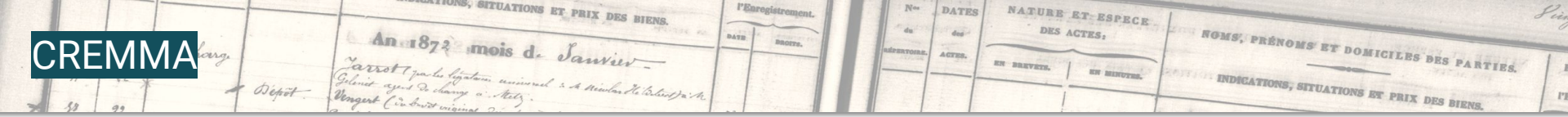
CREMMA : Une infrastructure mutualisée pour la reconnaissance d'écritures manuscrites et la patrimonialisation numérique

Alix Chagué

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 ;

Traces6 (DAHN)	CREMMA (DIM MAP)
• 2 GPUs NVIDIA RTX 6000 • 23 To de stockage HDD/SSD • 128Go de RAM	• 2 GPUs NVIDIA A100 • 52 To de stockage HDD • 256 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 ;

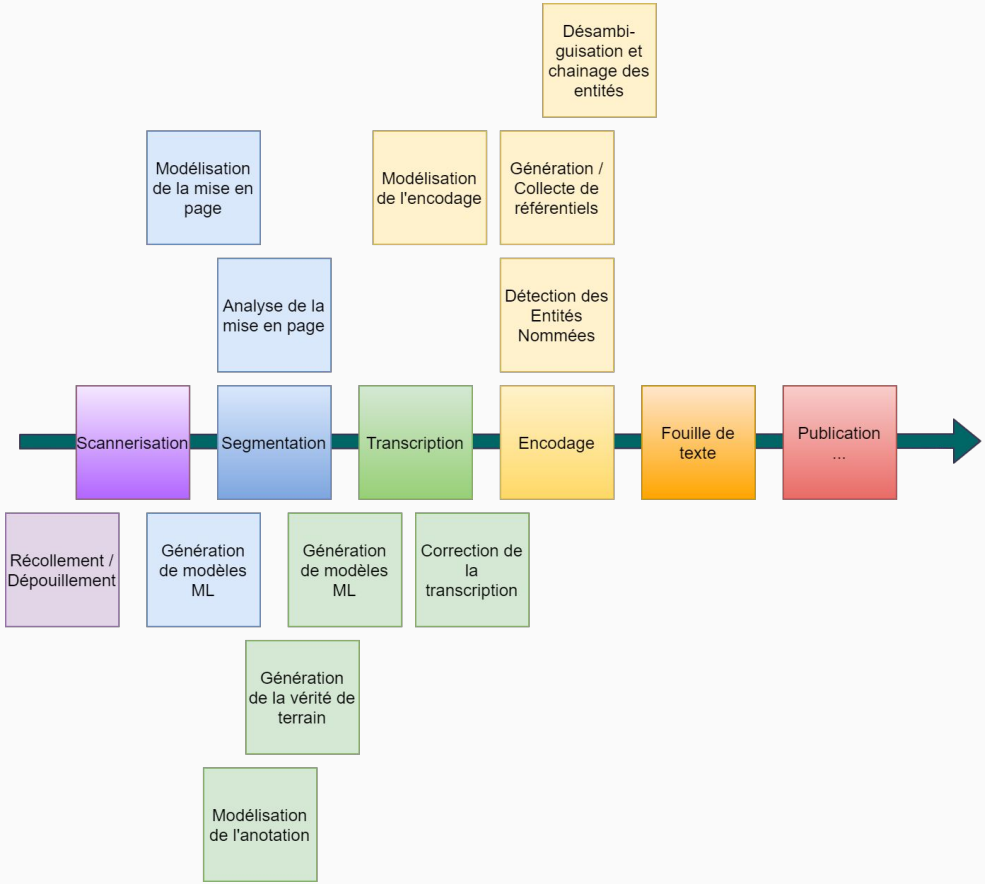
Plus vite

Plus grand

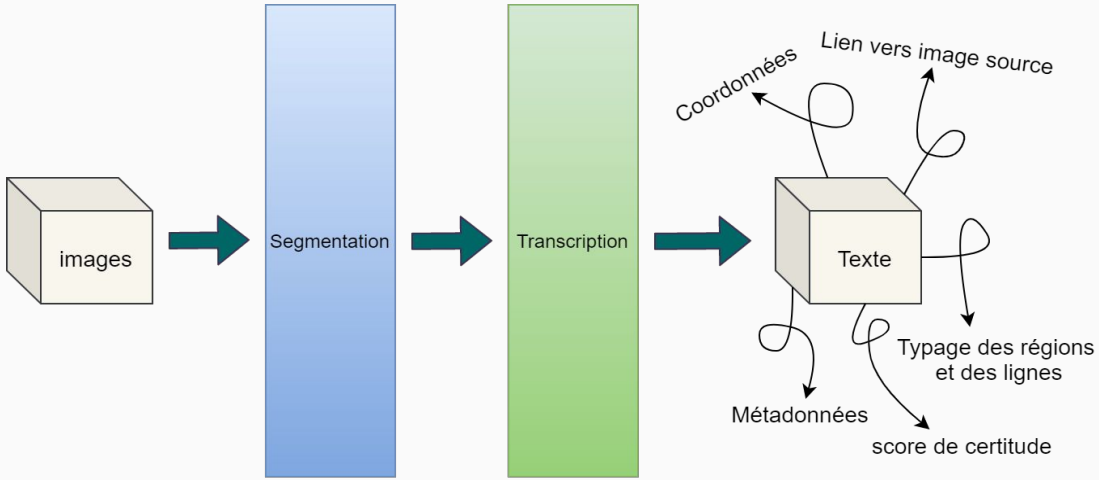
Plus précis

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 ;

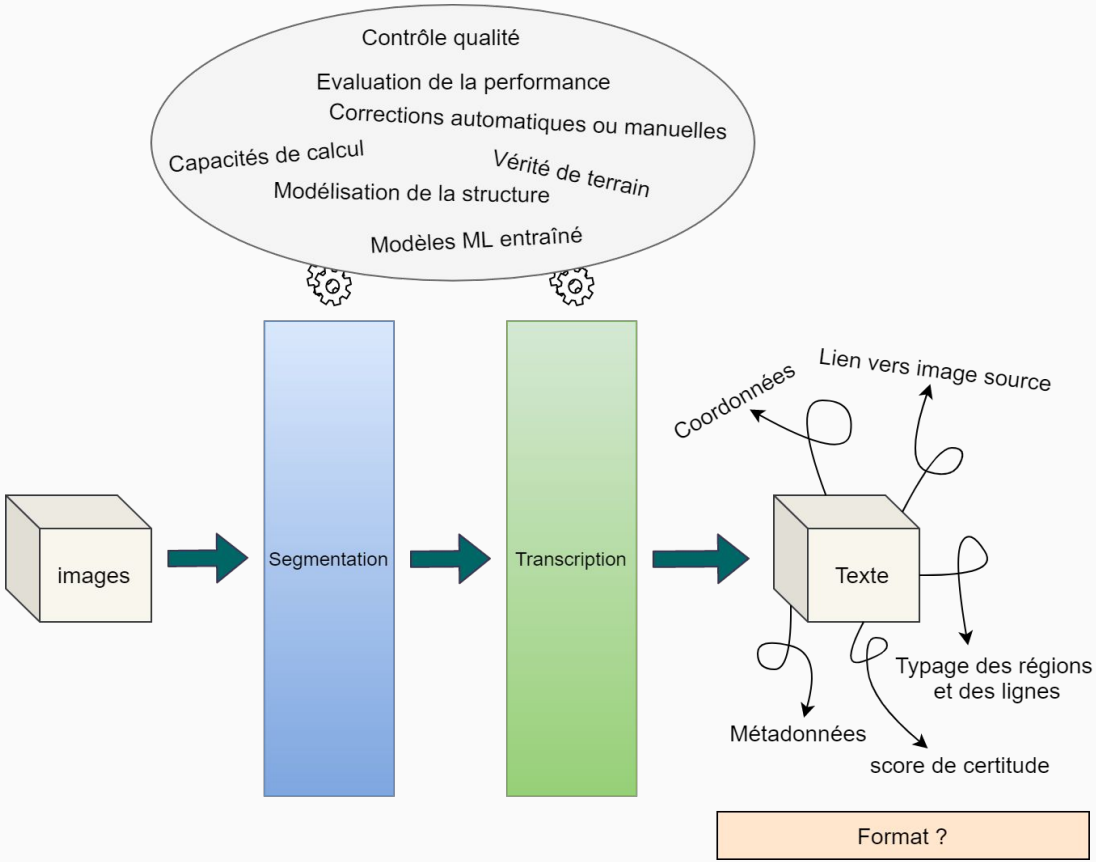


La Reconnaissance d'Écritures Manuscrites (HTR) : principes



- 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

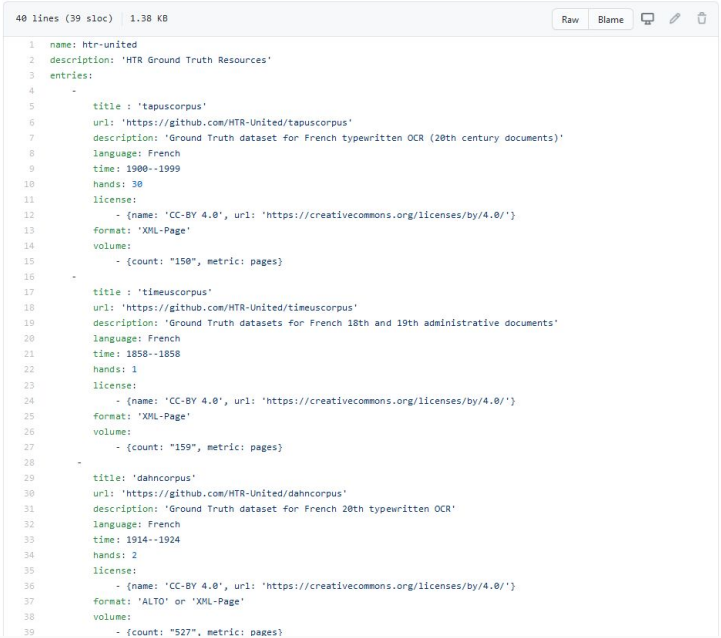
La Reconnaissance d'Écritures Manuscrites (HTR) : défis



- 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

```yaml
name: htr-united
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}
    -
        title : 'timeuscorpus'
        url: 'https://github.com/HTR-United/timeuscorpus'
        description: 'Ground Truth datasets for French 18th and 19th administrative documents'
        language: French
        time: 1858--1858
        hands: 1
        license:
            - {name: 'CC-BY 4.0', url: 'https://creativecommons.org/licenses/by/4.0/'}
        format: 'XML-Page'
        volume:
            - {count: "159", metric: pages}
    -
        title: 'dahncorpus'
        url: 'https://github.com/HTR-United/dahncorpus'
        description: 'Ground Truth dataset for French 20th typewritten OCR'
        language: French
        time: 1914--1924
        hands: 2
        license:
            - {name: 'CC-BY 4.0', url: 'https://creativecommons.org/licenses/by/4.0/'}
        format: 'ALTO' or 'XML-Page'
        volume:
            - {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 ?