

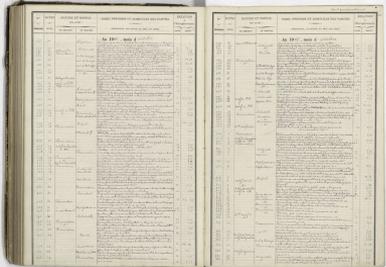
# LECTAURE

## Lecture automatique de répertoires de notaires

Datalab - BnF - 26/01/2021

Alix CHAGUÉ - Inria - [alix.chague@inria.fr](mailto:alix.chague@inria.fr)

Aurélia ROSTAING - Archives nationales - [aurelia.rostaing@culture.gouv.fr](mailto:aurelia.rostaing@culture.gouv.fr)



# LECTAUREP en quelques mots

Lecture et exploitation de registres de notaires assistées par apprentissage machine

N <sup>os</sup> DU RÉPERTOIRE	DATES DES ACTES	NATURE ET ESPÈCE DES ACTES :		NOMS, PRÉNOMS ET DOMICILES DES PARTIES  INDICATIONS, SITUATIONS ET PRIX DES BIENS	RELATION DE L'Enregistrement.	
		EN BREVETS	EN MINUTES		DATES	DROITS
733	11		obligation	An 1915, mois d'octobre Bourgoing (de) Baron, Manfrede Honoré Paul Alexis Camille, Lucie adèle de Labore ép. de R. Marbeuf s. à Arthur H. Wattiez, R. de la Plaine 13 à Doulogne, Seine, de 10000 <sup>fr</sup> prof. de 3 ans à 5%	12	125
734	11		Notoriété	Darbois Louis Ludovic Baptiste, Lafayette 110, y décède le 7 Nov 1914	12	3 70
735	12	Noti rectif ve		Billet Louis Constant Napoléon, Bd Margherita, 9, y décède le 7 août 1911 des fièvres de,	14	3 75
736	12	Procuration		Rouzeé Eugène Albert et Henriette Marie Calmé av. Nief et en l'h. pr vendre	14	3 75
737	12	cont. de mariage		Boisbeau Emile, r. Chavel 1, veuf de Louise Valentine Geoffroy et Marie Rey au même lieu, v. Charles Feo, Edmond Rose commis. St. aignets	14	104 98
738	12	cont. de mariage		Mauvoly Marie Louis Alphonse docteur médecin à Gap, H. les Alpes, et Françoise Marie Elisabeth Bertrand, r. du croisé, 1, Reoline dotal	14	291 38
739	12	cont. de mariage		Guillevie Yves Auguste Marie Martial, à Comberveire (Seine) quai de Comberveire, 61, et Louise Vincent, à Comberveire r. de Bécon 148, séparat <sup>on</sup> de biens	16	114 25
740	12		Depot judiciaire	Bony, Mathilde Modeste, ecb. R. du chevalier de la Barre 8, du test olog, 3 août 1911	23	9 38
741	12		— Id.	Niermann Blanche Louise Siffert, ép. Louis Eugène, quai saint-Nicolas, 16, du test. olog. du 27. sept. 1911	23	9 38
742	14	Decharo de Mandat		Duchessier Philiberte, r. de la Seine, séparat <sup>on</sup> de biens, au lieu de la Barre		



# 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

REGISTRE des Contrats de Mariage entre Epoux dont l'un est Commerçant, de l'article 67 du Code de Commerce.

roulé par extrait à la Chambre des Notaires, ainsi à Paris, en exécution (Ledit Registre tenu par ordre.)

N <sup>o</sup>	DATE DE LA CHAMBRE	NOM DE L'ÉPOUX	DATE DU CONTRAT	NOM ET PRÉNOMS	QUALITÉ	DOMICILE	ÉCRITURE	PARAPHE
1 <sup>er</sup>	26 août 1719	M. P. B...	1 août 1719	Marie-Anne de la Roche	épouse	Paris	Comm. de Paris	
2	11 août	M. de la Roche	10 août 1719	Barthelemy de la Roche	notaire	Paris		
3	18 août	M. de la Roche	17 août 1719	Barthelemy de la Roche	notaire	Paris		
4	25 août	M. de la Roche	24 août 1719	Barthelemy de la Roche	notaire	Paris		
5	1 <sup>er</sup> sept	M. de la Roche	30 août 1719	Barthelemy de la Roche	notaire	Paris		
6	8 sept	M. de la Roche	7 sept 1719	Barthelemy de la Roche	notaire	Paris		
7	15 sept	M. de la Roche	14 sept 1719	Barthelemy de la Roche	notaire	Paris		
8	22 sept	M. de la Roche	21 sept 1719	Barthelemy de la Roche	notaire	Paris		
9	29 sept	M. de la Roche	28 sept 1719	Barthelemy de la Roche	notaire	Paris		
10	6 oct	M. de la Roche	5 oct 1719	Barthelemy de la Roche	notaire	Paris		
11	13 oct	M. de la Roche	12 oct 1719	Barthelemy de la Roche	notaire	Paris		
12	20 oct	M. de la Roche	19 oct 1719	Barthelemy de la Roche	notaire	Paris		
13	27 oct	M. de la Roche	26 oct 1719	Barthelemy de la Roche	notaire	Paris		
14	3 nov	M. de la Roche	2 nov 1719	Barthelemy de la Roche	notaire	Paris		
15	10 nov	M. de la Roche	9 nov 1719	Barthelemy de la Roche	notaire	Paris		
16	17 nov	M. de la Roche	16 nov 1719	Barthelemy de la Roche	notaire	Paris		
17	24 nov	M. de la Roche	23 nov 1719	Barthelemy de la Roche	notaire	Paris		
18	1 <sup>er</sup> dec	M. de la Roche	30 nov 1719	Barthelemy de la Roche	notaire	Paris		
19	8 dec	M. de la Roche	7 dec 1719	Barthelemy de la Roche	notaire	Paris		
20	15 dec	M. de la Roche	14 dec 1719	Barthelemy de la Roche	notaire	Paris		
21	22 dec	M. de la Roche	21 dec 1719	Barthelemy de la Roche	notaire	Paris		
22	29 dec	M. de la Roche	28 dec 1719	Barthelemy de la Roche	notaire	Paris		

Fevrier 1742

Fevrier 1742

1	Blanchard	épouse de la Roche	18	Blanchard	épouse de la Roche
2	Blanchard	épouse de la Roche	19	Blanchard	épouse de la Roche
3	Blanchard	épouse de la Roche	20	Blanchard	épouse de la Roche
4	Blanchard	épouse de la Roche	21	Blanchard	épouse de la Roche
5	Blanchard	épouse de la Roche	22	Blanchard	épouse de la Roche
6	Blanchard	épouse de la Roche	23	Blanchard	épouse de la Roche
7	Blanchard	épouse de la Roche	24	Blanchard	épouse de la Roche
8	Blanchard	épouse de la Roche	25	Blanchard	épouse de la Roche
9	Blanchard	épouse de la Roche	26	Blanchard	épouse de la Roche
10	Blanchard	épouse de la Roche	27	Blanchard	épouse de la Roche
11	Blanchard	épouse de la Roche	28	Blanchard	épouse de la Roche
12	Blanchard	épouse de la Roche	29	Blanchard	épouse de la Roche
13	Blanchard	épouse de la Roche	30	Blanchard	épouse de la Roche

# Développements autour d'eScriptorium

JAN 2021

 fonctionnalités déjà disponible dans eScriptorium

 contributions de LECTAUREP à eScriptorium

 solutions logicielles développées hors eScriptorium

 serveur



déploiement

eScriptorium

PSL  SCRIPTA  
UNIVERSITÉ PARIS UNIVERSITÉ PARIS UNIVERSITÉ PARIS

Documentation et formation

Débogage courant

Affiner gestion des utilisateurs  
(profils / myriadisation)

Ajouter un scénario de lecture  
simple

- 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

Affiner gestion documentaire

Annoter la transcription (NER)

Gérer un export format  
XML TEI

Fonctionnalité de recherche  
(exacte ou floue) dans le texte



**KaMI :**

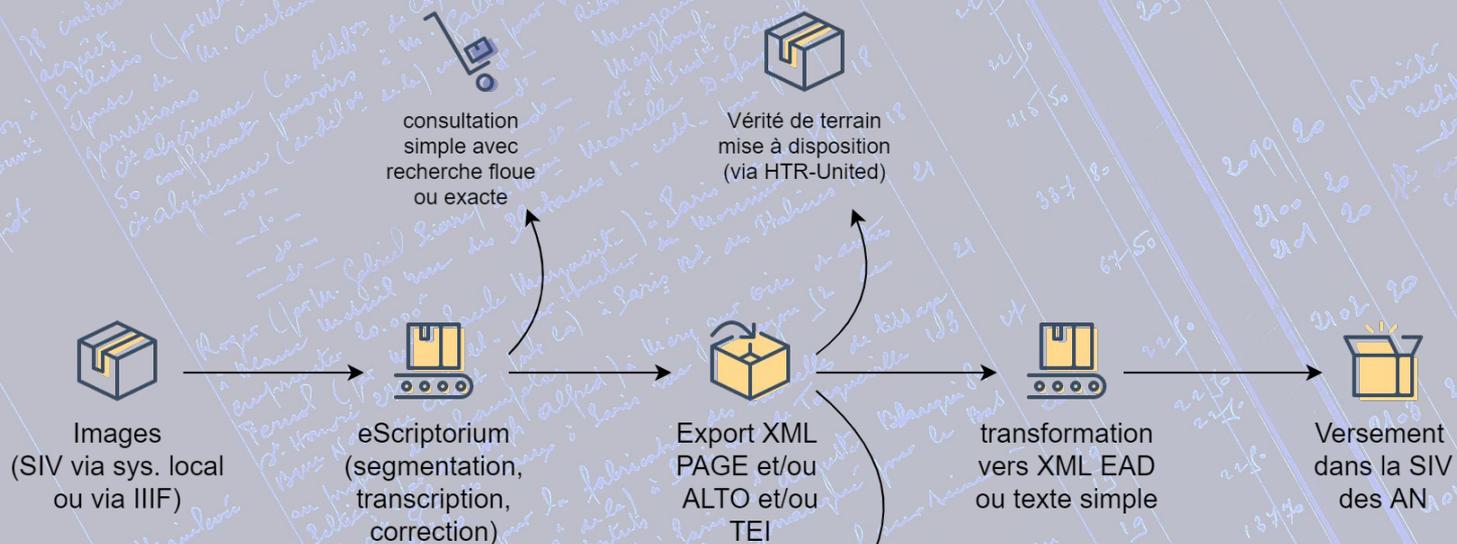
un dashboard pour évaluer les performances d'un modèle sur un ensemble de données



**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-unity.github.io/>

# Stratégie d'entraînement des modèles

## Segmentation :

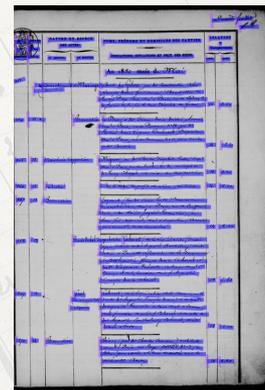
- ❖ Un modèle générique opérationnel (*lectau1\_8*) à affiner pour les cas particuliers (écritures serrées)

## 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.*



N° DATES	NATURE ESPÈCE DES ACTES	INDICATIONS SITUATIONS ET MOULDES DES PARTIES	RELATION Emplacement des sig.
253	Procuration	Le Comptable à Mairie-Contable Salvo par Mr Coqueret César	10 2.20
254	1 1	Procuration Le Duc de M. Ebene Marie Michel, avocat	4 2.20
255	13	Mariée d'opposition Monon Jean Marie Thérèse Victor Joseph	14 2.20
256	13	Autre Pare m.ère Agnès d'années autriches	14 2.20
257	13	Procuration Procuration de M. de la Roche de la Roche de la Roche	4 2.20
258	73	Proès Notaire Cantoulet de la Liquidation de la	18 5.50
259	13	État Procuration de M. de la Roche de la Roche de la Roche	18 1.10
270	14	Procuration Procuration de M. de la Roche de la Roche de la Roche	14 2.20

# 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 <sup>r</sup> accepter d <sup>l</sup> on à t. dep.ant à celle ci, par	de sa fille mineure, p <sup>r</sup> accepter d <sup>l</sup> on à ts dep.ant à celle ci, par	de sa fille mineure, p <sup>r</sup> accep <sup>r</sup> er d <sup>l</sup> on à tr dep. ant à celle ci, par	de sa fille mineure, <b>XX</b> accepter à <b>XX</b> a celle ci, par	de sa fille mineure, pr accepter don à tr dep. ant àfor elle ci, par	de sa fille mineure, p <sup>r</sup> accepter d <sup>l</sup> on à tr dep. ant à celle ci, par	de sa fille mineure, p <sup>r</sup> aco <sup>p</sup> ter d <sup>l</sup> on 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

An 1927, mois d'Avril  
Niveau 24, à la Société des Grands Magasins de la Samaritaine, de 1.000.000<sup>fr</sup> (indemnité pour défaut de renouvellement de bail, ainsi qu'avant défaut de renouvellement) par acte Me Labouret des 17 (16 et 17) février 1927

207 22 Dépôt Carlton, d'un certificat de coutume délivré par Pierre Pellerin, avocat au bureau de Londres, demeurant à Paris 56 rue La Boétie, concernant la S<sup>on</sup> de Walter) Sujet anglais

208 25 Suite du 1<sup>er</sup> Septembre 1928 Me Dittie en 2<sup>o</sup>

209 25 Spécimen de

3 7.500

3 45

1038,75

#207 - acte : 22 (avril 1927) - enregistrement : 27 (avril 1927)  
Dépôt : **Carlton**, d'un certificat de coutume délivré par **Pierre Pellerin**, avocat au bureau de Londres, demeurant à **Paris 56 rue La Boétie**, concernant la **S<sup>on</sup> de Walter**) Sujet anglais.  
45(.00)

# 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 ( $0 \neq 1$  ;  $A$  (ou  $a, \grave{a}, \acute{a}, \ddot{a}, \hat{a}$ )  $\neq B$  (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...



# 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*.

