



**HAL**  
open science

## Atelier HTR : eScriptorium

Ariane Pinche

► **To cite this version:**

Ariane Pinche. Atelier HTR : eScriptorium. Master. Université d'Avignon, France. 2023. hal-04309133

**HAL Id: hal-04309133**

**<https://hal.science/hal-04309133>**

Submitted on 27 Nov 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Atelier HTR : eScriptorium

Ariane Pinche <sup>1</sup>

<sup>1</sup>CNRS, <sup>2</sup>CIHAM ■ UMR 5648

30 novembre 2023, Université d'Avignon

# Table of Contents

- 1 ATR : définition
- 2 Les étapes de l'HTR
- 3 Présentation de Kraken et eScriptorium
- 4 Pour aller plus loin...
- 5 Références

# Qu'est-que l'ATR ?



Figure: Prédiction HTR

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



# Pourquoi utiliser ou créer modèle pour l'HTR ?

- 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.)

# Table of Contents

- 1 ATR : définition
- 2 Les étapes de l'HTR**
- 3 Présentation de Kraken et eScriptorium
- 4 Pour aller plus loin...
- 5 Références

- Chargement des images
  - ▶ Chargement d'une collection d'image en JPG ou tif en local
  - ▶ Chargement depuis un manifeste iiif (e.g collections issues de Gallica ou de e-Codices)
- Traitement des images (facultatif)
  - ▶ résolution 300dpi
  - ▶ couleur ou niveau de gris
  - ▶ possible binarisation pour réduire le bruit
  - ▶ imagerie multispectrale (dans le cas de documents très abimés)

# Les étapes de l'HTR

- Segmentation des zones de l'image



Figure: Bnf, fr. 412, fol.10r

# Les étapes de l'HTR

- Segmentation des lignes contenant du texte

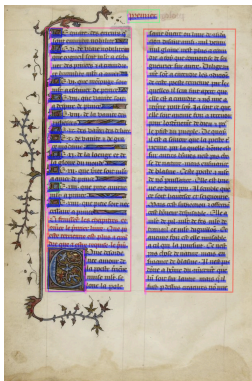
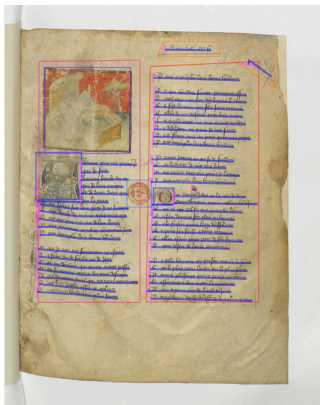


Figure: Bnf, fr. 1728, fol. 8v

# Les étapes de l'HTR

- Prédiction du texte qui se trouve sur l'image



1 ucunes gens ne prient ¶  
2 que le face  
3 Aucuns beaux diz et  
4 que le leur enuoye  
5 Et de ditter dient que  
6 A  
7 Mais sauue soit leur paix le ne sauroye  
8 lay la grace  
9 Faire beaux diz ne bons, mais touteuoye  
10 Puis que prie men ont de leur bonte  
11 Paine y mettray combien qu'ilgnoit soy  
12 Pour accomplir leur bonne uolente  
13 Mais le n ay pas seulement ne espace  
14 De faire diz, de soules ne de loye  
15 Car ma doulour qui toutes autres paise  
16 Mon sentiment loieux du tout desuoye  
17 Mais du grant quel qui me tient morne ¶coye  
18 Puis bien parler assez et aplante  
19 Si en diray uolentiers plus feroye  
20 6259  
21 Pour accomplir leur bonne uolente  
22 Et qui uouldra sauoir pourquoy efface  
23 Duel tout mon bien, uolentiers le droye  
24 Ce list la mort qui ferl sans merace  
25 Cellui de qui trestout mon bien auoye  
26 Laquelle mort ma mis et met en uoye  
27 De desespoir ne puis le nos sante  
28 De ce feray mes diz puis qu'on men proye  
29 Pour accomplir leur bonne uolente  
30 Princes prenez en gre se le failloye  
31 Car le ditter le ray mie hante  
32 Mais maint men ont pria ¶ le lottroye  
33 Pour accomplir leur bonne uolente  
34 u temps ladis en cite de Rome  
35 il.  
36 O  
37 ung en yot, Tel fu que quant un home  
38 Orent Rômainz maint noble ¶ bel usaige

Figure: Bnf, fr. 12779, fol.9r

- Export des données (txt, alto, page)

```
<Layout>
<Page WIDTH="4648" HEIGHT="3407" PHYSICAL_IMG_NR="8" ID="eSc_dummypage_">
  <PrintSpace HPOS="0" VPOS="0" WIDTH="4648" HEIGHT="3407">
    <TextBlock HPOS="693" VPOS="321" WIDTH="1701" HEIGHT="2451"
      ID="eSc_textblock_08b9f915" TAGREFS="BT3852">
      <Shape>
        <Polygon
          POINTS="693 413 693 2772 2394 2772 2254 321"/>
      </Shape>
      <TextLine ID="eSc_line_d939596f" TAGREFS="LT1299"
        BASELINE="746 476 2143 428" HPOS="743" VPOS="352"
        WIDTH="1400" HEIGHT="156">
        <Shape>
          <Polygon
            POINTS="2078 388 2050 388 2021 386 1993 383 1964 383 1936 380 1908 377 1876 374 1848 374 1820 371 1811
            />
          </Shape>
          <String
            CONTENT="fors de la uille. Tant fut lassault merueilleux et"
            HPOS="743" VPOS="352" WIDTH="1400" HEIGHT="156"/>
        </TextLine>
  </TextBlock>
</Page>
</Layout>
```

Figure: Exemple de fichier Alto

# Table of Contents

- 1 ATR : définition
- 2 Les étapes de l'HTR
- 3 Présentation de Kraken et eScriptorium**
- 4 Pour aller plus loin...
- 5 Références





## Kraken

- Outil d'analyse de mise en page et d'HTR
- fondé sur de l'apprentissage profond
- développé par Ben Kiessling dans le projet Scripta (PSL);
- Module Python, <https://github.com/mittagessen/kraken>;
- Doc: <https://kraken.re/master/index.html>
- Il peut être utilisé directement en ligne de commande ou via l'interface d'eScriptorium

## eScriptorium

- 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 mise en page et HTR;
- nécessite d'être déployé sur un serveur par une institution;
- Code: <https://gitlab.inria.fr/scripta/escriptorium>;
- Video demos: <https://escripta.hypotheses.org/escriptorium-video-gallery>.
- Tutoriel en ligne <https://openiti.org/assets/documents/eScriptorium-Tutorial.pdf>

Utiliser eScriptorium nécessite :

- D'ouvrir un compte sur une instance d'eScriptorium 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>

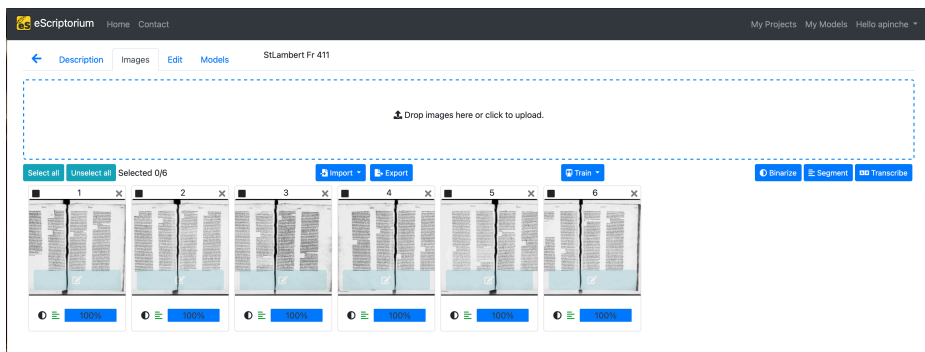


Figure: Interface d'eScriptorium

*eScriptorium* est une interface web qui permet :

- de segmenter la page d'un document (zones et lignes)
- de transcrire des documents
- d'entraîner un modèle HTR
- d'appliquer un modèle de HTR ou de segmentation à un document

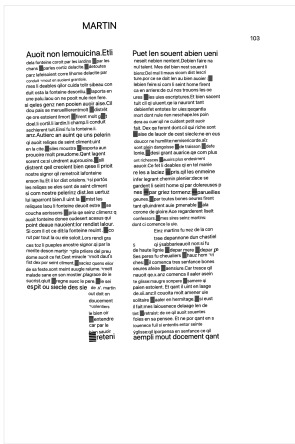
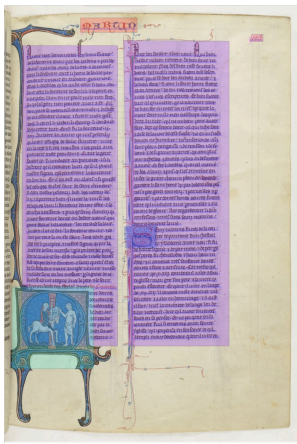


Figure: Segmentation et transcription d'un document à l'aide d'eScriptorium

# Exercice

- 1 S'identifier sur le serveur: <https://traces6.paris.inria.fr>
  - ▶ login : apinche\_formation - mdp : training1234
- 2 Créer un nouveau projet, puis un nouveau document
- 3 Importer des images en JPEG
  - ▶ Télécharger depuis Gallica 2 images :
    - ★ Bibliothèque nationale de France. Département des Manuscrits. NAF 23686, (Vies de saint en français, <https://gallica.bnf.fr/ark:/12148/btv1b8446925z>)
    - ★ Bibliothèque nationale de France. Département des Manuscrits, NAL 775 (latin, legenda aurea), <https://gallica.bnf.fr/ark:/12148/btv1b52509205f/f5.item>)
    - ★ Bibliothèque municipale de Metz. Ms. 399v <https://gallica.bnf.fr/ark:/12148/btv1b10557035g/f5.item>)
    - ★ Bibliothèque nationale de France. Département des Manuscrits. Latin 5335 (latin, Vita Sancti Martini), <https://gallica.bnf.fr/ark:/12148/btv1b8529515t/f14.item>)
    - ★ Bibliothèque nationale de France. Département Arts du spectacle, RESERVE 4-RF-87512 (latin, legenda aurea), <https://gallica.bnf.fr/ark:/12148/bpt6k1281158q>)

# Exercice : Testaments de Poilus

- S'identifier sur le serveur: <https://escriptorium.inria.fr>
  - ▶ login : apinche\_formation - mdp : training1234
- Ouvrir un nouveau projet
- Charger les images de votre choix, parmi le corpus de testament de poilus : <https://github.com/HTR-United/CREMMA-AN-TestamentDePoilus>
- Appliquer des modèles de transcription et choisir le plus adapté.
- Établissez vos normes de transcription pour la constitution d'un corpus d'entraînement
- Faites les corrections nécessaires



# Exercice : Légendier de saint Pétersbourg

- 1 S'identifier sur le serveur: <https://escriptorium.inria.fr>
  - ▶ login : `apinche_formation` - mdp : `training1234`
- 2 Ouvrir un nouveau projet, créer un nouveau document
- 3 Téléchargez deux images de votre choix à l'adresse suivante :  
<https://gallica.bnf.fr/ark:/12148/btv1b8446925z>
- 4 Appliquer des modèles de transcription et choisir le plus adapté.
- 5 Établissez vos normes de transcription pour la constitution d'un corpus d'entraînement
- 6 Faites les corrections nécessaires.
  - ▶ L'édition de la *Vie de saint Benoit* est disponible au lien suivant :  
[https://scd-resnum.univ-lyon3.fr/out/theses/2021\\_out\\_pinche\\_a.pdf](https://scd-resnum.univ-lyon3.fr/out/theses/2021_out_pinche_a.pdf) (p.494).
  - ▶ Un clavier avec les principaux caractères spéciaux du français médiéval est disponible au lien suivant : <https://github.com/HTR-United/cremma-medieval/blob/main/CremmaLab.json>

# Table of Contents

- 1 ATR : définition
- 2 Les étapes de l'HTR
- 3 Présentation de Kraken et eScriptorium
- 4 Pour aller plus loin...**
- 5 Références

# Pour aller plus loin...

- 1 Installer une instance eScriptorium:  
<https://gitlab.com/scripta/escriptorium/> - Video sur le sujet : <https://www.canal-u.tv/chaines/enc/21-fondue-a-lightweight-htr-infrastructure-for-geneva>
- 2 Utiliser Kraken en ligne de commande :  
<https://github.com/mittagessen/kraken>
  - ▶ `ketos train -r 0.0001 -lag 20 -s '[1,120,0,1 Cr3,13,32 Do0.1,2 Mp2,2 Cr3,13,32 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 S1(1x0)1,3 Lbx200 Do0.1,2 Lbx200 Do.1,2 Lbx200 Do]' -augment -device cuda:0 -preload -f alto -e val.txt -t train.txt --batch-size 16 -u NFC -o modelCremma-Medieval`
- 3 Améliorer l'analyse de la mise en page en entraînant un modèle avec YALTAI : <https://github.com/PonteIneptique/YALTAi>, CLÉRICE, Thibault, « You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine », 2022, [En ligne : <https://hal-enc.archives-ouvertes.fr/hal-03723208>].

# Table of Contents

- 1 ATR : définition
- 2 Les étapes de l'HTR
- 3 Présentation de Kraken et eScriptorium
- 4 Pour aller plus loin...
- 5 Références**

- 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://htr-unique.github.io/index.html>].
- Thibault Clérice, « You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine », 2022.// CLÉRICE, Thibault et PINCHE, Ariane, *Choco-Mufin, a tool for controlling characters used in OCR and HTR projects*, 2021, [En ligne: <https://github.com/Pontelneptique/choco-mufin>].
- CLÉRICE, Thibault et PINCHE, Ariane, *HTRVX, HTR Validation with XSD*, 2021, [En ligne: <https://github.com/HTR-United/HTRVX>].
- GABAY, Simon, CAMPS, Jean-Baptiste, PINCHE, Ariane, [et al.], « 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, 2021, [En ligne: <https://hal.archives-ouvertes.fr/hal-03336528>].
- GABAY, Simon, PINCHE, Ariane, LEROY, Noé, [et al.], « Données HTR incunables du 15e siècle », eds. Alix Chagué et Thibault Clérice, *HTR United*, 2022 [En ligne: <https://github.com/Gallicorpora/HTR-incunable-15e-siecle>].
- HODEL, Tobias, SCHOCH, David, SCHNEIDER, Christa, [et al.], « General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example », *Journal of Open Humanities Data*, vol. 7 / 0, Ubiquity Press, juillet 2021, p. 13.
- 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. 19-19.
- PINCHE, Ariane, « CREMMA Medieval, an Old French dataset for HTR and segmentation », eds. Alix Chagué et Thibault Clérice, *HTR United*, 2021, [En ligne: <https://github.com/HTR-United/cremma-medieval>].
- PINCHE, Ariane, *CREMMALAB | Constitution de corpus en ancien français pour l'HTR*, 2022, [En ligne: <https://cremmalab.hypotheses.org/>].
- WILLS, Tarrin, « The Medieval Unicode Font Initiative », *Medieval Unicode Font Initiative*, février 2016, [En ligne: <https://skaldic.org//m.php>].