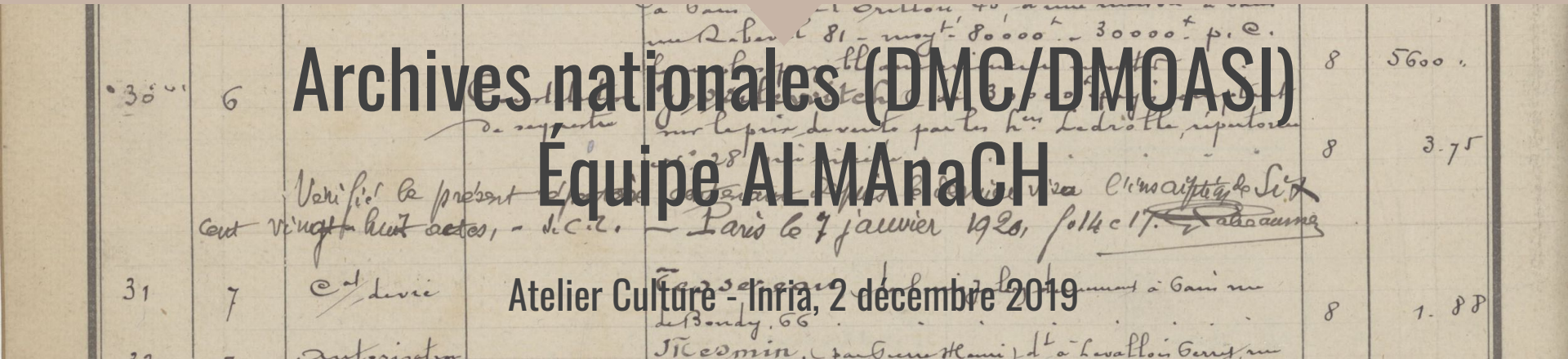
LECTAUREP
Lecture Automatique de Répertoires



Archives nationales (DMC/DMOASI)
Équipe ALMAnaCH

Atelier Culture - Inria, 2 décembre 2019

Un partenariat pluridisciplinaire

Archives nationales

Département du Minutier Central (DMC)
Département de la maîtrise d'ouvrage du système d'information (DMOASI)

- ★ Virginie Grégoire (DMC)
- ★ Danis Habib (DMC)
- ★ Marie-Françoise Limon-Bonnet (DMC)
- ★ Gaetano Piraino (DMOASI)
- ★ Aurélia Rostaing (DMC)
- ★ Frédéric Zamarreno (DMOASI)

ALMAnaCH (Equipe Inria *Automatic Language Modelling and Analysis & Computational Humanities*)

- ★ - - Marie-Laurence Bonhomme
- ★ ++ Alix Chagué
- ★ Eric de la Clergerie
- ★ - - Marie Puren
- ★ - - Charles Riondet
- ★ Laurent Romary
- ★ ++ Lionel Tadjou

SCRIPTA (EPHE)

- ★ Marc Bui
- ★ Daniel Stökl Ben Ezra
- ★ ++ El Hassane Gargem
- ★ Benjamin Kiessling
- ★ ++ Robin Tissot

Le corpus d’archives

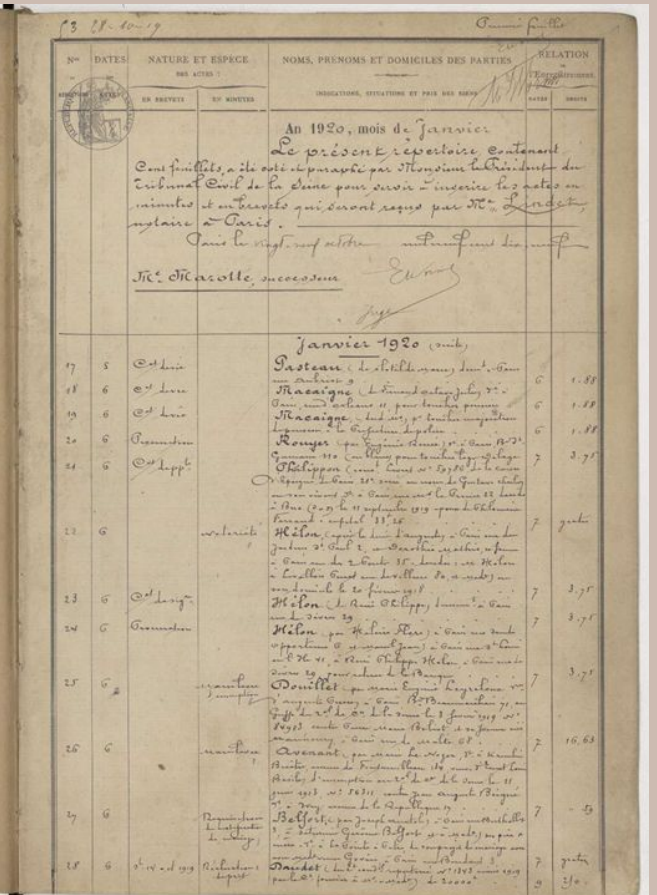
Les répertoires de notaires de Paris (1803-1944)

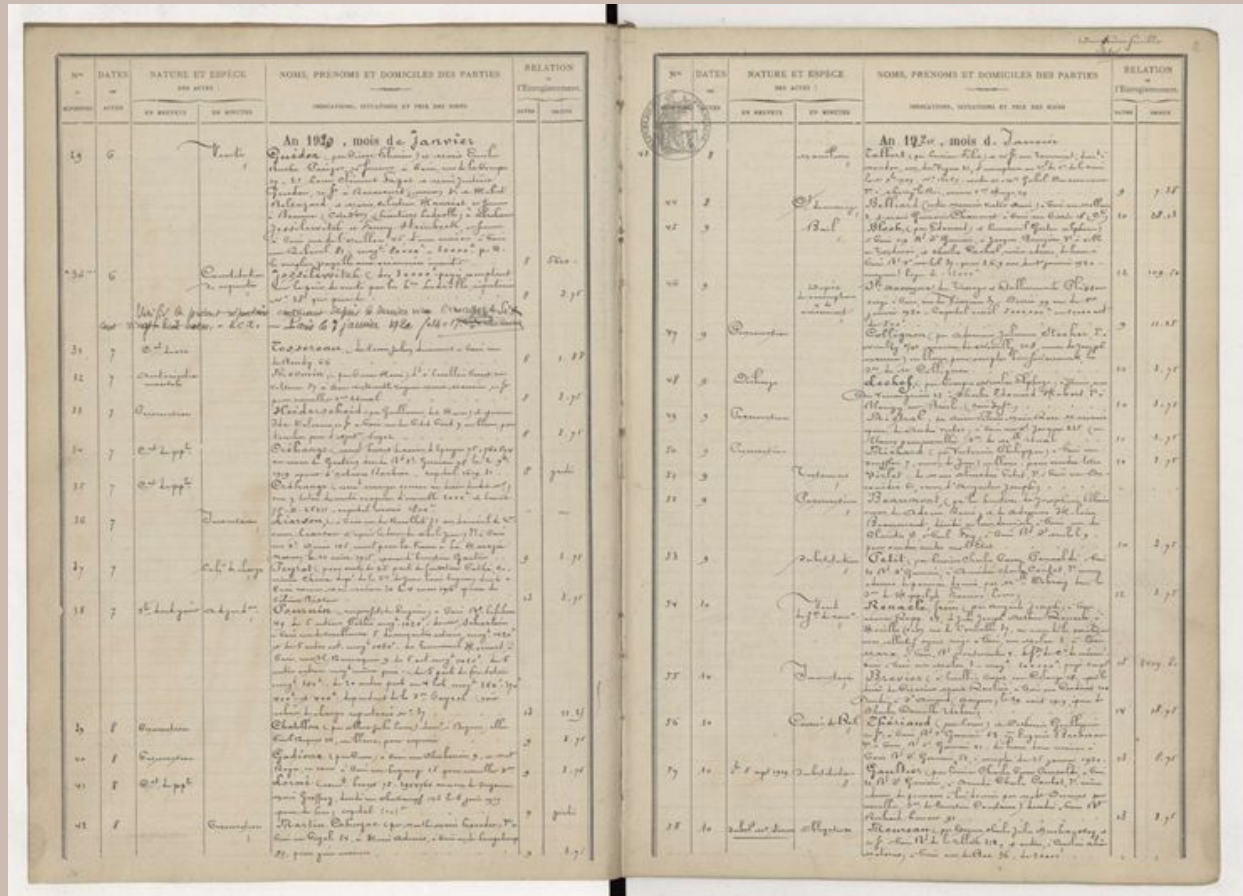
Plus de 900 notaires

Environ 2000 registres (300 à 500 pages chacun)

Plusieurs milliers de mains de clercs différentes

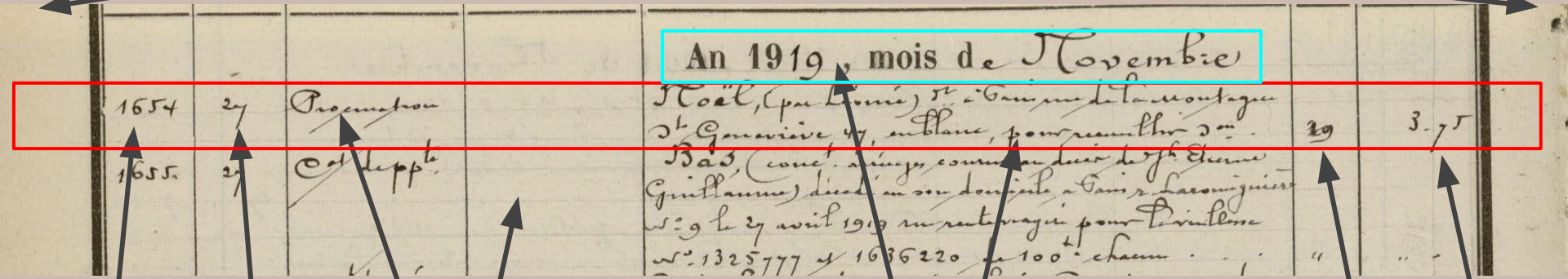
Le corpus d'archives





La reconstitution des actes

Vide (sauf accident dans colonnes adjacentes)



Nº de l'acte
Nombre entre 1 et 3000 (estimation)

Typologie de l'acte
Chaîne de caractères
Vocabulaire contrôlé

Date d'enregistrement (jour)
Nombre entre 1 et 31

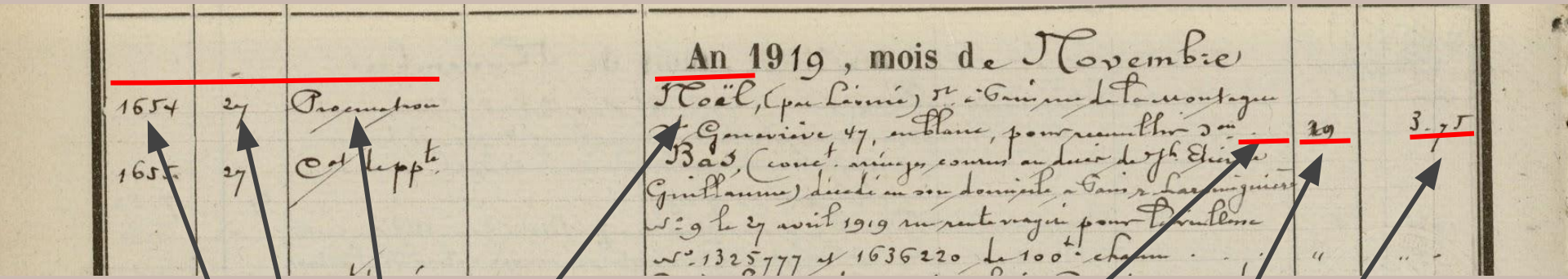
Date de l'acte (jour)
Nombre entre 1 et 31

Date de l'acte (année et mois)
Écritures mixtes (imprimées et manuscrites)

Description de l'acte
Nom et adresse des signataires, prix de vente d'un bien, date d'un décès, etc.

Taxes acquittées
Chiffres, chaînes de caractères (gratis, etc.)

Détection des unités d'informations



Limite supérieure de l'acte :

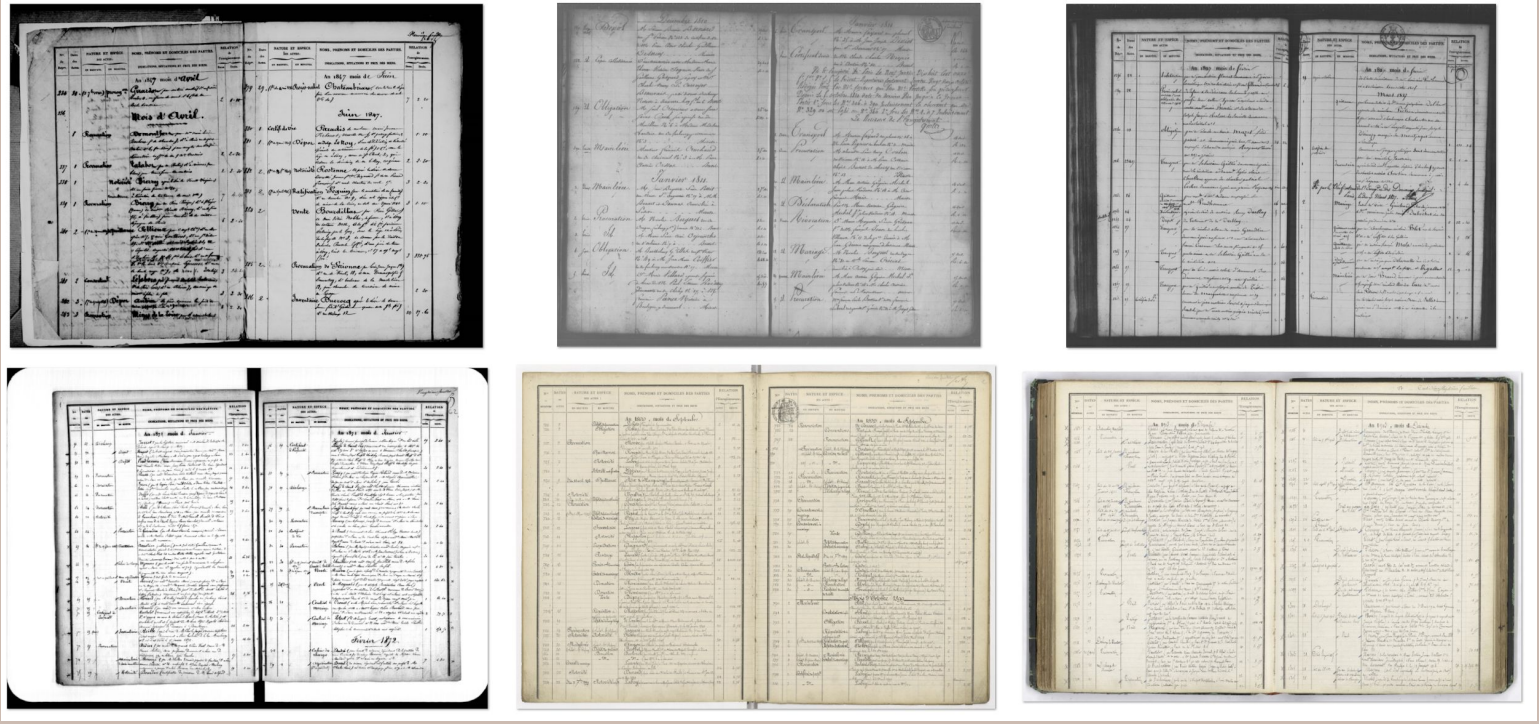
- *N° d'ordre*
- *Date de l'acte*
- *Typologie*
- *Nom du signataire en gras*

Limite inférieure de l'acte :

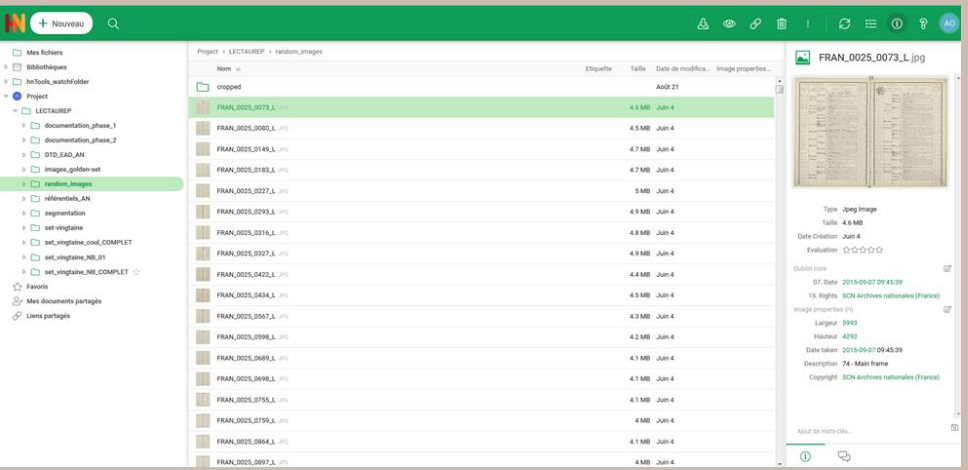
- *Description de l'acte : dernière ligne parfois incomplète*
- *Date de l'enregistrement de l'acte*
- *Taxes acquittées*

Le corpus d’images numériques

Près d’un demi-siècle de campagnes de reprographie analogique et numérique



Les sous-corpus d'images numériques



Des dizaines de mains différentes

Un *golden set* : 10000 doubles pages de 41 registres (1789-1875) numérisés en noir et blanc et en couleur, référence pour les phases 2 et 3

Un *random set* : 1000 doubles pages aléatoires de quatre campagnes de numérisation récentes en couleur (années 1880-années 1930)

Basculement du stockage sur Sharedocs (Huma-Num)

Objectifs du projet

Objectifs du projet

Enjeux :

- pour le public des archives : *service de recherche en texte intégral*
- pour les chercheur·ses : *exploitations statistiques sur l'ensemble du corpus*
- pour les services publics d'archives et les institutions patrimoniales : *outil mutualisé, en réseau*

Approche :

- Reconnaissance automatique de structure et d'écriture manuscrite (HTR)
- Indexation et publication sur une plateforme dotée d'une recherche avancée

Chaîne de traitements

1. Cadrage et redressement des tableaux
2. Analyse de la mise en page des répertoires
3. Détection des lignes de texte et segmentation (polygones)
4. Reconnaissance d'écriture manuscrite (HTR + *Word Spotting*)
5. Extraction d'informations (personnes, adresses, lieux, professions, types d'actes, mots clés)
6. Mise à disposition du public et de la communauté des chercheur·ses
7. Brique participative (correction de la segmentation, des transcriptions et de l'indexation)

Objectifs des phases 1 et 2

Phase 1 (2018) :

- état de l'art, banc d'essai et premiers développements à partir de l'exemple de l'étude Marotte
- établissement de préconisations pour le déploiement de la chaîne de traitements

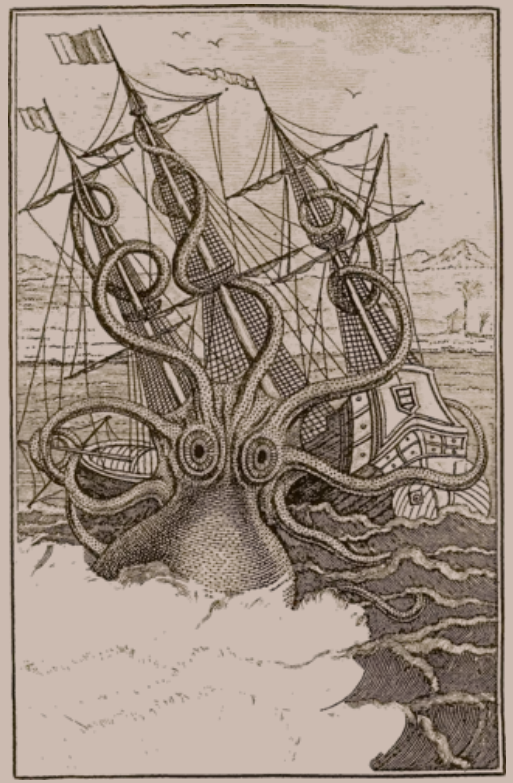
Phase 2 (mai 2019 - novembre 2019) :

4 volets :

1. optimisation des outils, méthodes et flux de travaux pour l'analyse de la structure des documents et leur transcription
2. indexation de la transcription et alignement sur des référentiels
3. étude des scénarios et cas d'usage dans la perspective de développer une plateforme de transcription
4. élaboration de recommandations pour la mise en place d'une plateforme technique et d'une infrastructure d'hébergement

Kraken et eScriptorium

Kraken



En quelques mots :

- un logiciel en ligne de commande (CLI) pour la détection de lignes de texte (segmentation) et la reconnaissance automatique de texte manuscrit (transcription)

Avantages :

- développé par Benjamin Kiessling (Scripta, EPHE-Inria)

Documentation :

- http://kraken.re

eScriptorium

En quelques mots :

- une plateforme web dotant **Kraken** d'une interface graphique
- destinée à gérer l'essentiel de la chaîne de traitements des documents

Avantages :

- proximité avec l'équipe de développement
- entièrement gratuit, *open source* et agnostique
- modèles entraînés maîtrisés, accessibles et partageables

Freins :

- développement en cours
- peu documenté

eScriptorium a été choisie dès le début de la phase 2 comme cadre pour l'ensemble des développement produits dans le cadre du projet LECTAUREP.

Changement d'outils

Transkribus

- facile d'utilisation
- travail collaboratif
- formats d'export variés
- customisation faible

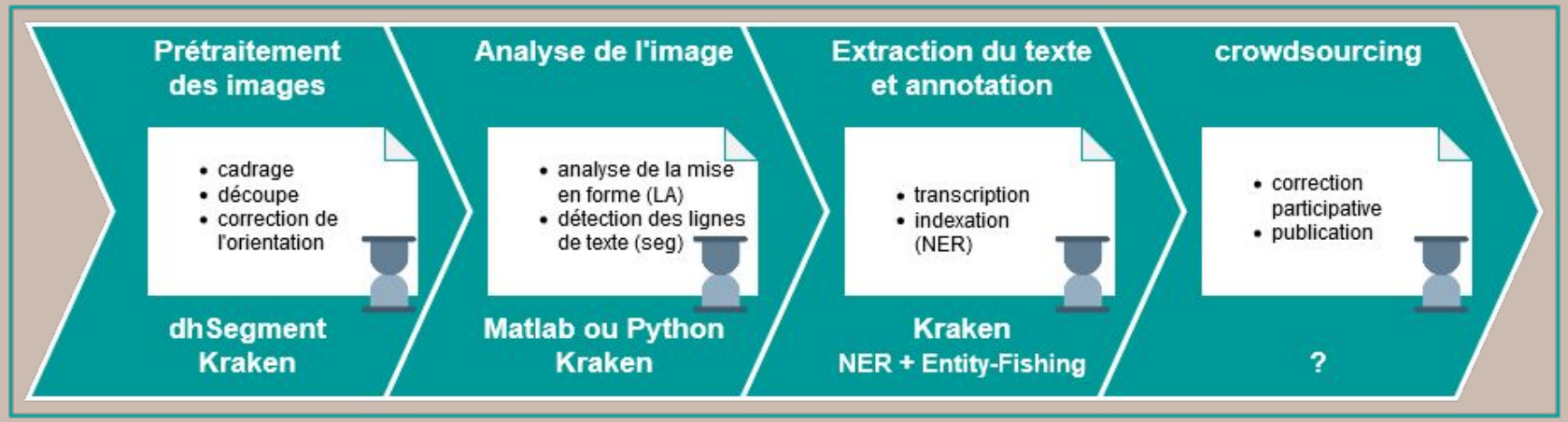
Kraken + eScriptorium

- unité logicielle
- facile d'utilisation
- customisation forte (*open source*)
- maîtrise du stockage des données et des modèles
- développements prometteurs

- pas de maîtrise du stockage des données et des modèles entraînés
- modèles non *open source*
- avenir incertain

- travail collaboratif limité à ce stade
- pas d'instance à jour disponible

Une chaîne complète dans eScriptorium



A terme, eScriptorium intégrera les modules de prétraitement et d'analyse d'images, de gestion des métadonnées, de segmentation, de transcription, d'indexation, accompagnés de fonctionnalités (*features*) permettant d'en faire une plateforme ouverte pour le *crowdsourcing*.

Analyse de la mise en page

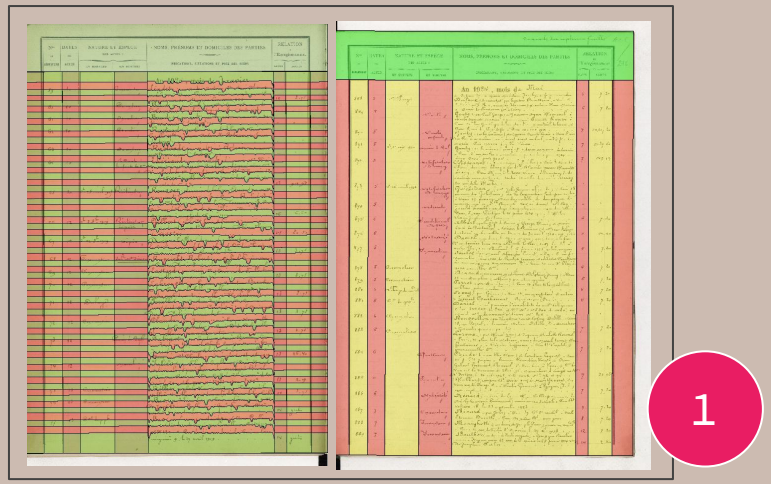
Analyse de la mise en page

Conclusion de la phase 1 :

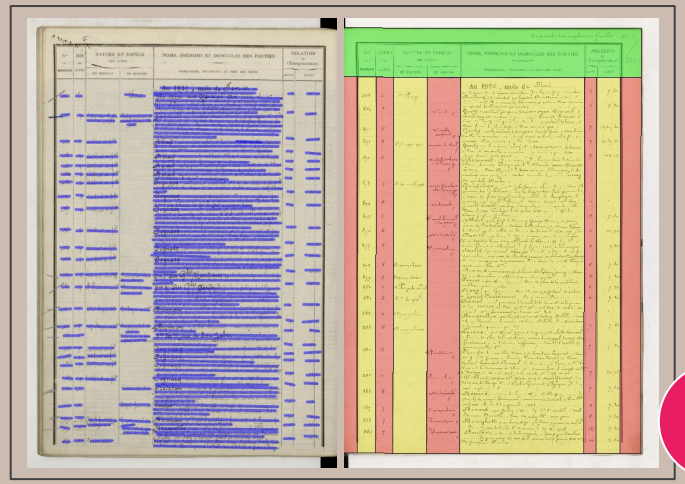
- algorithme de recadrage (*seam carving*) pour détecter la structure du tableau puis transcrire les cellules

Exploration de la phase 2 :

- détecter les colonnes, détecter les segments de texte sur la page, déduire les lignes à partir des coordonnées des segments

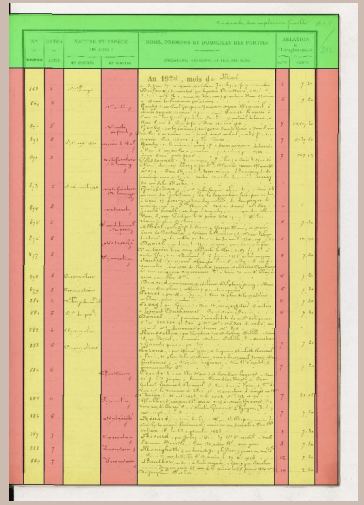


1



2

Détection automatique des colonnes



Etapes :

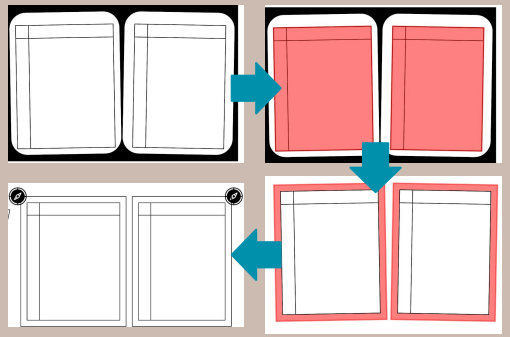
- détecter les lignes horizontales et verticales du tableau
- segmenter en colonnes et segmenter l'en-tête
- supprimer les séparateurs du tableau

Point fort :

- fonctionne bien grâce à la régularité de la mise en page

Points de blocage :

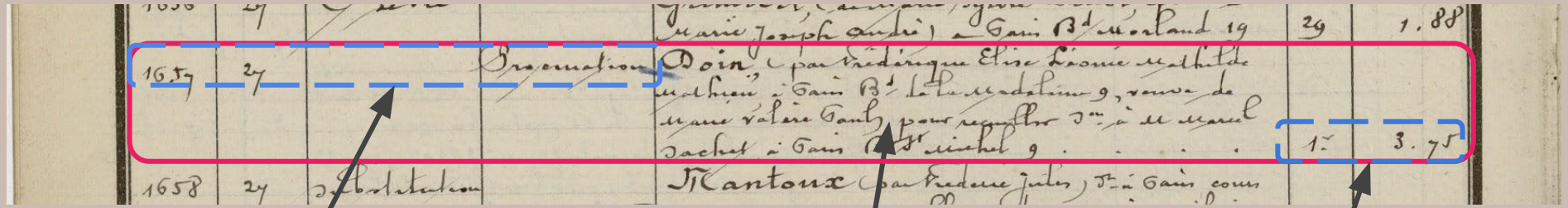
- hétérogénéité des numérisations
- fonctionne assez mal pour gérer le texte débordant des colonnes



Solution :

- avoir recours à de l'apprentissage automatique pour la détection de masques permettant de cadrer et corriger l'orientation à des fins d'homogénéité

Détection des unités d'informations



Début de l'acte : *écriture à la première ligne de l'acte*

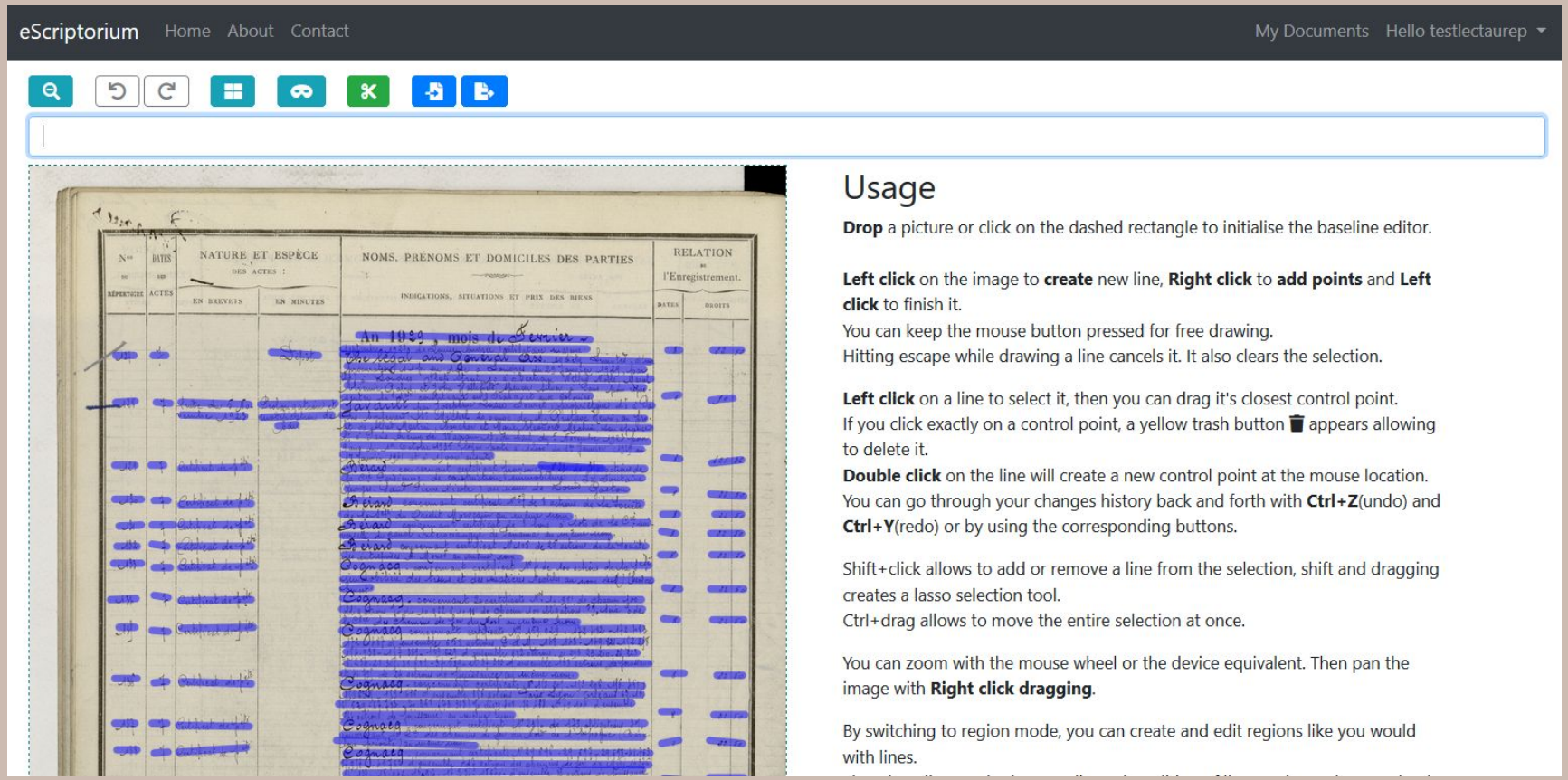
Colonne centrale : Description de l'acte

Fin de l'acte : *écriture à la dernière ligne de l'acte*

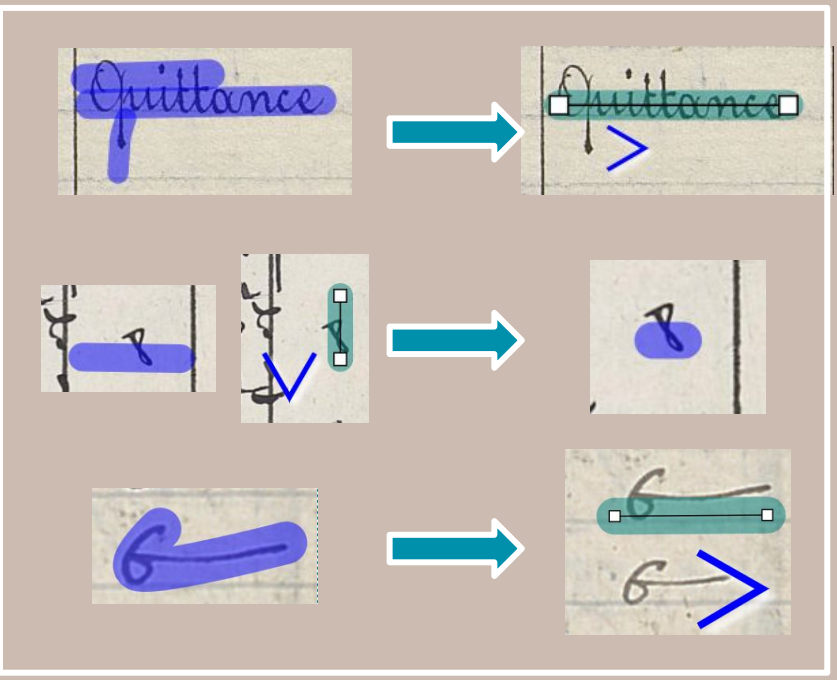
Les résultats d'analyse de la phase 1 sur les répertoires de l'études Marotte ne sont pas généralisables.

Segmentation

Edition des lignes de base (*baselines*) avec eScriptorium



Acquisition des données d'entraînement



exemples d'erreurs de segmentation

Plusieurs temps de formation et de calage :

- pour créer les données d'entraînement
- pour entraîner les modèles

2 campagnes de segmentation :

- 3 versements
- 20 n&b + 22 couleur
- 42 n&b et couleur

6 modèles entraînés :

- variation des paramètres et des corpus de données d'entraînement
- résultats entre 50 % et 60 % d'exactitude
- cible : 70 %

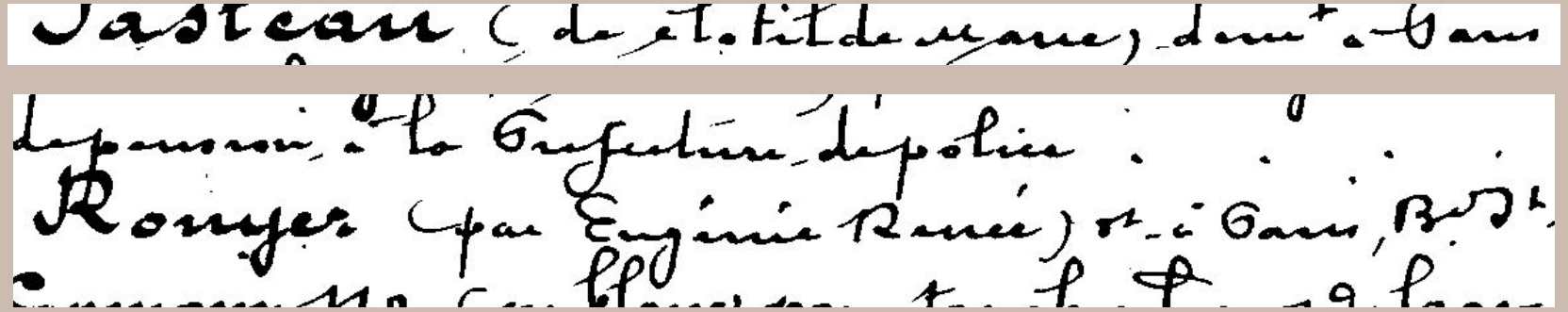
Transcription

HTR : résultats de la phase 1

- Entraînement de 2 modèles d'HTR à partir de pages transcrites manuellement (un seul scribe, étude Marotte) :
  - M1 : 40 pages
  - M2 : 50 pages (1 million de mots)
- Taux d'erreur par caractères (CER) sur un échantillon test du même répertoire:
  - M1 : **13,5 %**
  - M2 : **10,4 %**
- Test du modèle M2 sur d'autres registres (écriture et qualité différentes) :
  - taux d'erreur autour de 40 %
  - nécessité de données d'entraînement plus hétérogènes.

Avancées de la phase 2

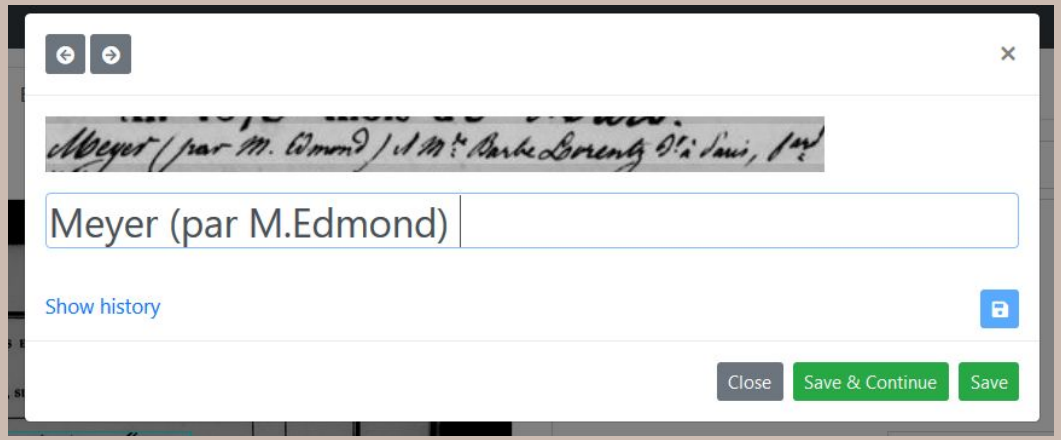
- 1 modèle entraîné avec Kraken à partir des données d'entraînement produites sur Transkribus :
  - pipeline : Transkribus → XML PAGE → HTML → Kraken
  - 19,36% CER
  - imprécision des segments récupérés avec la pipeline :



- pas davantage d'entraînement en l'attente d'une segmentation opérationnelle et d'une mise à jour de l'interface d'eScriptorium

Transcription dans eScriptorium

- interface ergonomique pour la transcription et l'édition du texte
- eScriptorium et Kraken gèrent désormais les polygones, ce qui améliore la qualité des données d'entraînement



Perspectives

Développements à venir

À approfondir (Kraken / eScriptorium) :

☛ Découpage et redressement éventuel des doubles pages après détection des zones des tableaux (module basé sur dhSegment ou Kraken) ;

☛ Structuration des analyses à partir des indices de mise en page et de mise en forme ;

☛ Détection automatique des mains d'écriture pour adapter le modèle de transcription ;

☛ Entraînement de modèles de transcription spécifiques à certaines mains d' écriture ;

☛ Mise en production de l'interface de traitement des images.

Développements à venir

Plusieurs mains de scribes par répertoire, plusieurs pratiques de mise en page

Plus de 1800 répertoires

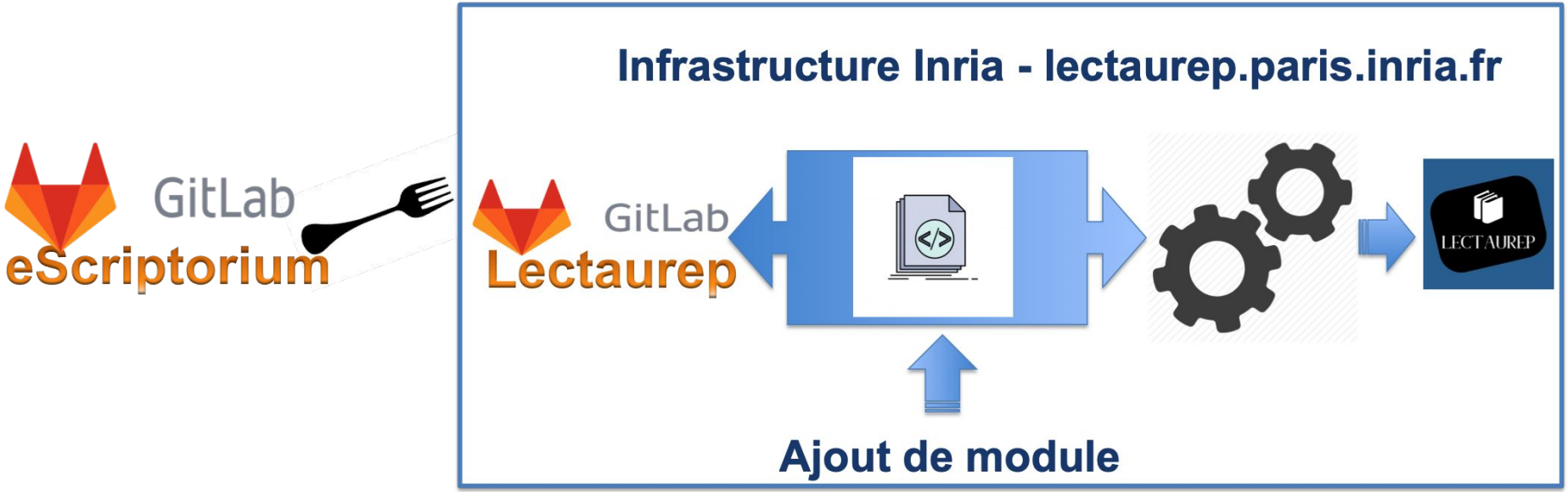
Un modèle à entraîner pour chaque main…

☛ Une interface collaborative adossée à eScriptorium pour entraîner puis corriger les données de segmentation, de transcription, voire d'indexation obtenues par automatisation (plateforme de myriadisation) ;

☛ Si possible, des outils de visualisation et de traitement des données ;

☛ Si possible, des fonctionnalités de reconnaissance d'entités nommées et de liage de ces entités à des référentiels internes ou externes aux Archives nationales.

Une plateforme collaborative adossée à une instance eScriptorium pour LECTAUREP



Merci !

des questions ?