



**HAL**  
open science

## LectAuRep (2018-2021) :Projet de lecture automatique de répertoires de notaires

Aurélia Rostaing, Hugo Scheithauer

### ► To cite this version:

Aurélia Rostaing, Hugo Scheithauer. LectAuRep (2018-2021) :Projet de lecture automatique de répertoires de notaires. Segmenter et annoter les images : déconstruire pour reconstruire, Nov 2022, Paris, France. hal-03855439

**HAL Id: hal-03855439**

**<https://hal.science/hal-03855439v1>**

Submitted on 16 Nov 2022

**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.

# Segmenter et annoter les images *déconstruire pour reconstruire*

## LectAuRep (2018-2021)

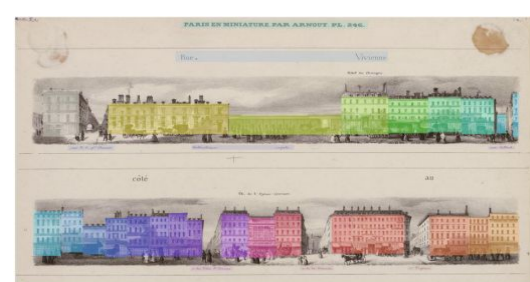
### Projet de lecture automatique de répertoires de notaires

Aurélia Rostaing (Archives nationales)

Hugo Scheithauer (ALMAnaCH, Inria)



INHA, salle Vasari, 15 novembre 2022



Jules Arnoux, *Paris en miniature*, "Rue Vivienne, côté au levant, Rue Vivienne, côté du couchant", pl. 246, musée Carnavalet, G. 22998, segmentation des façades - Projet Richelieu. Histoire du quartier (détail).

(BnF



PSL



N° REPERTOIRE	BITS NATURE ET ESPÈCE DES ACTES	INDICATIONS, SITUATIONS ET PAIX DES BIENS	RELATION de l'Enregistrement.	
			DATES	DRITES
1357	8	CONFÉ DE PP <sup>te</sup>		
1359	8	d°		
1360	8	d°		
1361	8	(M <sup>re</sup> D'elare) Mariage		
1362	8	Securition (M <sup>re</sup> D'elare)		
1363	8	Mariage d°		

# 1803-années 1940 : une mise en page structurée

N <sup>o</sup> de séparatice	DATES ACTES	NATURE ET ESPÈCE DES ACTES :		NOMS, PRÉNOMS ET DOMICILES DES PARTIS  INDICATIONS, SITUATIONS ET DROITS DES BIENS	RELATION de l'Enregistrement.	
		EN BREVETS	EN MINUTES		DATES	DROITS
86	28		<del>Acte judiciaire</del>	An 1801, mois de Janvier (Luce), Bellevue (du testament de Victor Edouard) de son en- son domicile à Paris, rue du Rivoli n. 11. 19 Janvier 1801.	2 fév	9.38
87	28		<del>Acte judiciaire</del>	Montcaux (du testament olographe de Julie Boye) de son domicile à Paris en 1801, domicile sa rue Victor Lévy, n. 11. 11 Janvier 1801.	4 fév	28.58
88	28		<del>Acte de charges</del>	de Calleraud. Perigord, à la requête de Napoléon Louis Eugène Alexandre Charles Emmanuel Comte), à Paris, boulevard des Capucines n. 11. Sans paiement à la vente de propriété à Paris, rue de Valenciennes n. 11. 11 1801. - mise à prix 100,000 fr.	4 fév	11.10
89	28	Substitution		Quelques années donnée par M. George Alton, de Berlin, à M. de France, de prescription à lui donnée par son veuve Hilkenfeld, née).	6 fév	3.71
94	29		<del>Obligation</del>	Aine (par Louis Auguste) à Paris, place Vendôme, n. 11. Jacques Doucet, buye de la Seine n. de 1800. avec cautionnement sur fonds de commerce place Vendôme n. 11, et autres fonds. 1801.	10 fév	50.71

# Développer l'exploration des données textuelles contenues dans des images de répertoires

## HTR (segmentation linéaire)

> faciliter la recherche en texte intégral (ex. *Jacques Doucet*)

## HTR + KWS/REN (segmentation linéaire discontinue et segmentation par zones)

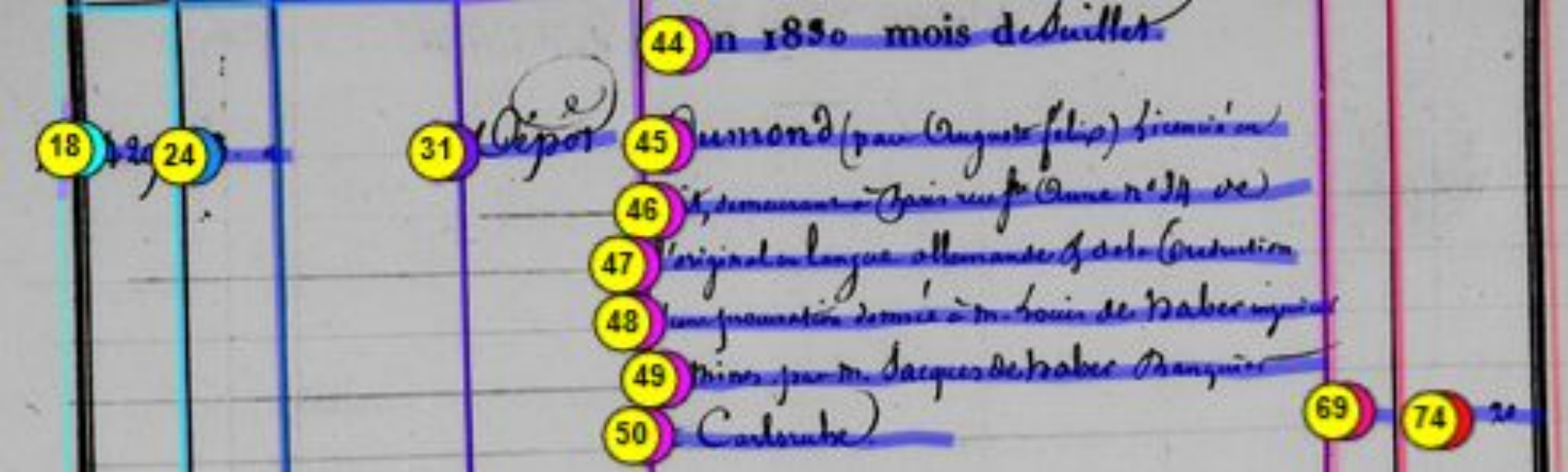
> favoriser la constitution de corpus de recherche à l'intérieur de ce corpus (ex. : *le milieu de l'édition, les photographes, cinéastes... dans le quartier/la rue de Richelieu entre telle et telle date*).

# Segmenter les images numériques

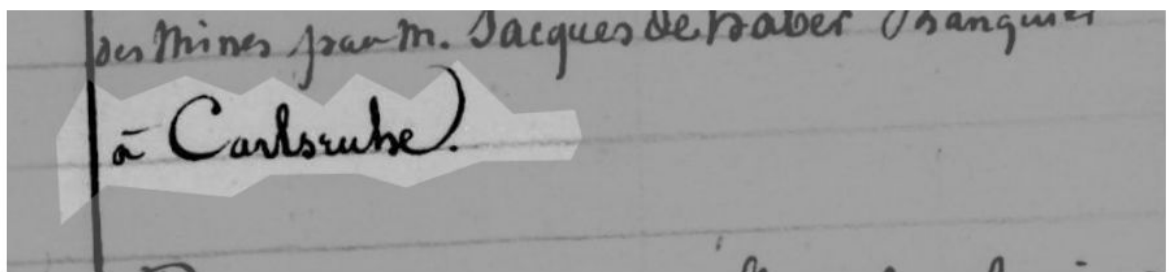
Segmenter le texte de chaque colonne, segmenter les zones de la page (*colonnes et unités documentaires*) pour en extraire le **texte**, la **hiérarchie interne**, et isoler des **entités correctement corrélées** entre elles.

17 r°-48 r°      Liste chronologique des actes pour la période du 2 janvier au 31 décembre 1901      Permalien      Télé

82	28	<del>Cahier de Charges</del>	de Calleyrand - Perigord (à la requête de Napoléon Louis Eugène Alexandre Arque Emmanuel Comte), à Paris, boulevard des Invalides, 37, - pour parvenir à la vente de propriété à Paris, rue de Crivelle, 36, 37 <sup>bis</sup> et 37, - mise à prix 789,000 <sup>f</sup> .	4 fév.	11.10
83	28	Substitution	Kugelmann (donnée par M <sup>rs</sup> George Aron, de Berlin, à - en blanc - de procuration à lui donnée par M <sup>rs</sup> veuve Lohmefeld, née).	6 fév.	3.71
84	29	<del>Obligation</del>	Aine (par Louis Eugène) à Paris, place Vendôme, 1, à Jacques Doucet, vue de la Paix, 21, - de 10,000 <sup>f</sup> ... avec nantissement sur fonds de commerce place Vendôme 21, - et autres fonds.	10 fév.	62.
85	30	<del>Acceptation de bail</del>	La Française, Electrique (par la Société anonyme) siège à Paris, rue de Crivelle, 44, - du bail de l'union à Paris, rue de Crivelle, 44, - fait par M <sup>rs</sup> Cardozo, suivant acte reçu par M <sup>rs</sup> Bertrand et M <sup>rs</sup> Fontana, le 31 juillet 1901, - agréement par M <sup>rs</sup> Gustave Steverlynck, - de Lille, - et autres.	10 fév.	18.71



Line #50



à Carlsruhe.

# Conserver la structure logique des répertoires de notaires après la transcription

429.3.11

An 1830 mois de Juillet  
Dépot Dumond (par Auguste Félix) Licencié en  
droit, demeurant à Paris rue Ste Anne n°34 de  
l'original en langue allemande & de la Traduction  
d'une procuration donnée à M. Louis de Haber ingénieur  
des Mines par M. Jacques de Haber Banquier  
à Carlsruhe.

12 2 20

46 droit, demeurant à Paris rue Ste Anne n°34 de  
47 l'original en langue allemande & de la Traduction  
48 d'une procuration donnée à M. Louis de Haber ingénieur  
49 des Mines par M. Jacques de Haber Banquier  
50 à Carlsruhe.

44 An 1830 mois de Juillet  
45 Dumond (par Auguste Félix) Licencié en  
46 droit, demeurant à Paris rue Ste Anne n°34 de  
47 l'original en langue allemande & de la Traduction  
48 d'une procuration donnée à M. Louis de Haber ingénieur  
49 des Mines par M. Jacques de Haber Banquier  
50 à Carlsruhe.

56 Procuration  
57 Audtorisation  
58 Certificat de ppte  
59 Procuration  
60 Mainlevée  
61 de signification  
62 Mainlevée  
63 Mainlevée  
64 Mainlevée  
65 Mainlevée  
66 Obligation  
67 Dépôt  
68 Transport  
69 Vente  
70 Constitution  
71 deSequestre  
72 Décharge  
73 Procuration  
74 Substitution  
75 Quittance  
76 Vente  
77 An 1914, mois d. Fevrier  
78 à Paris 34 rue Sedaine  
79 Lamour (par Jn B^e) employé à la Caserne de la G^de Républicaine  
80 à Paris rue St Etienne du Mont en blanc pour recueillir SS^on  
81 Boettcher (par Henry Emile) Ing^t conseil à Paris 39 Bd St Martin  
82 à Léontine Jeanne Blanchot safe dt avec lui pour gérer Comce  
83 Boileau-Desombres (parlaSté) à Paris 73 rue Lafayette d'inxn c/  
84 Sophie Marie Louise Dessoliers épse de Edouard Lavelaine de Maubeuge  
85 Gront (par Jean Eugène André) direstt d'usme à Bahigny route des  
86 Petits Ponts à André Valern Fortin ind^el à Paris 34 rue Sedaine

Extrait d'une transcription,  
exportée en texte brut

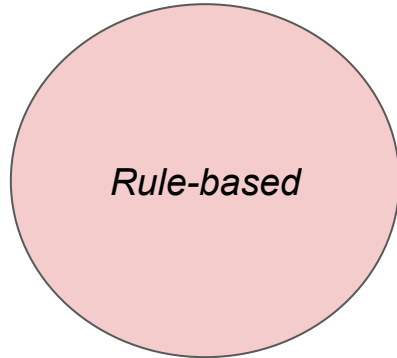
Le texte se retrouve  
complètement mis à plat,  
sans information de  
structure.

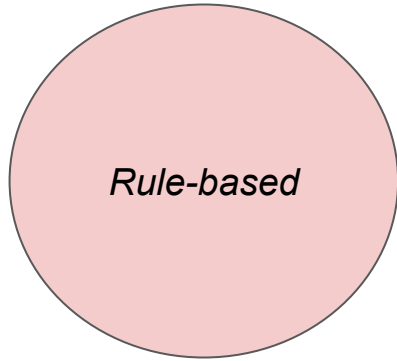


# Pourquoi reconstruire la structure ?

- Aller plus loin que l'exploration du texte intégral,
- Traiter le texte comme une base de données (par exemple, restructurer en XML-TEI, XML-EAD),
- Accéder aux niveaux atomiques d'information,
- Cibler l'extraction d'information,
- Recherches à facettes,
- Exploration de corpus,
- etc.

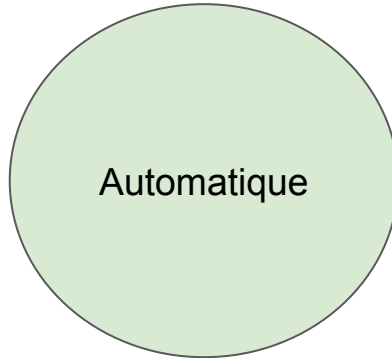
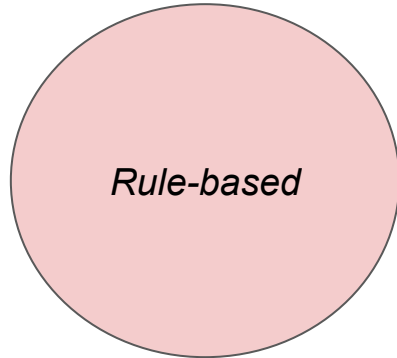
# Quelles stratégies pour restructurer les transcriptions ?

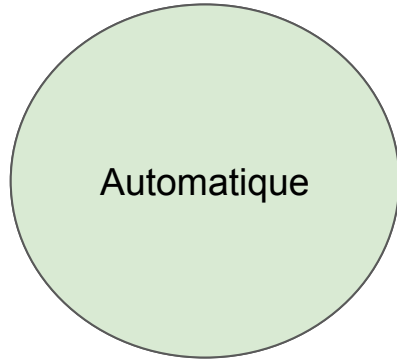




Trouver des marqueurs structurants  
et définir un ensemble de règles.

# Quelles stratégies pour restructurer les transcriptions ?





Utiliser des outils d'apprentissage machine :

- Détection d'objets, classification de pixels
- Classification de tokens
- etc.

# Détection des différentes régions de texte dans un répertoire de notaire avec Kraken et visualisation sur eScriptorium

N <sup>os</sup> du RÉPERTOIRE	DATES	NATURE ET ESPÈCE DES ACTES		INDICATIONS, SITUATIONS ET PRIX DES BIENS	RELATION DE l'Enregistrement.	
		ARTIES	ARTIES		DATES	DROITS
1358	8	Cord <sup>l</sup> de pp <sup>te</sup>		An 1927, mois d'Août Baladilhe (concern <sup>t</sup> 5 extraits d'inscri <sup>on</sup> au total de 1419 <sup>r</sup> de rente f <sup>rs</sup> 30 au même nom que ci-dessus	11	22 50
1359	8	d <sup>o</sup>		Baladilhe (concern <sup>t</sup> divers extraits d'inscri <sup>on</sup> de rente f <sup>rs</sup> au nom de Séphore Desvallières Jeanne Elisabeth Georgina, f <sup>e</sup> de Emile / dumoulin	11	22 50
1360	8	d <sup>o</sup>		Baladilhe (concern <sup>t</sup> divers extraits d'inscri <sup>on</sup> de rente f <sup>rs</sup> au même nom que ci-dessus	11	22 50
1361	8	(10 <sup>me</sup> N <sup>o</sup> Delarue)	Mainlevée	Blonde et Bigaud (par la 1 <sup>re</sup> / de Paris 114 rue de Mécon d'inscri <sup>on</sup> et Raphaël Bodique et 4 <sup>es</sup> de Paris 24 p <sup>o</sup> de Marech <sup>e</sup> St Honoré		
1362	8	Procuration	(10 <sup>me</sup> N <sup>o</sup> Delarue)	Wathiez (par Fernand / de Paris 113 rue du Chemin Vert en blanc pour vendre		
1363	8	Mainlevée	d <sup>o</sup>	Tricotel (par Georges / de Paris 8 bd Boissonnière - d'inscri <sup>on</sup> et André Guicaud et 4 <sup>es</sup> de Paris 207 rue de Quince		

Kiessling, B. 2019. "Kraken – A Universal Text Recognizer for the Humanities." Digital Humanities Book of Abstracts. Utrecht.

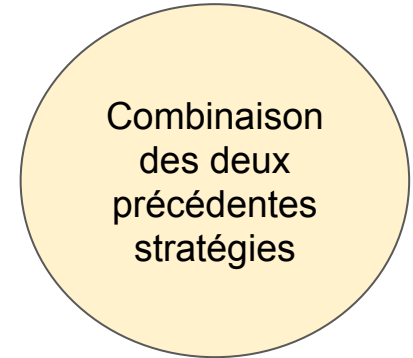
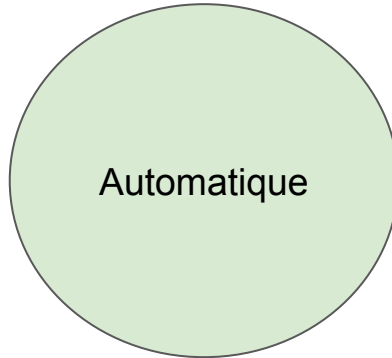
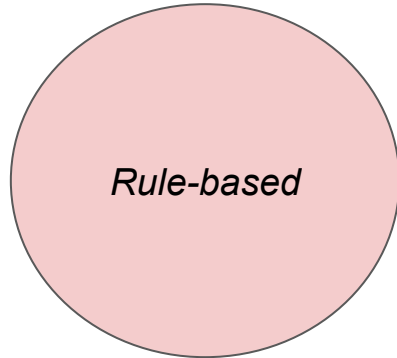
Thibault Clérice. You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine. 2022. fffal-03723208f

# GROBID (GeneRation Of Bibliographic Data)

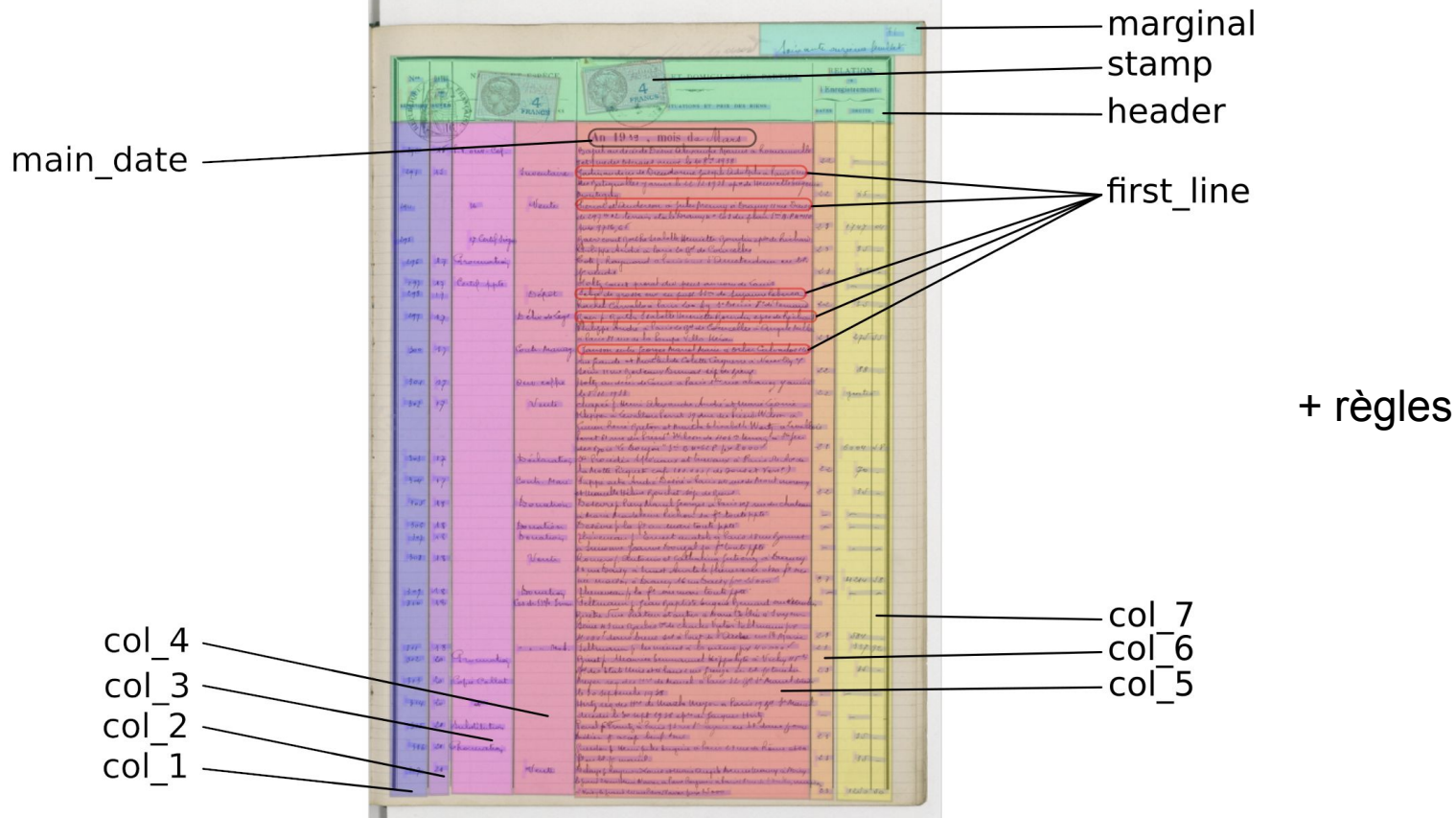
<https://github.com/kermitt2/grobid>

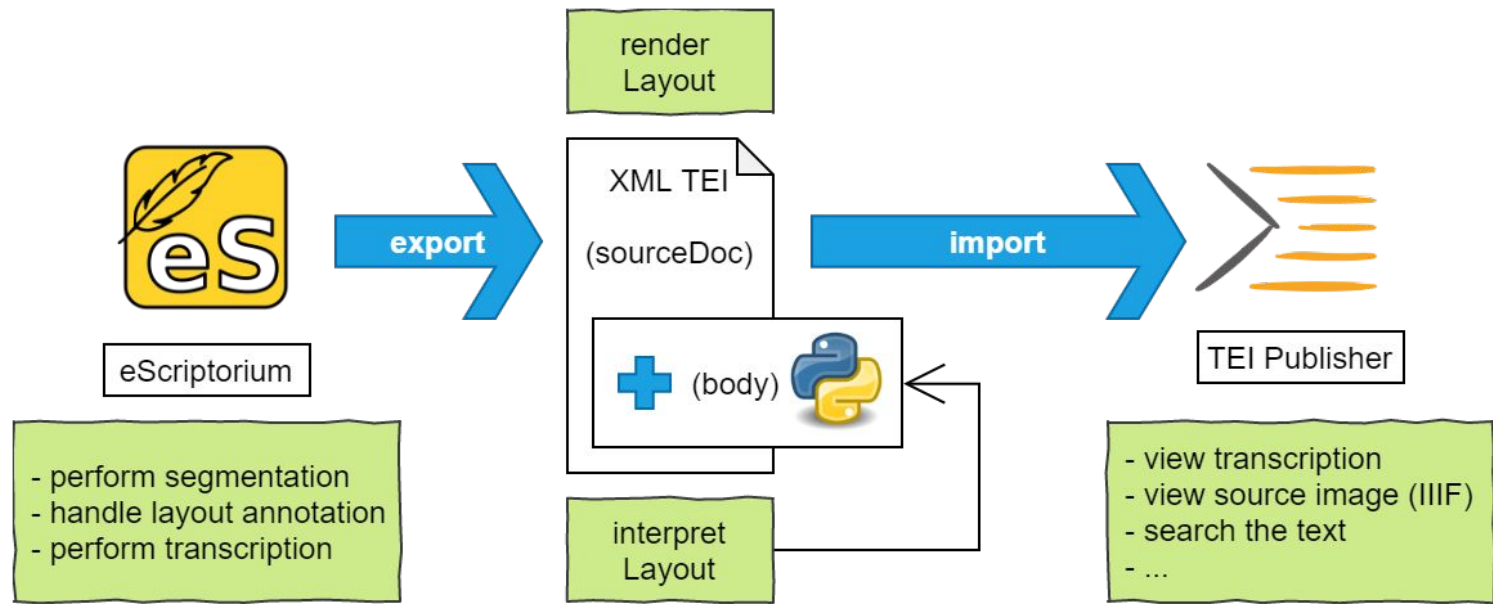
- Librairie d'apprentissage machine, *open source*, pour **re-structurer** des documents **PDF** en **XML-TEI**, créée par Patrice Lopez en 2008.
- Utilise des *features* lexicales et visuelles : token, token suivant, version normalisée du token original, n-grammes, police, gras, italique, etc.

# Quelles stratégies pour restructurer les transcriptions ?









La chaîne de traitement LEPIDEMO, qui vise à restructurer les transcriptions des répertoires de notaires en XML-TEI

```

<text>
  <body>
    <div type="main">
      <table cols="8" rows="37">
        <row n="0" role="label">
          <cell n="1" role="label1">Numéros du répertoire</cell>
          <cell n="2" role="label2">Dates des actes</cell>
          <cell n="3" role="label3">Actes en brevets</cell>
          <cell n="4" role="label4">Actes en minutes</cell>
          <cell n="5" role="label5">Noms, prénoms et domiciles des
parties ; indication, situations et prix des biens</cell>
          <cell n="6" role="label6">Date de l'enregistrement</cell>
          <cell n="7" role="label7">Droits de l'enregistrement</cell>
          <cell n="8" role="label8">Autres</cell>
        </row>
        ...

```

Extraits de la restructuration en XML-TEI  
d'un répertoire de notaire avec LEPIDEMO

```

<row>
  <cell n="1" role="col1"><lb facs="#eSc_line_389127d2"/>736</cell>
  <cell n="2" role="col2"/>
  <cell n="3" role="col3"><lb facs="#eSc_line_501bcee7"/>(Substt Me L.
Baudrier</cell>
  <cell n="4" role="col4"><lb facs="#eSc_line_b1fae513"/>Dépôt de
testament</cell>
  <cell n="5" role="col5"><lb facs="#eSc_line_dd8a6530" n="1"/>Devillers (de
Fèlicie Josèphe) fe divorcée, à Paris, rue Viollet le Duc 5</cell>
  <cell n="6" role="col6"><lb facs="#eSc_line_a58155df"/></cell>
  <cell n="7" role="col7"><lb facs="#eSc_line_3875f920"/>"</cell>
  <cell n="8" role="misc"><lb facs="#eSc_line_d9aa26b7" n="1"/>18</cell>
</row>
...

```

N <sup>os</sup> DU RÉPERTOIRE	DATES	NATURE ET ESPÈCE DES ACTES		INDICATIONS, SITUATIONS ET PRIX DES BIENS	RELATION DE l'Enregistrement.	
		DES ACTES	DES BIENS		DATES	DROITS
1358	8	Certif. de pp <sup>te</sup>		An 1927, mois d'Avril Bouladilhe (concernant 5 extraits d'inscrip <sup>on</sup> au total de 1419 <sup>x</sup> de rente f <sup>ce</sup> au même nom que ci-dessus	11	22 50
1359	8	d <sup>o</sup>		Saladilhe (concernant divers extraits d'inscrip <sup>on</sup> de rente f <sup>ce</sup> au nom de Séphore Desvallières Jeanne Elisabeth Georgina, f <sup>ce</sup> de Emile / dubonnie	11	22 50
1360	8	d <sup>o</sup>		Saladilhe (concernant divers extraits d'inscrip <sup>on</sup> de rente f <sup>ce</sup> au même nom que ci-dessus	11	22 50
1361	8	(son N <sup>o</sup> Delarue)	Mainlevée	Blonde et Boigaud (par la St <sup>e</sup> / de Paris 14 rue de Valenciennes d'inscrip <sup>on</sup> et Raphaël Boigaud et St <sup>e</sup> de Paris 24 f <sup>ce</sup> du Marché St Honoré		
1362	8	Procuration	(son N <sup>o</sup> Delarue)	Wathoz (par Fernand / de Paris 113 rue du Chemin Vert en blanc pour vendre		
1363	8	Mainlevée	d <sup>o</sup>	Tricotel (par Georges / de Paris 8 bd Poissonnière d'inscrip <sup>on</sup> d'André Lécuyer et St <sup>e</sup> de Paris 207 rue de Clichy		
1364	9	Certif. de vie	d <sup>o</sup>	Bourcier (concernant Louis / f <sup>ce</sup> Paris 43 imp <sup>te</sup> du Puitsseau		

Reconstituer la structure des répertoires pour cibler l'extraction d'information

# Visualiser et explorer les répertoires restructurés avec TEI Publisher



te*i*  
Publisher

Language  
English
Login

FRAN\_0025\_3056\_L-0-tei

Numéros du répertoire	Dates des actes	Actes en brevets	Actes en minutes	Noms, prénoms et domiciles des parties ; indication, situations et prix des biens	Date de l'enregistrement	Droits de l'enregistrement	Autres
198	9	Procuration		Lamour (par Jn Bte) employé à la Caserne de la Gde République à Paris rue St-Etienne du Mont en blanc pour recueillir Sson	10	3.75	
				Boettcher (par Henry Emile) Ingr Conseil à			

INDICATIONS, SITUATIONS ET PRIX DES BIENS

An 1914, mois de Janvier

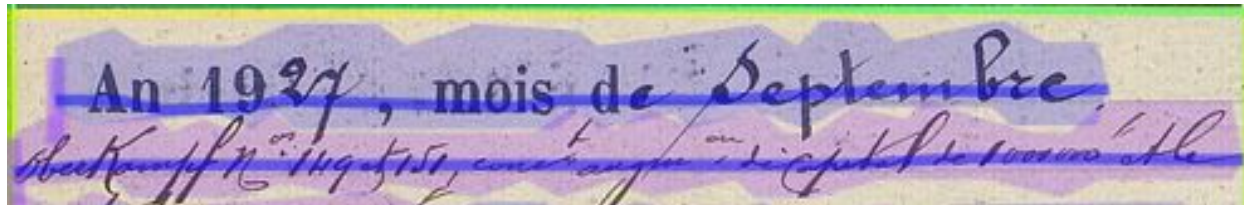
a Paris et me Madame  
 Lamour (par J<sup>n</sup> B<sup>te</sup>) employé à la Caserne de la G<sup>de</sup> République  
 à Paris me St-Etienne du Mont en blanc pour recueillir S<sup>son</sup>  
 Boettcher (par Henry Emile) Ingr<sup>s</sup> conseil à Paris 39 St-Martin  
 à Lionne femme Blanchet sep. et avec lui pour leur comm<sup>un</sup>  
 Boileau Desombres (par J<sup>n</sup> B<sup>te</sup>) à Paris 13 me Lafayette d'ins<sup>c</sup> et  
 Sophie Marie Sime Dessoliers ep<sup>x</sup> de Edouard Lavallaine de Manbeuge  
 Gronot (par Jean Eugène André) Direct<sup>r</sup> d'usine à Bobigny route de  
 Petit Ponts à André Valen Fortin int<sup>r</sup> à Paris et me Madame  
 Valter (par Léon J<sup>n</sup> B<sup>te</sup>) à Paris 39 me Grange aux Belles d'ins<sup>c</sup>  
 et Jean Louis Perreol à Montigny de Bretonneux (d. s. o.)  
 Billema (par Lucien) sieg<sup>r</sup> et à Paris me d'Annale 13 bis d'ins<sup>c</sup>  
 et Claude Emile Selvaux à Paris 13 bis d'ins<sup>c</sup>

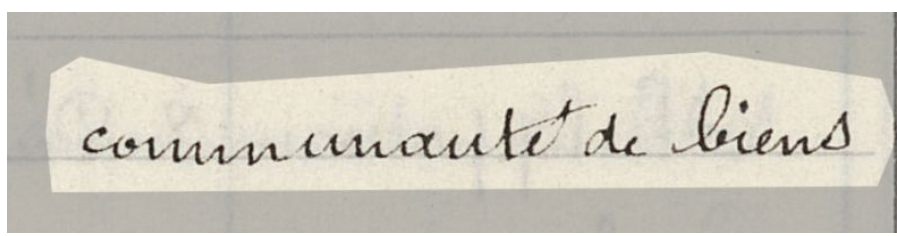
En conclusion



Le défi technique posé par LectAuRep ne consiste pas en l'utilisation de l'HTR ou de la REN, mais dans **l'infrastructure, la capacité de calcul et la chaîne de traitements** que supposerait un passage à l'échelle sur des sous-corpus choisis (une étude, une même année pour chacune des 122 études...).

Pour ce projet, la **qualité de la segmentation** par ligne ou par zone a un impact sur celle de l'HTR et de la REN, qui dépend aussi de **l'algorithme générant les polygones de points (masques) associés aux segments**, en particulier quand les lignes d'écriture sont serrées.





Une mise en commun au service du patrimoine écrit (archives et GLAM), avec une prise en compte du patrimoine numérisé

> mutualisation de données (vérité terrain de 239 images)

<https://github.com/HTR-United/lectaurep-mariages-et-divorces> ;

<https://github.com/HTR-United/lectaurep-bronod> ;

<https://github.com/HTR-United/lectaurep-repertoires>

HTR  
United

> outil de calcul paramétrable **KaMI** du CER / WER ([code source](#))

> modèles et méthodes documentés

<https://lectaurep.paris.inria.fr/>

 **hypotheses**

 **GitLab**

 **GitHub**

> produits avec un logiciel libre

 **PSL**

 **eS**

Merci pour votre attention !

**Contacts :**

- aurelia.rostaing[at]culture.gouv.fr
- hugo.scheithauer[at]inria.fr