

Reconnaissance automatique de caractères avec l'OCR et l'HTR

12 mars 2026

Denisa-Florina BUMBA (IGE, projet ERC Philiumm, CNRS, Université Paris Cité)
denisa-florina.bumba@cnr.fr

Marine TIGER (Cellf, CERES)
marine.tiger@sorbonne-universite.fr



Définitions

- ***Optical Character Recognition (OCR)*** = reconnaissance de caractères pour les imprimés
- ***Handwritten Text Recognition (HTR)*** = reconnaissance de texte manuscrit
- ... mais aussi *Handwritten Mathematical Expression Recognition (HMER)*
- ***Automatic Text Recognition (ATR)*** = englobe les outils de reconnaissance automatique de texte
- Exemple d'un projet HTR: FoNDUE à l'université de Genève
- Autres projets HTR avec des modèles entraînés sur des manuscrits médiévaux : [CREMMA](#) (8e-15e s.), [TRIDIS](#) (11e-16e s.)

Définitions

- Gain de temps
- Traiter de plus large et vaste ensemble de document
- Transcrire différents types de documents de périodes différentes
- Ajouter de la valeur aux documents historiques

Définitions

Différents défis dans la reconnaissance de documents historiques:

- Constituer un corpus suffisamment vaste pour entraîner un modèle:
<https://htr-united.github.io/catalog.html>
- Nettoyer le corpus: tâches d'encre, encre qui traverse la page, plusieurs mains, etc...

Préparer son corpus

Collecte des images

Qualité des images :

- formats images recommandés : JPG, JPEG, PNG
- 300 DPI (Dot per inch) : qualité optimale
- < 200 DPI : qualité insuffisante pour une bonne reconnaissance
- > 400 DPI : l'augmentation de la qualité devient négligeable par rapport au poids des fichiers

Manifeste IIIF et la BnF :

- Entièreté du document : eScriptorium
- Sélection d'une plage de téléchargement d'images :

scripts Python pour la récupération des images au format IIIF :
<https://github.com/Jean-Baptiste-Camps/IIIF-Crawler> (liens Gallica, e-codices, bvm, etc)
https://gitlab.com/eman8/Ecolired/travail_sur_les_images/-/tree/main/extract_IIIF (uniquement pour les liens Gallica)

ex. : <https://gallica.bnf.fr/ark:/12148/btv1b85775308/>

Prétraitement des images

- Amélioration du contraste du texte dans l'image
- Réduction du bruit (tâches d'encre, autre)
- Découpage des images

Précautions concernant la binarisation : dans certains cas ce traitement peut empirer les résultats particulièrement pour les documents avec une écriture pâle ou à un éclairage inégal.

scripts pour le prétraitement et le découpage des images :
https://gitlab.com/eman8/Ecolired/travail_sur_les_images/-/tree/main/image_preprocessing

https://gitlab.com/eman8/Ecolired/travail_sur_les_images/-/tree/main/decoupage_images

Quel outil OCR/HTR choisir ?

Outil	Interface utilisateur	Prétraitement d'image	Prise en main	Formats d'entrée	Formats de sortie	Modèles OCR/HTR intégrés	Possibilité de réentraînement modèle layout	Possibilité de réentraînement modèle HTR/OCR
eScriptorium	X	-	+	PDF, JPG, JPEG, PNG, TIFF	XML PAGE, XML ALTO, TXT	HTR United	X	X
Transkribus	X	-	+	PDF, JPG, JPEG, PNG	XML PAGE, TXT, DOCX, Transkribus PDF	X	X	X
ocr4all	X	X	+	PDF, JPG, JPEG, PNG	XML PAGE, TXT	X	-	X
Tesseract	-	-	++	JPG, JPEG, PNG, TIFF	TXT, hOCR, XML ALTO, PDF (interrogeable)	X	X	X
Kraken	-	-	++	JPG, JPEG, PNG	TXT, hOCR, XML ALTO, XML PAGE, JSON	X	X	X
docTr	-	-	+++	PDF, JPG, JPEG, PNG	JSON ou XML	X	X	X

Étape 1 : Segmentation & mise en page

- Deux niveaux de segmentation : zones de texte & lignes
- Il est important de choisir le bon modèle ou réentraîner à partir d'un modèle de segmentation de base.

ex. : https://github.com/Gallicorpora/Segmentation-and-HTR-Models/tree/main/Segmentation_model

- Standard recommandé pour l'annotation des segments de texte : [SegmOnto](#)

Note : Une annotation au préalable des zones de textes (savoir qu'il s'agit d'une zone de titre, d'une colonne, de la zone principale de texte) peut beaucoup faciliter l'extraction des informations concernant la mise en page en post-traitement.

Cependant, si le focus pour vous est le texte uniquement et que la segmentation est majoritairement correcte, l'annotation des zones de textes n'est pas nécessaire.

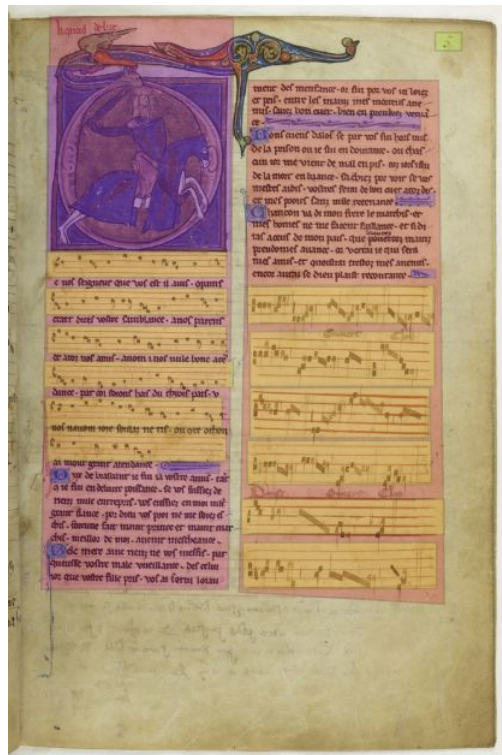
Exemple de création de zones en utilisant l'ontologie SegmOnto



source :
cremma-medieval, [bnf_fr_1728](https://nfdi.ub.uni-leipzig.de/cremma-fr/1728),
btv1b84473026_f5.jpg

ontologie : [SegmOnto](#)

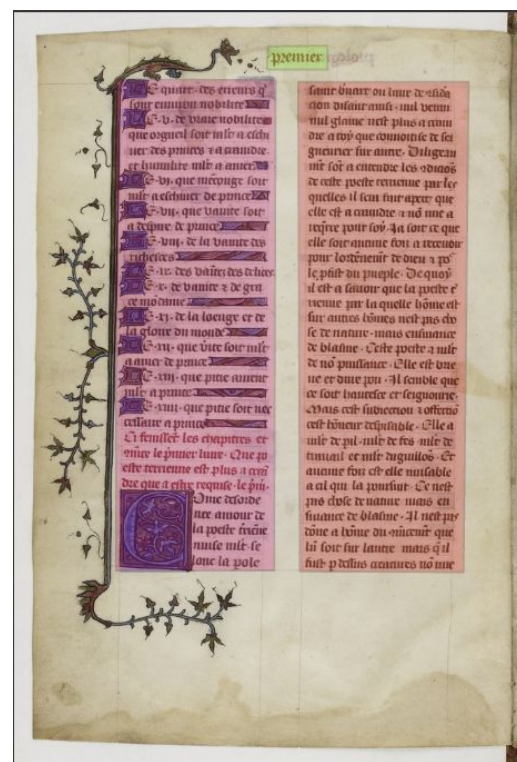
Autres exemples



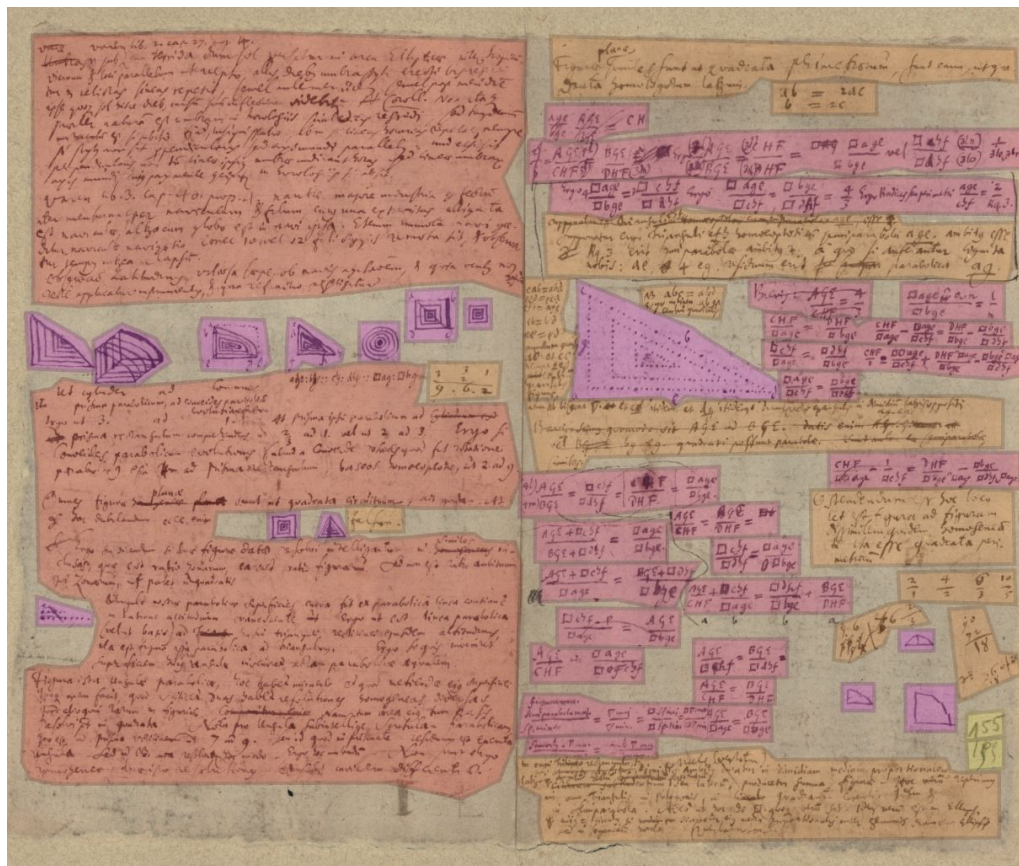
source :
cremma-medieval,
[bnf_fr_844-manuscritDuRoi](https://nbn-resolving.org/urn:nbn:fr:844-manuscritDuRoi)



source :
cremma-medieval,
[arsenal_3516_imagesDuMonde](https://nbn-resolving.org/urn:nbn:fr:3516_imagesDuMonde)

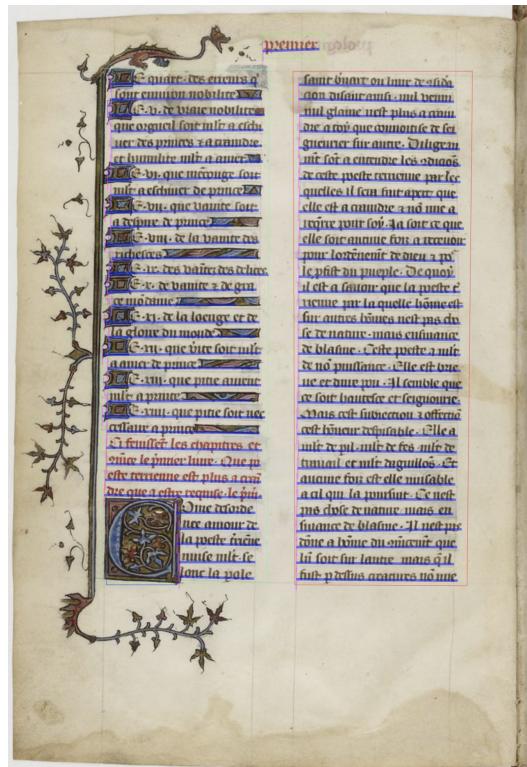


source :
cremma-medieval, [bnf_fr_1728](https://nbn-resolving.org/urn:nbn:fr:1728)

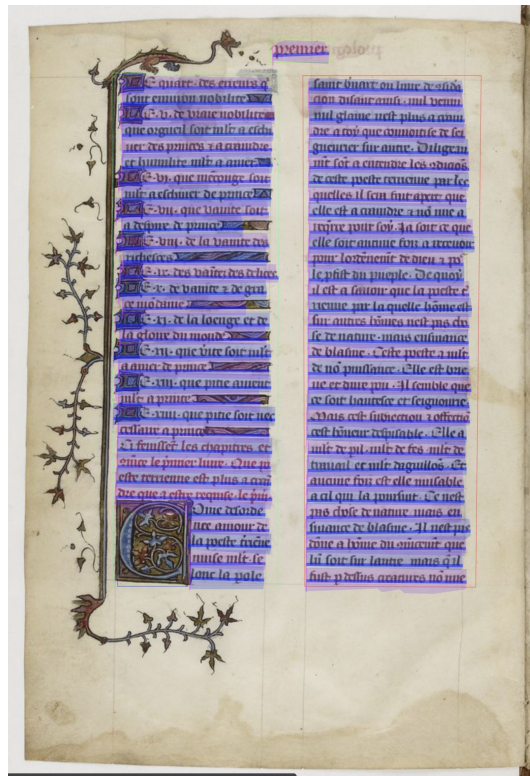


source :
manuscripts de Leibniz,
LH 35 12 2 0156v-0155r.jpg

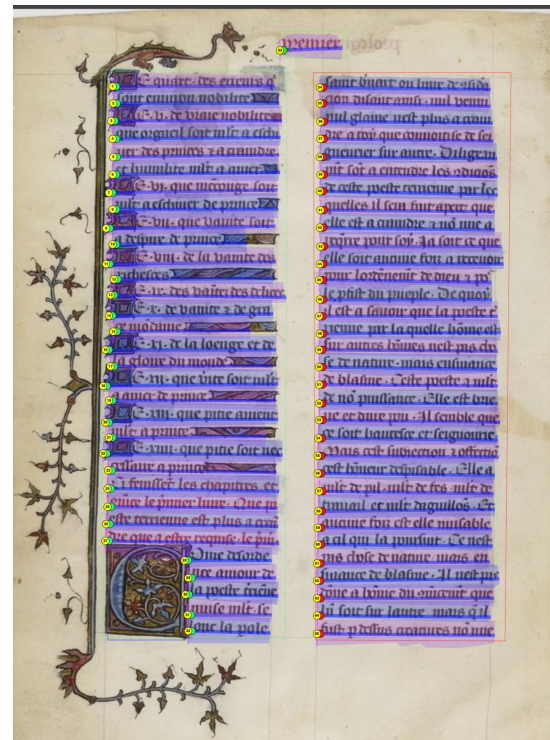
Lignes (baselines), masques, ordre de lecture



1. lignes (baselines)



2. masques



3. ordre de lecture

Étape 2 : Reconnaissance automatique du texte (OCR/HTR)

Une fois que les images ont été segmentées, la reconnaissance du texte peut commencer.

Facteurs à considérer dans le choix de modèle OCR/HTR :

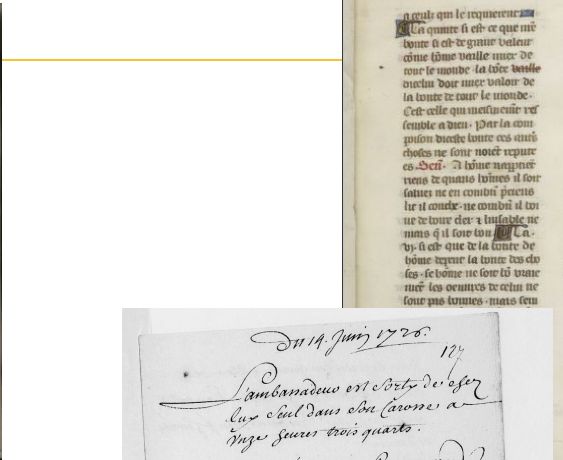
- support : manuscrit ou imprimé
- état de langue : certains modèles couvrent plusieurs siècles ; d'autres sont spécialisés sur une écriture en particulier d'une période précise.

Catalogue en ligne des modèles OCR/HTR : [HTR United](#)

Pour aller plus loin :

Pipeline complète à lancer dans le terminal : <https://github.com/Gallicorpora/application?tab=readme-ov-file>

Vidéo démo pipeline complète : <https://www.youtube.com/watch?v=iSpGQuKMvIY>



D'ailleurs
 D'ailleurs Rodal ouïl ses pas
 Biter.
 D'ailleurs mune d'au il a derendu
 egez m'illier pour y m'ni m'
 l'aroff qui l'on fait pour m'
 m'baradeau ouïl aroto m'
 l'anne d'une juve.
 Al'fol de l'alar m' du grand
 Gantio ouïl aroto m' b'vne
 juve.
 Au temple egez ote. Les m'andeur
 de l'amarthie ouïl a d'ine et dou
 il n'en b'ote que ouïl juve; m'
 Caroff a d'et egeaux aly est

a ceulz qui le requierent
P La quinte si est ce que nfe
 bone si est de grant ualeor
 cœm hōme uaille miez de
 tout le monde, la bōte uaille
 d'icelz qui doit miez ualeor de
 la bone de tout le monde.
 Cest celle qui meismēt nfe
 semble a dieu. Par la com-
 pōsion d'icelle bone ces aïds
 choses ne sont nœit repue-
 es. Señ, J. hōme nappōit
 riens de quans hōmes il soit
 saueuz ne en cōmbien pēs
 il li couche, ne cōmbien il boi-
 ue de boire cler / hūsaible ne
 mais q' il soit bon **P** La-
 ui, si est que de la bone de
 hōme depend la bone des ches-
 es, se hōme ne soit bō uraie
 māt les oeures de celui ne
 sont pas bonnes, mais sem-
 blables de turs. Sap, Enōis
 les sages sōt choses hōmies
 tes, en enōis le cōmū du

lure
 10

Les bōte temporez sont
 mauuais / r'iusables au
 pecheur. Cest fait p' le
 iuste sugert de dieu, gme
 aucun est mauuais a nfe
 aieit que les choses sont
 mauuaises a cell as q'iles
 li seignourne et sōt aiusi
 que uengiane leur creati.
 Sap, Ne chose n'est si
 bonne a hōme hōme il mais
 mes. eēc xij. Bonne est
 la sustance puis que le pe-
 chie n'est pas en la pēcie.
 Aieit que le sanc n'est pas
 bon au corps ou a il mate-
 re de enfante. Aussi n'est pas
 peccune bone a hōme qui
 est matiere de pechie.
 Le .viij. est que se hōme
 ne soit bon il est plus uil
 de ces autres creatures u
 aies, et en sōt mōte se il
 est fait mauuais il est
 plus chaifit d'icelles que le

Du 14. Juin 1726. 127
 L'ambassadeur est sorti de chez
 luy Seul dans son Carosse a
 vnze heures trois quarts.
 Il a été chez M. Gruyn garde
 du tresor Royal ou jl na pas
 Entré.
 dans la même Rue jl a dessendu
 chez vn sellier pour y voir vn
 Carosse que l'on fait pour vn
 ambassadeur, ou jl a resté vne
 bonne demie heure.
 a l'hôtel de Talars Rue du grand
 Chantier ou jl a resté vne bonne
 heure.
 au Temple chez M. le Comandeur
 de Caumartin ou jl a diné et d'ou
 Jl n'est sorti qu'a cinq heures ; vn
 Carosse a six cheuaux a luy est

mauvais nait rien tant
 seulement malice de passer
 mais de blasme et si est
 detour de mort tēporel et de
 pēdurable de quoy il est de
 pire gdition de (pens) de
 crapoz. Semblable chose

source : modèle
 CREMMA ([bnf_fr_1728](#))

13

Pratique avec eScriptorium

Lien vers les modèles à télécharger:

[CREMMA](#) (8e-15e s.) : latin et français, graphie gothique bâtarde, [transcription diplomatique](#)

[TRIDIS](#) (11e-16e s.) : latin, français et espagnol, graphie cursive, [abréviations développées](#)

[Manu Mc Fondue](#) (16e - 20e s.) : français, graphie cursive

Documentation eScriptorium pour la prise en main :

<https://escriptorium.readthedocs.io/en/latest/>

Lien vers eScriptorium:

<https://escriptorium.marceau-h.fr/>

<https://atelier-escriptorium.marceau-h.fr/>

Identifiants:

utilisateur: guest

mot de passe: ceres2025

Récapitulatif

Etapas importantes:

- Collecte du corpus
- Pré-traitement des images
- Choisir l'outil de transcription et un modèle adapté à ses besoins
- Segmentation du texte
- Transcription automatique du texte
- Récupération des fichiers transcrits

Liens pratiques

- pour convertir du format txt/docx vers XML TEI:

<https://obtic.huma-num.fr/teinte/>

- dépôt Zenodo avec plusieurs modèles d'HTR/OCR:

https://zenodo.org/communities/ocr_models/records?q=&l=list&p=1&s=10&sort=newest