Atelier HTR : eScriptorium

Ariane Pinche

Atelier HTR : eScriptorium

Ariane Pinche [1]

[1]CNRS, [2]CIHAM ▪ UMR 5648

30 novembre 2023, Université d'Avignon

Table of Contents

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

Les étapes de l'HTR

- 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

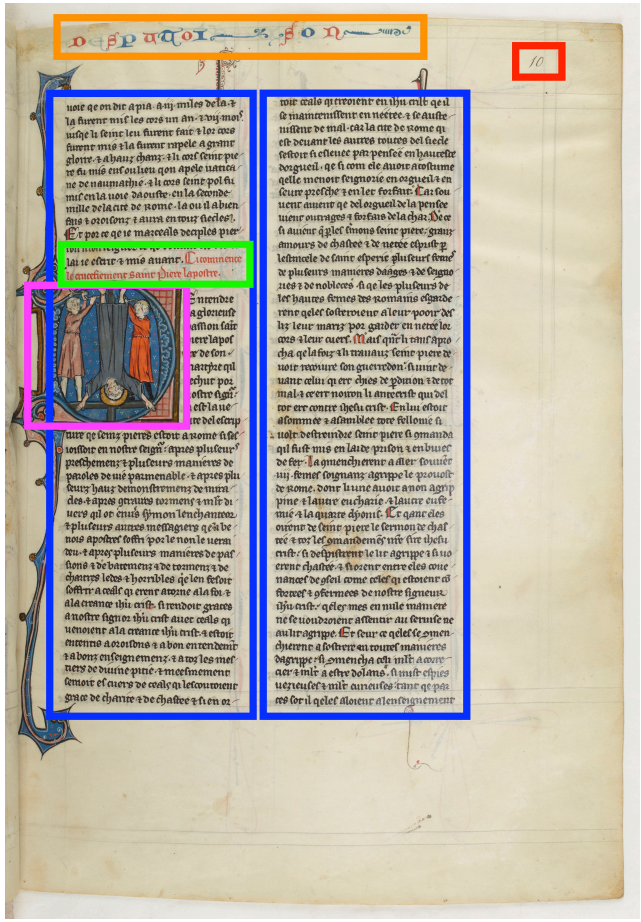


Figure: Bnf, fr. 412, fol.10r

Les étapes de l'HTR

- Segmentation des lignes contenant du texte

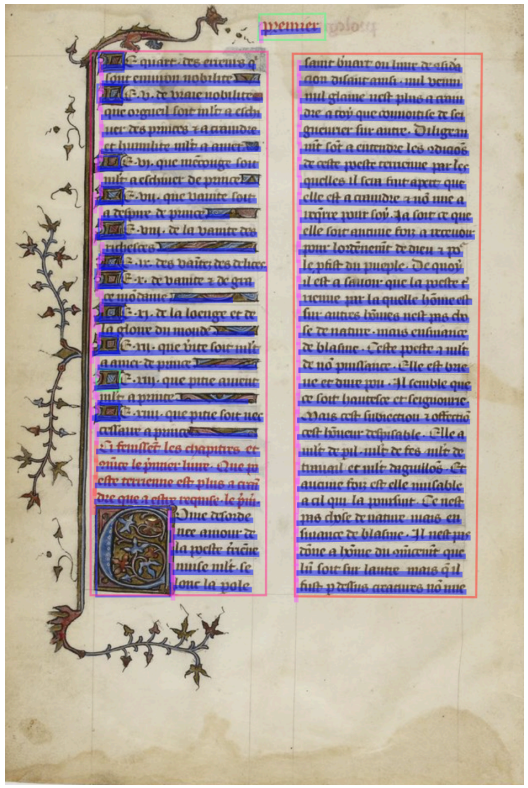
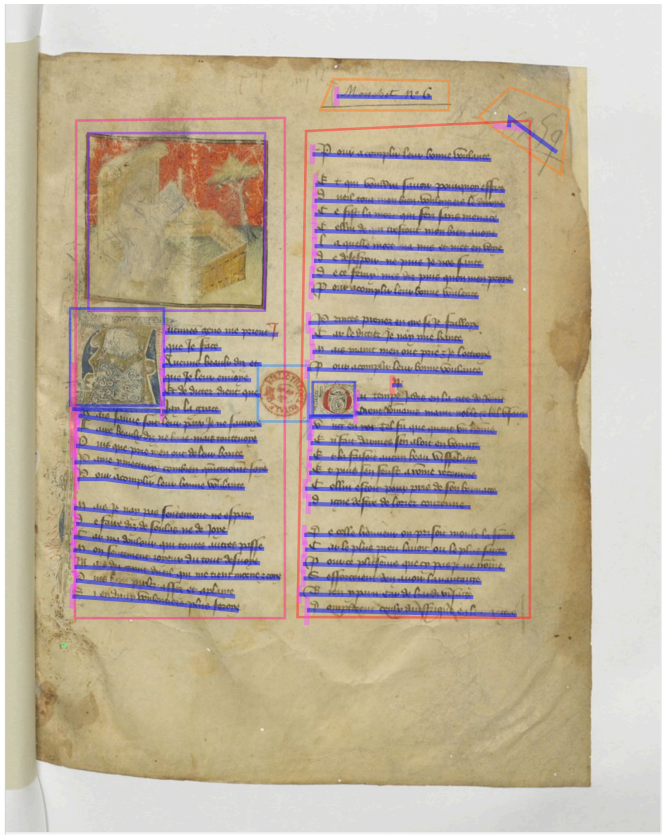


Figure: Bnf, fr. 1728, fol.8v

Les étapes de l'HTR

- Prédiction du texte qui se trouve sur l'image



```
ucunes gens me prient ꝑ.
que le face
Aucuns beaulx diz et
que le leur enuoye
Et de ditter dient que
A
Mais sauue soit leur paix le ne sauroye
lay la grace
Faire beaulx diz ne bons. mais touteuoye
Puis que prie men ont de leur bonte
Paine y mettray combien qũgnouĩt soye
Pour acomplir leur bonne uoulante
Mais le n ay pas sentement ne espace
De faire diz. de soulas ne de loye
Car ma doulour qui toutes autres passe
Mon sentement ioyeux du tout desuoye
Mais du grant dueil qui me tient morne ꝯcoye
Puis bien parler assez et aplante
Si en diray uoulentiers plus feroye
6259
Pour acomplir leur bonne uoulante
Et qui uouldra sauoir pourquoy efface
Dueil tout mon bien. uoulentiers le diroye
Ce fist la mort qui feri sans menace
Cellui de qui trestout mon bien auoye
Laquelle mort ma mis et met en uoye
De desespoir ne puis le nos sante
De ce feray mes diz puis quon men proye
Pour acomplir leur bonne uoulente
Princes prenez en gre se le failloye
Car le dittier le nay mie hante
Mais maint men ont prie ꝯ le lottroye
Pour acomplir leur bonne uoulante
u temps ladis en le cite de Rõme
.ii.
O
ung en yot. Tel fu que quant un hõme
Orent Rõmains maint noble ꝯ bel usaige
```

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

Les étapes de l'HTR

- 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>
```

Figure: Exemple de fichier Alto

Table of Contents

Kraken

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`

eScriptorium

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

eScriptorium

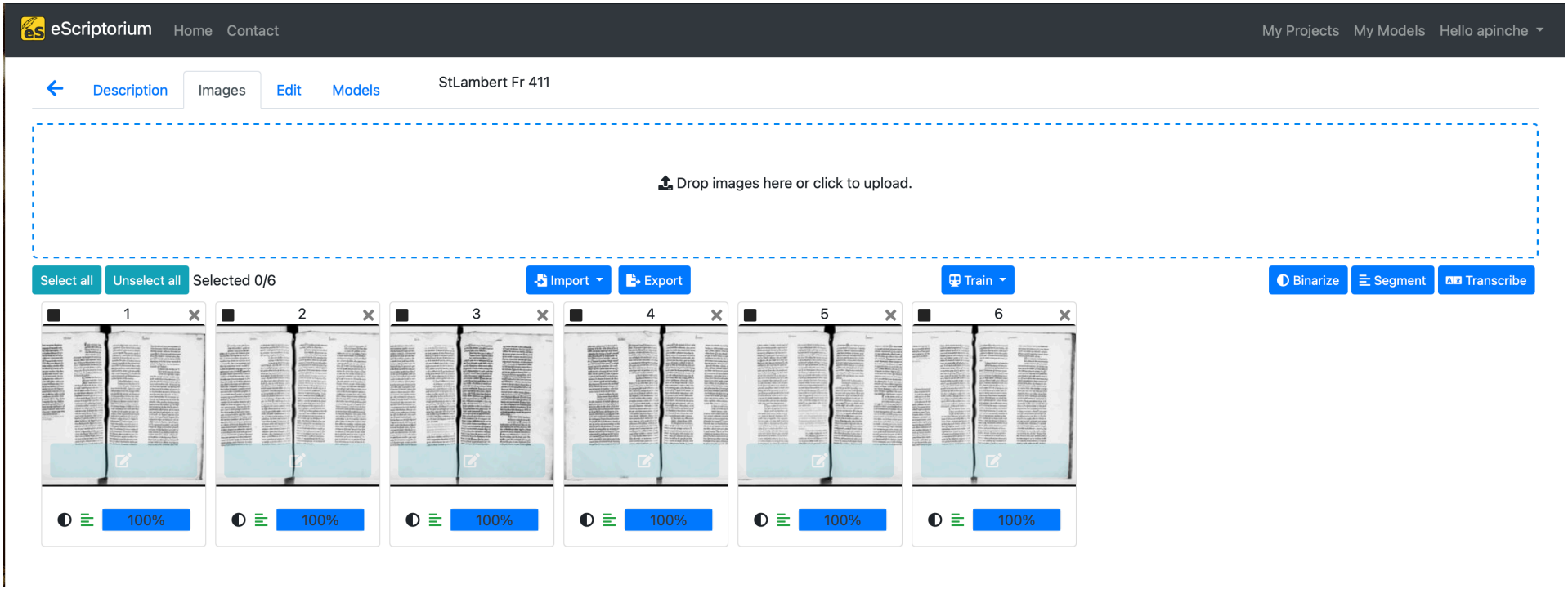


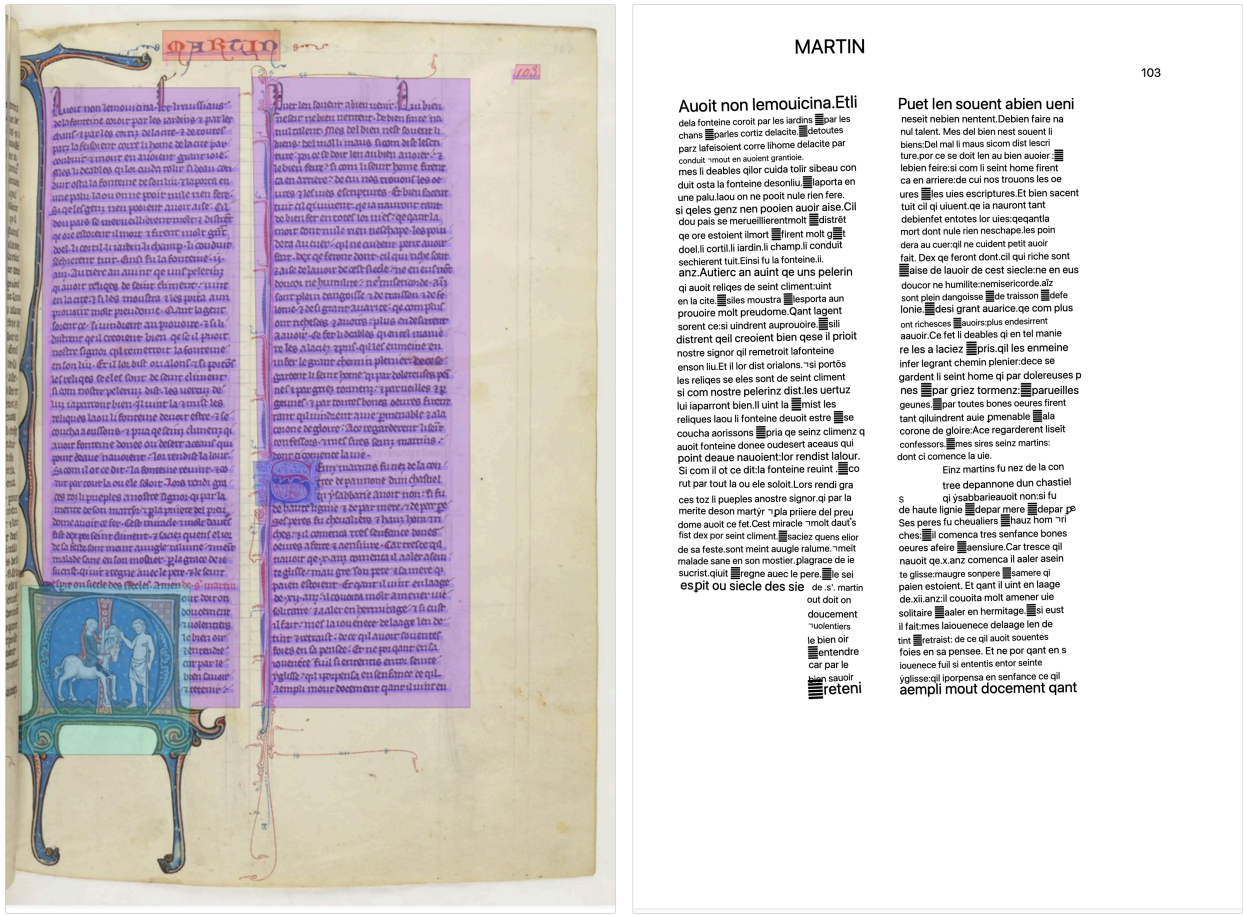
Figure: Interface d'eScriptorium

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

eScriptorium



MARTIN

103

Auoit non lemouicina.Etli
dela fonteine coroit par les iardins par les
chans parles cortiz delacite.detoutes
parz lafeisoient corre lihome delacite par
conduit mout en auoient grantioie.
mes li deables qilor cuida tolir sibeau con
duit osta la fonteine desonliu laporta en
une palu.laou on ne pooit nule rien fere.
si qeles genz nen pooien auoir aise.Cil
dou pais se merueillierentmolt distrêt
qe ore estoient ilmort firent molt o et
doel.li cortil.li iardin.li champ.li conduit
sechierent tuit.Einsi fu la fonteine.ii.
anz.Autierc an auint qe uns pelerin
qi auoit reliqes de seint climent:uint
en la cite.siles moustra lesporta aun
prouoire molt preudome.Qant lagent
sorent ce:si uindrent auprouoire.sili
distrent qeil creoient bien qese il prioit
nostre signor qil remetroit lafonteine
enson liu.Et il lor dist orialons.si portôs
les reliqes se eles sont de seint climent
si com nostre pelerinz dist.les uertuz
lui iaparront bien.Il uint la mist les
reliques laou li fonteine deuoit estre se
coucha aorissons pria qe seinz climenz q
auoit fonteine donee oudesert aceaus qui
point deaue nauoient:lor rendist lalour.
Si com il ot ce dit:la fonteine reuint .co
rut par tout la ou ele soloit.Lors rendi gra
ces toz li pueples anostre signor.qi par la
merite deson martyr pla priiere del preu
dome auoit ce fet.Cest miracle molt daut's
fist dex por seint climent saciez quens elior
de sa feste.sont meint augle ralume.meilt
malade sane en son mostier. plagrace de ie
sucrist.qiuit regne auec le pere.le sei
espit ou siecle des sie

Puet len souent abien ueni
neseit nebien nentent.Debien faire na
nul talent. Mes del bien nest souent li
biens:Del mal li maus sicom dist lescri
ture.por ce se doit len au bien auoier :
lebien feire:si com il seint home firent
ca en arriere:de cui nos trouons les oe
ures les uies escriptures.Et bien sacent
tuit cil qi uiuent.qe ia nauront tant
debienfet entotes lor uies:qeqantla
mort dont nule rien neschape.les poin
dera au cuer:qil ne cuident petit auoir
fait. Dex qe feront dont.cil qui riche sont
aise de lauoir de cest siecle:ne en eus
doucor ne humilite:nemisericorde.alz
sont plein dangoisse de traisson defe
lonie.desi grant auarice.qe com plus
ont richesces auoirs:plus endesirent
aauoir.Ce fet li deables qi en tel manie
re les a laciez pris.qil les enmeine
infer legrant chemin plenier:dece se
gardent li seint home qi par dolereuses p
nes par griez tormenz:parueilles
geunes.par toutes bones oeures firent
tant qiluindrent auie pmenable ala
corone de gloire:Ace regarderent liseit
confessors mes sires seinz martins:
dont ci comence la uie.

Einz martins fu nez de la con
tree depannone dun chastiel
qi ysabbarieauoit non:si fu
de haute lignie depar mere depar p
Ses peres fu cheualiers hauz hom ri
ches:il comenca tres senfance bones
oeures afeire aensiure.Car tresce qil
nauoit qe.x.anz comenca il aaler asein
te glisse:maugre sonpere samere qi
paien estoient. Et qant il uint en laage
de.xii.anz:il couoita molt amener uie
solitaire aaler en hermitage si eust
il fait:mes laiouenece delaage len de
tint retraist: de ce qil auoit souentes
foies en sa pensee. Et ne por qant en s
iouenece fuil si ententis entor seinte
yglisse:qil iporpensa en senfance ce qil
aempli mout docement qant

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. 399vhttps://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` (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

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

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-united.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/PonteIneptique/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 », éds. 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 », éds. 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].