

# Comment faire lire des gribouillis à mon ordinateur ?

Alix Chagué

► **To cite this version:**

| Alix Chagué. Comment faire lire des gribouillis à mon ordinateur ?. 2021. hal-03170345

**HAL Id: hal-03170345**

**<https://hal.archives-ouvertes.fr/hal-03170345>**

Submitted on 16 Mar 2021

**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.



ATIONS, SITUATIONS ET PRIX DES BIENS  
An 1872 mois de Janvier



BY

tuto@mate  
11/03/2021

# Comment faire lire des gribouillis à mon ordinateur ?

Alix Chagué  
Inria Paris (ALMAnaCH)

Procuration	Notar. major. Demourant a Paris, boulevard de la Ville-Aux-Loyaux, n° 26, le 27 Janvier 1872.	26	2.40
	Didier (par son avoué) contre le sieur Didier, demeurant à Paris, rue de Valenciennes, n° 14, en faveur pour lui-même.	27	16.40
Meinlovici	Blange (par M. Charles Edmond) signifié à l'adresse de l'ancien propriétaire, n° 14, rue de Valenciennes, le 27 Janvier 1872.	27	2.40
Notar. de Bourdon	Bourdon (notar. de Bourdon) signifié à l'adresse de l'ancien propriétaire, n° 14, rue de Valenciennes, le 27 Janvier 1872.	27	2.40

# ALMAnaCH contribue au projet eScriptorium (SCRIPTA-PSL)



*Inria*

ALMAnaCH

projet  
DAHN

projet  
CREMMA

projet  
LECTAUREP

- ❖ Transcription automatique de documents d'archive
- ❖ Édition numérique
- ❖ Indexation automatique et fouille de données
- ❖ Mise en production d'une infrastructure publique

NOMS, PRÉNOMS ET DOMICILES DES PARTIES.			INDICATIONS, SITUATIONS ET PRIX DES BIENS.	
An 1872 mois de Janvier				
du	des	NATURE ET ESPECES DES ACTES:	DATES	DES
REPertoire.	ACTES.	EN BREVETS.	EN MINUTES.	
76	29	Certificat de Propriété		
77	29			

NOMS, PRÉNOMS ET DOMICILES DES PARTIES.			INDICATIONS, SITUATIONS ET PRIX DES BIENS.	
An 1872 mois de Janvier				
du	des	NATURE ET ESPECES DES ACTES:	DATES	DES
REPertoire.	ACTES.	EN BREVETS.	EN MINUTES.	
76	29	Certificat de Propriété		
77	29			

# Définitions

NOMS, PRÉNOMS ET DOMICILES DES PARTIES.			INDICATIONS, SITUATIONS ET PRIX DES BIENS.	
An 1872 mois de Janvier				
du	des	NATURE ET ESPECES DES ACTES:	DATES	DES
REPertoire.	ACTES.	EN BREVETS.	EN MINUTES.	
76	29	Certificat de Propriété		
77	29			

NOMS, PRÉNOMS ET DOMICILES DES PARTIES.			INDICATIONS, SITUATIONS ET PRIX DES BIENS.	
An 1872 mois de Janvier				
du	des	NATURE ET ESPECES DES ACTES:	DATES	DES
REPertoire.	ACTES.	EN BREVETS.	EN MINUTES.	
76	29	Certificat de Propriété		
77	29			

# Définitions : OCR vs HTR ?

- ❖ **OCR** : Optical Character Recognition ➔ reconnaissance du texte en général (imprimé principalement)
- ❖ **HTR** : Handwritten Text Recognition ➔ reconnaissance du texte manuscrit (qui prend en compte les difficultés inhérentes à ces textes)

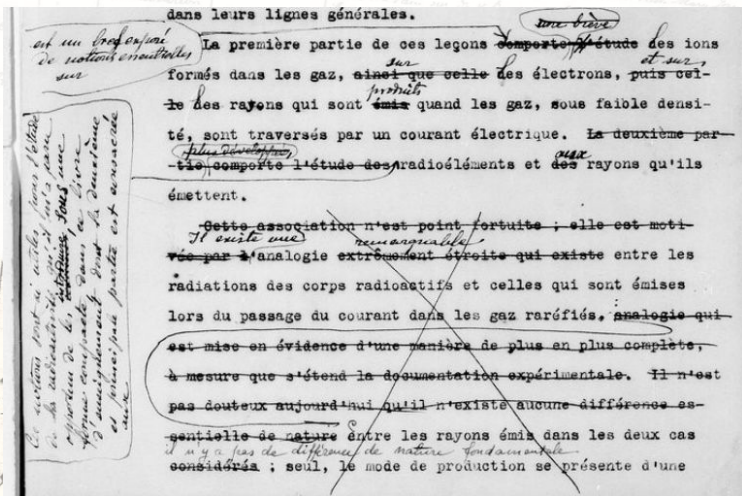
Il y avait en l'an de grâce 1845, dans ces années d'abondance et de paix où toutes les faveurs de l'esprit, du talent, de la beauté et de la fortune entouraient cette France d'un jour, une jeune et belle personne de la figure la plus charmante, qui attirait à elle, par sa seule présence, une certaine admiration mêlée de déférence pour

Etabli A. J. Mart (d. alt. rend. des. l. any  
se col. de 20.000 de la St. des) à St. Germain  
en Laye rue Stephanie Mony 10.

Entre Madame Adelaide Ague, surseinte vendutelle  
épouse assistée et autorisée du sieur Marie Camouier, et ce  
dernier, demeurant et domiciliés ensemble, à Paris, rue Saint-Denis,  
numéro cent quatre vingt-cinq; Demandeurs, Compromis,

Un logiciel « d'HTR » est capable de lire du texte imprimé !

# Définitions : OCR vs HTR ?



❖ Grandes catégories d'écriture :

- Imprimée
- "Tapuscrite"
- Manuscrite

❖ Elles peuvent être mélangées au sein d'un même document (épreuve corrigée à la main, formulaire pré-imprimé, ...)

- ❖ Différentes orientations du texte
- ❖ Différents sens de lecture du texte (GàD, DàG, HeB, BeH)
- ❖ Différents alphabets

An 8. Pracial

N° de Rép.	Date des Actes.	DÉNOMINATIONS des ACTES.	MINUTES ou BREVETS.	NOMS, PRÉNOMS et DOMICILES des PARTIES, et INDICATIONS des BIENS.	RELATION de l'Enregistrement.	
					DATE.	DROITS.
ffo	2.	Procuration	B. Randon	(Sous le nom de) Michel Pierre Lepetit, fargues, dem. à Paris, au n° 103. D'ou		

# De la REM, pour quoi faire ?

L'objectif de la **reconnaissance automatique d'écriture manuscrite (REM)** est principalement :

- ❖ d'accélérer la transcription ;
- ❖ Donc de la rendre moins pénible ;
- ❖ Donc de traiter de plus grands ensembles documentaires.

Mais la REM n'est pas une fin en soi :

- ❖ Édition numérique ?
- ❖ Analyse stylistique ou fouille de texte ?
- ❖ Accessibilité des ressources (moteur de recherche) ?
- ❖ etc.

# Quelques logiciels d'HTR



TESSERACT

Voir le recensement réalisé par Awesome-OCR :  
<https://github.com/kba/awesome-ocr>



# Schématisation du workflow

Charger une image  
matricielle

Exporter du texte

(version sur-simplifiée)

# Schématisation du workflow

Charger une image  
matricielle

Exporter du texte

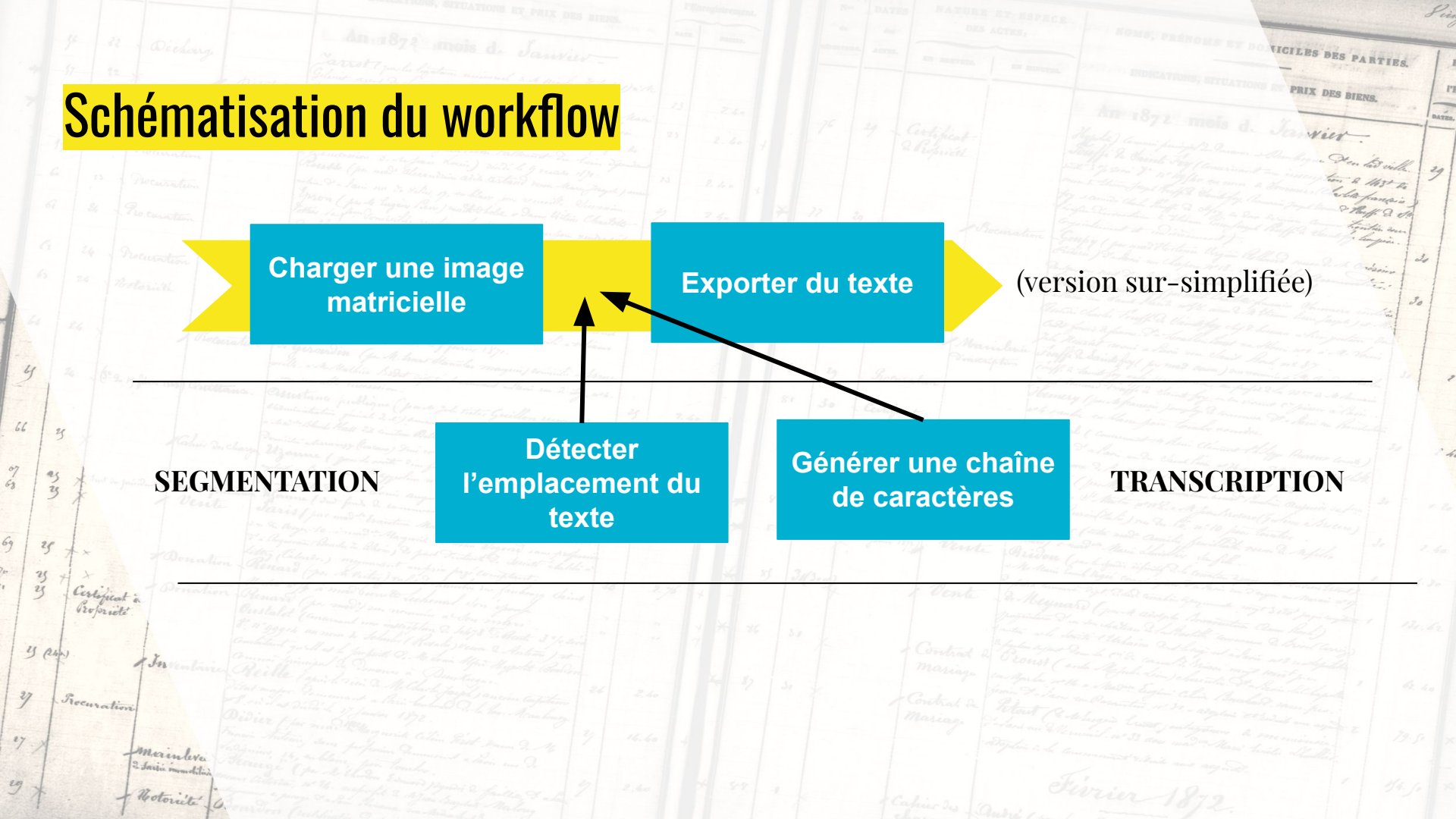
(version sur-simplifiée)

SEGMENTATION

Détecter  
l'emplacement du  
texte

Générer une chaîne  
de caractères

TRANSCRIPTION



# Schématisation du workflow

Charger une image  
matricielle

Exporter du texte

(version sur-simplifiée)

SEGMENTATION

Détecter  
l'emplacement du  
texte

Générer une chaîne  
de caractères

TRANSCRIPTION

Collecter des  
données  
d'entraînement

Entraîner un  
modèle

# Deux workflows à ne pas confondre

- A** On part de zéro :
- ❖ Il faut créer (ou collecter) des données d'entraînement (*vérité terrain*)
  - ❖ Puis entraîner un ou des modèles de segmentation et/ou de transcription)

La transcription doit être la plus parfaite possible mais pas nécessairement exhaustive

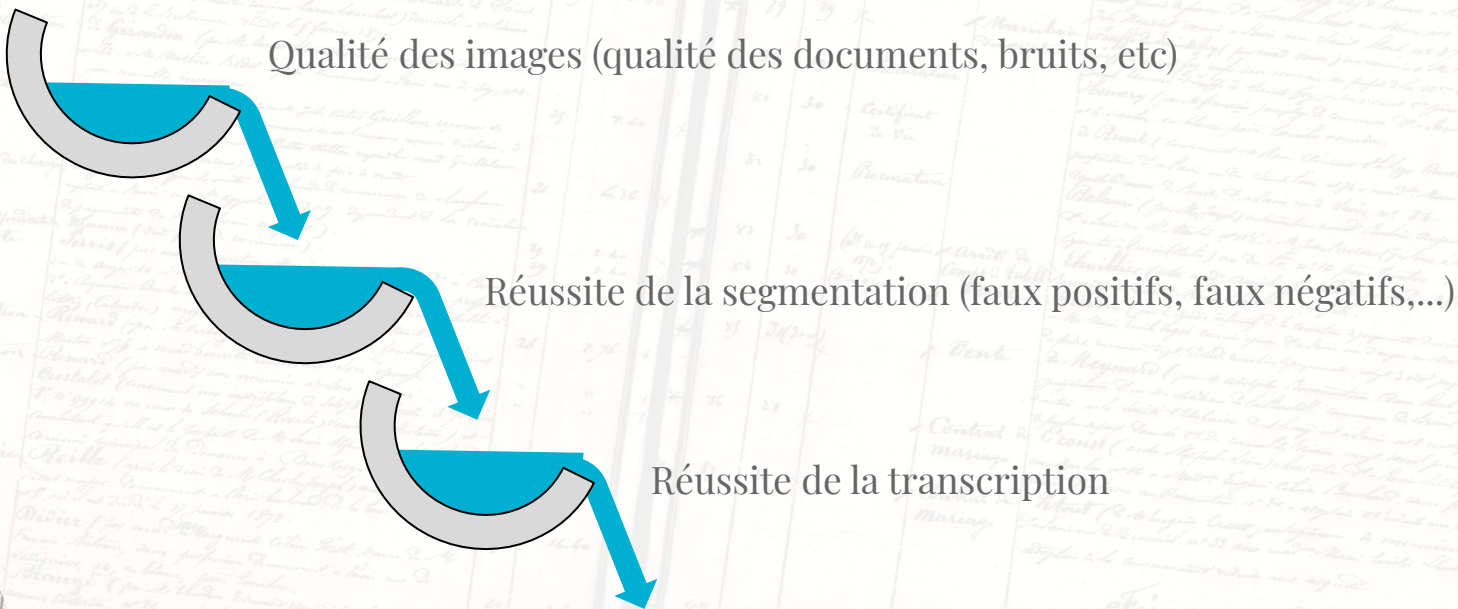
- B** On a déjà un modèle de transcription / segmentation :
- ❖ Simplement appliquer le(s) modèle(s) aux images et vérifier le résultat

On vise l'exhaustivité mais il ne faut pas nécessairement attendre une transcription parfaite

Si le modèle n'est pas suffisamment bon, on bascule dans le premier cas de figure

# Attention à ne pas négliger l'étape de la segmentation

La qualité de la segmentation est aussi importante que la transcription.



# Un gain de temps ?

La REM accélère la tâche de transcription mais pas celles :

- ❖ De la vérification des images
- ❖ De la vérification de la segmentation
- ❖ De la vérification de la transcription

=> elles ne sont pas prises en charge dans le cadre de la REM pure.

Plusieurs paramètres sont à considérer :

- ❖ La complexité des écritures et des documents ;
- ❖ Les objectifs d'exploitation de cette transcription ;
- ❖ L'existence d'autres corpus similaires déjà annotés ou de modèles correspondants au corpus.

**Il faut donc préparer la campagne d'HTR en amont en prenant en compte ces éléments !**

Ces tâches peuvent être confiées à des personnes aidées d'outils de pré/post-traitement.

Exemples :

- ScanTailor pour le découpage des images
- PySpellchecker pour le post-traitement, etc



# Pourquoi préparer une campagne

- ❖ Pour choisir le logiciel le plus adapté ;
- ❖ Pour connaître et anticiper les points de blocages ;
- ❖ Pour adapter la stratégie d'annotation aux difficultés et aux objectifs (et éviter d'inutiles retours en arrière) ;
- ❖ Pour pouvoir créer des jalons pour évaluer l'avancée de la campagne ;



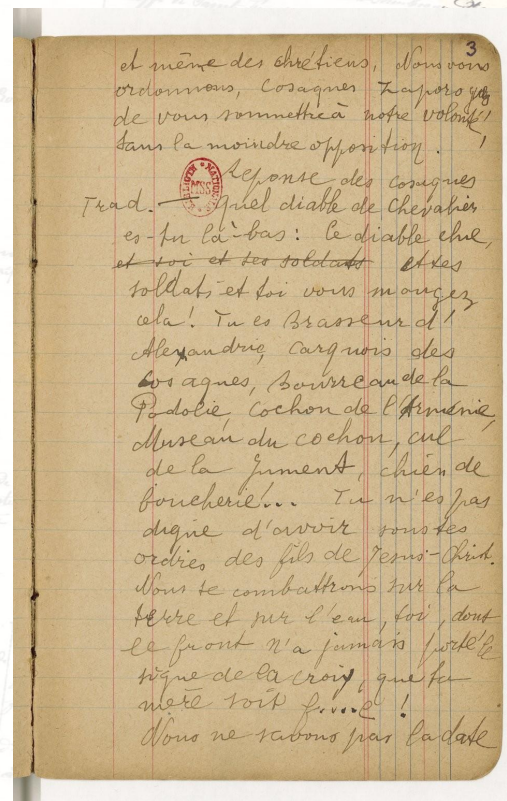
# Anticiper les difficultés : évaluer le corpus

- ❖ Quel type d'écriture ? Y a-t-il plusieurs écritures ? Combien ? Varient-elles beaucoup ?
- ❖ Quelle est la qualité des numérisations (résolution, cadrage, binarisé...) ?
- ❖ Quelle est la qualité des documents (déchirures, taches, plis, etc) ?
- ❖ Y a-t-il des abréviations dans le texte ? Y a-t-il des symboles particulier ?
- ❖ Y a-t-il autre chose que du texte (des illustrations, des dessins, des tableaux) ? Comment cela s'articule-t-il avec le texte ?
- ❖ Y a-t-il des décoration sur le texte (surlignage, soulignage, etc) dont je veux tenir compte ?
- ❖ Etc.

# Anticiper les difficultés : exemple

J'ai 1 000 pages d'un carnet personnel du XXe siècle et je veux pouvoir faire de la fouille de texte :

- ★ Des numérisations couleur en haute résolution
- ★ Une dizaine de pages gondolées et abîmées par l'eau
- ★ Une seule main d'écriture
- ★ Des abréviations et des mots mal formés
- ★ Quelques symboles "maisons"
- ★ Quelques dessins dans les marges
- ★ Des mots importants soulignés, des mots barrés



# Anticiper les difficultés : adapter les objectifs

En fonction des objectifs du projet de transcription, trier les critères :

- ❖ Ce qui dépend de la REM et ce qui n'en relève pas
- ❖ Ce qu'il est fondamentale de réussir à détecter (ce qui n'a pas besoin d'être segmenté ou transcrit)
- ❖ Quel degré de perfection est nécessaire

---

Si on reprend l'exemple :

- ★ Seul modèle
- ★ Des caractères pour transcrire les symboles maisons
- ★ Les mots soulignés ou barrés ne seront pas traités différemment
- ★ Dessins ignorés
- ★ Prévoir des difficultés à lire les mots mal formés (corriger avec un script ou à la main après)
- ★ Pages gondolées traitées à la main

# Définir des règles d'annotation

Seul-e ou en équipe : définir des règles d'annotation en amont !

- ❖ Expliciter les règles des cas “normaux” ;
- ❖ Définir des règles précises et univoques pour les cas particuliers ;
- ❖ Définir des comportements pour les cas “anormaux” imprévus ;

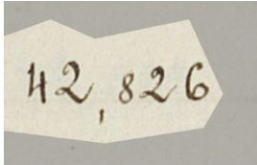
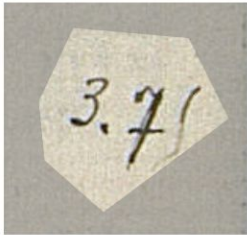
Pourquoi définir des règles ?

- ❖ Faciliter le contrôle qualité ;
- ❖ Assurer la constance de l'annotation dans le temps et entre les annotateurs ;
- ❖ Documenter l'annotation pour de futurs utilisateurs ;

# Définir des règles d'annotation : exemples

- ❖ Comment annoter les mots barrés ou illisibles ?
- ❖ Comment traiter les abréviations ? Les fautes ?
- ❖ Comment traiter les caractères spéciaux ?
- ❖ Quel tiret utiliser pour transcrire un tiret ?
- ❖ Comment distinguer manuscrit et imprimé ?

## Transcription des nombres

Exemple	Transcription	Explication
	42,826	Le scribe a employé une virgule pour séparer la partie entière et la partie fractionnaire.
	3.75	Le scribe a employé un point pour séparer la partie entière et la partie fractionnaire.
	2.20	Le scribe a employé un point pour séparer la partie entière et la partie fractionnaire, bien que l'alignement avec la base du tracé des chiffres n'est pas parfait.

# Annoter la mise en forme : exemple

- Les mots soulignés sont précédés d'un “\_”
- Les mots barrés ne sont pas distingués des autres
- Les mots illisibles sont remplacés par un “#”
- Les mots écrits à la machine sont distingués de ceux écrits à la main car ils sont systématiquement précédés d'un symbole : “·”

A l'aide d'un script, il serait facile de transformer ces indicateurs en annotations xml en phase de post-traitement.

§3. ~~#~~ Types d'appareillage-. L'intensité du courant au tra-

§3 # · Type · d'appareillage- · L'intensité · du · courant · au · tra-

# Remarques sur la gestion d'un projet de REM

**Objectif général :** obtenir la transcription de la totalité du corpus

Mais :

- Avec quel niveau de perfection ?
- Avec quels jalons ?

Quelques exemples de jalons dans le scénario B :

- ❖ Créer un set d'évaluation avancée représentatif du corpus et des difficultés identifiées (↔ sc. A & B)
- ❖ Annoter et contrôler  $x$  pages/lignes de vérité terrain (  $\times$  nombre de modèles nécessaires) (↔ sc. B)
- ❖ Entraîner un/des modèle(s) et tester la performance effective sur le set d'évaluation avancée (↔ sc. A & B)
- ❖ Développer/sélectionner les outils de post-traitement (correction/nettoyage/augmentation) (↔ sc. A & B)





# Des données intermédiaires à ne pas jeter

La vérité de terrain est un produit dérivé précieux et qui fait partie des données de la recherche !

- ❖ Elle peut être utile à d'autres projets disposant d'écritures similaires
- ❖ Plus on a de données disponibles, moins on perd de temps pour l'entraînement d'un modèle !

Un package minimal pour le partage :

- ❖ Images (ou liens vers les images)
- ❖ Transcription dans un format adapté (XML ALTO ou XML PAGE)
- ❖ Documentation sur les règles d'annotations et les écritures

# HTR-United : une solution de partage des données

- ❖ Une organisation Github pour partager vos données sous la forme de repository si vous ne savez pas où les publier
- ❖ Un point de centralisation des signalement de sets de données (un fichier de référencement dans le projet principal)

```
name: htr-unity
description: 'HTR Ground Truth Resources'
entries:
-
  title: 'tapuscorpus'
  url: 'https://github.com/HTR-United/tapuscorpus'
  description: 'Ground Truth dataset for French typewritten OCR (20th century documents)'
  language: French
  time: 1900-1999
  hands: 30
  license:
    - {name: 'CC-BY 4.0', url: 'https://creativecommons.org/licenses/by/4.0/'}
  format: 'XML-Page'
  volume:
    - {count: "150", metric: pages}
```

HTR  
United

HTR-United

Repositories Packages People Teams Projects Settings

Q Find a repository... Type Language Sort

htr-unity

HTR Ground Truth Resources for French

modern french ground-truth htr handwriting-text-recognition

xml-into

HTML 3 1 7 11 Updated 2 days ago

htr-unity.github.io

https://htr-unity.github.io/

CSS 0 0 1 10 Updated 12 days ago

timeuscorpus

Ground Truth datasets for French 10th and 19th HTR produced by the ANR project TIME US

dataset french htr

CC-BY-4.0 0 0 0 10 Updated on 21 Jan

tapuscorpus

Ground Truth dataset for French typewritten OCR

dataset french htr

CC-BY-4.0 0 0 0 10 Updated on 21 Jan

HTR  
United

<https://htr-unity.github.io/>



# Liens vers de la documentation

Tutoriel pas à pas :

<https://lectaurep.hypotheses.org/documentation/prendre-en-main-escriptorium>

Démo vidéo :

<https://escripta.hypotheses.org/escriptorium-video-gallery>



# Le projet CREMMA

Mutualiser l'investissement infrastructurel pour rendre une solution open-source comme eScriptorium plus accessible



*Inria*

ALMANaCH

+ ENC ; EFEO ;  
EPHE ; IRHT ;  
LAMOP

Financement DIM MAP (IDF)

- RAM
- GPU
- Stockage
- Maintenance

Déploiement sur  
Serveur Inria

eScriptorium  
SCRIPTA  
PSL

An 1872 mois de Janvier

		L'Eregistrement.	
		DATE	DROITS.
34	32		
37	32		
39	32	23	2.40
39	15	23	2.60
6	15	23	2.40
61	28	27	2.60

		NATURE ET ESPECE DES ACTES:		NOMS, PRÉNOMS ET DOMICILES DES PARTIES.	
		EN BREVETS.	EN MINUTES.		
76	29	Certificat de Propriété		An 1872 mois de Janvier	
77	29				

**Merci !**

		DATE	DROITS.
35	35	26	2.40
35 (24)		26	2.40
37		27	16.60
37		27	2.40

		DATE	DROITS.
35	35	26	2.40
35	35	26	2.40
37		27	16.60
37		27	2.40

Fevrier 1872