

Des images au texte : comment apprendre à des ordinateurs à lire des manuscrits ?

A. Pinche

10 mai 2022



1 Introduction

- 2 Qu'est-ce que l'HTR, comment et pourquoi l'utiliser ?
- 3 Performances Kraken : Cas d'étude
- 4 Constituer et partager des modèles et des données d'entraînement
- 5 Présentation de Kraken et eScriptorium

Qu'est-ce que l'HTR



Figure – Prédiction HTR

- Prédiction d'un contenu texte
- à partir d'une image de la source par une
- intelligence artificielle entraînée par un humain
- dans un processus alternant
 - phases d'interventions humaines
 - et phases de calcul

Recherche et HTR

- Développement des outils mis à disposition : Transkribus, eScriotorium, Kraken
- Des projets pionniers : le projet Himanis (2015), le projet ANR Horae (2017) dirigés par Dominique Stutzmann .
- Une technologie qui fait partie des attendus :
 - Le projet ANR LiBeR
 - Le projet Biblissima+ a consacré un de ses clusters à cette problématique : cluster 3 "*Intelligence artificielle, reconnaissance de formes et d'écritures manuscrites*"

ENC et HTR

Initiative de recherche collective (INRIA, LAMOP, EPHE, IRHT, DIM MAP) et d'infrastructure :

- CREMMA
- CREMMALab.

Ces projets ont déjà permis :

- La mise à disposition de données pour entraîner un modèle d'HTR pour les manuscrits médiévaux entre le 13^e et le 15^e siècle (cremma-medieval)
- La mise à disposition du modèle Bicerin - accuracy de 95,49 % (CER).
- La mise à disposition des comptes-rendus des séances du séminaire : "Création de modèle(s) HTR pour les documents médiévaux en ancien français et moyen français, X^e-XV^e siècles".
- Organisation d'un colloque sur les documents anciens et la reconnaissance automatique des écritures manuscrites les 23 et 24 juin 2022, le programme est disponible au lien suivant :
<https://cremmalab.hypotheses.org/colloque-htr-programme>

1 Introduction

2 Qu'est-ce que l'HTR, comment et pourquoi l'utiliser ?

3 Performances Kraken : Cas d'étude

4 Constituer et partager des modèles et des données d'entraînement

5 Présentation de Kraken et eScriptorium

Différences entre OCR et HTR

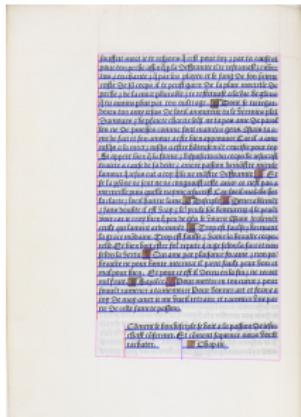
OCR	HTR
Performance : Taux d'erreur sur les caractères inférieur à 2 %, fonctionne uniquement sur les documents imprimés	Performance : Taux d'erreur sur les caractères entre 5 et 10 %, fonctionne sur les documents manuscrits
Outils : Abby (adobe), mais commercial, pas de code ouvert ; Tesseract 4 (gratuit,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

Mise en place d'une chaîne de reconnaissance automatique du texte

Étape 1 : Segmenter



A. Pinche



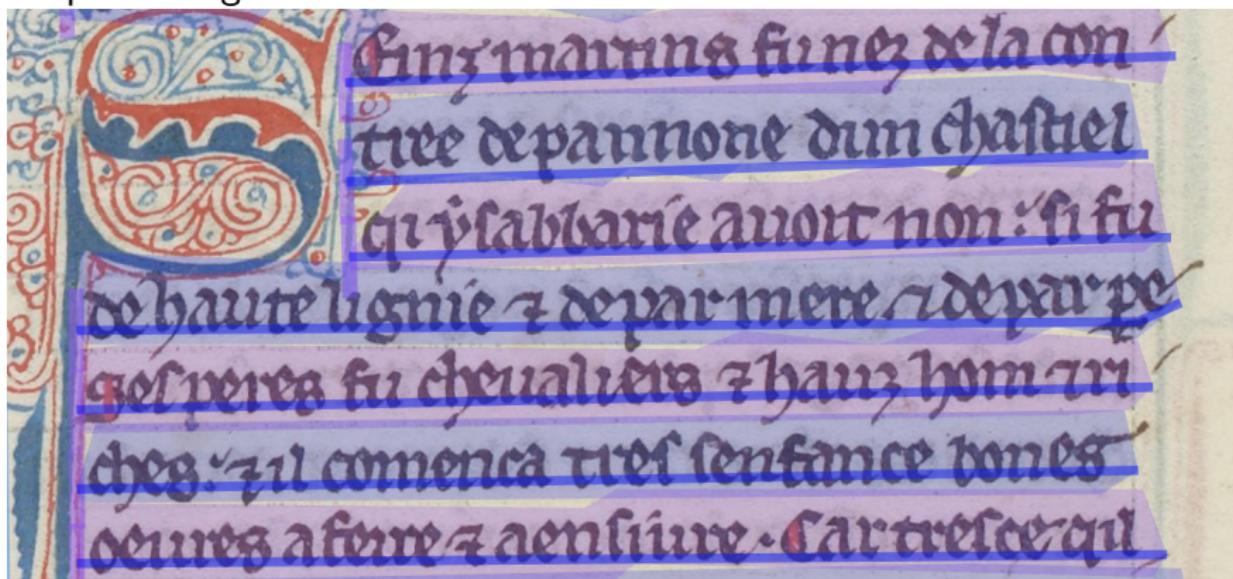
HTR et mss. médiévaux



10 mai 2022

Mise en place d'une chaîne de reconnaissance automatique du texte

Étape 1 : Segmenter



Mise en place d'une chaîne de reconnaissance automatique du texte

Étape 2 : Transcrire

The screenshot shows a digital transcription interface. At the top, there are three small icons: a blue square with a white circle, a red square with a white circle, and a green square with a white circle. To their right is the text "Line #1". On the far right, there is a close button "x" and a "My IP" button.

The main area displays a photograph of a medieval manuscript page with text written in a Gothic script. Below this is a transcription window containing the text "A ncois quil fussent bien volāt".

Below the transcription window is a keyboard labeled "Centralab" with a dropdown arrow. To its right is a "Change keyboards" button. To the left of the keyboard, there is a vertical list of names: "A r", "L es", "C on", "D ou", "C elu", "L e r", and "A me".

At the bottom of the interface, there is a footer with the text "celui qui paraît victorieux" followed by several small navigation icons.

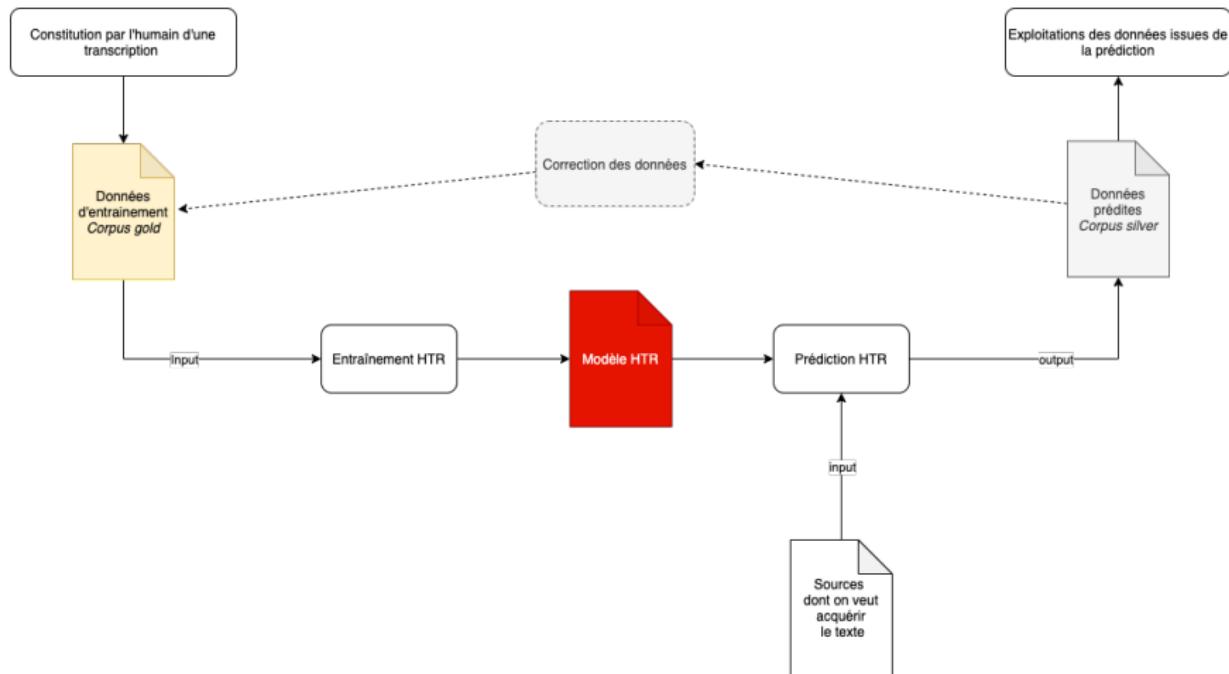
	lir	P orroit veoir & esgardeir
L es	ir	L e urai soleil q raire cleir
C on	it	C est ih'c ēs li douz li pis
D ou	it	Qj en lui a son esgart mis
C elu	gn	E nlesgardeir se renouele
L e r		A utresi com fait li oisele
A me		

Mise en place d'une chaîne de reconnaissance automatique du texte

Étape 2 : Transcrire

- Utiliser un modèle préexistant (voir HTR-united) ;
- Personnaliser (*fine tuner*) un modèle préexistant.
- Transcrire un corpus à la main pour entraîner un nouveau modèle ;

Comment entraîner et utiliser un modèle HTR ?



Pourquoi utiliser ou créer modèle pour l'HTR ?

- Un modèle performant existe déjà. Où le trouver ? Où trouver les données pour l'entraîner ? HTR-united
- Pour accélérer la phase d'acquisition du texte. La prédiction peut servir :
 - de base à une édition : niveau de précision haut, supérieur à 95 % d'*accuracy*
 - à de la mise à disposition de texte brut : niveau de précision moyen, entre 90 % et 95 %
 - de base à des analyses quantitatives : niveau de précision faible, supérieur à 80 % (voir EDER, Maciej, « Mind your corpus : systematic errors in authorship attribution », *Literary and Linguistic Computing*, vol. 28 / 4, décembre 2013, p. 603-614.)

1 Introduction

2 Qu'est-ce que l'HTR, comment et pourquoi l'utiliser ?

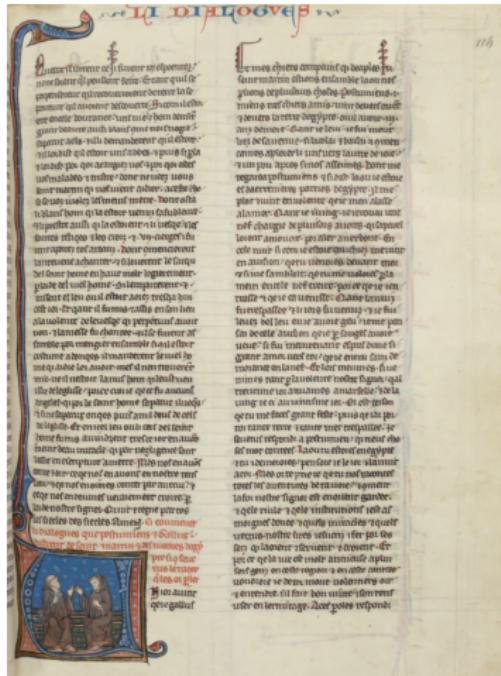
3 Performances Kraken : Cas d'étude

4 Constituer et partager des modèles et des données d'entraînement

5 Présentation de Kraken et eScriptorium

Performances Kraken : Cas d'étude

Modèle *saintMartin* entraîné sur le manuscrit BnF, fr. 412



A. Pinche

Performances Kraken : Cas d'étude

Modèle *saintMartin* entraîné sur le manuscrit BnF, fr. 412

- Train : 10 folios - soit 1680 lignes transcrrites
- Accuracy : 95.38% sur une même main

Type d'erreur	Nombre total	Tx d'erreurs/ligne
Insertions	1 883	0.76
Délations	725	0.29
Substitutions	1 823	0.73

Performances Kraken : Cas d'étude

Table des erreurs les plus fréquentes

Nb d'erreurs	Vérité de terrain	Prévision
762	[SPACE]	[]
473	[]	[SPACE]
162	[i]	[]
77	[.]	[]
73	[n]	[]

Exemple de prédiction du modèle dans eScriptorium

The screenshot shows the eScriptorium application interface. At the top, it says "Line #3". Below that is a dark gray text area containing medieval Latin text in a Gothic script. The text reads: "donerent grant clarite. et grant lumiere. tres qarat. qeli seinz euesques fu enterrez. Nil langues ne porroient mie dire se ie les auoie". Below this text area, there is a white input field containing the predicted transcription: "qatant:qeli seinz euesques fu enterrez.Nil". At the bottom left, there is a small note: "by apinche (eScriptorium) on Tue Oct 19 2021 16:02:29 GMT+0200". At the very bottom, there are several small icons for navigating through the text.

Performances Kraken : Cas d'étude

Modèle *Cremma-medieval*

- entraîné sur onze manuscrits différents
- 18385 lignes transcrites
- modèle Bicerin :
 - 22629 Characters
 - 1020 erreurs
 - 95,49% d'accuracy, mais sur le corpus complet qui comporte des mains différentes et des manuscrits compris entre le 13^eet le 14^esiècle

Performances Kraken : Cas d'étude

Type d'erreur	Nombre total
Insertions	317
Délations	229
Substitutions	474

Table des erreurs les plus fréquentes

Nb d'erreurs	Vérité de terrain	Prévision
160	[SPACE]	[]
153	[]	[SPACE]
35	[u]	[v]
34	[i]	[]
14	[u]	[n]
73	[v]	[u]

Limite et amélioration du modèle

Le modèle Bicerin possède une élasticité limitée, il peut arriver qu'il atteigne ses limites, car le document proposé est trop différent de son corpus d'entraînement.

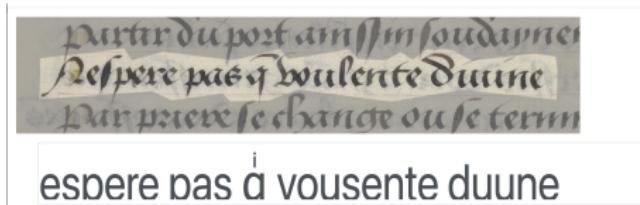


Figure – Prédiction Philadelphia, university of pennsylvania, ms codex 909 par Bicerin, accuracy : 80 %



Figure – Prédiction Chantilly, ms. 734 par Bicerin, accuracy : 83%

Limite et amélioration du modèle

Résolution du problème en personnalisa^tn un modèle existant voire 5 pages de la nouvelle source (environ 350 lignes de transcription)

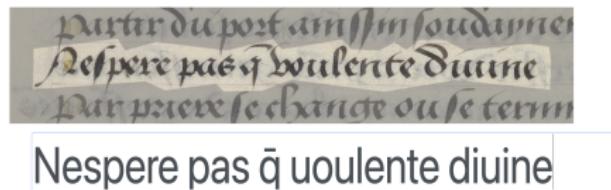


Figure – Prédiction Philadelphia, university of pennsylvania, ms codex 909 avec un modèle personnalisé à partir de Bicerin, accuracy : 97%



Figure – Prédiction Chantilly, ms. 734 avec un modèle personnalisé à partir de Bicerin, accuracy : 91%

- 1 Introduction
- 2 Qu'est-ce que l'HTR, comment et pourquoi l'utiliser ?
- 3 Performances Kraken : Cas d'étude
- 4 Constituer et partager des modèles et des données d'entraînement
- 5 Présentation de Kraken et eScriptorium

Constituer des données d'entraînement

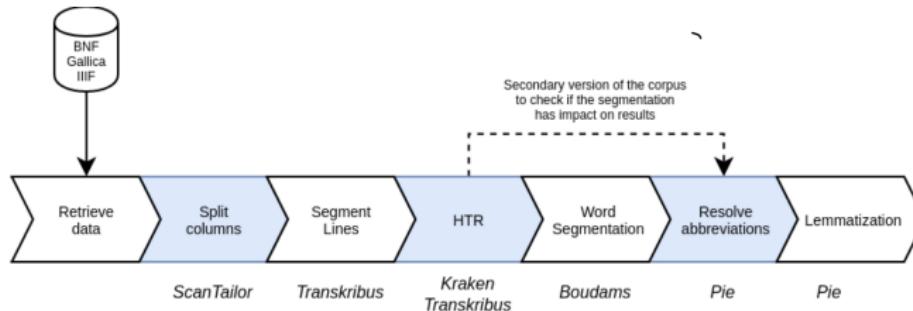
- Identifier les besoins :

- Produire un corpus lisible sur un document particulier ;
- Produire une transcription proche de la source avec le moins d'interprétation possible et la plus grande "élasticité" possible pour le modèle ;
- Gérer la tension entre le besoin de modèle(s) générique(s) et le besoin d'un modèle spécifique pour chaque projet.

Constituer des données d'entraînement

- Identifier les besoins :

- Mettre en place une chaîne de production textuelle avec une ou plusieurs étapes :
 - Voir 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 / Supplement₂, octobre2021, p. ii49-ii71.



Constituer des données d'entraînement

- L'harmonisation des données permettra d'échanger des données et des modèles HTR. Comment faire ?
 - Gérer la tension entre le besoin de modèle(s) générique(s) et le besoin d'un modèle spécifique pour chaque projet
 - Établir des référentiels et des pratiques communes.
 - Partager ses données et ses modèles.

Définir des normes de transcription

- Définir des méthodes de transcription adaptées à une problématique de recherche et à l'apprentissage machine.
- Définir le degré de précision recherché dans la transcription
- Utiliser un set de caractères prédéfini et documenter ses choix
- Exemple : voir les compte-rendus des séances du séminaire : "Création de modèle(s) HTR pour les documents médiévaux en ancien français et moyen français entre le Xe-XIVe siècle"

Définir des normes de transcription

Listes des caractères préconisés pour la transcription des abréviations dans les textes médiévaux français :

	Signe	code caractère
<i>Les tildes</i>		
COMBINING TILDE	˜	U+0303
COMBINING VERTICAL TILDE	ߑ	U+033E
<i>Les abréviations par lettres suscrites</i>		
COMBINING LATIN SMALL LETTER A	܂	U+0363
COMBINING LATIN SMALL LETTER C	܄	U+0368
COMBINING LATIN SMALL LETTER E	܅	U+0364
COMBINING LATIN SMALL LETTER I	܆	U+0365
COMBINING LATIN SMALL LETTER M	܈	U+036B
COMBINING LATIN SMALL LETTER N	܉	U+1DE0
COMBINING LATIN SMALL LETTER O	܊	U+0366
COMBINING LATIN SMALL LETTER R	܃	U+036C
COMBINING LATIN SMALL LETTER S	܄	U+1DE4
COMBINING LATIN SMALL LETTER T	܅	U+036D
COMBINING LATIN SMALL LETTER U	܇	U+0367
COMBINING LATIN SMALL LETTER X	܋	U+036F
COMBINING LATIN SMALL LETTER Y	܌	U+F03C
COMBINING LATIN SMALL LIGATURE AR ABOVE	܍	U+EFAA
COMBINING UR ABOVE	܎	U+1DD1
<i>Les abréviations par signes spéciaux</i>		
LATIN SMALL LETTER CON	܏	U+A76F
COMBINING US ABOVE	ܐ	U+1DD2
LATIN SMALL LETTER H WITH STROKE	ܑ	U+0127
LATIN SMALL LETTER L WITH STROKE	ܒ	U+0142
LATIN SMALL LETTER P WITH STROKE	ܓ	U+A751
LATIN SMALL LETTER P WITH FLOURISH	ܔ	U+A753
LATIN SMALL LETTER LONG S WITH DIAGONAL ST	ܕ	U+1E9C
TIRONIAN SIGN ET	ܖ	U+204A
DIVISION SIGN	ܗ	U+00F7
COMBINING LONG STROKE OVERLAY	ܘ	U+0336

Utiliser des ontologies

- Repérage des différentes zones du document : utiliser un vocabulaire contrôlé, comme SegmOnto.



Figure – Bnf, fr. 412, fol.10r

Créer et développer des sets de données partageables

- Décrire ses données et ses métadonnées : moteur HTR, interface de transcription, type d'écriture, volume des données, normes de transcription.
- Assurer un contrôle qualité
- Exemple : **HTR-united** et ses outils de contrôle qualité :
 - **HTRUC** permet de contrôler que le fichier avec les métadonnées du corpus comporte toutes les rubriques requises.
 - **ChocoMufin** contrôle les caractères qui sont utilisés dans la transcription.
 - **htrvx** vérifie le schéma XML et l'utilisation de l'ontologie Segmonto pour la segmentation
 - **htr-united-metadata-generator** génère des métadonnées, spécifiquement des métriques du corpus : nombre de documents, nombres de zones, de lignes transcrrites, des caractères.

- 1 Introduction
- 2 Qu'est-ce que l'HTR, comment et pourquoi l'utiliser ?
- 3 Performances Kraken : Cas d'étude
- 4 Constituer et partager des modèles et des données d'entraînement
- 5 Présentation de Kraken et eScriptorium

Kraken, moteur HTR



- Outil d'analyse de mise en page et d'HTR ;
- Fondé sur de l'apprentissage profond (IA) ;
- Développé par Ben Kiessling dans le projet Scripta (PSL) ;
- Module Python, <https://github.com/mittagessen/kraken> ;
- Doc : kraken.re.

eScriptorium, interface de transcription



A. Pinche

MARTIN

Auoit non lemoicina.Etli

dela fonteine coroit par les lardins par les
chans paries cortiz delacite detoutes
parz lafisoient corre lihome delacite par
conduit r'mout en auoient grantiole.

mes li deables qilor cuida toli sibeau con
duit esta la fonteine desonlu.■aporta en
une palu.lau on ne poit nule rien fere.
si qleas genz her poiont avoir aise.Cil
dou pais se merueillentmolt ■distret
qe ore estoient ilmot ilmot molt g;
doel.li corbill.i lardin.li chang.li conduit
sechierent tut.Einsi fu la fonteine.■

anz.Autierc auant un qnt de uns pelein
qz auoit reliques de saint climent.unt
en la cite de soleus mostrera l'espresa aun
prouesse molt prouidene.Qnt lagent
sorent ce si uidrent auprouoie.lesil
distren qnt creolent bien qnse il proit
nostr signor qz remetrot lafonteine
ensiu lly.Et il lor dist oratons. se portos
les reliques se etre sont de saint climent
si com nostre pereinie dist lez uertuz
lui aparront bien il uant la mist les
religies lou i fontaine deuict estre lese
coucha aorsisons dorez qz seint climent z
aut fontaine donez uersedes auseaus qui

point deue nauliontiers rendist lajour.
Si com il ot ce dictia fonteine reunit **co**
rut par tout la ou ele soloit.Lors rendi gra
ces toz li pueples anostre signor,qi par la
merite deson maryer **la plus priere** del preu
dome auoit ce fet.Cest miracle **rmolt** d'aut
fist dex par saint clement **saciez** quens elior
de sa feste,sont meint auugle rumeur,**rmolt**
malade,sane en son mestier,plagacie de le
sucrist,quit **regne** avec le pere **ele sei**
se en pit ou singulier des eis.

Puet len souent abien ueni

neſſel neblen nenten. Deben faire na
nul tenu. Mais le bens nest souüst li
bienſt. Del mai il mame ſicom del leſci
ture, por ce ſo de lant en au bieuſt
leſien feſte. Com il ſaint home firent
ca en arriereſte cui nos trouvons les oe
ues ſes ues eſcriptures. Et ſaint
tut cil qui uulent, ca ne lauron tant
deſeinſt entolles por ues eſcriptures
mort dont niale n'ien ſauve les poſt
dura au cui d'orſt pell aux
ſtai. Dex qe feront don cil qui riche ſont
aile ſe lauoir de cest siecle ne en eus
doucer ou humilité conſiderante, aile
ſont plus dangereux de traiſon, aile
forſte. Des grant auſſirica que com plus
ont numeroſes, aile ſe endormisſent
auco. Ce fel le deſtous qil en tel marie
re les a laicet aile qil les emmeine
infer legnat cheſtin plener deſce
ſe gardien li ſaint home qj par doloreſſes p
nes aile grize tormezz: aile
geuen ſar toutes boies ſeuves firent
tan quidourent ſale prenable aile
corone de gloire. Aice regarberant ſeit
confefſion. Aimes ſens merz: batinc

Einz martins fu nez de la con-
tree depannone dun chastiel

5 qj ysebarreiaut nonsi fu de haute lignie qj deparre mene p[er]d[re] p[er]d[re] Ses peres fu cheueux lezau hom 'n' ches qj commeca tres sentance bofes oeures aleire esmeure. Car trece quil auoit q[ue] x an[ns] comenza il aeler asen[ir] te glisse meugre sonpre q[ue] amere q[ue] pa[er] est enonci. Et q[ue] il un[ir] en laige de xil[an]z il couulta molt amener uie solitaire q[ue] amur en hermige q[ue] eust il fait mes laisounne delage len de tint etractat de ce q[ue] il souuent folie en sa pensie. Et ne por q[ue] ant en s'iouence ful si entretir ente seirrie yglise q[ue] loprence en sentance ce q[ue] aempli mout docement q[ue] ant

eScriptorium, interface de transcription

- Logiciel libre qui permet de segmenter un document, de détecter les lignes, de transcrire, d'entraîner un modèle HTR et de l'appliquer à ses sources
- Développé dans le cadre du projet Scripta (PSL) ;
- Se branche sur Kraken pour l'analyse de la mise en page et la phase HTR ;
- Cette interface peut être utilisée pour créer un corpus d'entraînement ou corriger des prédictions HTR.
- Code : <https://gitlab.inria.fr/scripta/escriptorium> ;
- démos vidéos : <https://scripta.hypotheses.org/escriptorium-video-gallery>.

eScriptorium

Utiliser eScriptorium nécessite :

- d'ouvrir un compte sur une instance d'eScriptorium (auprès de l'INRIA ou de l'ENC) ou d'installer une instance locale d'eScriptorium
- d'avoir accès aux fichiers images de ses sources : des fichiers locaux ou téléchargés directement depuis un site institutionnel en utilisant un manifeste IIIF :

<https://gallica.bnf.fr/iiif/ark:/12148/btv1b84259980/manifest.json>

eScriptorium

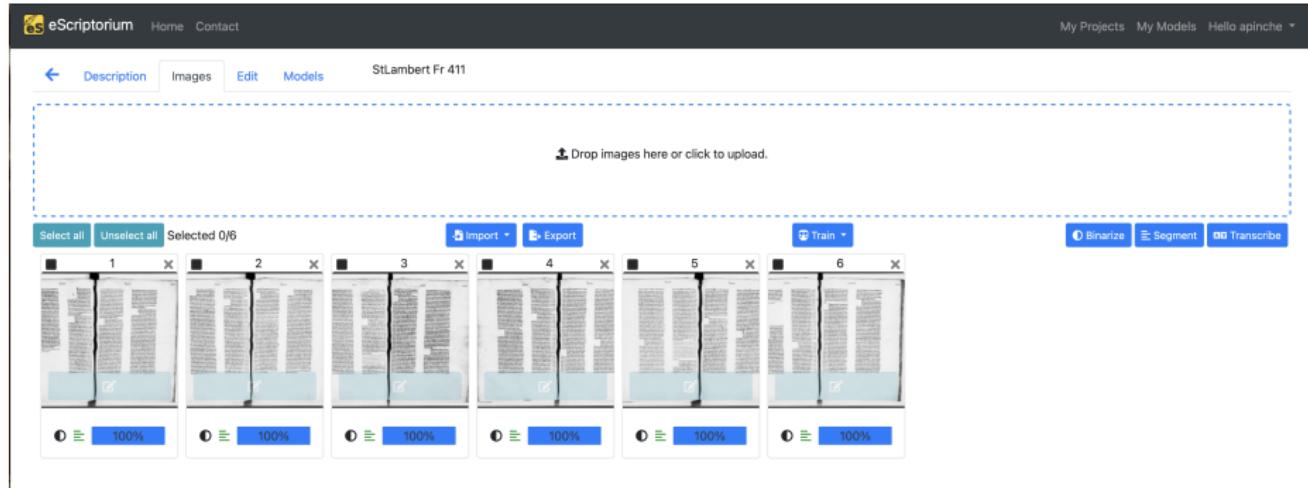


Figure – Interface d'eScriptorium

eScriptorium

eScriptorium est une interface web qui permet :

- de segmenter la page de votre manuscrit et détecte les lignes
- de transcrire des documents pour le corpus d'entraînement
- d'entraîner un modèle HTR
- d'appliquer un modèle de HTR ou de segmentation à un document

eScriptorium



MARTIN

Figure – Segmentation et transcription d'un document à l'aide d'eScriptorium

Bibliographie

- 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-SCRIPTORIUM>].
- 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 Analysis and Recognition (ICDAR 2021)*, Lausanne, Switzerland.
- Kahle, P., Colutto, S., Hackl, G., and Mühlberger, G. (2017), "Transkribus - A Service Platform for Transcription, Recognition and Retrieval 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 », Utrecht, 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>].