

HTR en théorie et en pratique : Transkribus



Ljudmila PETKOVIC

Sorbonne Université, Faculté des Lettres, UFR Littératures françaises et comparée
Centre d'étude de la langue et des littératures françaises (CELLF), UMR 8599
Observatoire des textes, des idées et des corpus (ObTIC)

ljudmila.petkovic@sorbonne-universite.fr

Formation OCR · HTR et Transkribus
M2 « Bibliothèques et archives de l'univ. d'Angers »
Angers, le 1 février 2024, année 2023-2024

Petite enquête

- Avez-vous entendu parler de la méthode d'**OCR** (ou bien d'**HTR**) ?
- Qu'en savez-vous ?
- Envisagez-vous d'utiliser ces méthodes dans le cadre de vos recherches ?

N'hésitez pas à noter vos idées et commentaires dans ce [Framapad](#).

Objectif de cette formation

- 0 (re)découverte des notions principales en lien avec cette formation
- 1 sensibilisation à la question de la transcription automatique de textes (OCR · HTR)
- 2 découverte d'un outil d'HTR : Transkribus
- 3 transcription automatique d'un corpus-test
- 4 aperçu des projets d'HTR menés par des *GLAM*¹

¹ angl. **Galleries, Libraries, Archives and Museums** : institutions patrimoniales

Matériel

Le matériel de cette formation (diapositives et corpus jouet) se trouve sur le [dépôt GitHub](#), ainsi que sur Moodle.

0 Concepts de base

Archivage

Processus de collection, d'organisation, de description, de préservation, de diffusion et d'accès aux archives (collections de documents anciens, classés à des fins historiques).

(Lepron, 2023)

- les archives sont définies, collectées, triées et classées selon une logique rationnelle et informationnelle, selon une visée a priori plus fonctionnelle
- mise en jeu des valeurs d'information, juridiques, et secondairement historiques

(Davalon, 2014)

Patrimonialisation

Processus de création, de fabrication de patrimoine (héritage du passé existant aujourd'hui), reconnu en tant que bien collectif. Le phénomène s'est développé en France au XIX^e siècle, notamment pour la sauvegarde des monuments historiques.
→ valorisation patrimoniale.

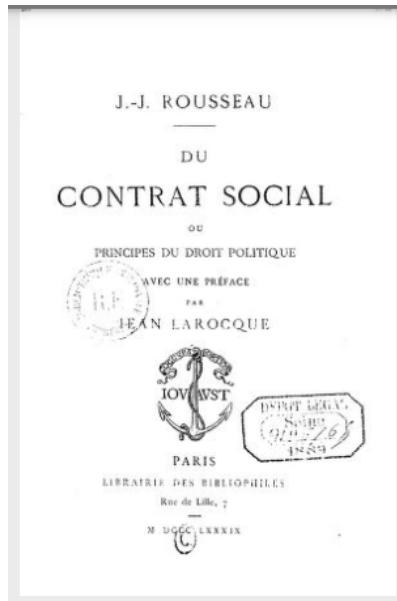
(Ressources de géographie pour les enseignants, 2019)

- concerne des objets qui se trouvent être encore présents tandis que leurs homologues ont disparu
- le choix de les patrimonialiser suppose un intérêt social pour eux
- mise en jeu des valeurs cognitives, sociales, voire identitaires
- nouvel usage des archives : patrimonialisation numérique

Archives numérisées

L'archive **numérisée**, à l'inverse de l'archive numérique, n'est pas nativement électronique. Il s'agit du résultat de la reproduction d'une archive initialement physique au moyen d'un outil numérique, tel qu'un scanner ou un appareil photo.

(Barbier & Mandret-Degeilh, 2018)



Exemple d'une page au format PDF issue de [Gallica](#).

Dématérialisation (rétro-conversion)

- convertir les documents papier dans un format logique pivot les rendant de nouveau rééditables et donc réutilisables
- au-delà de la simple transcription automatique de textes
- réhabilitation de la structure interne du document, la plus proche possible de celle qu'il avait à l'origine

(Toumit *et al.*, 2000)

Archives numériques

[...] tout type de document produit aujourd'hui sous une forme électronique et dématérialisée.
[...] toute archive « nativement » (c'est-à-dire dès sa création) **numérique**. Cette archive est faite à la fois de *données* (le contenu du document à proprement parler) et de *métadonnées* (les informations sur le document, telles que sa date de création, son auteur, etc.).

(Barbier & Mandret-Degeilh, 2018)

- exemples :
 - archives numériques conservées par les Archives nationales
 - contenu généré par l'utilisateur·trice (angl. *user-generated content, UGC*)
 - issu des réseaux de communication (Internet, email, réseaux sociaux)

Algorithme

- ensemble de règles indiquant à l'ordinateur comment effectuer une tâche
 - algorithme d'OCR : « *diviser l'image d'un caractère de texte en sections et distinguer les régions vides et non vides. En fonction de la police ou du type d'écriture utilisé pour la lettre, la somme de contrôle de la matrice résultante est ensuite étiquetée (initialement par une personne) comme correspondant au caractère de l'image*

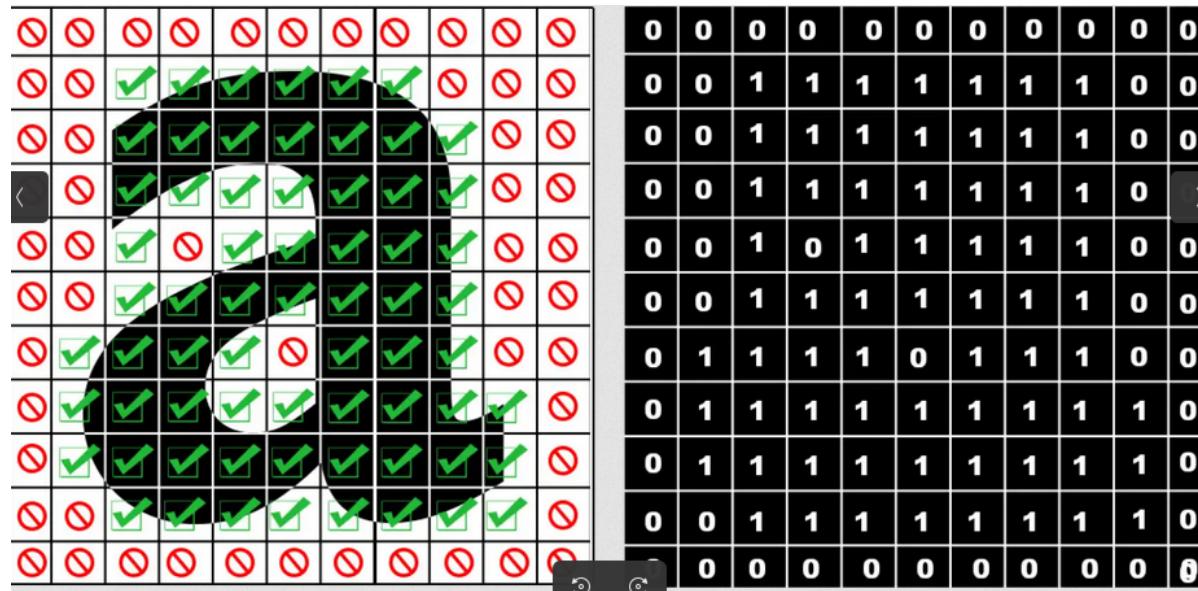
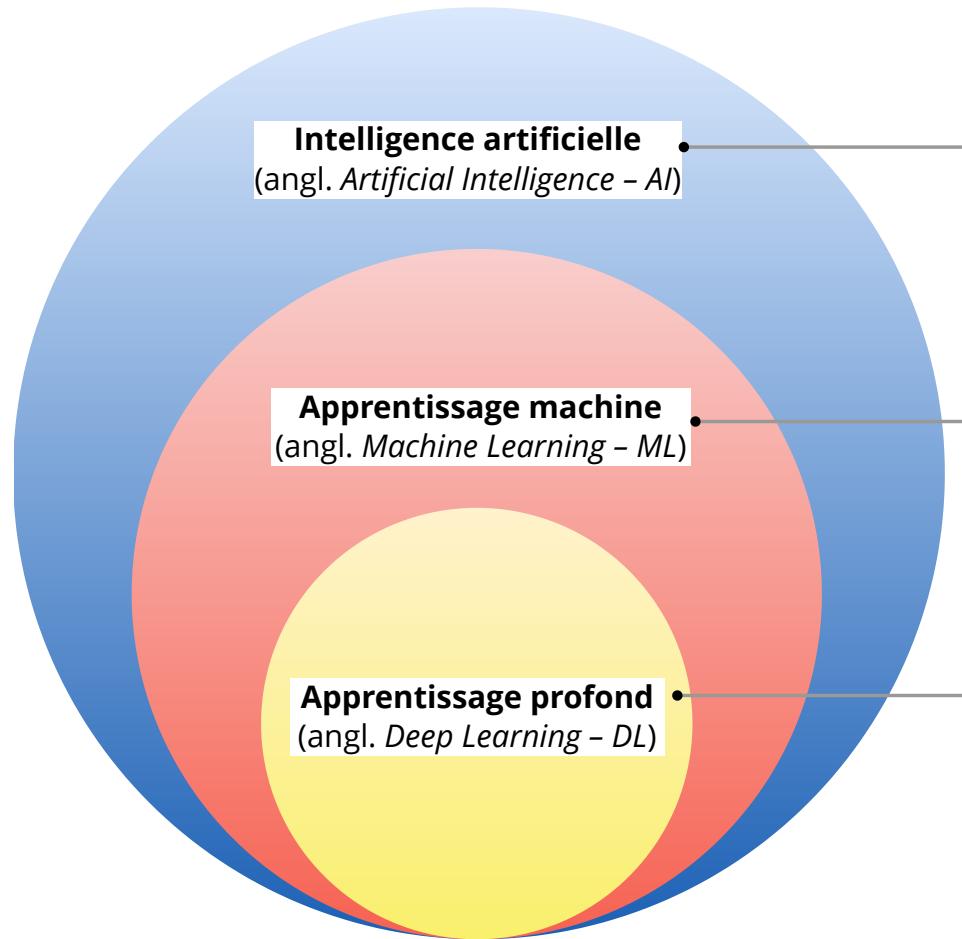


Illustration d'un algorithme d'OCR ([Anderson, 2021](#)).

Intelligence artificielle (IA)



ensemble des techniques inspiré par le domaine des neurosciences, permettant d'apporter une forme d'intelligence aux machines, afin de résoudre des problèmes plus ou moins complexes.

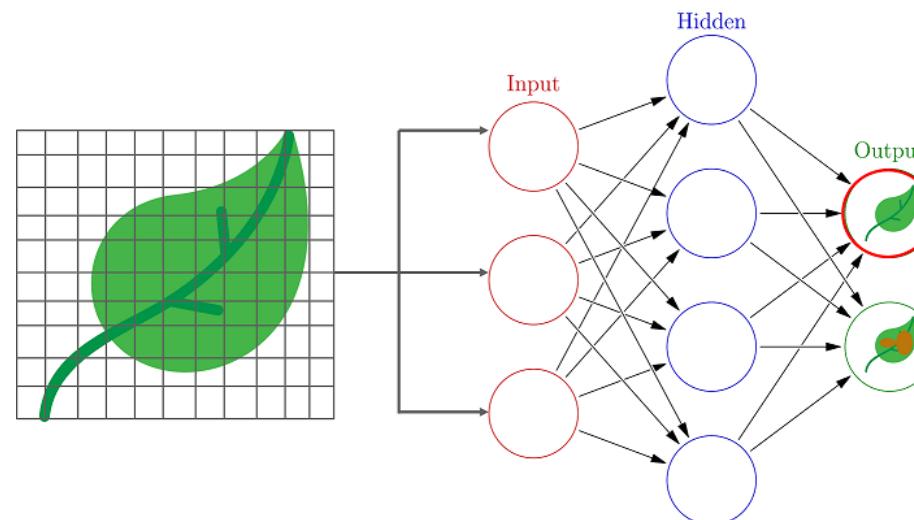
ensemble des techniques inspiré par le domaine des statistiques, dans lequel l'ordinateur formalise et apprend des règles à partir des données. Néanmoins, elles nécessitent souvent l'intervention humaine quant à l'analyse des données.

ensemble des techniques inspiré par le domaine des neurosciences, dans lesquels l'ordinateur détermine par lui-même les schémas d'apprentissage à partir des données. Dans ce schéma, l'intervention humaine est plus limitée.

Relations entre les termes « IA », « apprentissage automatique » et « apprentissage profond » (adapté de [Cendre, 2021](#)).

Principes de l'apprentissage profond

- résoudre un problème en le divisant en plusieurs tâches
- tâches distribuées à des algorithmes de ML, organisés en **couches** successives
- les couches contiennent des neurones et aident à faire circuler l'information
- chaque couche travaille sur le résultat de la précédente → **réseau de neurones**

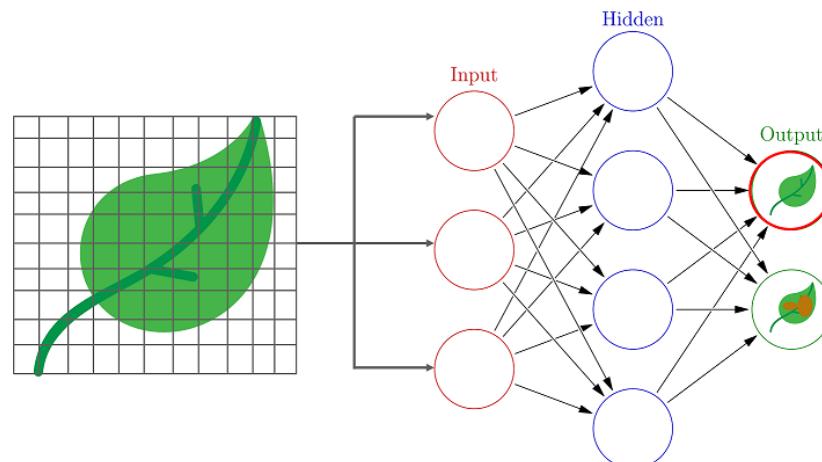


Structure d'un réseau de neurones pour la tâche de classification ([Lateef, 2023](#)).

Réseau de neurones artificiel profond

- algorithmes produisant une donnée de sortie à partir des données d'entrée
 - * couche d'entrée (accepte les données d'entrée : images, documents...)
 - couche cachée (effectue des calculs nécessaires pour produire la sortie)
 - * couche de sortie (prédit une donnée de sortie : classe, caractère...)

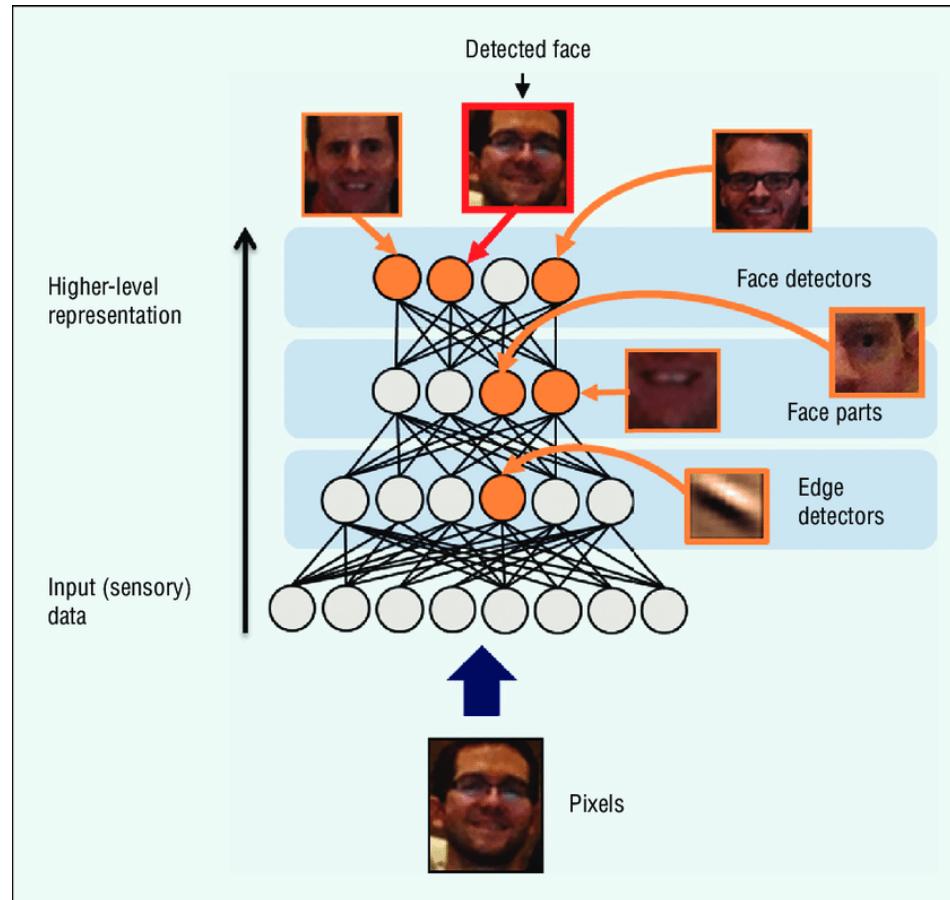
✿ Types de réseaux : classiques, convolutifs, récurrents etc.



Structure d'un réseau de neurones pour la tâche de classification ([Lateef, 2023](#)).

Apprentissage profond

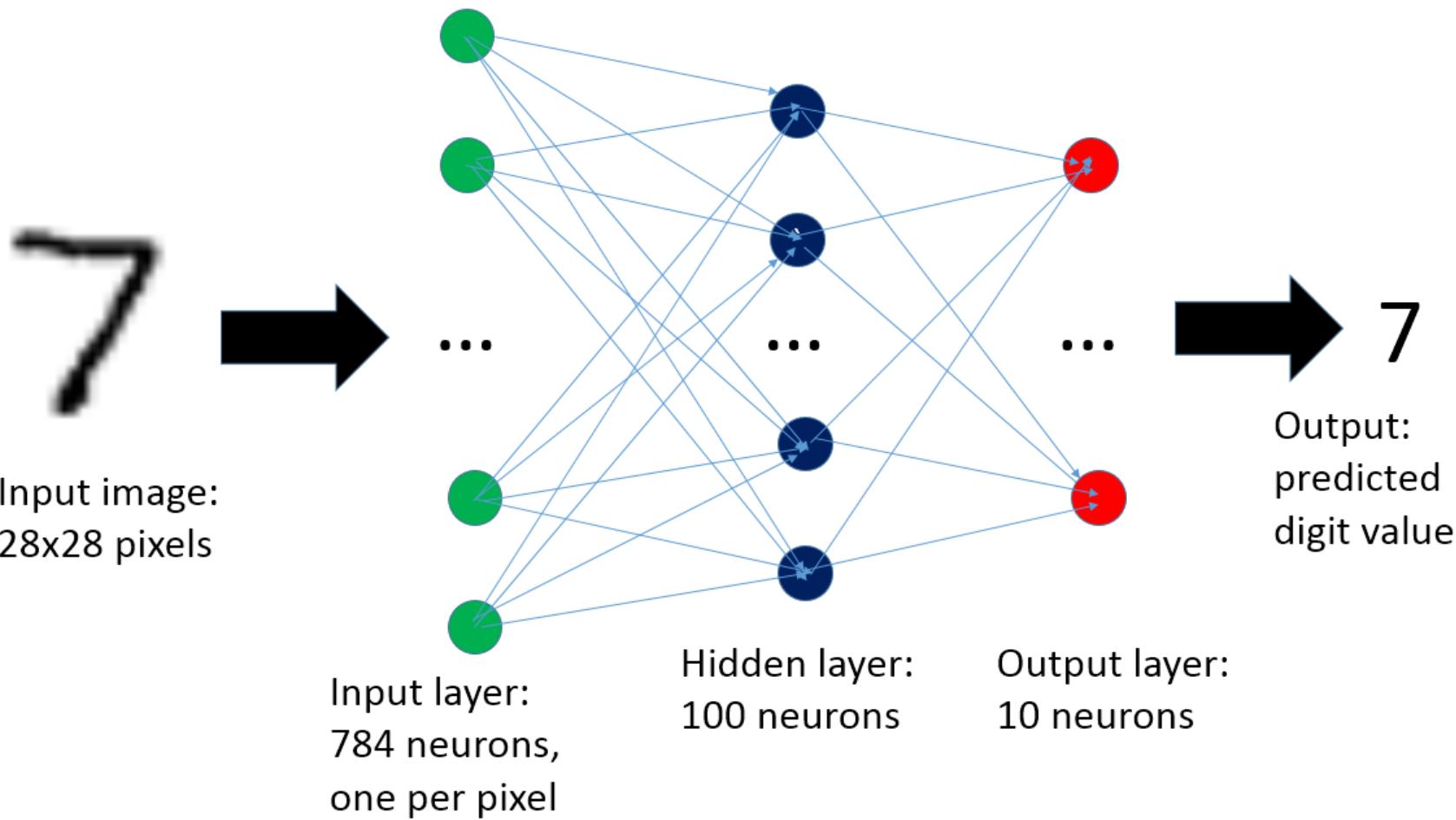
- cas de figure : reconnaissance de visages



L'illustration du mécanisme du réseau de neurones pour la tâche de reconnaissance de visage ([Lehman & Clune, 2014](#)).

Formation OCR · HTR et Transkribus, M2 « Bibliothèques et archives de l'UA », 01/02/2024, Ljudmila PETKOVIC

Apprentissage profond : reconnaissance de caractères



L'illustration du mécanisme du réseau de neurones pour la tâche de reconnaissance de caractères ([Canali, 2016](#)).

Les projets en humanités numériques

Quatre principales étapes :

Les projets en humanités numériques

Quatre principales étapes :

1 Acquisition d'un objet d'étude

Les projets en humanités numériques

Quatre principales étapes :

- 1** Acquisition d'un objet d'étude
- 2** Traitement d'un objet d'étude

Les projets en humanités numériques

Quatre principales étapes :

- 1** Acquisition d'un objet d'étude
- 2** Traitement d'un objet d'étude
- 3** Exploitation d'un objet d'étude

Les projets en humanités numériques

Quatre principales étapes :

- 1** Acquisition d'un objet d'étude
- 2** Traitement d'un objet d'étude
- 3** Exploitation d'un objet d'étude
- 4** Publication des résultats

Les projets en humanités numériques

Quatre principales étapes :

- 1** Acquisition d'un objet d'étude
- 2** Traitement d'un objet d'étude
- 3** Exploitation d'un objet d'étude
- 4** Publication des résultats



Les frontières entre ces étapes sont très poreuses et l'ordre des opérations ne respecte pas toujours cette progression logique.

Les projets en humanités numériques

Quatre principales étapes :

1 Acquisition d'un objet d'étude

2 Traitement d'un objet d'étude

3 Exploitation d'un objet d'étude

4 Publication des résultats



Les frontières entre ces étapes sont très poreuses et l'ordre des opérations ne respecte pas toujours cette progression logique.

1 Transcription automatique de textes

Enjeux de l'acquisition de données

Format natif → format numériquement et pleinement exploitable

- texte lisible, analysable et requêtable par la machine
 - utilisation de la puissance des ordinateurs pour effectuer la transcription
 - traitement des documents en masse, redistribution au format numérique
 - éviter la transcription manuelle et chronophage de documents volumineux
- ≠ projets de production participative (angl. *crowdsourcing*) :
- [From The Page](#), [Transcribathon](#), [Transcrire](#), [T-PEN](#) etc.

Méthodes de transcription automatique de textes

La transcription automatique de textes comprend les méthodes suivantes :

1. OCR (angl. *optical character recognition*)
2. HTR (angl. *handwritten text recognition*)

Ces méthodes sont parfois regroupées sous le nom d'ATR (angl. *automatic text recognition*) ([Scheithauer, 2023](#))

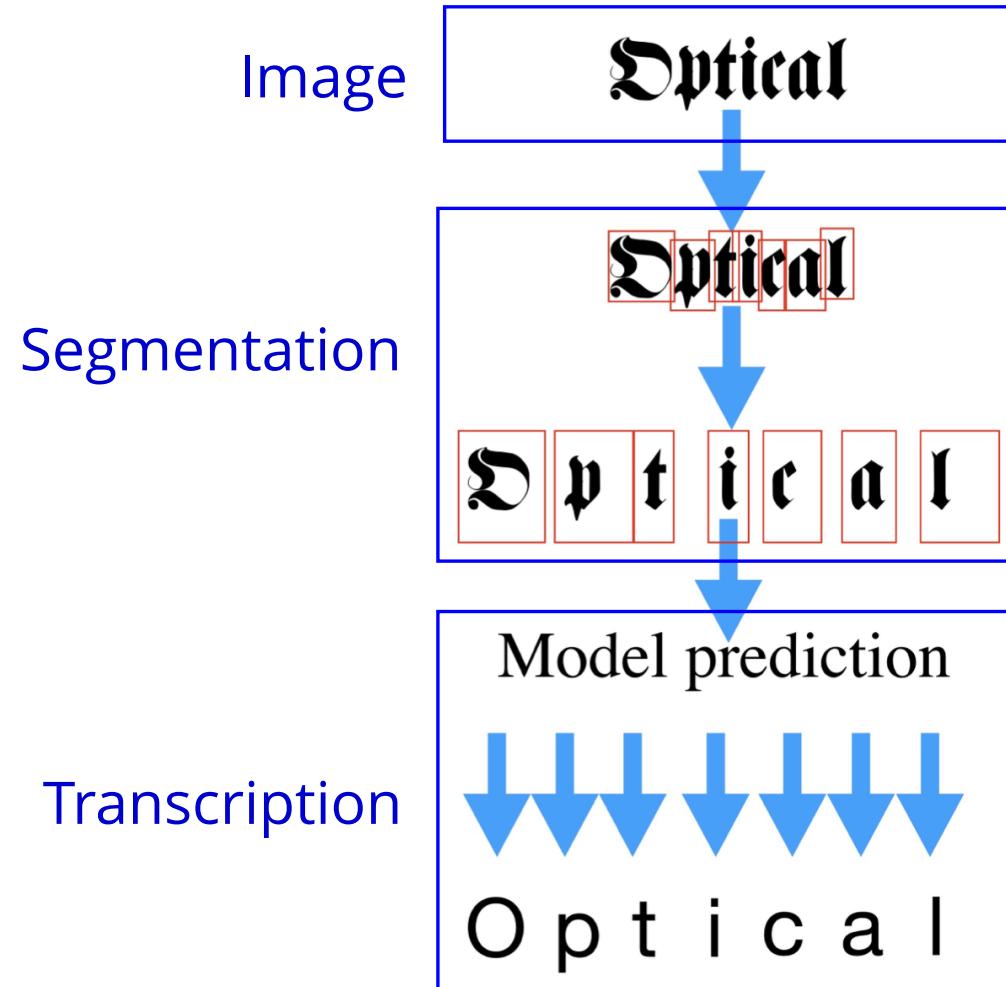
OCR

- reconnaissance optique de caractères
- technique généralement appliquée aux textes **dactylographiés**

Processus qui consiste à convertir un ensemble de signes graphiques, le plus souvent alphanumériques (mais aussi les ponctuations, espacements...), encodés sous la forme d'une image, en mode texte. L'OCR désigne à la fois un **processus** (d'OCR) et un **logiciel** (d'OCR).

(Camps & Perreux, 2021)

OCR

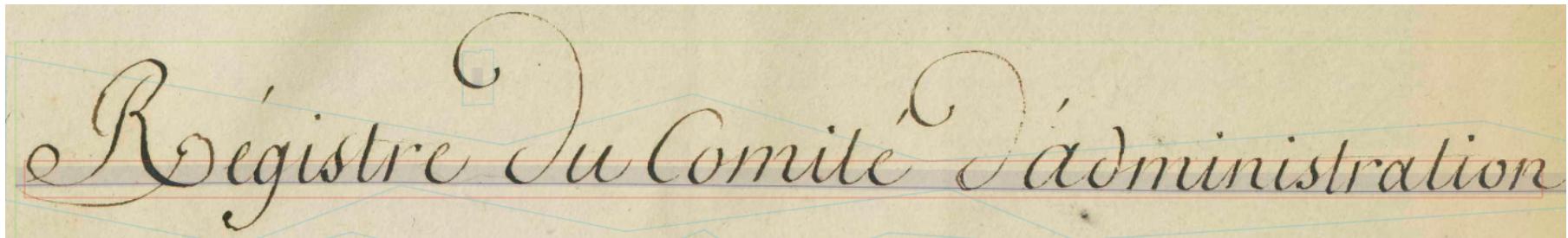


Mécanisme de l'OCR (adapté de [Gabay & SciCoS, 2021](#)).

- reconnaissance de l'écriture manuscrite (angl. *handwritten text recognition*)
- technique applicable aux textes **manuscrits**, mais aussi **dactylographiés**

Comment peut-on segmenter les caractères en écritures cursives, étant donné que les mêmes sont écrits attachés de manière fluide ? Dans ce cas, la segmentation au niveau des caractères s'avère impossible.

(adapté de Gabay & SciCoS, 2021)



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

Cas de figure

1. On part de zéro

- créer (ou collecter) des données d'entraînement
- entraîner un ou des modèles de segmentation et/ou de transcription
- appliquer le(s) modèle(s) entraîné(s) aux images

2. On a déjà un modèle de segmentation et/ou de transcription

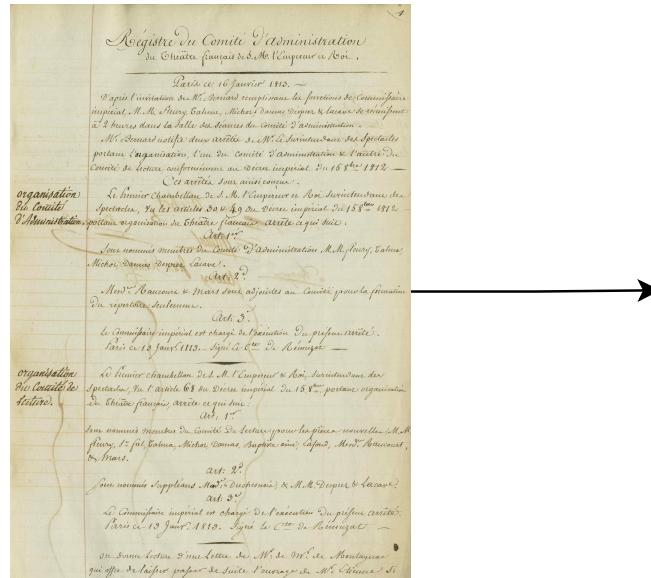
- appliquer le(s) modèle(s) existant(s) aux images et vérifier le résultat
- on vise l'exhaustivité, sans pour autant attendre une transcription parfaite
- si le modèle n'est pas suffisamment performant, on bascule dans le 1.
 - ré-entraînement du modèle

1. Entraîner un modèle

- créer des **données d'entraînement** (« vérité terrain », angl. *ground truth*)
 - transcriptions obtenues par l'annotation manuelle d'un échantillon de différentes pages de documents
- entraîner un modèle sur les pages annotées
- appliquer le modèle aux pages non vues lors de l'entraînement (**données de test**)
 - *in-domain* : données provenant des mêmes données que celles présentes dans les données d'entraînement (p. ex. d'autres lignes d'un même document imprimé)
 - *out-of-domain* : données provenant d'une source différente de celles présentes dans les données d'entraînement (p. ex. d'autres lignes d'un autre document imprimé)
- détecter les lignes de textes et effectuer une transcription automatique

(Gabay, 2020)

2. Workflow (pipeline) classique d'HTR



1. Acquisition des données

- Import des documents
- Pré-traitement (optionnel)

2. Analyse de la mise en page

- Détection des zones de texte
- Classification des zones de texte
- Ordre de lecture du texte

modèle de segmentation

Region 1

- organisation
- di lomcire
- d'Aeministration
- organisation
- dir coutié de
- Lecture.

Region 2

- 1
- 2 Registre du comité d'administration
- 3 du cheâtre français de S. M. l'Empereur et Roi.
- 4
- 5 Paris ce 16 Janvier 1813.
- 6 D'après l'invitation de Mr. Dernard remplissant les fonctions de commissaire
- 7 Eimpercal, M. M. fleury valina, Michor damas, Despiez & lacare se remussent
- 8 la 2 heures dans la Salle des séances du comité d'administation.
- 9 Mr. Derard notifie deux arrêtés de Mr. Le surintendant des spectautes
- 10 portant l'aganisation, l'un du comité d'asministration & l'autre du
- 11 comité de Lecture conformément au Decret impérial du 15 8be 1812
- 12 Ces arrêtés sont ainsi concus

3. Reconnaissance de texte

- Identification du modèle
- Extraction du texte par zone d'intérêt
- Post-traitement (correction du texte)

modèle de transcription

Chaîne de traitement d'HTR adaptée de Vidal-Gorène (2023).

2.1 Acquisition des données

- téléchargement des documents disponibles en ligne (format PDF, JPG, PNG...)
- numérisation des documents non disponibles en ligne (idéalement, scans PDF)
- import des documents (avec ou sans pré-traitement d'images)

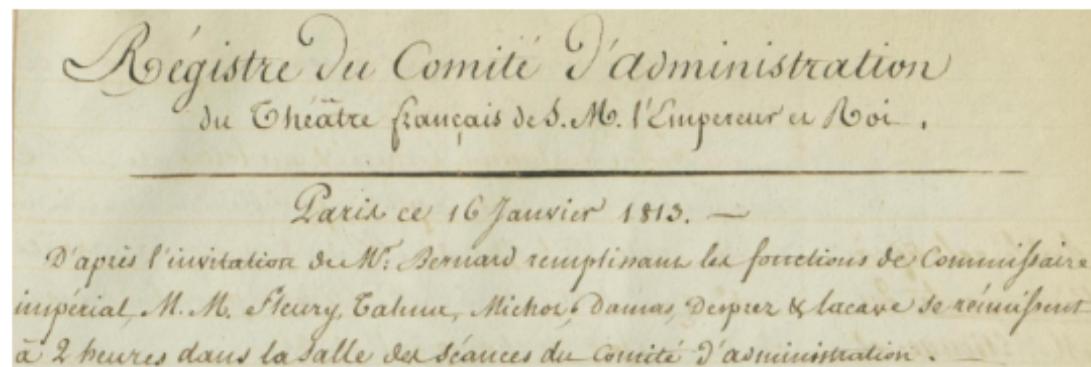


Image originale.

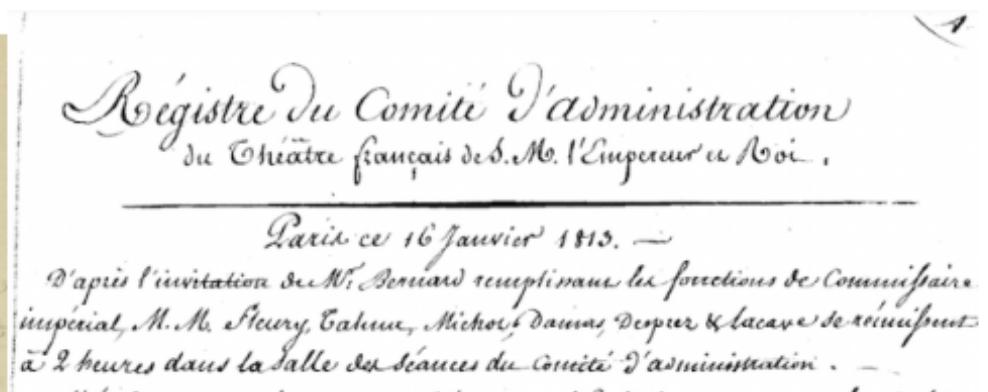
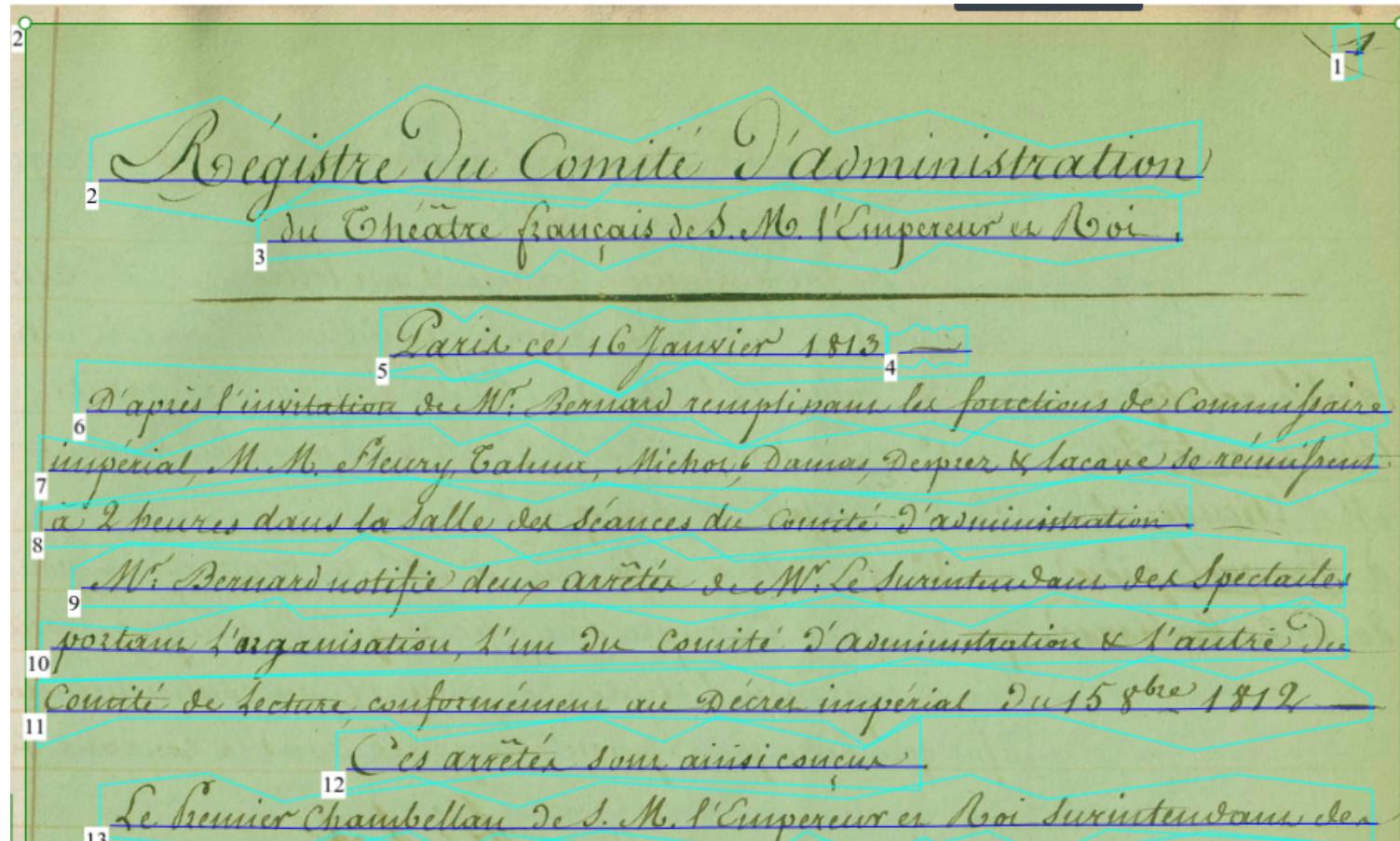


Image pré-traitée (binarisation).

2.2 Analyse de la mise en page

- * ligne de base (angl. *baseline*) : polyligne sous la ligne de texte manuscrite
- zones (régions) de texte



2.3 Reconnaissance de texte

- le résultat de la transcription automatique de texte n'est pas parfait

Region 1	
1	organisation
2	di lomcite
3	d'Aeministration
4	organisation
5	dir coutité de
6	Lecture.
Region 2	
1	
2	Registre du comité d'administration
3	du cheâtre français de S. M. l'Empereur et Roi.
4	
5	Paris ce 16 Janvier 1813.
6	D'apriés l'invitation de Mr. Dernard remplissant les fonctions de commissaire
7	Eimpercal, M. M. fleury valina, Michor damas, Despiez & lacare se remussent
8	la 2 heures dans la Salle des séances du comité d'administiation.
9	Mr. Derard notifie deux arrêtés de Mr. Le surintendant des spectaites
10	portant l'aganisation, l'un du comité d'asministration & l'autre du
11	comité de Lecture conformément au Decret impérial du 15 8be 1812
12	Ces arretés sont ainsi concus

2.3 Résultat de transcription

Back Save 0 unsaved changes Registre_R416 - #7 < 7 482 > In Progress ...

Region 1

- 1 organisation
- 2 di lomcrite
- 3 d'Aeministration
- 4 organisation
- 5 dir coutié de
- 6 Lecture.

Region 2

- 1
- 2 Registre du comité d'administration
- 3 du cheâtre français de S. M. l'Empereur et Roi.
- 4
- 5 Paris ce 16 Janvier 1813.
- 6 D'apriés l'invitation de Mr. Dernard remplissant les fonctions de commissaire
- 7 Eimpercal, M. M. fleury valina, Michor damas, Despiez & lacare se remussent
- 8 la 2 heures dans la Salle des séances du comité d'administiation.
- 9 Mr. Derard notifie deux arrêtés de Mr. Le surintendant des spectaite
- 10 portant l'aganisation, l'un du comité d'asministration & l'autre du
- 11 comité de Lecture conformément au Decret impérial du 15 8be 1812
- 12 Ces arretés sont ainsi concus

Logiciels OCR · HTR

- [Transkribus](#)
 - [Kraken + interfaces eScriptorium](#)
 - [Tesseract](#)
 - [Calamari](#)
 - [ocular](#)
- ...

Cf. la [liste](#) exhaustive des outils de transcription de textes créée par [awesome-OCR](#).

2 Transkribus

Transkribus

page web

- outil d'IA pour la reconnaissance automatique de texte et l'édition numérique des documents d'archives (documents historiques)
- potentiel tant pour la recherche que pour l'archivage

Transkribus Lite	Transkribus Expert Client
interface intuitive (navigateur web)	téléchargement requis
fonctionnalités de base	fonctionnalités avancées
interface en différentes langues (dont français)	interface en anglais

Envergure du logiciel

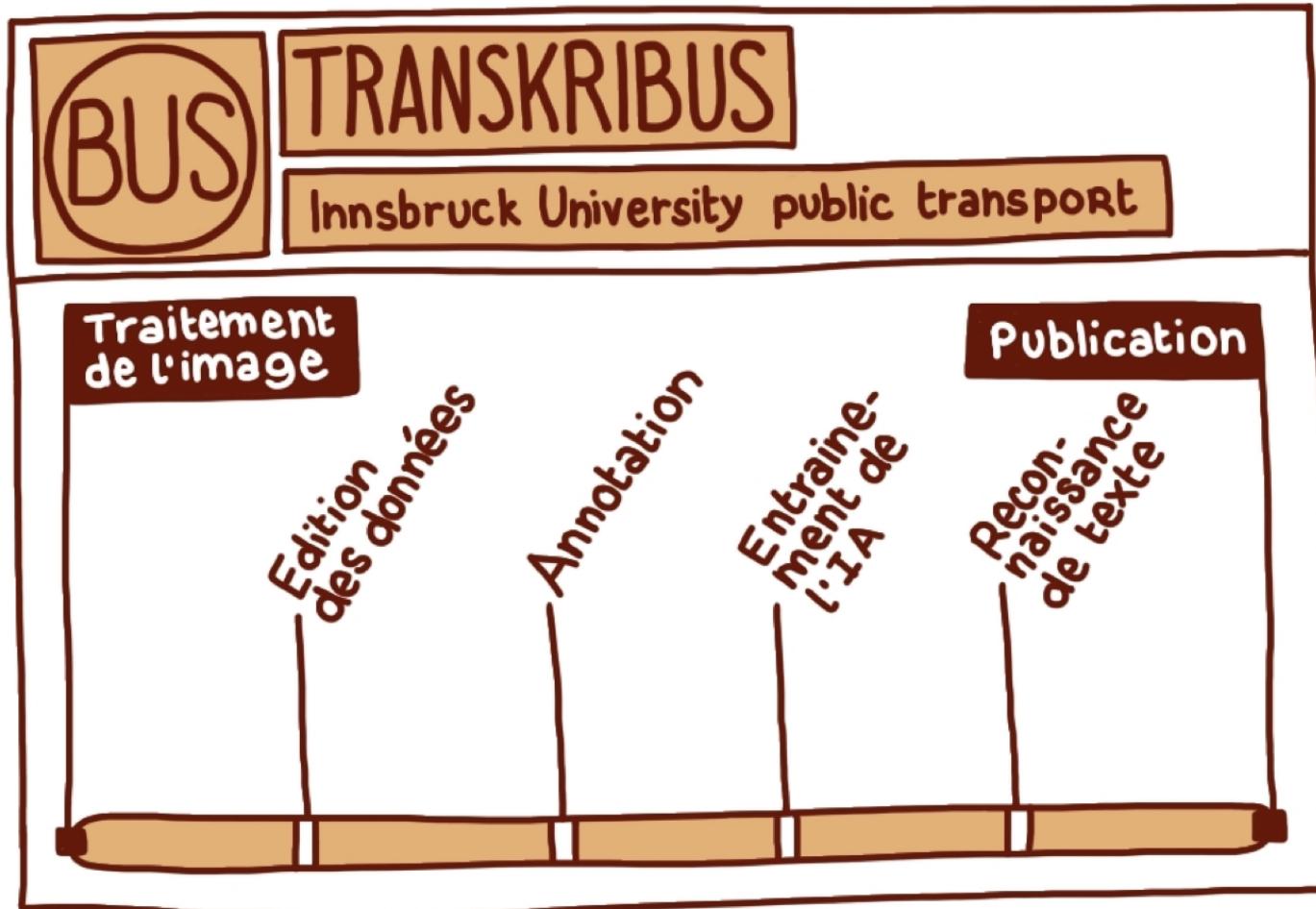


Illustration reprise de [Gautier et al. \(2022\)](#).

Choix du modèle de transcription

The screenshot shows the Transkribus interface with three examples of transcription models:

- Example 1:** ANNNUAIRES_PROPRIETAIRES_ADR_PARIS_1898_1923 - PRINT
Annuaire des Propriétaires Adressés à Paris, 1898-1923
By: Gabriela Elgarrista, Frédérique Mélanie-Becquet (LATTICE CNRS), Carmen Brando (EHESS)
French x Latin alphabet 0.30 (CER)
VIEW & TRY MODEL
- Example 2:** ID: 37758
en leur compagnie. Nous causons souvent
us. Je dois leur dire comment vous donnez
TRANSKRIBUS FRENCH MODEL 1 - HANDWRITTEN
French – General Model
By: Transkribus Team
French Latin alphabet 7.8 (CER)
VIEW & TRY MODEL
- Example 3:** ID: 20124
Donné à Fontainebleau au mois d'Octobre 1691.
Rejeté au Parlement de Tournay le 26 Février 1692.
LOUIS, PAR LA GRACE DE DIEU, ROI DE FRANCE ET DE NAVARRE :
s brefs & à venir. SALUT. Nous avons par notre Ordinance
FRENCH_18THC_PYLIA - PRINT
French 18th century print
By: Entangled Histories project
French Latin alphabet 0.91 (CER)
VIEW & TRY MODEL

Recherche du **modèle** public de Transkribus :

- 120 modèles
- 20 langues
- XI^e-XXI^e siècle
- imprimés, manuscrits ou dactylographiés
- 26 alphabets
- *CER (angl. character error rate) : 0% – 13%*
→ taux d'erreur de caractère
- 🚩 combien de % des caractères avaient été mal transcrits ?

3 Mise en pratique

Étapes d'utilisation de Transkribus

0. créer un compte sur READ-COOP (**pré-requis**)
1. importer des documents
2. effectuer des analyses de mise en page
3. transcrire automatiquement des documents à l'aide de modèles existants
4. entraîner un modèle spécifique pour vos documents (**vérité terrain nécessaire**)
5. exporter des transcriptions dans différents formats

Capsules vidéo

▶ playlist YouTube

YouTube FR

Rechercher

Accueil

Shorts

Abonnements

Vous

Historique

IHN · L1HN001 | HTR en pratique : Transkribus Lite

Ljudmila Petkovic

6 vidéos 69 vues Dernière modification le 9 nov. 2023

Tout lire Aléatoire

0. Créer un compte sur READ-COOP
Ljudmila Petkovic • 30 vues • il y a 2 mois
1:12

1. Télécharger un document de Gallica
Ljudmila Petkovic • 32 vues • il y a 2 mois
1:01

2. Importer des documents
Ljudmila Petkovic • 22 vues • il y a 2 mois
1:30

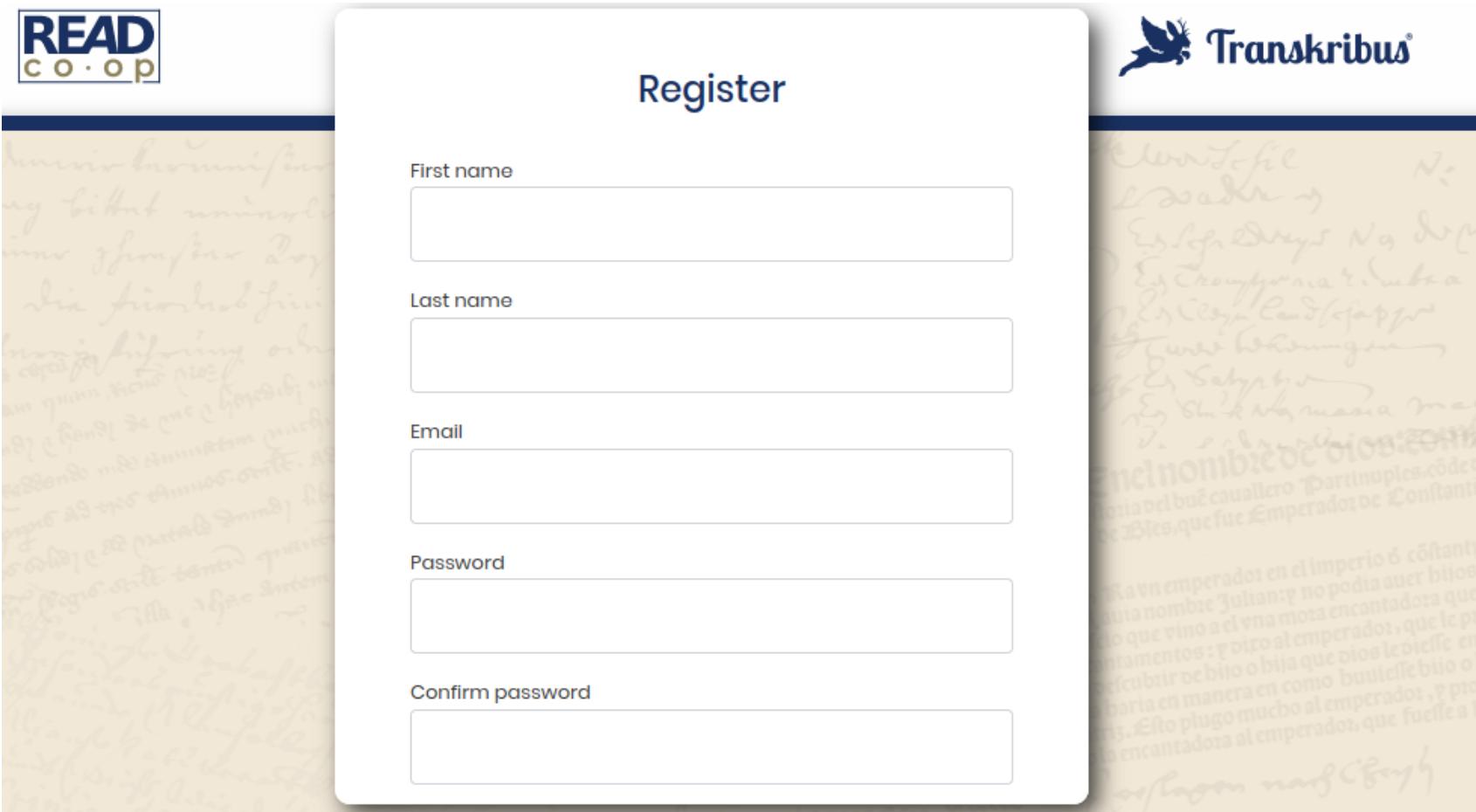
3. Analyse de mise en page + transcription automatique
Ljudmila Petkovic • 27 vues • il y a 2 mois
1:44

4. Export des transcriptions
Ljudmila Petkovic • 19 vues • il y a 2 mois
0:25

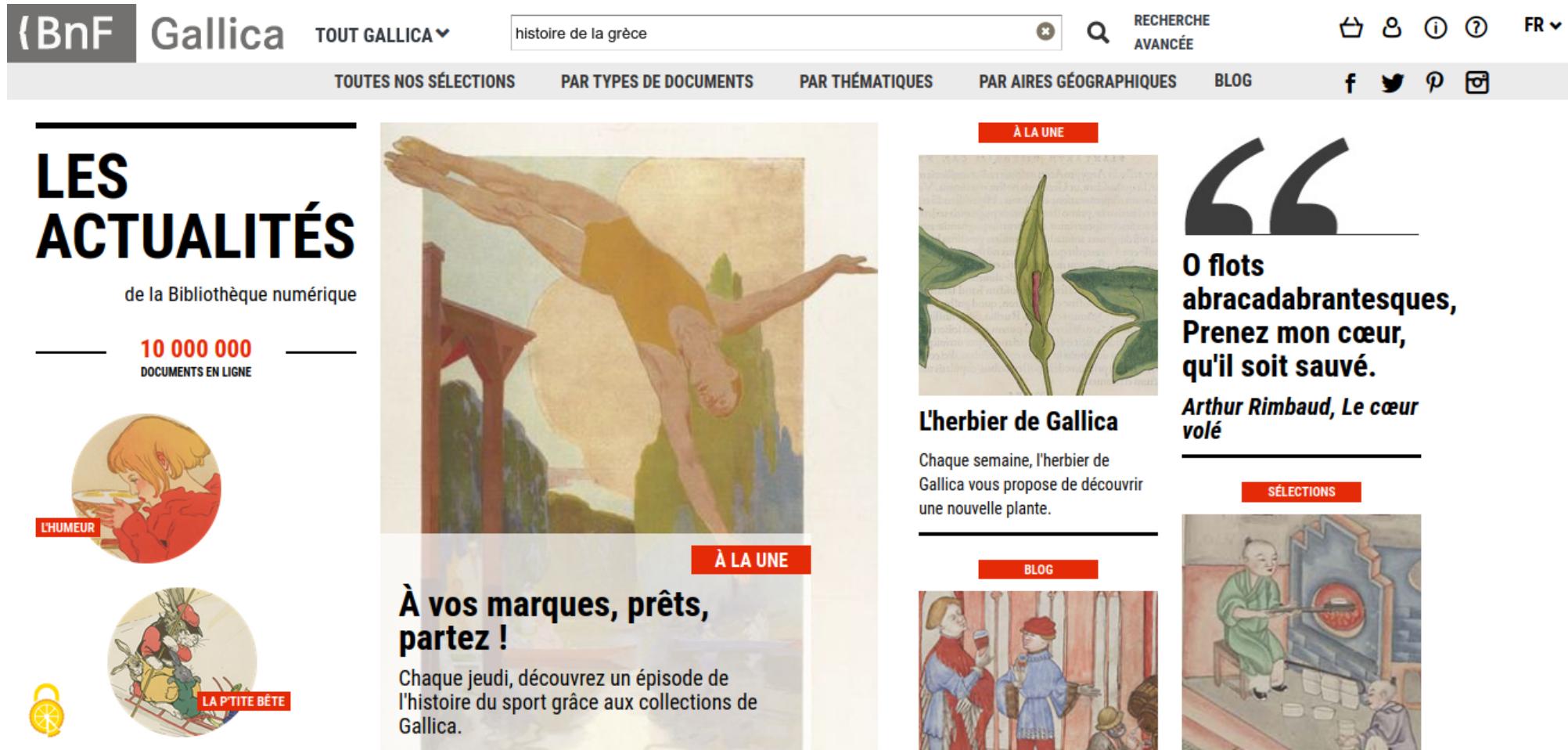
5. Recherche et comptage des mots
Ljudmila Petkovic • 24 vues • il y a 2 mois

0. Crée un compte sur READ-COOP [lien YouTube](#)

- inscription et connexion sur le [site web](#) de READ-COOP



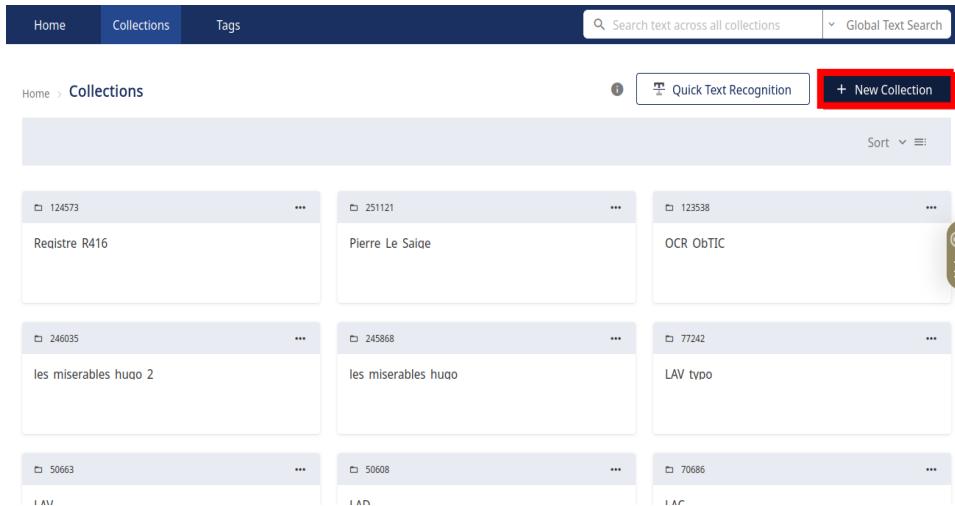
* Télécharger un document de Gallica lien YouTube



The screenshot shows the Gallica digital library homepage. At the top, there's a search bar with the query "histoire de la grèce". Below the search bar are navigation links: TOUT GALLICA, RECHERCHE AVANCÉE, and a language switcher (FR). The main content area features several sections:

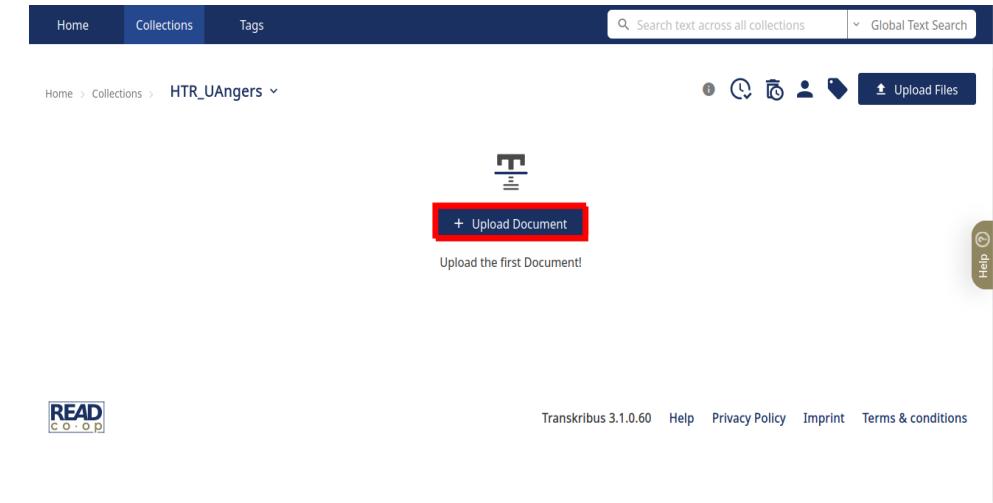
- LES ACTUALITÉS**: A large section with a red banner stating "10 000 000 DOCUMENTS EN LIGNE". It includes two circular icons: one with a girl eating (L'HUMEUR) and another with a rabbit (LA PTITE BÊTE).
- À LA UNE**: A large image of a painting of a diver. Below it, a red banner says "À vos marques, prêts, partez !". A text box below the banner reads: "Chaque jeudi, découvrez un épisode de l'histoire du sport grâce aux collections de Gallica."
- À LA UNE**: An image of a plant.
- L'herbier de Gallica**: A text box stating: "Chaque semaine, l'herbier de Gallica vous propose de découvrir une nouvelle plante."
- À LA UNE**: An image of a person working at a furnace.
- O flots abracadabrantiques, Prenez mon cœur, qu'il soit sauvé.**: A quote by Arthur Rimbaud from his poem "Le cœur volé".
- SÉLECTIONS**: An image of a person working at a furnace.

1. Importer des documents lien YouTube



The screenshot shows the 'Collections' section of the Transkribus interface. At the top, there are navigation tabs for 'Home', 'Collections' (which is selected), and 'Tags'. A search bar at the top right contains the placeholder 'Search text across all collections' and a 'Global Text Search' dropdown. Below the search bar are two buttons: 'Quick Text Recognition' and '+ New Collection', with the latter being highlighted by a red box. The main area displays a grid of collection cards. Each card includes a small thumbnail, a unique identifier (e.g., 124573, 251121, 123538), the collection name (e.g., 'Registre R416', 'Pierre Le Saïque', 'OCR ObTIC'), and three vertical ellipsis dots for more options. A 'Sort' dropdown menu is visible at the top of the list.

Créer une collection.



The screenshot shows the 'HTR_UAngers' collection page within Transkribus. The top navigation bar is identical to the previous screenshot. The main content area features a large, central 'Upload Document' button with a blue background and white text, which is also highlighted by a red box. Above the button, there is a placeholder text 'Upload the first Document!'. In the bottom left corner, the 'READ co-op' logo is displayed. At the very bottom of the page, a footer navigation bar includes links for 'Transkribus 3.1.0.60', 'Help', 'Privacy Policy', 'Imprint', and 'Terms & conditions'.

Y importer les documents.

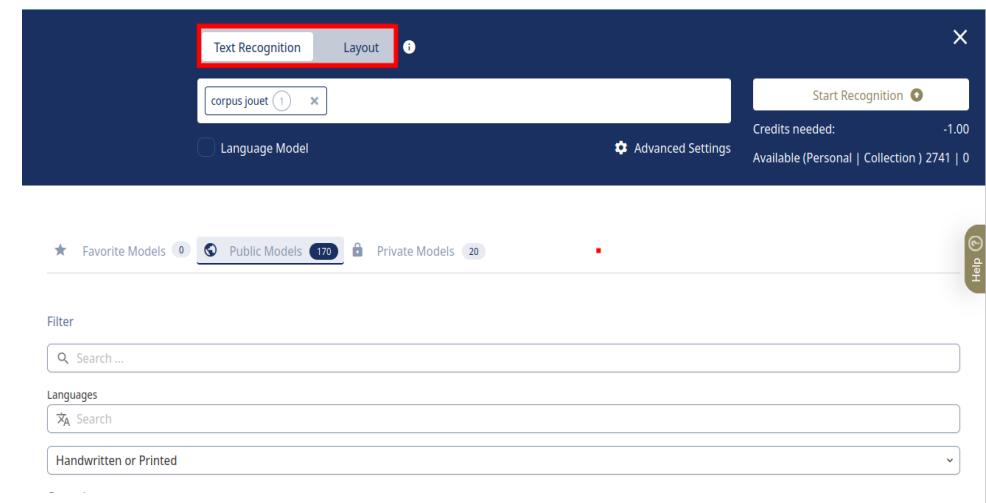
2-3. Analyse de mise en page + transcription automatique

Lien YouTube



The screenshot shows the Transkribus web interface. At the top, there's a navigation bar with 'Home', 'Collections' (which is selected), and 'Tags'. Below the navigation is a search bar with 'Search text across all collections' and 'Global Text Search'. The main area shows a collection named 'corpus jouet'. There are five pages displayed as thumbnails, with the fifth one checked. Below the thumbnails are buttons for 'Recognize' (which is highlighted with a red box) and 'Train Model'. At the bottom, there's a 'READ co-op' logo and links for 'Transkribus 3.1.0.60', 'Help', 'Privacy Policy', 'Imprint', and 'Terms & conditions'.

Module de transcription.



The screenshot shows the Transkribus interface focusing on the 'Text Recognition' and 'Layout' tabs, which are highlighted with a red box. A search bar contains the text 'corpus jouet'. Below it is a 'Language Model' section with a checkbox. On the right, it says 'Credits needed: -1.00' and 'Available (Personal | Collection) 2741 | 0'. At the bottom, there are sections for 'Favorite Models' (0), 'Public Models' (170), 'Private Models' (20), and filters for 'Search' and 'Handwritten or Printed'.

Segmentation et transcription.

2.3 Liste des processus

💡 Pensez à rafraîchir la page pour voir si la transcription s'est terminée.

The screenshot shows the Transkribus web application interface. At the top, there is a navigation bar with links for 'Home', 'Collections', 'Tags', and a search bar. On the right side of the top bar are icons for 'Desk', 'Models', 'Sites', 'Jobs', and a user profile. Below the top bar, the main content area has a breadcrumb navigation 'Home > Jobs'. There are three search input fields: one for general text search, one for 'Search' (with a magnifying glass icon), and one for 'Job ID...' (with a magnifying glass icon). A table lists various jobs with columns for 'Title', 'Type', 'User', 'State', 'Date created', and 'Date started'. The table includes several entries for 'Text Recognition' and 'Create Document' tasks, all completed by the user 'ljudmila.petkovic@gmail.com'. A 'Help' button is located on the right side of the table.

Title	Type	User	State	Date created	Date started
Registre_R416	Text Recognition	ljudmila.petkovic@gmail.com	WAITING	Jan 31, 2024, 15:04	
GrosFichiers - PUECH	Text Recognition	ljudmila.petkovic@gmail.com	FINISHED	Nov 22, 2023, 14:20	Nov 22, 2023, 14:
GrosFichiers - PUECH	Text Recognition	ljudmila.petkovic@gmail.com	FINISHED	Nov 22, 2023, 14:15	Nov 22, 2023, 14:
GrosFichiers - PUECH	Text Recognition	ljudmila.petkovic@gmail.com	FINISHED	Nov 22, 2023, 14:13	Nov 22, 2023, 13:
GrosFichiers - PUECH	Create Document	ljudmila.petkovic@gmail.com	FINISHED	Nov 22, 2023, 14:10	Nov 22, 2023, 14:
Registre_R416	Export Document	ljudmila.petkovic@gmail.com	FINISHED	Nov 13, 2023, 15:04	Nov 13, 2023, 15:

5. Export des transcriptions lien YouTube

Back Save 0 unsaved changes Registre_R416 - #7 7 482 In Progress ...



The screenshot shows a historical document page from 'Registre_R416 - #7' with handwritten text and numbered annotations. To the right, the transcription is presented in two regions:

Region 1:

- organisation
- di lomcite
- d'Aeminstiration
- organisation
- dir coutité de
- Lecture.

Region 2:

- Registre du comité d'administration
- du cheâtre français de S. M. l'Empereur et Roi.
-
- Paris ce 16 Janvier 1813.
- D'apriés l'invitation de Mr. Dernard remplissant les fonctions de commissaire
- Eimpercal, M. M. fleury valina, Michor damas, Despiez & lacare se remussent
- la 2 heures dans la Salle des séances du comité d'administiation.
- Mr. Derard notifie deux arrêtés de Mr. Le surintendant des spectaites
- portant laganisation, l'un du comité dasministration & l'autre du

Search: search Tout surligner Respecter la casse Respecter les accents et diacritiques Mots entiers X

6. Recherche et comptage des mots [lien YouTube](#)

- ces fonctionnalités ne sont pas disponibles dans Transkribus lite
- pour une utilisation plus avancée, télécharger Transkribus Expert Client

Une alternative à Transkribus : *eScriptorium*

code source

Plusieurs instances :

- [SCAI](#) (*Sorbonne Center for Artificial Intelligence*) de Sorbonne Université
- [FONDUE](#) (Université de Genève)
- [Inria](#) Paris

etc.

NB : Pour accéder à l'interface eScriptorium, il faut disposer d'un compte individuel.

Étape(s) suivante(s)

La transcription automatique de textes n'est pas une fin en soi.

- édition numérique
 - fouille de textes (analyse stylistique, modélisation de sujets, etc.)
 - établissement d'un moteur de recherche
- ...

4 Projets d'HTR

Patrimonialisation numérique

Acteurs majeurs de patrimonialisation numérique dans le domaine d'HTR :

- institutions patrimoniales (les *GLAM*) : BnF, Archives nationales, Bodmer Lab...
- institutions d'ESR : ENS, EPHE, École des Chartes, INRIA...

(Gautier *et al.*, 2022)

Bibliothèque nationale de France (BnF) transcrit systématiquement ses contenus imprimés et les rend disponibles dans sa bibliothèque numérique Gallica.



Interface de Gallica, exemple d'un document historique OCRisé.

Projets et infrastructures

- [OCR-D](#) OCR pour les documents historiques imprimés en allemand
 - [Archives municipales de Belfort](#) OCR · HTR en collaboration avec [TEKLIA](#)
 - [HIMANIS](#) angl. *H*istorical *M*ANuscript *I*ndexing for user-controlled *S*earch
 - [LECTAUREP](#) **L**ECTure **A**utomatique de **R**EPERTOIRES
 - [CREMMA](#) Consortium Reconnaissance d'Écritures Manuscrites des Matériaux Anciens
 - [FoNDUE](#) **F**ORMes **N**umerisees et **D**etection **U**nifiee des **E**critures
- ...

?

Questions et discussion libre

Références

- **Anderson, M.** (2021). « [OCR algorithms: a complete guide](#) ». itransition.
- **Barbier, J., & Mandret-Degeilh, A.** (2018). 8. Les archives numériques et numériséS. Dans *Le travail sur archives*.
- **Camps, J.-B. & Perreaux, N.** (2021). [Reconnaissance optique des caractères et des écritures manuscrites](#). Projet E-NDP [diapositives].
- **Chagué, A.** (2021). « [Comment faire lire des gribouillis à mon ordinateur ?](#) » [diapositives]. Inria Paris (ALMAnaCH).
- **Canali, L.** (2016). [A neural network scoring engine in PL/SQL for recognizing handwritten digits](#) [blog]. CERN Database Services.
- **Cendre, R.** (2021). [Classification par méthodes d'apprentissage supervisé et faiblement supervisé d'images multimodales pour l'aide au diagnostic du lentigo malin en dermatologie](#) [thèse]. Université de Bourgogne.
- **Davalon, J.** (2014). [Une patrimonialisation des archives ?](#). Dans *L'archive dans quinze ans*.
- **Gabay et SciCoS** (2021). [FoNDUE \(FOrmes Numerisees et Detection Unifiee des Ecritures\) · An Handwritting Text Recognition infrastructure for Geneva](#). Université de Genève.
- **Gabay, S.** (2020). « [OCR](#) ». Formation OCR/*GROBID-dictionaries* 2020, Strasbourg.
- **Lateef, Z.** (2023). « [What Is A Neural Network? Introduction To Artificial Neural Networks](#) ». edureka!.
- **Lehman, J. & Clune, J.** (2014). [An Anarchy of Methods: Current Trends in How Intelligence Is Abstracted in AI](#). *Intelligent Systems, IEEE* 29(6): 56-62.
- **Lepron, S.** (2023). « [Archivage, stockage, quelles différences ?](#) ». Ab Antiquo.

Références II

- **Galleron, I.** (2021). « Introduction aux humanités numériques (L1HN001) [*diapositives en interne*]. Sorbonne Nouvelle.
- **Gautier, D., Huguet, A., Massot, M.-L., Tricoche, A., Carlin, M., Moreux, J.-P., & Rostaing, A.** Compte-rendu de la journée d'étude « Point HTR 2022 » *Transkribus / eScriptorium : Transcrire, annoter et éditer numériquement des documents d'archives* [*rapport de recherche*]. Bibliothèque nationale de France – Site François-Mitterrand. CAPHES - UMS 3610 CNRS/ENS; AOROC.
- **Moufflet, J.-F.** (2021). *5 ans d'expérimentation de la technologie HTR aux Archives nationales · Quels usages pour le futur ?*. Futurs Fantastiques – BnF.
- **Ressources de géographie pour les enseignants.** (2019). « Patrimonialisation ». Géoconfluences, ENS-Lyon.
- **Scheithauer, H.** (2023). *Acquisition, intégration et redistribution de données structurées dans les GLAM : harmonisation des pratiques*. [*thèse*] PSL · EPHE.
- **Toumit, J. Y., Emptoz, H., Jean-Michel, R. O. Y., & Drézen, F.** (2000). *Le texte mathématique : Du document papier... à la version HTML*. Dans *CIFED'2000 : colloque international francophone sur l'écrit et le document* (p. 319). PPUR presses polytechniques.
- **Vidal-Gorène, C.** (2023). *La reconnaissance automatique d'écriture à l'épreuve des langues peu dotées*. Programming Historian.