



Transkribus, Kraken, eScriptorium, Tesseract

Johanna Cordova, Ljudmila Petkovic

Paris, le 18 novembre 2021

Corpus

Registre du Comité d'administration du Théâtre français de S. M. l'Empereur et Roi.

- Rédigé par Nicolas Bernard, commissaire du théâtre
- Paris, le 16 janvier 1813
- Texte manuscrit
- **Document** : collection d'images formant un ensemble
- Échantillon de 10 pages

I Transkribus

1. Pré-traitement des données

Pré-traitement des données

- **Sélectionner les pages à pré-traiter** (de bonne qualité)
 - Exclure les pages vides, non pertinentes, le dos de livre etc.
- **Découper les scans PDF**
 - afin de ne transcrire que le véritable contenu textuel d'une page
 - sinon, le moteur d'OCR / HTR va générer des symboles inutiles
 - Impact de la couverture du document, des traces de l'écriture de la page précédante / suivante, des taches etc. sur la qualité de transcription

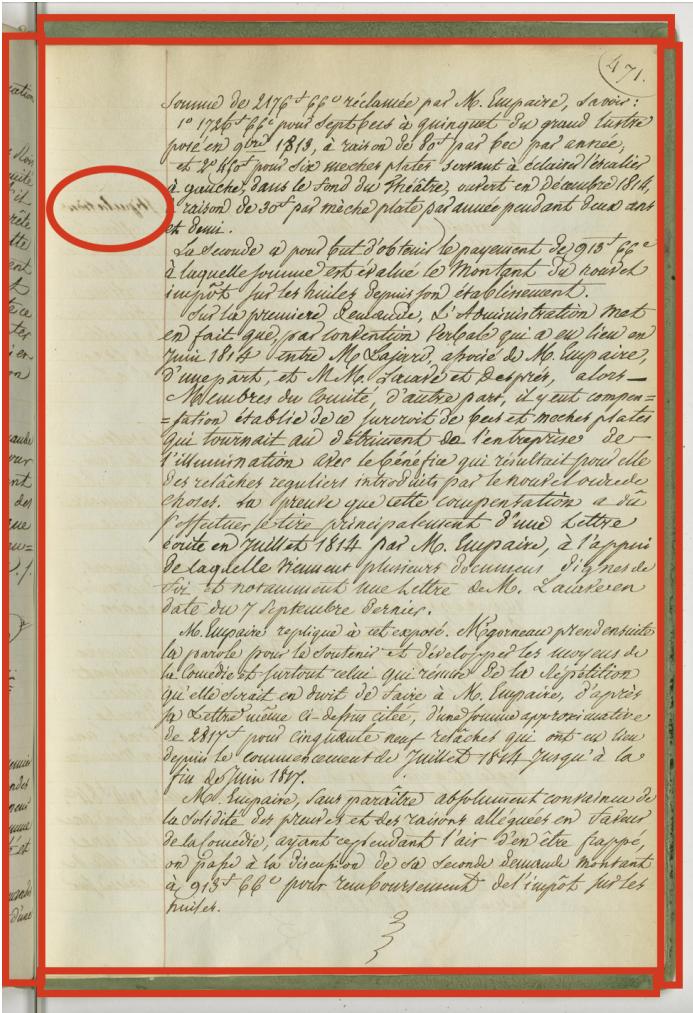


Figure 1a : PDF original.

471.

Somme de 2176 francs réclamée par M. Luyaire, Savoie :
1^{er} 176 francs pour sept places à quinze francs par grand théâtre
posé en juillet 1813, à raison de 80 francs par place par annéé,
et 2^{me} 176 francs six meches plates servant à éclairer l'étable
à gauche dans le fond du Théâtre, auroit été décaupé 1814,
à raison de 30 francs par meche plate par annéé pendant deux ans
et demi.

La somme a pour but d'obtenir le paiement de 913 francs
à laquelle somme est évalué le Montant du recouvrement
impôt sur les huiles depuis son établissement.

Sur la première Recouvrement, l'Administration met
en fait que par convention Perbale qui a eu lieu en
juin 1814 entre M. Luyaire, auroit été M. Luyaire,
d'une part, et M. Luyaire et Desprez, auroit —
Mémoires du Comité, d'autre part, il y eut compen-
sation établie de 10 francs de peu et moches plates
qui tournaient au détriment de l'entreprise de
l'éclaircation. Sur le bénéfice qui résultait pour elle
des relâches régulières introduites par le Comité pour
chos. Sa preuve que cette compensation a été
effectuée principalement d'une lettre
écrite en juillet 1814 par M. Luyaire, à l'appui
de laquelle viennent plusieurs documents, d'après
lesquels il résulte une lettre de M. Luyaire en
date du 7 septembre Bernier.

M. Luyaire réplique à tel exposé. M. Goyonau prend ensuite
la parole pour le bâtonnier et développe les motifs de
la contestation, surtout celui qui remonte à la législation
qui elle était en droit de faire à M. Luyaire, d'après
la Lettre même à ce sujet, d'une somme approximative
de 2817 francs pour cinquante-neuf relâches qui ont eu lieu
depuis le commencement de juillet 1814 jusqu'à la
fin de juillet 1817.

M. Luyaire, sans paraître abîmément convaincu de
la validité des preuves et des raisons alléguées en faveur
de la contestation, ayant cependant l'air bien être frappé,
on passe à la discussion de la seconde demande portant
à 913 francs pour remboursement de l'impôt sur les
huiles.

2

Figure 1b : PDF découpé.

Pré-traitement des données

- Charger toutes les pages, pour garder les références (page de titre)
- Normalisation du contrast **sans** binarisation lors du chargement des images du document ([Michael et al, 2018](#))
 - ≠ Kraken — binarisation comme une étape indépendante et facultative

I Transkribus

2. Choix du modèle de transcription

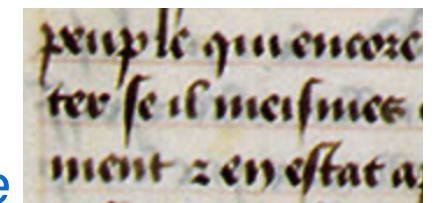
2.1 New France 17th-18th Centuries

Choix du modèle de transcription

Plusieurs paramètres :

- Langue : française
- Alphabet : latin

- Type d'écriture : *coulée*



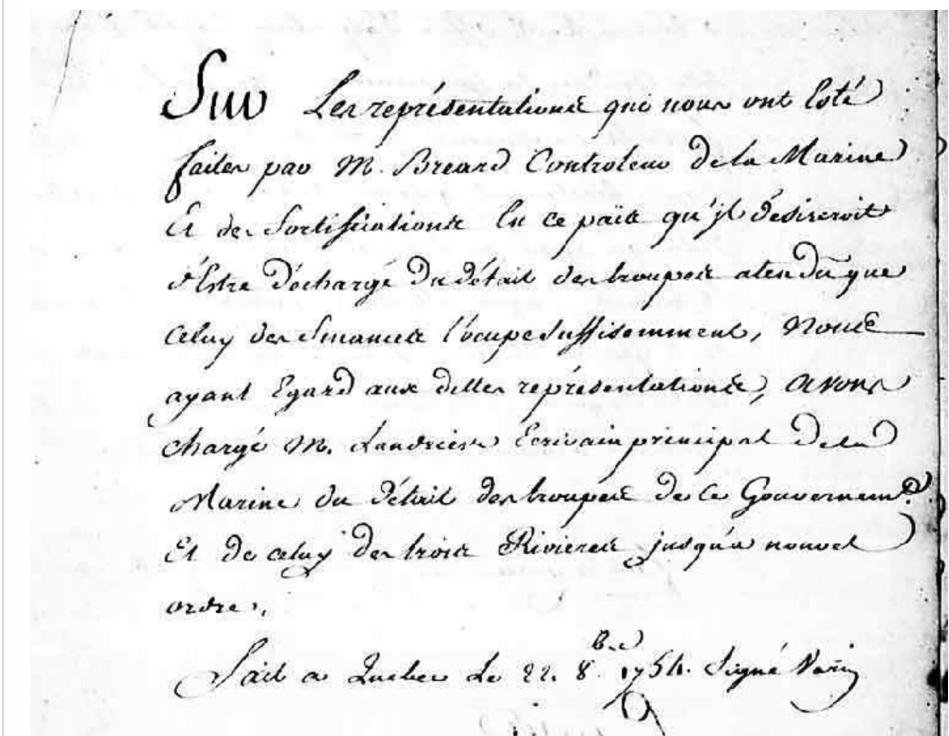
- Type de documents : administratif, ~ordonnances, chancellerie
- Époque : début du XIX^e s.
- Moteur de reconnaissance de caractères : HTR+
- CER (taux d'erreur de caractère, angl. *character error rate*)

Choix du modèle de transcription

Modèles d'HTR accessibles au public (cf. le site de Transkribus) :

- French — General Model (8.5% CER)
- Charter Scripts (German, Latin, French) (6.32% CER)
- French and Latin Chancery documents (5.33% CER)
- French Handwriting 19th century (7.73% CER)
- French Livre Rouge (8% CER)
- New France 17th-18th centuries (4.12% CER)
- Ordonnances des Intendants (4.18% CER)

New France 17th-18th centuries



Source : Projet « Nouvelle-France numérique »

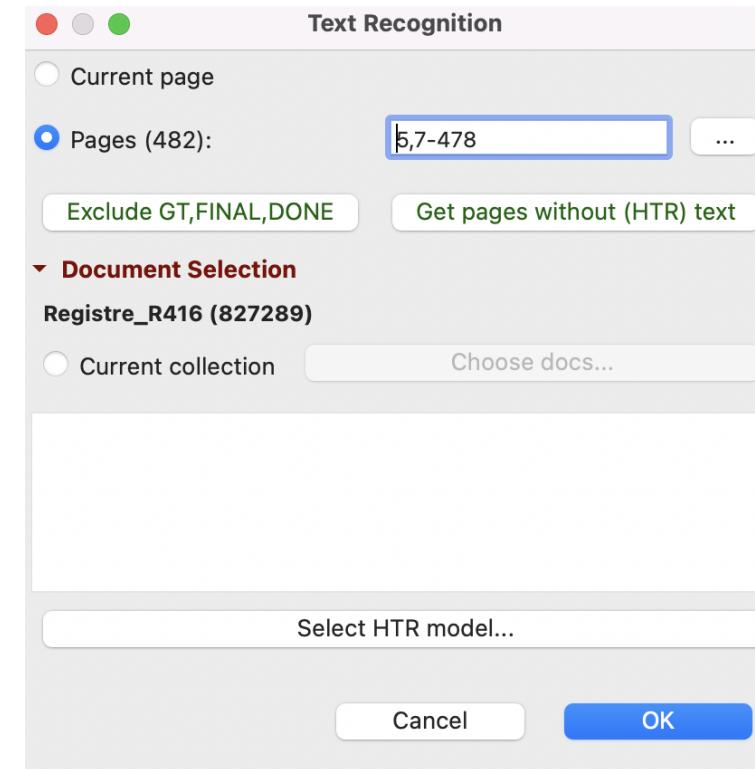
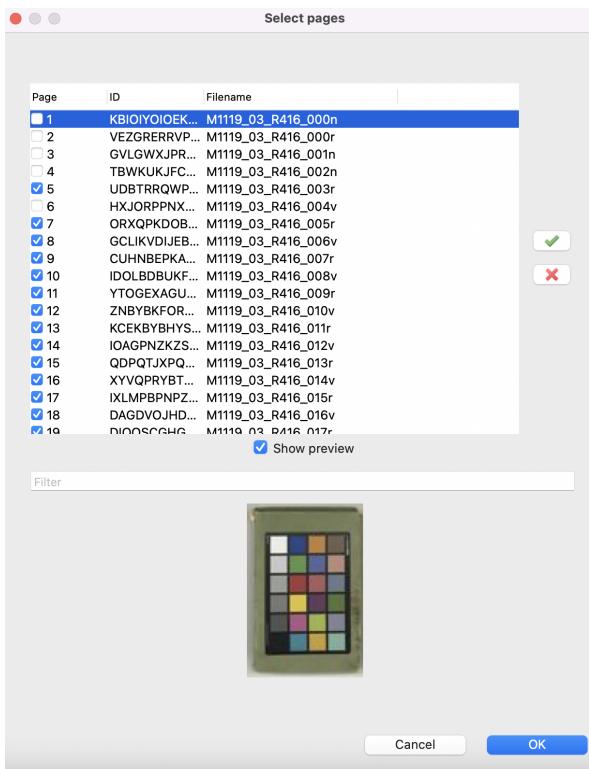
- Combinaison d'écritures françaises (ronde, bâtarde et coulée)
- 1 600 pages (296 403 mots) de correspondance et de registres des administrateurs coloniaux de la Nouvelle-France du XVII^e/XVIII^e s.
- CER de 4.12% (données de validation)
- Pas adapté aux écrits des notaires ou des greffiers

I Transkribus

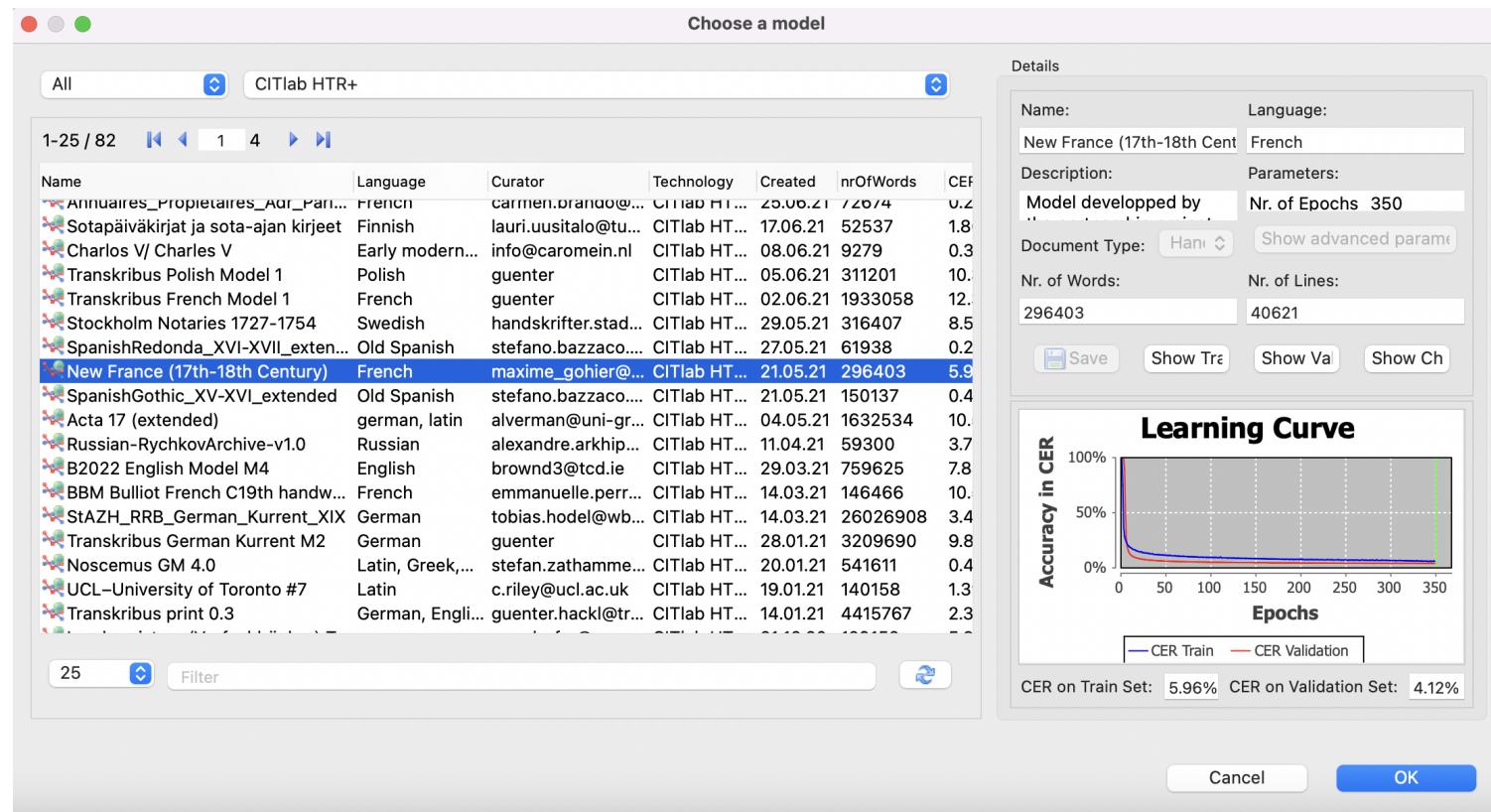
3. Lancement de l'HTR

Lancement de l'HTR

- Sélectionner les pages découpées au préalable pour l'HTR



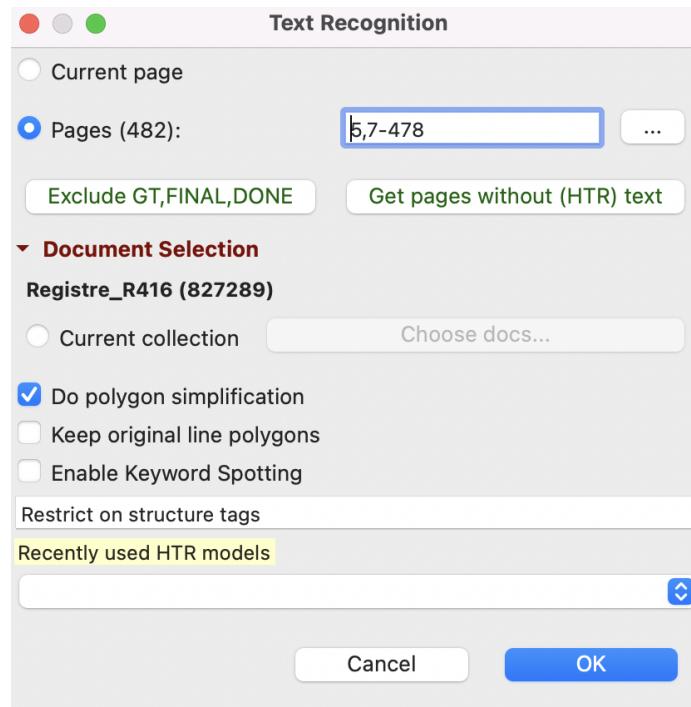
Lancement de l'HTR



Le modèle « New France » dans Transkribus.

Lancement de l'HTR

- La simplification des polygones réduit la complexité des segments de ligne, en économisant de la bande passante et de l'espace de stockage



I Transkribus

4. Résultats de l'HTR

Résultats de l'HTR

- Durée de transcription : environ 3h pour 480 pages

Registre du Comité d'Administration
du Théâtre français de S.M. l'Empereur et Roi.

Paris ce 16 Jauvier 1813.

D'après l'invitation de Mr. Bernard remplissant les fonctions de Commissaire impérial, M. M. Fleury, Cahier, Michot, Damas, Despize & Lacave se rennissent à 2 heures dans la Salle des séances du Comité d'administration.

Mr. Bernard notifie deux arrêtés de Mr. le Surintendant des Spectacles portant l'organisation, l'un du Comité d'administration & l'autre du Comité de lecture conformément au Décret impérial du 15^e 8^{me} 1812.

Ces arrêtés sont ainsi concus.

Le Premier Chambellan de S. M. l'Empereur et Roi surveille dans les spectacles, vu les articles 30 & 49 du Décret impérial du 15^e 8^{me} 1812.

Portant organisation du Théâtre français, arrêté ce qui suit :

Art. 1^{er}

organization
du Comité
d'Administration.

2-2 Registre du Comité d'administration
2-3 du Théâtre français de S. M. l'Empereur et Roi.
2-4
2-5 Paris ce 16 Jauvier 1813.
2-6 D'après l'invitation de Mr. Dernard remplissant les fonctions de Commissaire impérial, M. M. Fleury, Valins, Michot, Damas, Despize & Lacave se rennissent à L'heures dans la Salle des séances du Comité d'administration.
2-9 Mr Dernard notifie deux arrêtés de Mr. Le Surintendant des spectacles portant l'organisation, l'un du comité d'administration & l'autre de

II Kraken

1. Choix du modèle de transcription

Choix du modèle de transcription

Choix du modèle de transcription

- Modèles d'HTR accessibles au public (les *nouveaux* et les *anciens*, ALMAnaCH Inria)
- Les résultats de transcription ne sont pas tout à fait satisfaisants
- Modèles entraînés sur quels textes ?
- Les modèles obsolètes manifestent un problème d'initialisation du package `nnpack`

Binarisation

- Conversion d'une image en noir et blanc (image *binaire*)
- Distinguer le texte (ou tout autre élément d'image requis) de l'arrière-plan
- Comment « supprimer » la couleur de l'arrière-plan (sombre) sans perdre également les informations du texte ? (même avec de l'ajustement de la luminosité)
- Impossible de filtrer l'arrière-plan sans effacer ou éclaircir simultanément le texte
- L'arrière-plan introduit un « bruit » qui gêne la reconnaissance



Binarisation kraken

- Algorithme `nlbin` (non requis avec le segmenteur de ligne de base `-bl`)

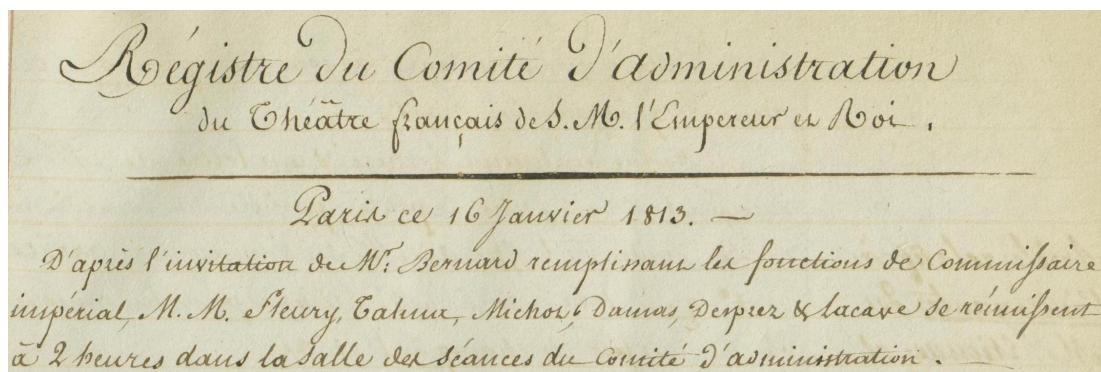


Figure 2a : Image non binarisée

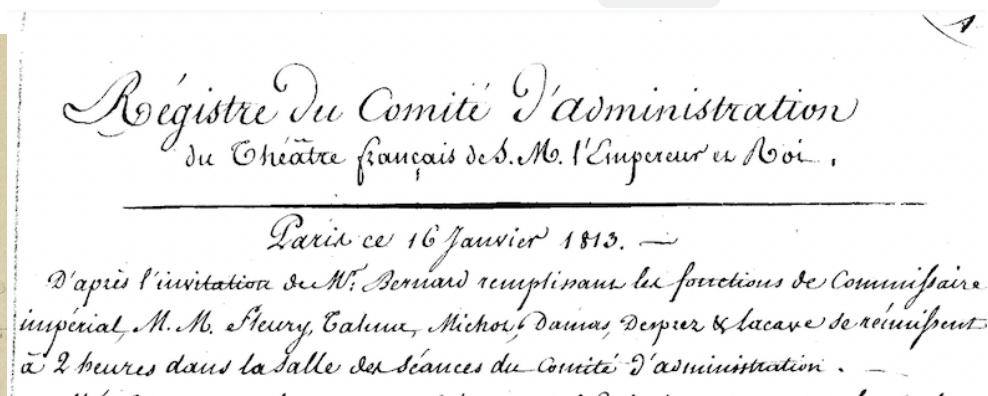


Figure 2b : Image binarisée

riant_ftmrs15_12.mlmodel (avec binarisation)

- Ancien modèle
- Génère meilleurs résultats par rapport aux nouveaux modèles de kraken

Régitre du Comité d'Administration
d'ou Tonéatre Français de S. M. Emvereur et Roi

2

Paris ce 16 Jauvier 1d

d'après l'inviration deMr Benard suupténant, les fouteions de Commipsaire
Anpérias M. M. FSeury Tahue Nicnos damors demuer & Lacave de nomciprent
a heures dans laSaile des Séances du Contite d adninntration
d Bernard notitie du arrete deMr Le surentendans desoictaite
oortans l'organisation d'un du Comite d'Aduntration à d'autie d
conité de Lechere confomeuient au denret inpérias d abre d d
'arretes dous airie doncu

Dganigation de Frenrier Chambellau de S. M. d'Ernvereur et Roi ouventeudant de
Vix Comile Svertactes, du les articles 30 & 49 du Décres unvérenl du1 5. 8bres 1812
d'Achunestrion, portaux orgouisation du Théatre Français, Arrite cíque Suit
drtefer

mane me Pourt, nommés meuibres du Comité d'Aduinistration M. M. Feury, Taluie
d

Michot Damars Denret Lacaue
dt, 2e

riant_ftmrs15_12.mlmodel (sans binarisation)

- le début du texte est déplacé ([...] Registre du Comite d'Administration
[...])

inpérial, M. M. Feury, Tahuu, Michos dumas, desier & Lacave se recrifent
Le Prerrier Chambellau deS. M. l'Empereur et Roi surinteudant des
d'Adhninistration. portaur orgouisation du Théatre Francais, Arrête cequi suit.
Flesdes Raucourt & Mars sout adjouites au Comité pour laPocuation
dis Coutité de spectartes, du l'Article 68 du Décret inpérial du15, 8bre portant orgouisation
Registre du Comite d'Administration
du Cnéatre français deS. M. l'Empereur et Roi,
Organisation
Paris ce 16 Janvier 1813.~
d'après l'invitation duNr Bernard receiptinant lesfouctions de Commisaire
à 2 heures dans laSalle desséances du Conrité d'Adminsstration-
Mr Beruardnotifie deux arrêtés deMr lesurintendant desspectarles
portaut l'organisation, l'un du Comité d'Adiniuntration & 'autre du
Comité de Lecture, conformément au décret inpérial du15 8bre 1 812-
Ces arrêtés sour arrisiconcur.
Svertacles, du les articles 30 & 49 du décres inpérial du15 8bre 1812
Sous nommés meurbres duComité d'Aduinstration M. M. Feury, Talure
le Commipaire inpérial est charge del'exécution

III eScriptorium

1. Introduction

Introduction

- Infrastructure web gratuite et à code source ouvert
- Spécialisée pour l'HTR des documents anciens
- Développée par PSL | INRIA
- Basée sur les techniques de l'apprentissage machine ([encog](#))
- Cadre de programmation fourni par l'application eScript
- Connexion au [serveur](#) ou installation en local / Docker
- Cf. le [tutoriel](#) en français

III eScriptorium

2. Préliminaires

Préliminaires

Connexion au serveur

- Instance eScriptorium en ligne
- Nom d'utilisateur : guest
- Mot de passe : GuestAccount2021!
- Tableau de bord pour la gestion des documents créés par et partagés avec un compte

Préliminaires

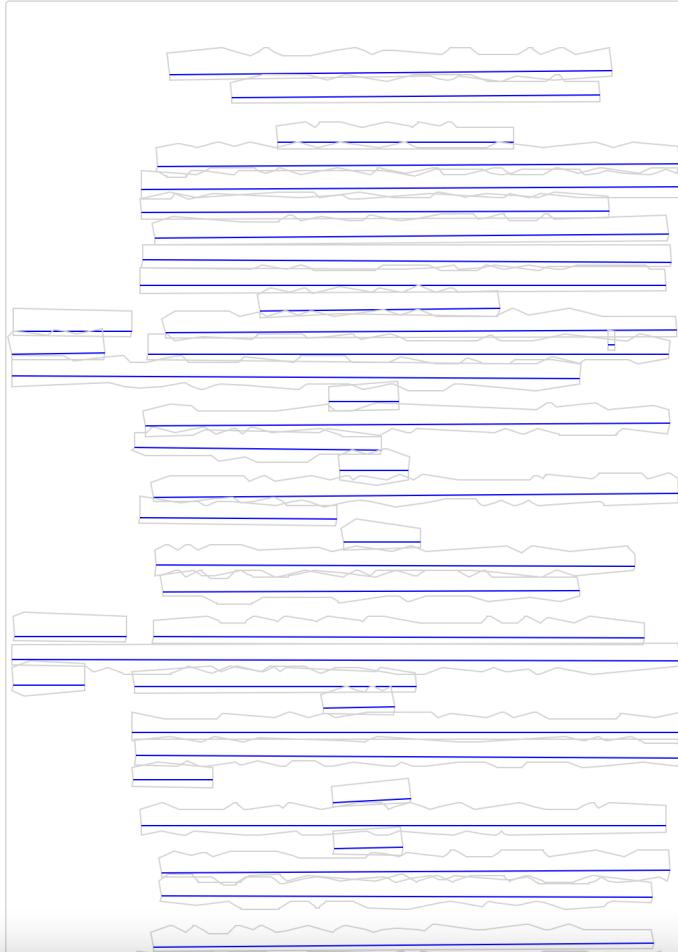
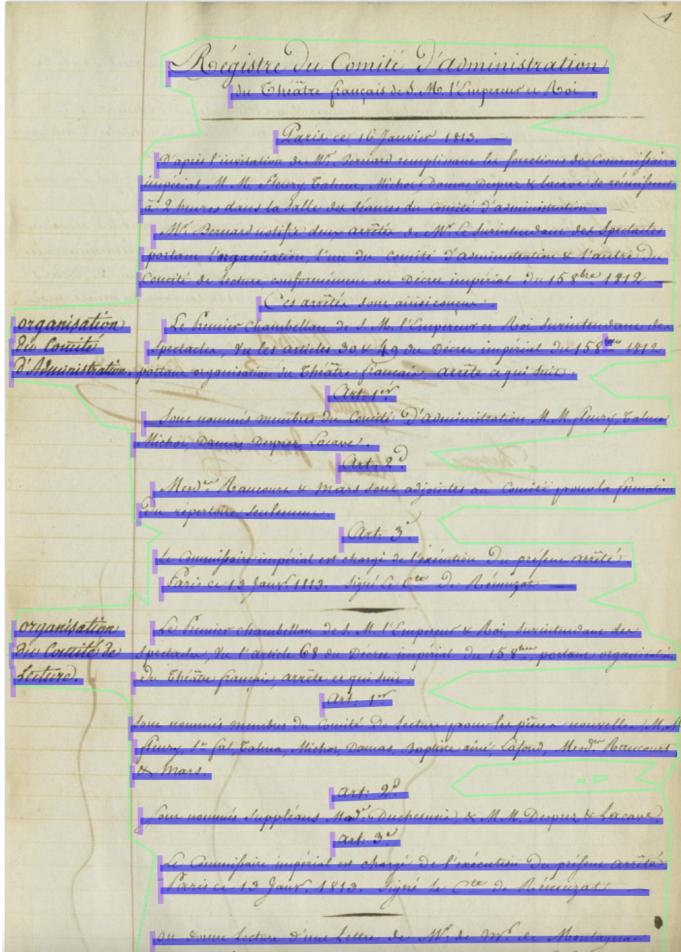
- Création d'un nouveau document
- Chargement des images

La binarisation n'est plus une étape obligatoire et pourrait même diminuer la qualité [des images à transcrire]

- Segmentation automatique : `blla.mlmodel` par défaut

Segmentation

- `blla.mlmodel` (par défaut)



Les lignes (violet) et les zones (vert).

Transcription (sans binarisation)

Registre du Comite d'Administration
du Théâtre Français deS. M. l'Empereur et Roi,
Paris ce 16 Janvier 1813.
d'après l'invitation de Mr Bernard recevant les fonctions de Commissaire
impérial, M. M. Fleury, Tahau, Michot dumas, desprer & Lacaue se recrissent
à 2 heures dans la Salle des séances du Comité d'Administration-
Me Deruard informe deux arrêtés & le Surintendant des spectacles
portant l'organisation, l'un du Comité d'Administration & l'autre du
Comité de Lecture, conformément au décret impérial du 15bre 1 812-
Ces arrêtés sont ainsi concus.

Organisation

Le Preurier Chambellau deS. M. l'Empereur et Roi Surintendant des
du Comité

S'ertactes, du les articles 30 a 49 du décrets impérial du 15bre 1812
20

d'Administration. portant organisation du Théâtre Français, Arrête ce qui suit.

Art. 1er

Sous nommés meubres du Comité d'Administration M. M. Feury, Talmre
Michot, Damas, Denrer, Lacane.

Art. 2e

IV Évaluation

1. Mesures d'évaluation
2. Outil d'évaluation
3. Comparaison des modèles HTR

Mesures

Paramètres de qualité : **reconnaissance des caractères**, reconnaissance de la mise en page

2 mesures principales pour évaluer la reconnaissance de caractères :

- *Character error rate* (CER) : pourcentage de caractères erronés dans le document
- *Word error rate* (WER) : pourcentage de mots qui contiennent des erreurs

Types d'erreurs : substitution, insertion, délétion

Exemple	Texte du manuscrit	Sortie d'OCR
	Arrêté de M. — . . .	

Outil d'évaluation

Étapes pour l'évaluation :

- Corriger manuellement une partie des sorties OCR (*ground truth*)

Outil d'édition de *ground truth* : Aletheia

(<https://www.primaresearch.org/tools/Aletheia/Editions>)

- Comparer les pages corrigées et leur version océrisée

Outil de comparaison : **ocrevalUAtion** (<https://github.com/impactcentre/ocrevalUAtion>)

Comparaison des modèles (segmentation)

	Transkribus	Kraken	eScri
Détection marge	✓	✗	<span style="color:

Comparaison des modèles (reconnaissance de caractères)

	Transkribus	Kraken (avec binarisation)	Kraken (sans binarisation)	eScriptorium
CER	10,70	41,67	55,8	36,96
WER	35,00	80,23	84,86	72,59
WER (sans ordre)	30,70	63,20	53,9	54,28

 image-20211117162024725

V Conclusion

- dépôt [GitHub](#)