Alix Chagué, Lucas Terriel, Laurent Romary (Inria - ALMAnaCH)

Des images au texte : LECTAUREP, un projet de reconnaissance automatique d'écriture

2020

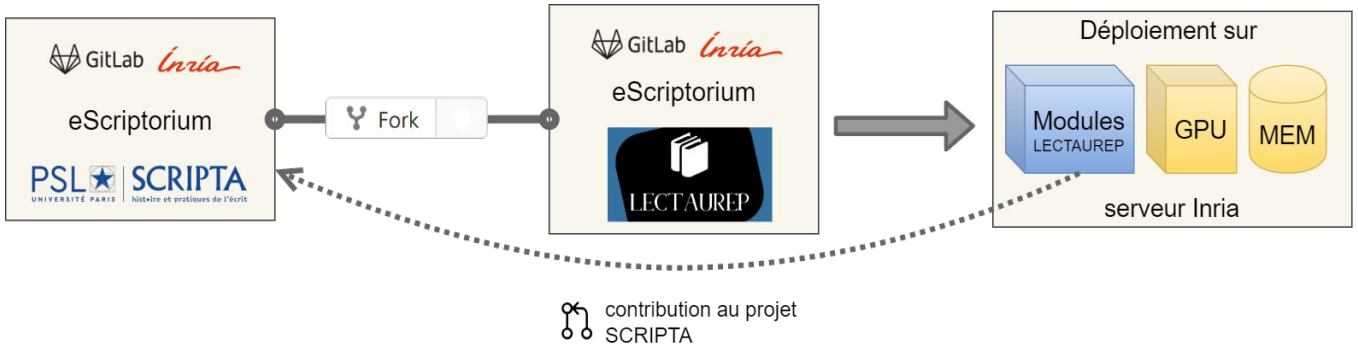
CC BY

#dhnord2020
http://dhnord2020.meshs.fr

ANALYSE AUTOMATIQUE DES IMAGES POUR L'ACQUISITION DU TEXTE

Le Département du Minutier Central des Archives nationales conserve plus de 3300 répertoires de notaires parisiens. C'est **un corpus impossible à explorer et à exploiter sans l'aide de la machine**. Depuis 2018, le projet LECTAUREP (partenariat Archives nationales - Inria) ambitionne d'utiliser les technologies de *machine learning* pour traiter ce fond et en extraire le texte.

Il s'agit de rendre un patrimoine commun accessible, appropriable et "enrichissable" par de nouveaux publics grâce au traitement des numérisations de la collection de répertoires. L'interprétation de ces images fait intervenir en premier lieu des technologies de **reconnaissance automatique d'écriture** (REM) et d'**extraction de structure logique** (ESL). Ce travail implique en outre nécessairement l'établissement d'une **chaîne de traitement des données** ainsi que leur **modélisation**.



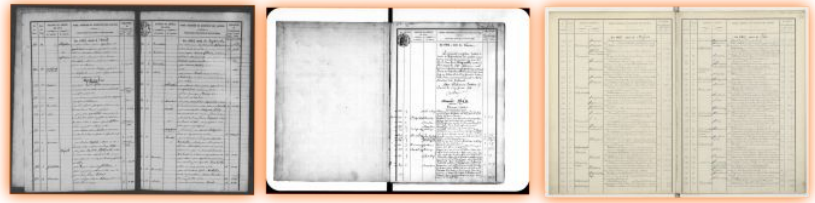
eScriptorium/Kraken : solution *open source* pour la REM

Un *fork* d'eScriptorium pour LECTAUREP, c'est-à-dire une version adaptée de la plateforme originale développée par l'équipe SCRIPTA PSL, est déployé sur un serveur doté d'une puissance de calcul (GPU) et d'un espace mémoire suffisants. Les modules développés pour répondre aux besoins du projet LECTAUREP peuvent ensuite être intégrés au fur et à mesure dans la plateforme eScriptorium initiale.

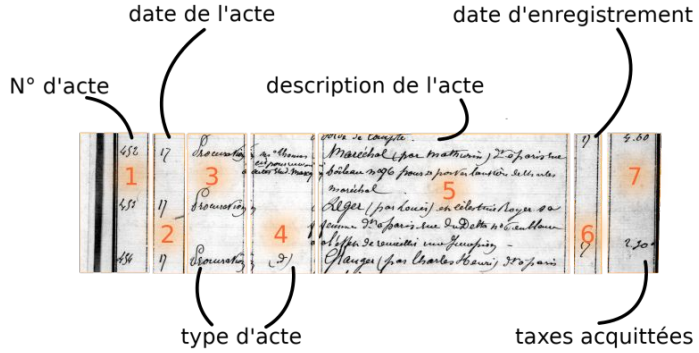
Quelles images ? Que cherche-t-on ?

Le corpus mobilisé par LECTAUREP rassemble la production de 917 notaires parisiens sur une période allant de 1803 à 1940 ; soit des milliers de mains d'écritures différentes réparties sur près de 720 000 pages de répertoires numérisés à l'issue de multiples campagnes menées sur plusieurs décennies.

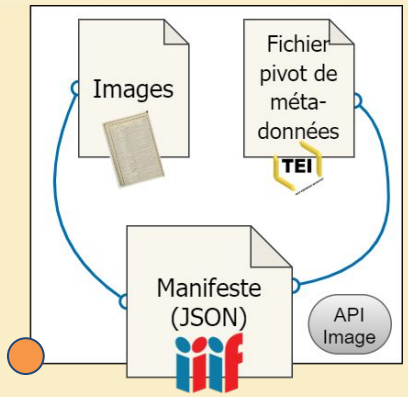
Pourtant nos images n'ont pas seulement en commun le contexte de production de ces archives, elles possèdent la même mise en page.



Chaque page de répertoire contient un tableau à 7 colonnes. L'analyse des images doit permettre d'identifier l'emplacement de ces colonnes pour reconstruire la structure logique de tableau.



DÉVELOPPEMENTS TECHNIQUES ET CONCEPTUELS AUTOUR D'eSCRIPTORIUM



Kraken Benchmark

Permet de générer des métriques plus robustes pour mesurer la performance d'un modèle de segmentation ou de transcription (*Word Error Rate, Character Error Rate, Word Accuracy*).

Par extension, permet d'identifier, étant donné l'échec d'un modèle, des cas hors normes ou de nouvelles mains d'écritures ; on évite ainsi de passer par une lourde étape de classification.

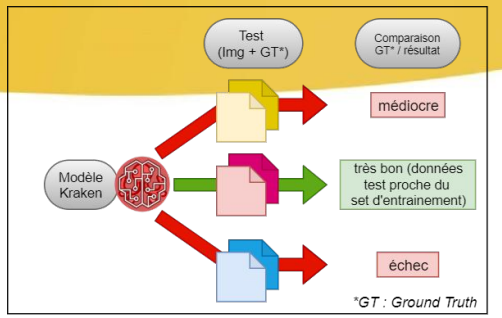
Aspyre

Permet d'assurer le rapatriement des données produites sur d'autres plateformes (ex: Transkribus) vers eScriptorium de manière à créer des modèles entièrement libres et distribuables.

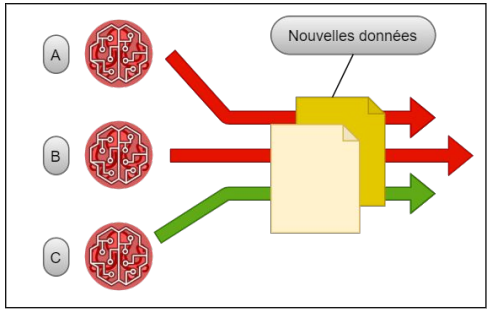
Choppy

Permet d'interpréter l'emplacement du texte sur l'image pour recomposer la structure logique des tableaux (colonnes, unités minutes, etc.).

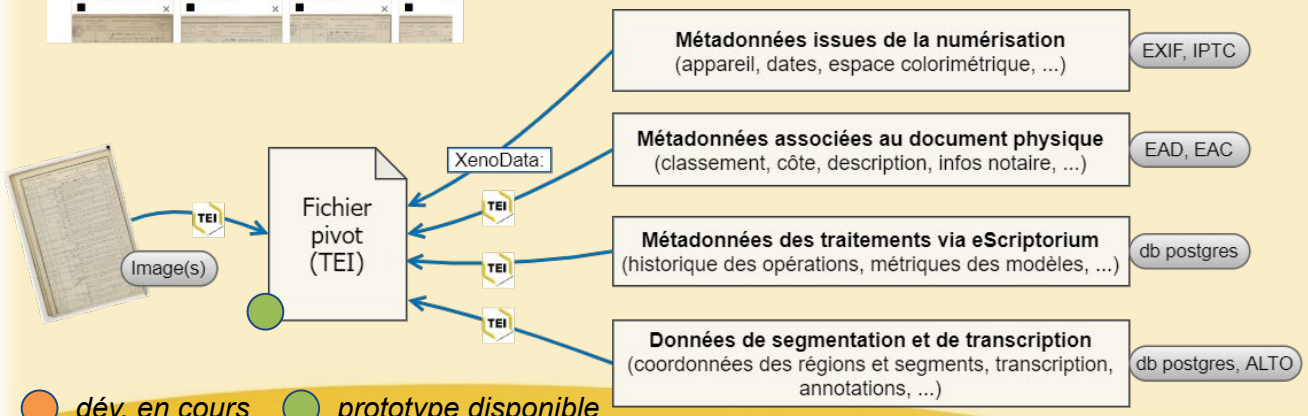
Usages de Kraken Benchmark



Exemple : Le modèle n'est pas généralisable mais fonctionne bien sur un type de données



Exemple : le modèle C peut être utilisé pour traiter les nouvelles données (après affinage)



Rationalisation de la gestion des métadonnées et modélisation

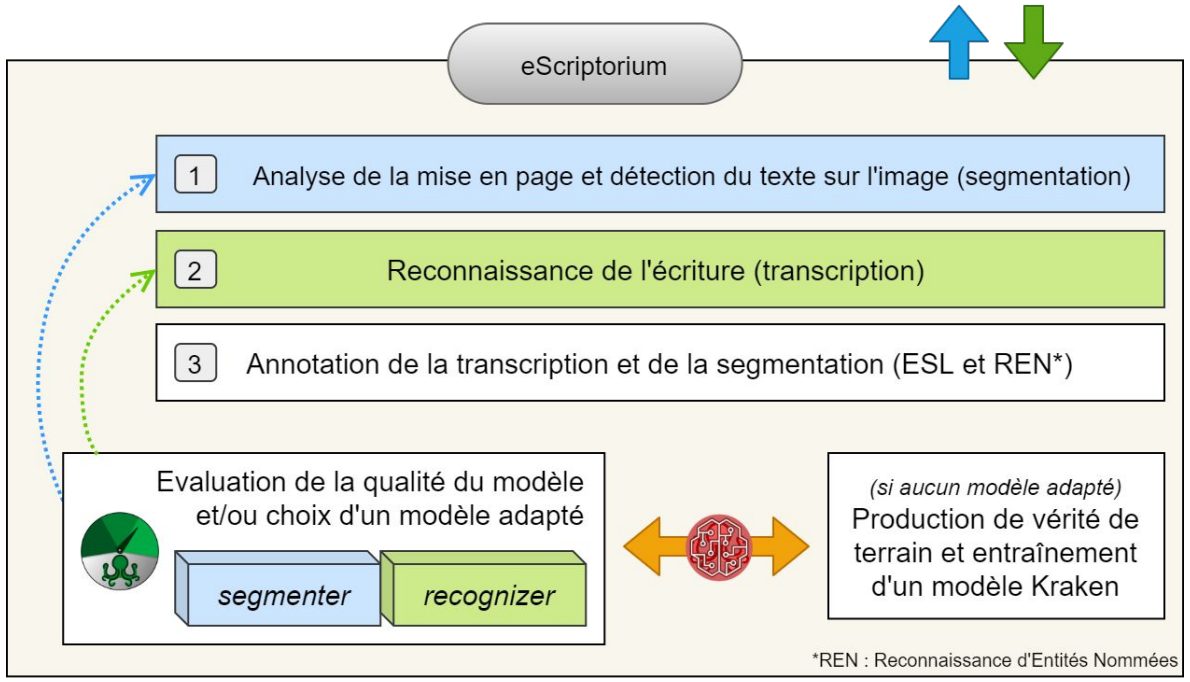
Un fichier XML TEI centralise les métadonnées associées aux documents et celles générées durant le traitement des images avec eScriptorium. Ce fichier est indexé dans le manifeste IIIF qui permet de servir images et métadonnées à notre plateforme de REM et ESL. eScriptorium renvoie ensuite des données vers ce fichier pivot.

Chaque type de métadonnées intégré dans le fichier pivot est modélisé en TEI ou dans un *namespace* spécifique (<XenoData>), quel que soit le format de sauvegarde initial.

dév. en cours *prototype disponible*

Chagué, Terriel, Romary, ALMAnaCH (Inria), *Des images au texte*

VERS LA DÉFINITION D'UNE INFRASTRUCTURE DE RECHERCHE ET DE SERVICE POUR LA REM



Un travail de pérennisation à tous les niveaux

- ➔ Mise en *open source* de tous les développements informatiques
- ➔ Large ouverture des documentations et composants pédagogiques
- ➔ Travail inter-institutionnel sur le renforcement des capacités de calcul et de leur hébergement
- ➔ Vers une coordination des établissements concernés dans la gestion des accès
- ➔ Perspective d'un service relié à Huma-Num et DARIAH à terme

Récapitulatif de la chaîne de traitement

1. Téléchargement des images et métadonnées dans eScriptorium depuis un serveur IIIF ;
2. Analyse de la mise en page et détection des segments de texte à l'aide d'un modèle Kraken sélectionné grâce à Kraken-Benchmark ;
3. Transcription du texte à l'aide d'un modèle Kraken sélectionné grâce à Kraken-Benchmark ;
4. Annotation de la structure logique et des entités nommées dans la transcription obtenue ;
5. Injection de toutes les données dans le fichier pivot TEI lié à l'image sur le serveur IIIF et/ou export dans différents formats standardisés et à jour.

Liens et ressources

- ★ Blog LECTAUREP : **https://lectaurep.hypotheses.org/**
- ★ Serveur Inria pour eScriptorium : **http://traces6.paris.inria.fr/**
- ★ Code source du fork eScriptorium : **https://gitlab.inria.fr/almanach/lectaurep/escriptorium/**
- ★ Code source de Kraken-Benchmark : **https://gitlab.inria.fr/dh-projects/kraken-benchmark**
- ★ Code source d'Aspyre : **https://gitlab.inria.fr/dh-projects/aspyre-gt**

LECTAUREP : Lecture Automatique des Répertoires