraken (disponible sous forme de package Python).
- Code libre d'accès sur Github et possibilité de déploiement local (ligne de commande) ou sur d'autres serveurs (local host ou serveur distant).
- Exemple à partir des registres du bailliage de Lille (1429-1441) par Hubert Ghommer (ADN, B 6237 et suivants).

`https://traces6.paris.inria.fr/`

Paléographie et *deep-learning*

Introduction

L'HTR, pourquoi et pour qui ? Etat des lieux et principes généraux

Mettre en oeuvre l'HTR : développement d'une chaîne de traitement

Les limites de l'HTR et ses développement à venir

Bibliographie

Au delà de la transcription, ouvrir les données de la recherche

- Partager et mutualiser les modèles entrainés de transcription + mettre à disposition des vérités de terrain : cf. initiative HTR-United (`https://htr-united.github.io/`).
- Contexte des plans nationaux pour la Science Ouverte (2018, renouvellement en 2022) : ouverture des publications financées + obligation d'ouverture des données et des codes informatiques produits.

Paléographie et *deep-learning*

Introduction

L'HTR, pourquoi et pour qui ? Etat des lieux et principes généraux

Mettre en oeuvre l'HTR : développement d'une chaîne de traitement

Les limites de l'HTR et ses développement à venir

Bibliographie

Les coûts multiples de l'HTR et de l'IA appliquée aux SHS

- Coût en terme de puissance de calcul et enjeux de mutualisation des GPU.
- Coût humain et social : dangers des dérives du *digital labor* (Antonio Casilli).

Figure – A. Casilli, *En attendant les robots : Enquête sur le travail du clic*, Paris, Seuil, 2019

HTR et « lecture distante » : vers un biais de numérisation ?

Un renouvellement du *distant reading* (cf. Franco MORETTI, "Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850)", *Critical Inquiry*, 36-1 (2009), p. 134-158).

Cependant, biais de numérisation des corpus :

- Constitution et numérisation de corpus invisibilisent certaines thématiques, méthodes ou groupes socio-culturels (cf. Ian Milligan, "Illusionary Order : Online Databases, Optical Character Recognition, and Canadian History, 1997-2010", *Canadian Historical Review*, n° 94 (4), décembre 2013, p. 540-569).

- La numérisation ne remplace pas une campagne réfléchie et suivie d'archivage numérique.

En guise de conclusion :

- Déplacement du rôle du paléographe en amont du traitement des sources.
- Nouveau rôle de contrôle et entrainement des modèles.
- Un besoin de structuration institutionnelle (rôle de l'IRHT, du Cremma-Lab et de l'ENC ?) et de « décentralisation paléographique » ?

- CAMPS, Jean-Baptiste et PERREAUX, Nicolas, Reconnaissance optique des caractères et des écritures manuscrites - Projet E-NDP , [En ligne : https ://outils.lamop.fr/lamop/mp3/E-Ndp/JBC-NP e-NDP OCR-et-HTR.pdf].
- CHAGUÉ, Alix, CLÉRICE, Thibault et ROMARY, Laurent, HTR-United : Mutualisons la vérité de terrain ! , 2021, [En ligne : https ://hal.archives-ouvertes.fr/hal-03398740].
- CHAGUÉ, Alix, Comment faire lire des gribouillis à mon ordinateur ? , Tuto@Mate, 2021, [En ligne : https ://mate-shs.cnrs.fr/actions/tutomate/tuto31-lire-des-gribouillis-chague/].
- CHAGUÉ, Alix, Prendre en main eScriptorium , LECTAUREP, [En ligne : https ://lectaurep.hypotheses.org/documentation/prendre-en-main-escriptorium].
- CHAGUÉ, Alix et CHIFFOLEAU, Floriane, An accessible and transparent pipeline for publishing historical egodocuments , WPIP21 - What's Past is Prologue : The NewsEye International Conference, Virtual, Austria, 2021, [En ligne : https ://hal.archives-ouvertes.fr/hal-03173038].
- CHAGUÉ, Alix, CLÉRICE, Thibault et CHIFFOLEAU, Floriane, HTR-United, a centralization effort of HTR and OCR ground-truth repositories mainly for French languages, 2021, [En ligne : https ://github.com/HTR-United/htr-united]. DUVAL, Frédéric, Transcrire le français médiéval : de l' Instruction de Paul Meyer à la description linguistique contemporaine , Bibliothèque de l'École des chartes, vol. 170 / 2, Persée - Portail des revues scientifiques en SHS, 2012, p. 321-342.
- Gabay, S., Camps, J.-B., Pinche, A., and Jahan, C. (2021), SegmOnto : common vocabulary and practices for analysing the layout of manuscripts (and more), 16th International Conference on Document Analysisand Recognition (ICDAR 2021), Lausanne, Switzerland.
- Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G.(2017), "Transkribus - A Service Platform for Transcription, Recognition andRetrieval of Historical Documents", 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, volume 04, pages 19–24.
- KIESSLING, Benjamin, Kraken - an Universal Text Recognizer for the Humanities , Ultrecht, CLARIAH, 2019, [En ligne : https ://dev.clariah.nl/files/dh2019/boa/0673.html].
- KIESSLING, B., TISSOT, R., STOKES, P., [et al.], EScriptorium : An Open Source Platform for Historical Document Analysis , 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), vol. 2, 2019, p. 19.
- PINCHE, Ariane et CLÉRICE, Thibault, HTR-United/cremma-medieval : 1.0.1 Bicerin (DOI), Zenodo, 2021, [En ligne : https ://zenodo.org/record/5235186].
- PINCHE, Ariane, Projet CREMMALAB , CREMMALAB, [En ligne : https ://cremmalab.hypotheses.org/23].