



HAL
open science

CREMMA : Une infrastructure mutualisée pour la reconnaissance d'écritures manuscrites et la patrimonialisation numérique

Alix Chagué

► To cite this version:

Alix Chagué. CREMMA : Une infrastructure mutualisée pour la reconnaissance d'écritures manuscrites et la patrimonialisation numérique. Sciences du patrimoine - sciences du texte. Confrontation des méthodes, Ecole nationale des chartes, May 2021, Paris, France. hal-03541887

HAL Id: hal-03541887

<https://inria.hal.science/hal-03541887>

Submitted on 25 Jan 2022

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.



Distributed under a Creative Commons Attribution 4.0 International License

Séminaire Sciences du Patrimoine, Sciences du Texte
Ecole nationale des chartes
Jeudi 20 mai 2021



CREMMA :

Une infrastructure mutualisée pour la reconnaissance d'écritures manuscrites et la patrimonialisation numérique

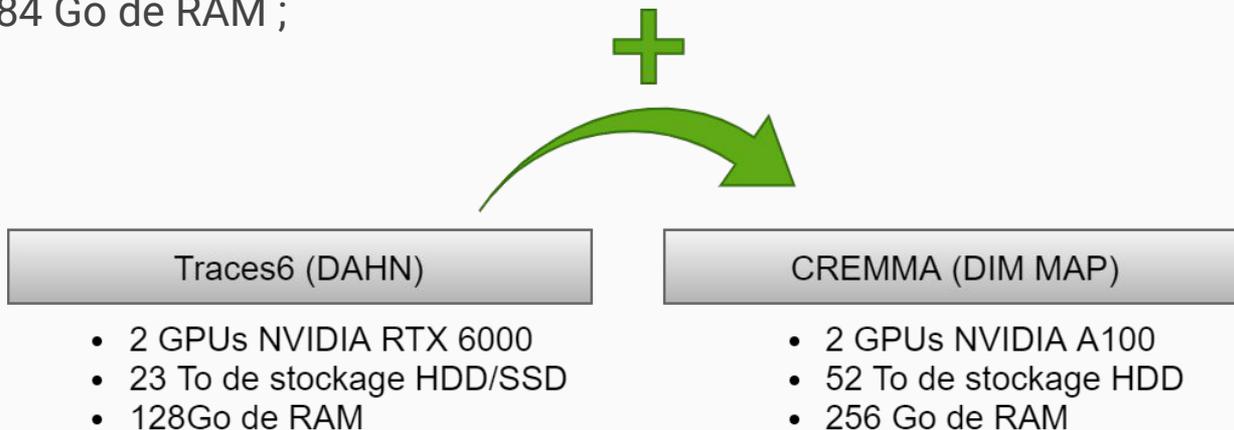


Alix Chagué
ALMAnaCH - Inria

- Consortium Reconnaissance d'Écritures Manuscrites des Matériaux Anciens
- Projet de 2 ans fondé sur le financement d'un équipement informatique par le DIM MAP
- Une communauté de partenaires d'Ile-de-France :
 - ALMAnaCH - Inria ;
 - Ecole nationale des chartes ;
 - Institut de Recherche et d'Histoire des Textes ;
 - LaMOP - Université Paris I ;
 - Ecole française d'Extrême-Orient ;
 - EPHE.

CREMMA : un gros achat matériel

- Une machine pour le déploiement d'une application de REM : eScriptorium ;
- 50 000 € de budget pour un équipement ajouté à l'existant ;
 - 4 GPUs ;
 - 85 To de stockage ;
 - 384 Go de RAM ;



CREMMA : “un bête achat matériel” ?

- Mettre sur le papier des questions méthodologiques, infrastructurelles, etc. autour de la transcription automatique ;
- 2 ans pour développer des méthodologies et des instruments communs pour mieux comprendre l'écrit patrimonial ancien et améliorer la transcription automatisée ;
- Acquisition, maintenance, test et mise à disposition des partenaires d'un serveur applicatif dédié à eScriptorium basée sur Kraken ;
 - <https://gitlab.inria.fr/scripta/escriptorium>
 - <http://kraken.re/>
- Démarche entièrement en mode open source et gratuit ;
- Un engagement à la mise en commun des vérités de terrains pour la création de modèles.

Approches numériques des textes anciens

Le numérique change la manière dont on obtient et travaille avec les textes :

- “à l’ancienne” :
 - transcription et lecture simultanées mais plus lentes à générer, à traiter et à partager ;
- avec le machine learning :
 - accélérer la transcription, déporter la lecture, (semi-) automatiser le traitement et le partage ;
 - maintien d’une traçabilité des traitements ;



Plus vite



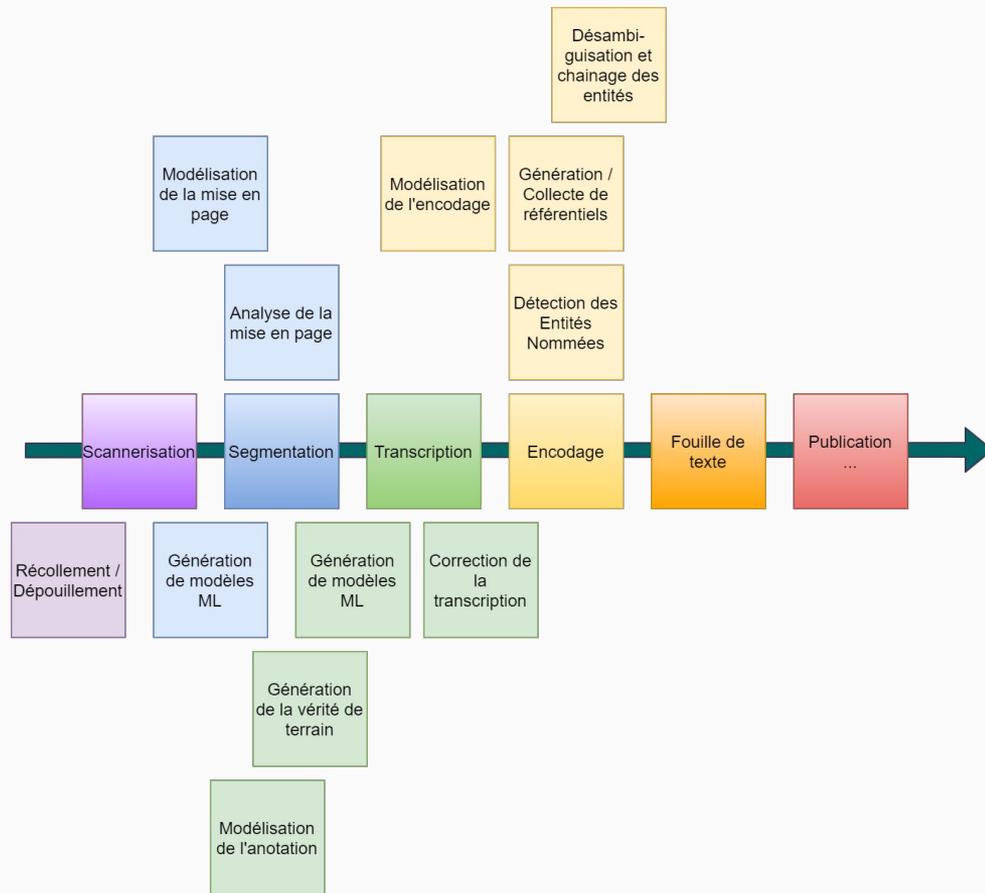
Plus grand



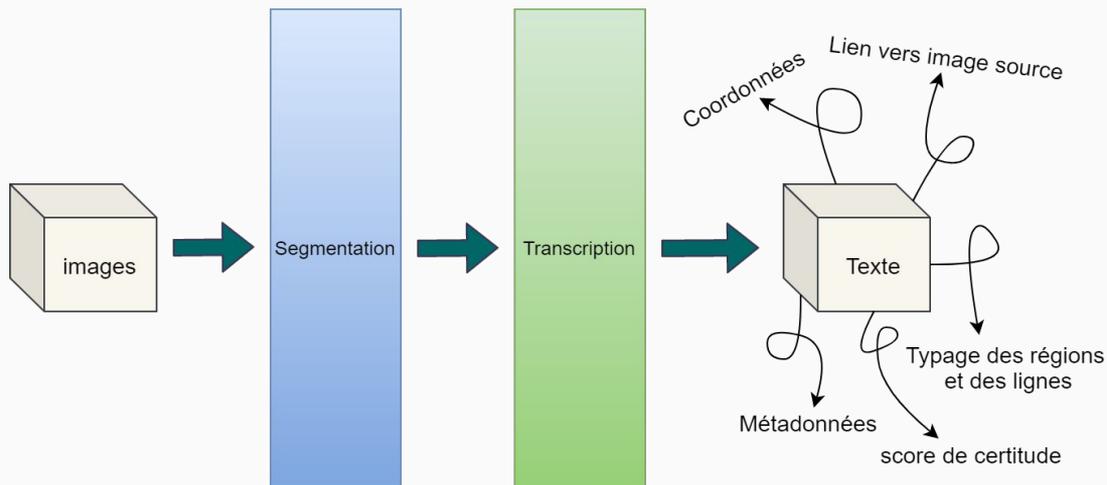
Plus précis

Un écosystème de “numérisation”

- transformation de l'image d'un document en un contenu exploitable numériquement ;
- la REM n'est qu'une étape parmi d'autres, mais elle est cruciale ;
- des questions stratégiques et techniques à tous les niveaux ;
- rarement (jamais ?) un logiciel tout en un ;
- des données générées à chaque étape ;

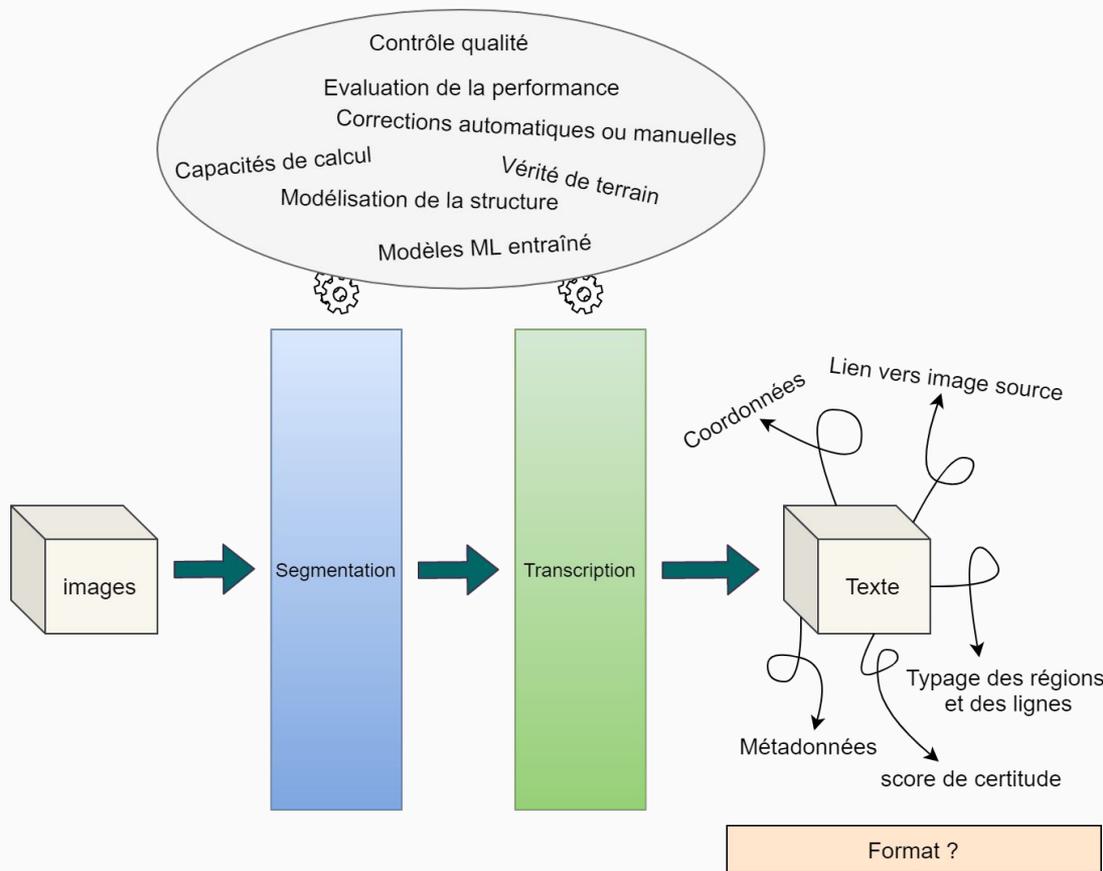


La Reconnaissance d'Écritures Manuscrites (HTR) : principes



- Des images en entrée, du texte en sortie
- 2 étapes étroitement liées, prises en charge par la plupart des logiciels
- En fait beaucoup plus que seulement du texte :
 - lien fin vers l'image
 - métadonnées
 - couches supplémentaires d'informations

La Reconnaissance d'Écritures Manuscrites (HTR) : défis



- Questions de matériel (puissance de calcul, espace de stockage, maintenance et administration système) ;
- Formation au(x) logiciel(s), partage des bonnes pratiques et de l'expertise ;
- Partage des données et des modèles de segmentation et de transcription ;
- Accès aux données et aux modèles déjà existants ;

Mutualisation et partage des données : HTR-United

- Une organisation Github publique ;
- Un annuaire de jeux de données mis à disposition de la communauté ;
- Un laboratoire pour la standardisation de la description des jeux de données de REM et l'établissement des bonnes pratiques ;
- Prise en compte que les données de vérité de terrain sont essentielles pour travailler plus efficacement et plus globalement sur les écritures manuscrites ;
- Les données sont tout terrain, pas les modèles ;



```
40 lines (39 sloc) 1.38 KB
Raw Blame
1 name: htr-unity
2 description: 'HTR Ground Truth Resources'
3 entries:
4
5     title: 'tapuscorpus'
6     url: 'https://github.com/HTR-United/tapuscorpus'
7     description: 'Ground Truth dataset for French typewritten OCR (20th century documents)'
8     language: French
9     time: 1900--1999
10    hands: 30
11    license:
12      - {name: 'CC-BY 4.0', url: 'https://creativecommons.org/licenses/by/4.0/'}
13    format: 'XML-Page'
14    volume:
15      - {count: "150", metric: pages}
16
17    title: 'timeuscorpus'
18    url: 'https://github.com/HTR-United/timeuscorpus'
19    description: 'Ground Truth datasets for French 18th and 19th administrative documents'
20    language: French
21    time: 1858--1858
22    hands: 1
23    license:
24      - {name: 'CC-BY 4.0', url: 'https://creativecommons.org/licenses/by/4.0/'}
25    format: 'XML-Page'
26    volume:
27      - {count: "150", metric: pages}
28
29    title: 'dahncorpus'
30    url: 'https://github.com/HTR-United/dahncorpus'
31    description: 'Ground Truth dataset for French 20th typewritten OCR'
32    language: French
33    time: 1914--1924
34    hands: 2
35    license:
36      - {name: 'CC-BY 4.0', url: 'https://creativecommons.org/licenses/by/4.0/'}
37    format: 'ALTO' or 'XML-Page'
38    volume:
39      - {count: "527", metric: pages}
```

- “Données pour tous : moyens et enjeux d’une vulgarisation réussie”
 - des collaborations
 - une infrastructure
 - des données
 - des bonnes pratiques
- vers une démocratisation réussie ?

N ^o	DATES	NATURE ET ESPECE DES ACTES;	NOMS, PRÉNOMS ET DOMICILES DES PARTIES.	
			INDICATIONS, SITUATIONS ET PRIX DES BIENS.	
			An 1872 mois de Janvier	
76	29	Certificat de Propriété		29
77	29	Procuration		30
78	29	Décharge		30
79	29	Mariage		30
80	29	Procuration		30
81	30	Certificat de Vie		30
82	30	Procuration		30
83	30	Arrêt de Compt. d'Intell. Vente		30
84	30	Vente		30
85	30	Contrat de mariage		30
86	30	Contrat de mariage		30
87	30	Contrat de mariage		30
88		Capit. des chanc.		30

Merci!

Fevrier 1872.