



HAL
open science

Présentation du projet Lectaurep (Lecture automatique de répertoires)

Alix Chagué, Aurélia Rostaing

► To cite this version:

Alix Chagué, Aurélia Rostaing. Présentation du projet Lectaurep (Lecture automatique de répertoires). Atelier sur la transcription des écritures manuscrites - BnF DataLab, Jan 2021, Paris, France. hal-03122019

HAL Id: hal-03122019

<https://hal.science/hal-03122019>

Submitted on 26 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

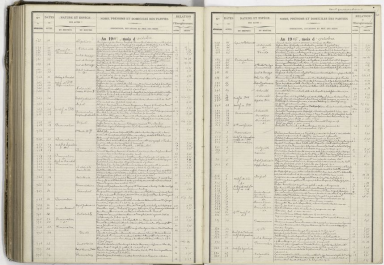
LECTAUREP

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

N ^{os} 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, Manfred de Honore Paul Alexis Camille, Lucie adèle de Labore ép. de R. Marbeuf s. à Arthur de Wattiez R de la Plaine 13 à Doulogne, Seine, de 10000 ^{fr} 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		Rouzeel 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 Jean Edmond Rose commis. St. aignets	14	104 98
738	12	cont. de mariage		Mauvoly Marie Louis Alphonse docteur médecin à Gap, H ^{tes} 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 Courbevoie (Seine) quai de Courbevoie, 61, et Louise Vincent, à Courbevoie r. de Bécon 148, séparat ^{on} 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	Decharg. de Mandat		Duchessier Philiberte, r. de la Seine, Paris, 14, y décède le 14 août 1911		

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 ^o	DATE DE LA CHAMBRE	NOM DE L'ÉPOUX	DATE DU CONTRAT	NOM ET PRÉNOMS	QUALITÉ	DOMICILE	ÉCRITURE	PARAPHE
1 ^{er}	26 août 1719	M. P. de ...	1 août 1719	de
2	11 août	M. de ...	10 août 1719	de
3	18 août	M. de ...	1 août 1719	de
4	août	M. de ...	11 août 1719	de
5	août	M. de ...	1 août 1719	de
6	août	M. de ...	1 août 1719	de
7	11 août	M. de ...	11 août 1719	de
8	août	M. de ...	11 août 1719	de
9	août	M. de ...	11 août 1719	de
10	août	M. de ...	11 août 1719	de
11	11 août	M. de ...	11 août 1719	de
12	août	M. de ...	11 août 1719	de
13	août	M. de ...	11 août 1719	de
14	août	M. de ...	11 août 1719	de
15	août	M. de ...	11 août 1719	de
16	août	M. de ...	11 août 1719	de
17	août	M. de ...	11 août 1719	de
18	août	M. de ...	11 août 1719	de
19	août	M. de ...	11 août 1719	de


Fevrier 1742

Fevrier 1742

1	de	12	de
2	de	13	de
3	de	14	de
4	de	15	de
5	de	16	de
6	de	17	de
7	de	18	de
8	de	19	de
9	de	20	de
10	de	21	de
11	de	22	de
12	de	23	de
13	de	24	de
14	de	25	de
15	de	26	de
16	de	27	de
17	de	28	de
18	de	29	de
19	de	30	de

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 DE 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 :

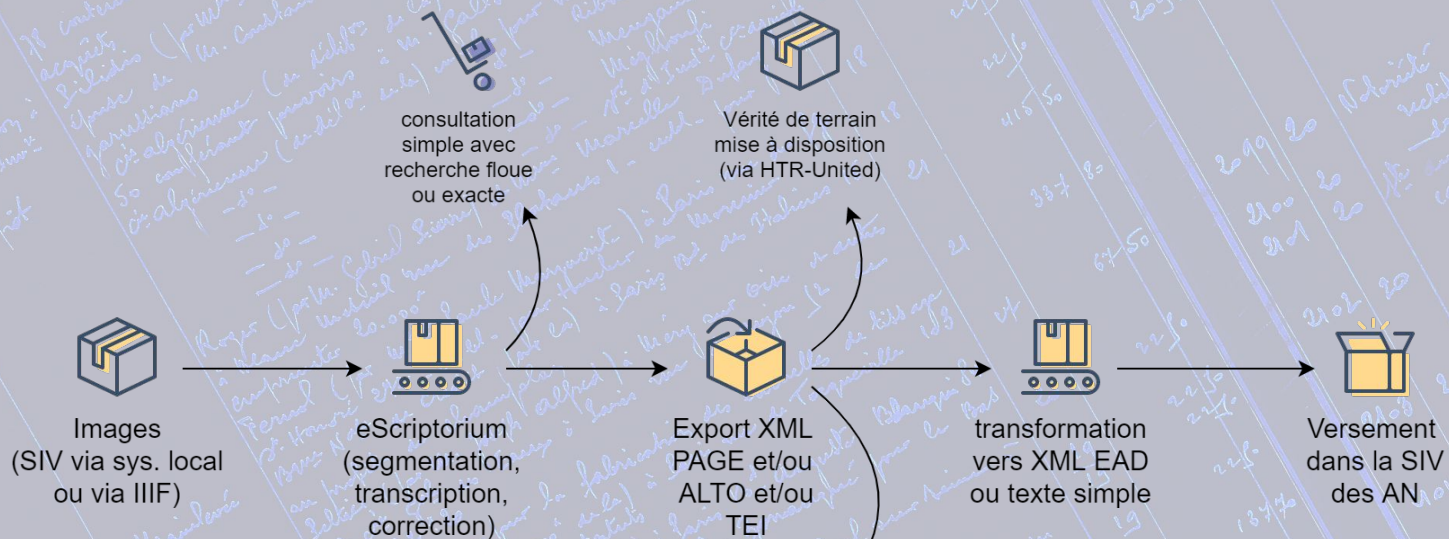
un dashboard pour évaluer les performances d'un modèle sur un ensemble de données



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/>

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 ^{ar} accepter d ^{on} à t. dep.ant à celle ci, par	de sa fille mineure, p ^{ar} accepter d ^{on} à ts dep.ant à celle ci, par	de sa fille mineure, p ^{ar} accep ^{er} 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 ^{ar} accepter d ^{on} à tr dep. ant à celle ci, par	de sa fille mineure, p ^{ar} acopter d ^{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^{fr} (indemnité pour défaut de renouvellement de bail, ainsi qu'au cas de 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^{on} de Walter) Sujet anglais.

208 25 Suite du 1^{er} Septembre 1928 Me Dittie en 2^o

209 25 Spécimen de

7.500

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^{on} 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*.

