Présentation du projet Lectaurep (Lecture automatique de répertoires)

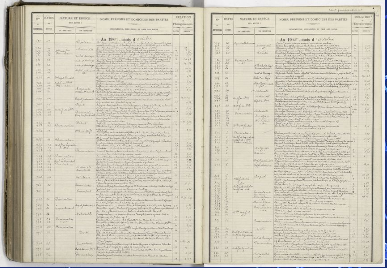
Alix Chagué, Aurélia Rostaing

LECTAUREP

Lecture automatique de répertoires de notaires

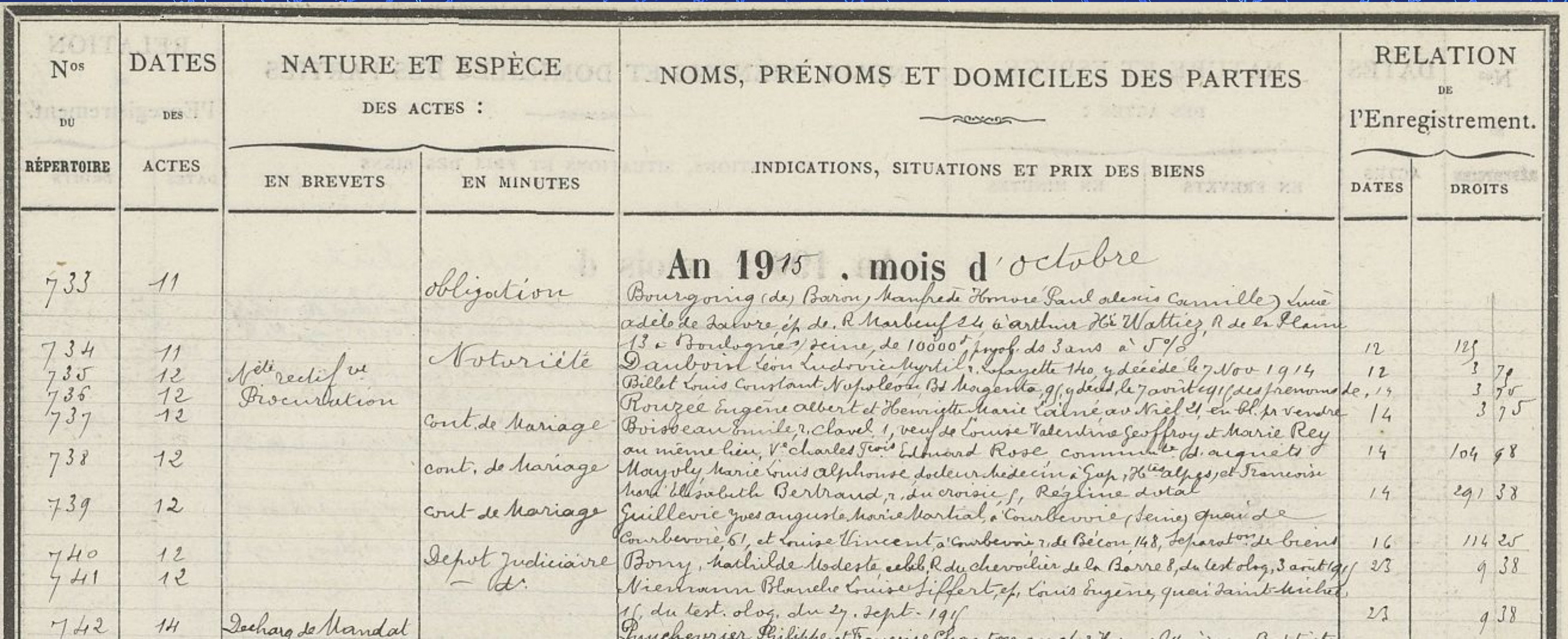
Datalab - BnF - 26/01/2021

Alix CHAGUÉ - Inria - alix.chague@inria.fr
Aurélia ROSTAING - Archives nationales - aurelia.rostaing@culture.gouv.fr



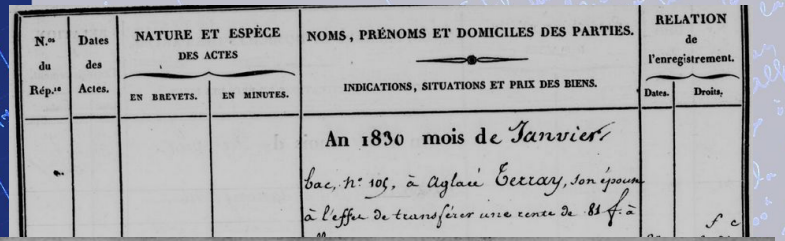
LECTAUREP en quelques mots

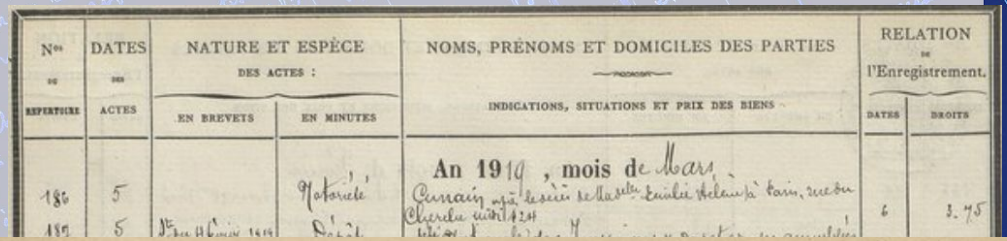
Lecture et exploitation de registres de notaires assistées par apprentissage machine

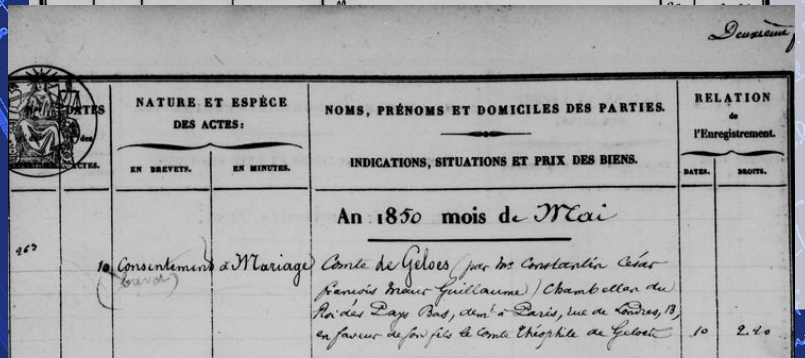


LECTAUREP en quelques chiffres

- Un ensemble physiquement cohérent (~ 2000 registres préimprimés, ~ 1803-1944), échantillonné en deux jeux d'images numériques
- Ecritures, abréviations et mises en page nombreuses et variées (des milliers de mains)
- Une variété de supports numériques

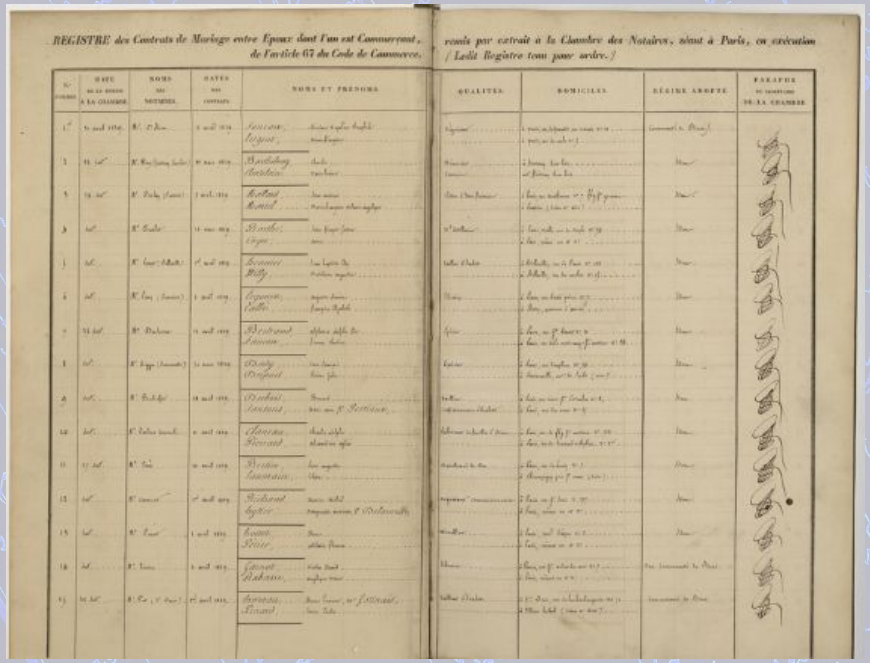


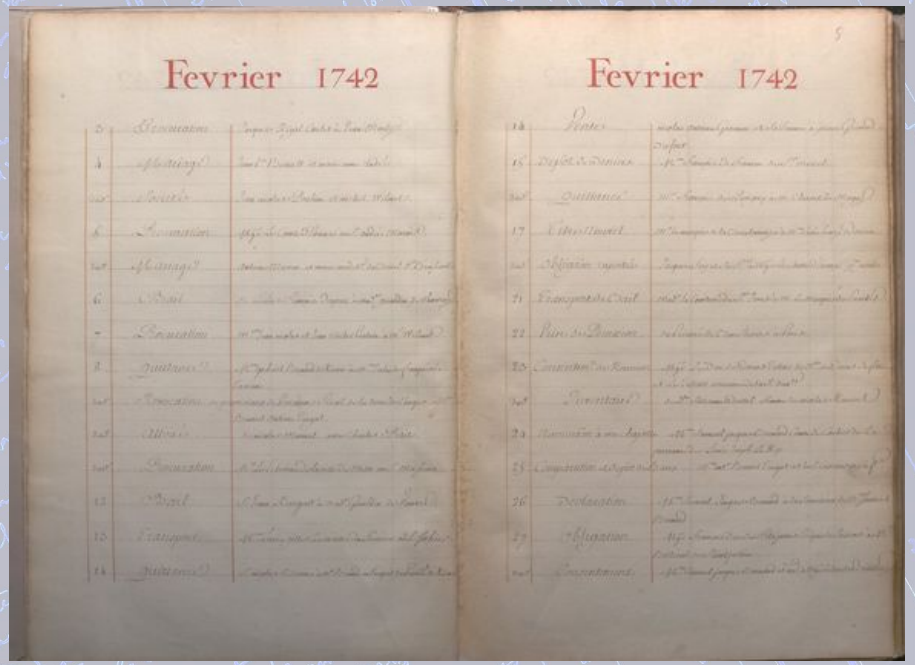




Plan B : réduire et simplifier le corpus

Contrats de mariage de négociants (41 registres, 1829-1934) ; Me Bronod (9 reg., 1719-1765) : écriture homogène et soignée, moins abrégée, plus aérée





Développements autour d'eScriptorium

déploiement

JAN 2021

- fonctionnalités déjà disponible dans eScriptorium
- contributions de LECTAUREP à eScriptorium
- solutions logicielles développées hors eScriptorium
- serveur

eScriptorium

PSL SCRIPTA

- administration des utilisateurs
- segmentation manuelle/automatique et édition
- transcription manuelle/automatique et édition
- import de segmentation/transcription
- association de métadonnées à des groupes d'images
- chargement d'images externes (sys. local / IIIF / PDF)
- export d'images/segments/transcription
- import/entraînement/affinage/export de modèles pour la segmentation ou la transcription
- annotation des segments
- partager des collections d'image et leur transcription

Formats gérés : XML ALTO, XML PAGE, TXT

Documentation et formation

Débogage courant

Affiner gestion des utilisateurs (profils / myriadisation)

Ajouter un scénario de lecture simple

Affiner gestion documentaire

Annoter la transcription (NER)

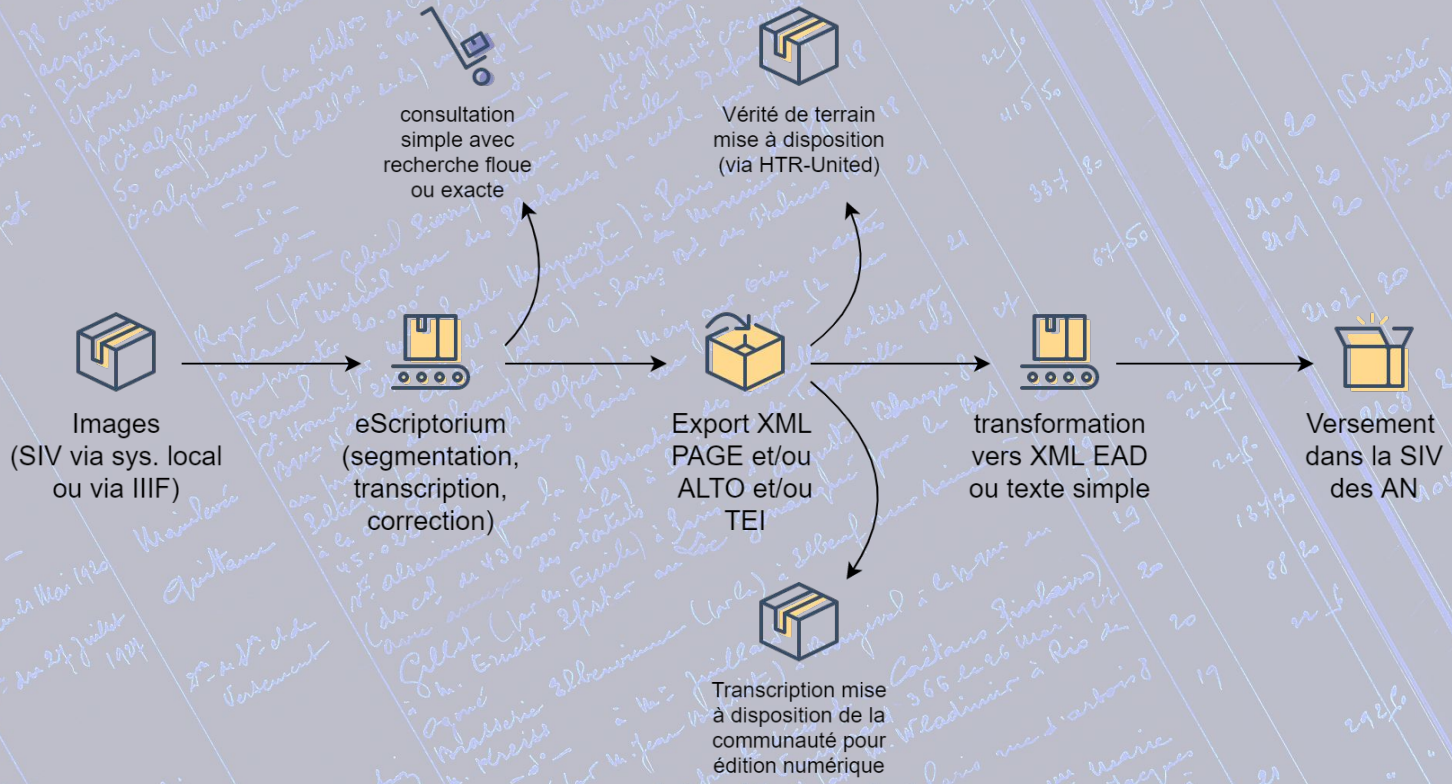
Gérer un export format XML TEI

Fonctionnalité de recherche (exacte ou floue) dans le texte

KaMI :
un dashboard pour évaluer les performances d'un modèle sur un ensemble de données

Aspyre :
un adaptateur pour la compatibilité des données à importer avec eScriptorium lorsque c'est nécessaire

Visualisation de la chaîne complète

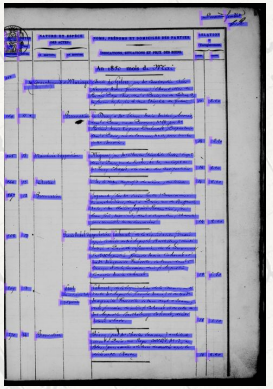


*HTR-United est un projet de Commons pour les données d'entraînement pour l'HTR : https://htr-united.github.io/

Stratégie d'entraînement des modèles

Segmentation :

- Un modèle générique opérationnel (*lectau1_8*) à affiner pour les cas particuliers (écritures serrées)

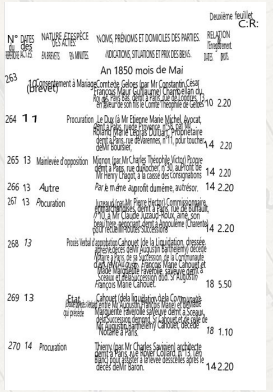


Transcription :

On ne peut pas entraîner un seul modèle pour l'ensemble du corpus !

- Un modèle générique (objectif : 20 % CER max.)
- Un affinage du modèle générique pour chaque nouvelle main d' écriture ou quand aucun modèle ne fonctionne (objectif : 10 % CER max.)

L'objectif est de réduire la quantité de données nécessaire pour l'entraînement d'un modèle spécifique à chaque écriture.



Production des données d'entraînement

La performance du modèle dépend de la qualité des données d'entraînement :

- On ne peut pas se contenter d'une transcription / image : il faut **plusieurs annotateurs pour une même image**, et une comparaison des transcriptions !
- Il est nécessaire d'établir **un guide d'annotation** et d'identifier les points de désaccord des annotateurs !
- Il faut **former** les annotateurs aux règles d'annotation qui ont été décidées
- Il faudra **former** les contributeurs à ces règles d'annotation

mainlevée	Mainlevée	mainlevée	mainlevée	mainlevée	mainlevée	mainlevée
de sa fille mineure, p^r accepter d^on à t. dep.ant à celle ci, par	de sa fille mineure, p^r accepter d^on à ts dep.ant à celle ci, par	de sa fille mineure, p^r acceper d^on à tr dep. ant à celle ci, par	de sa fille mineure, XX accepter à XX a celle ci, par	de sa fille mineure, pr accepter don à tr dep. ant àfor elle ci, par	de sa fille mineure, p^r accepter d^on à tr dep. ant à celle ci, par	de sa fille mineure, p^r acopter dôn à ts dep.ant à celle ci, par
(suite du 13.10.23)	(suite du 13.10.23)	(suite du 13.10.23)	(suite du 13.10.23)	(suite du 13.10.23)	(suite du 13.10.23)	(suite du 13.10.23)

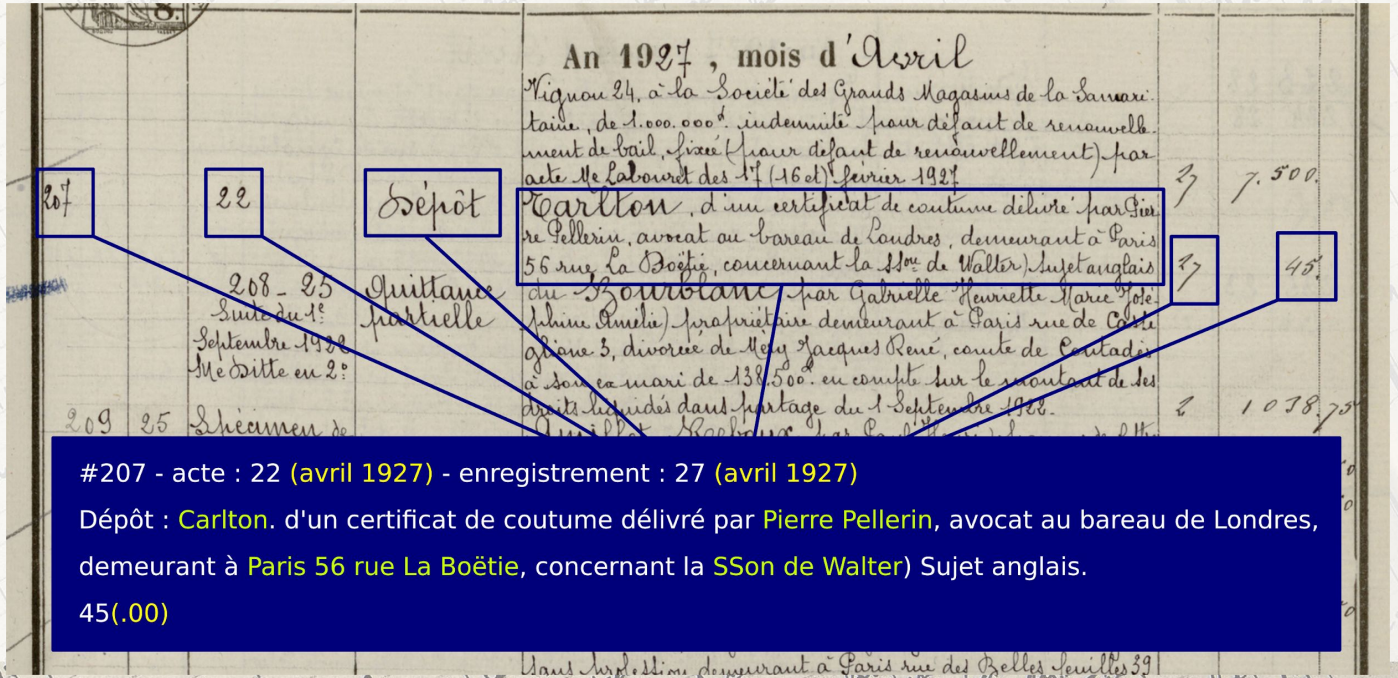
Extrait des résultats d'une expérimentation permettant d'illustrer les désaccords insoupçonnés !

Production des données d'entraînement
(segmentation & transcription)

- Commencée avec le premier confinement, en interne, surtout en alternance d'autres tâches *(4 agents, EPT : ~ 1,5 ; correction de la segmentation de 300 pages du golden set ; transcription de 700 pages du golden set : ~ 10/15 mains, ~ 1830/1836/1850/1901/1907, 6 notaires, 2 études)* ;
- Poursuivie à partir du 16/09, en interne, surtout en discontinu *(415 pages du random set par 5 agents, EPT 2 ; + correction de la segmentation et transcription de 254 doubles pages de registres de contrats de mariage par 8 personnes, EPT 2 ; commencé : Bronod, années 1740, 1 agent)* ;
- Une équipe aux compétences hétérogènes, avec de grands débutants et des Lectaurépiens plus chevronnés : une préfiguration de la phase participative, sans les outils d'accompagnement et de formation requis ;
- Des conventions qui doivent être évolutives en raison de la diversité des systèmes d'abréviations du corpus ;
- Un effet d'entonnoir à la phase cruciale de contrôle, correction et validation des données de vérité terrain (segmentation, transcription).

Exploration des transcriptions

- Un outil de recherche dans le texte intégral (exacte, floue ou par mots-clefs)
- Reconstitution des unités logiques et des informations (unité de l'acte, dates, sommes)
- Annotation des entités nommées, adresses, etc., alignement et visualisation



Objectifs et besoins de Lectaurep

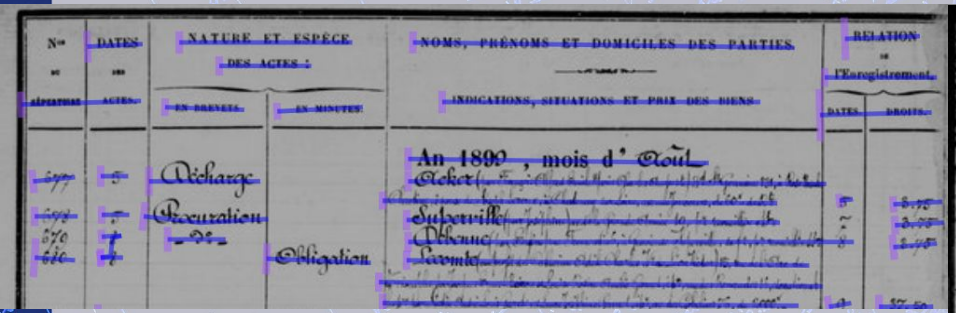
- **Minimiser les corrections manuelles (temps de correction < temps de saisie manuelle) pour optimiser la recherche floue -> outils d'analyse et métriques fiables**
- **CER rectifié** : apprécier le taux d'erreur par caractère en le réduisant à sa plus simple expression *(0 ≠ 1 ; A (ou a, a, à, á, ä, â) ≠ B (ou b))* afin d'évaluer l'efficacité d'une recherche floue (patronymes, métiers, mots matière...) ;
- **CER + WER** : disposer des deux valeurs pour évaluer la répartition des erreurs (concentrée sur certaines chaînes de caractères ou diffuses ?) ;
- **Brique participative + animation de communauté** : pour produire la vérité terrain (segmentation, transcription) nécessaire au projet ;
- **Fonctionnalités** : copier-coller des segments et des contenus, faire des corrections en masse au vu de l'original (cf. Open Refine), indiquer les modèles dans le fichier alto...

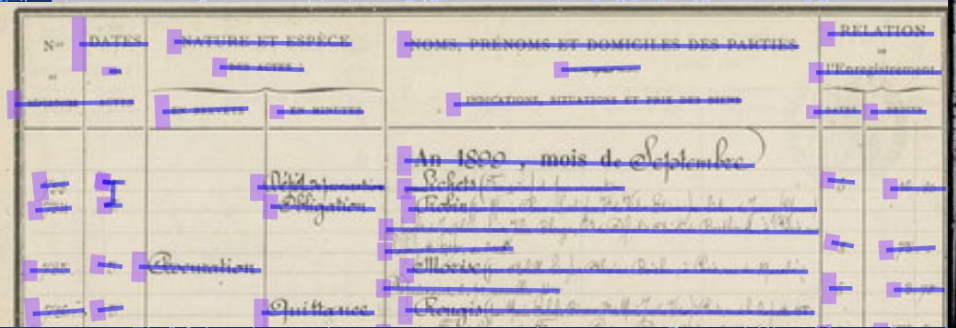
Premières hypothèses de Lectaurep

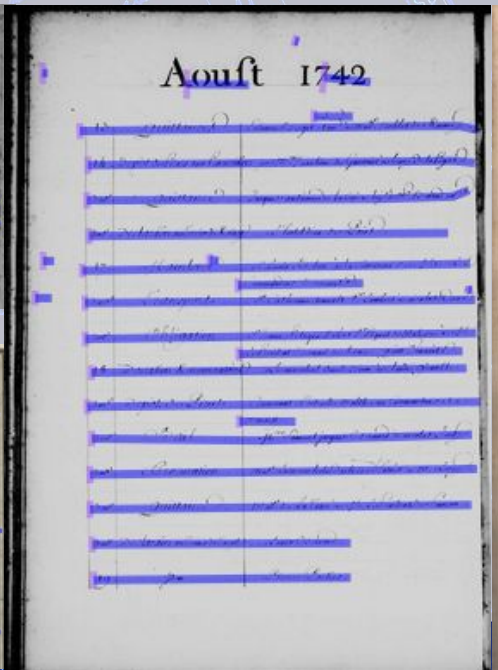
Segmentation

- numérisation NB d'après microfilm ~ couleur d'après originaux ? *(> HTR à la demande ?)*
- pas d'inconvénient à partir d'une double page plutôt que de son découpage en deux pages

HTR : ?







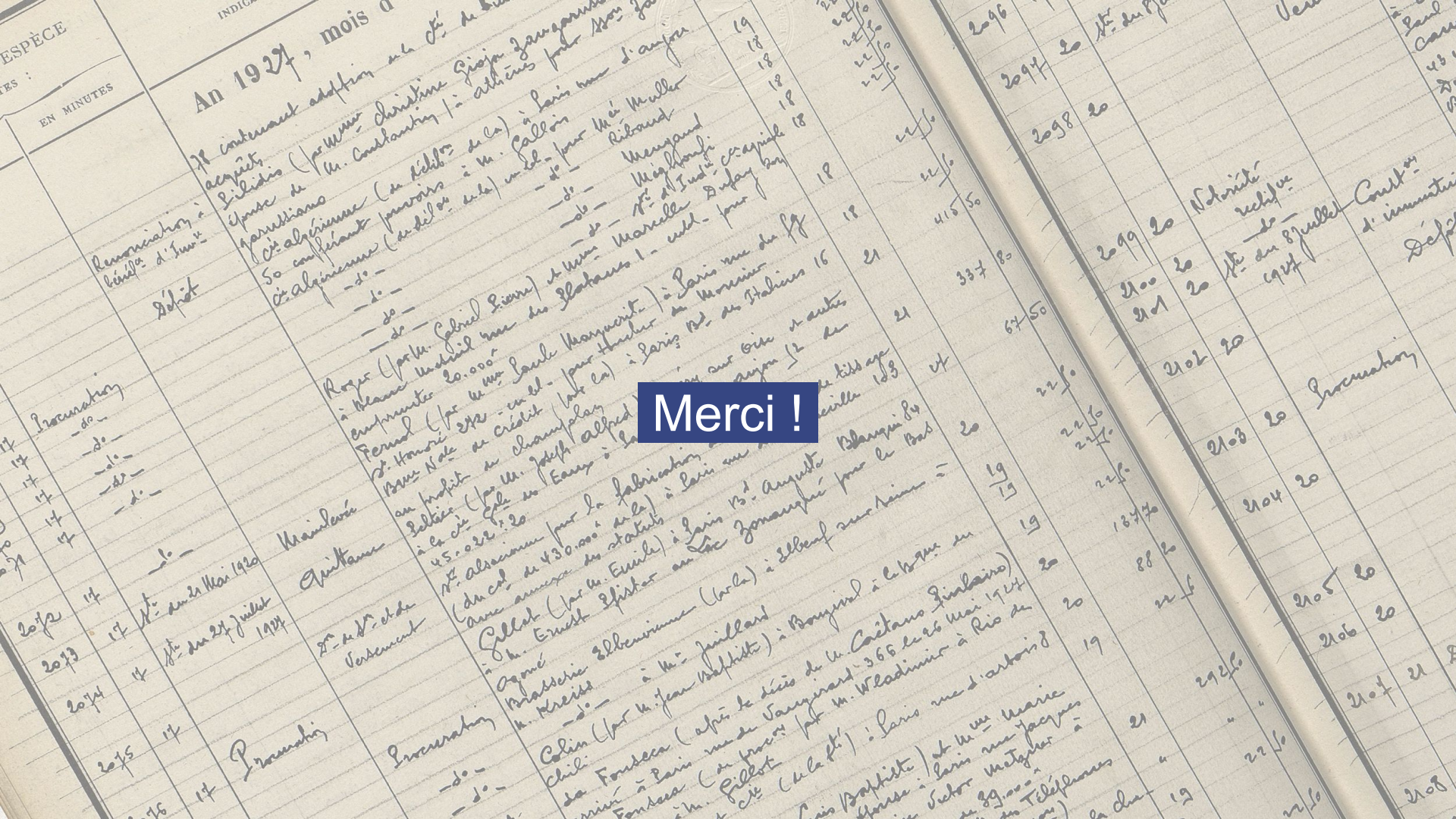
Recommandations pour une gestion de projet d'HTR

Calibrer le projet en amont, de A à Z, pour le maîtriser.

- 📖 **Corpus** : homogène, simple, limité *(écritures, abréviations, mise en page et autres aspects matériels)* ;
- 🛰️ **Suivi de projet** : campagne d'HTR ~ campagne de numérisation ? *(récolement page à page, conventions - CCH faisant l'unanimité...)* ;
- **RH** : équipe projet à TP, restreinte (?), formée et testée ;
- 👀 **Matériel** : grand écran (voire deux grands écrans), surtout avec des doubles pages ;
- **Logiciel** : utilisable en l'état, sans devoir développer des fonctionnalités supplémentaires ;
- ⏳ **Calendrier du projet** : **évaluer** le temps d'obtention des **modèles** de segmentation/transcription/annotation sémantique, des données de vérité terrain en **testant** un **échantillon représentatif éventuellement "maison"** (segmentation, HTR manuels/automatiques, annotation ; temps de **formation**, de **correction**) ;
- 📅 **Données** (vérité terrain, modèles...) : **plan de gestion et de documentation** pendant et après leur production ; modalités de réutilisation.

Enjeux interprofessionnels (GLAM)

- Mêmes problématiques, mêmes types de fonds (registres à colonne, mixtes - manuscrit / imprimé...)
- Documenter, formaliser et harmoniser les pratiques (grilles projet types à décliner ; référentiels et standards de segmentation et de transcription, nécessaires pour offrir des données interopérables à la paléographie computationnelle) ;
- Cartographier et documenter les projets et les supports d'HTR (logiciels, serveurs, corpus, financements, RH : Biblissima+, DIM MAP Cremma...) ;
- Écrire un manuel de référence sur l'HTR ;
- Ecrire un guide *Ecrire un cahier des charges d'HTR*.


Merci !