



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-03855439>

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 (ALMAAnaCH, 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).



N° REPERTOIRE	BITS NATURE ET ESTES DES ACTES	INDICATIONS, SITUATIONS ET PAIX DES BIENS	RELATION de l'Enregistrement.	
			DATES	DRITES
1357	8	CONFÉ DE PP ^{te}	11	20 50
1359	8	d°	11	20 50
1360	8	d°	11	20 50
1361	8	Mariage (M ^{lle} Delave)		
1362	8	Securition (M ^{lle} Delave)		
1363	8	Mariage d°		

An 1927, mois d'Avril

Saladille (concernant divers contrats d'inscriptions au état de 1819 de rente f^{te} au même nom que ci-dessus)

Saladille (concernant divers contrats d'inscriptions de rente f^{te} au nom de Salade - ex-catholique - pour Elisabeth Georgine, f^{te} de Louis / de Madame)

Blonde et Boigaud (sur la M^{lle} de Paris 111 rue de Valenciennes - d'inscriptions et d'inscriptions de rente en f^{te} de l'Etat)

M^{lle} Boigaud (sur l'ancien / de Paris 111 rue de Valenciennes - en f^{te} pour rendre)

Escotot (sur l'ancien / de Paris 111 rue de Valenciennes - d'inscriptions)

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.

82	28	Cahier de Charges	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 ^{bis} et 37, - mise à prix 789,000 ^f .	4 fév.	11.10
83	28	Substitution	Kugelmann (donnée par M ^{lre} George Aron, de Berlin, à - en blanc - de procuration à lui donnée par M ^{lre} veuve Lohmefeld, née).	6 fév.	3.71
84	29	Obligation	Aine (par Louis Eugène) à Paris, place Vendôme, 1, à Jacques Doucet, vue de la Paix, 21, - de 10,000 ^f ... avec nantissement sur fonds de commerce place Vendôme 21, et autres fonds.	10 fév.	62.
85	30	Acceptation de bail	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 ^{lre} Cardozo, suivant acte reçu par M ^{lre} Bertrand et M ^{lre} Fontana, le 31 juillet 1901, - agréement par M ^{lre} Gustave Steverlynck, - de Lille, - et autres.	10 fév.	18.71

18 24

31 Depot

44 n 1850 mois de juillet.

45 du monde (par Auguste Felix) s'écrit en

46 it, demeurant à Paris rue de la Harpe n° 24 de

47 l'original en langue allemande de date (certificat)

48 une présentation de mise à M. Louis de Baber ingénieur

49 mines par M. Jacques de Baber (Banquier

50 à Carlsruhe)

69 74

Line #50

des mines par M. Jacques de Baber (Banquier
à Carlsruhe.)

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

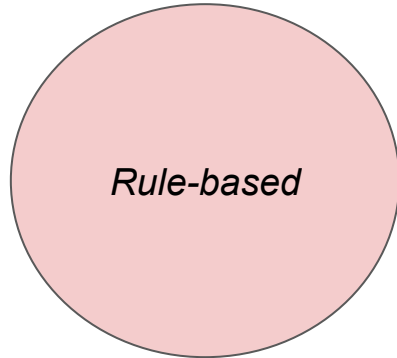
Extrait d'une transcription,
exportée en texte brut

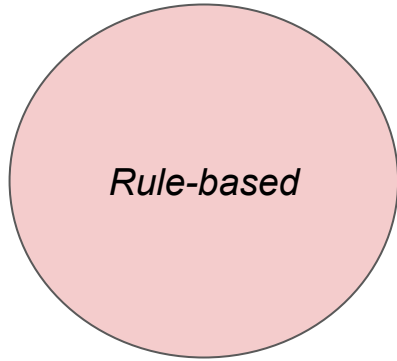
Le texte se retrouve
complètement mis à plat,
sans information de
structure.

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.

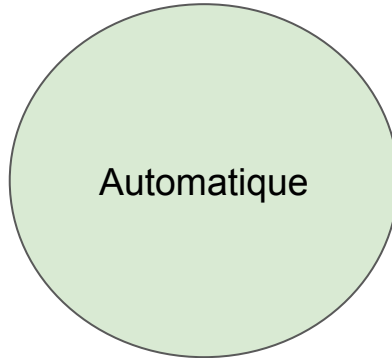
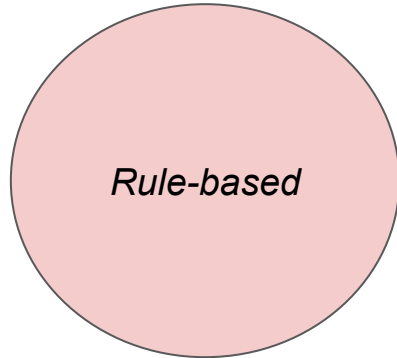
Quelles stratégies pour restructurer les transcriptions ?

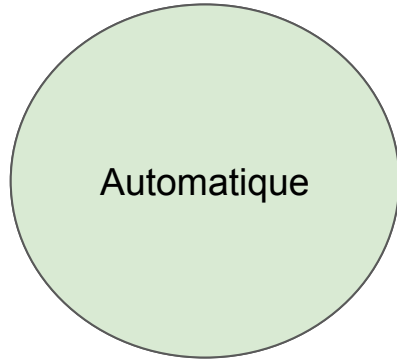




Trouver des marqueurs structurants
et définir un ensemble de règles.

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 ^{os} 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 ^l de pp ^{te}		An 1927, mois d'Août Baladilhe (concern ^t 5 extraits d'inscri ^{on} au total de 1419 ^r de rente f ^{rs} 30 au même nom que ci-dessus	11	22 50
1359	8	d ^o		Baladilhe (concern ^t divers extraits d'inscri ^{on} de rente f ^{rs} au nom de Séphore Desvallières Jeanne Elisabeth Georgina, f ^e de Emile / dumoulin	11	22 50
1360	8	d ^o		Baladilhe (concern ^t divers extraits d'inscri ^{on} de rente f ^{rs} au même nom que ci-dessus	11	22 50
1361	8	(10 ^{me} N ^o Delarue)	Mainlevée	Blonde et Bigaud (par la 1 ^{re} / de Paris 114 rue de Mécon d'inscri ^{on} et Raphaël Bodique et 4 ^{es} de Paris 24 p ^{ce} du Marché St Honoré		
1362	8	Procuration	(10 ^{me} N ^o Delarue)	Wathiez (par Fernand / de Paris 113 rue du Chemin Vert en blanc pour vendre		
1363	8	Mainlevée	d ^o	Tricotel (par Georges / de Paris 8 bd Boissonnière - d'inscri ^{on} et André Guicaud et 4 ^{es} de Paris 207 rue de Quince		

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

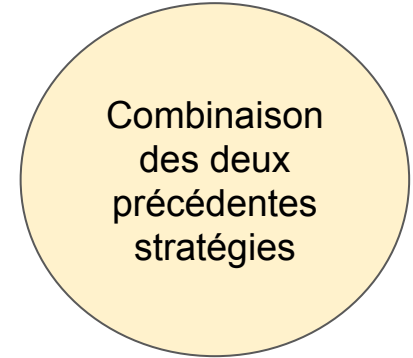
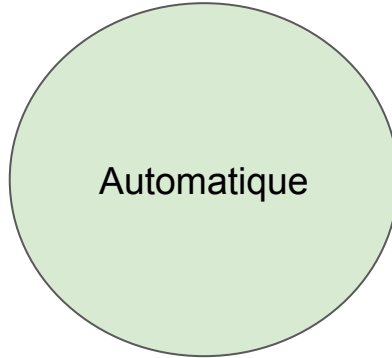
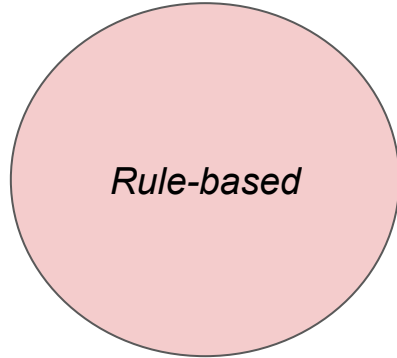
Thibault Clérice. You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine. 2022. fffal-03723208f

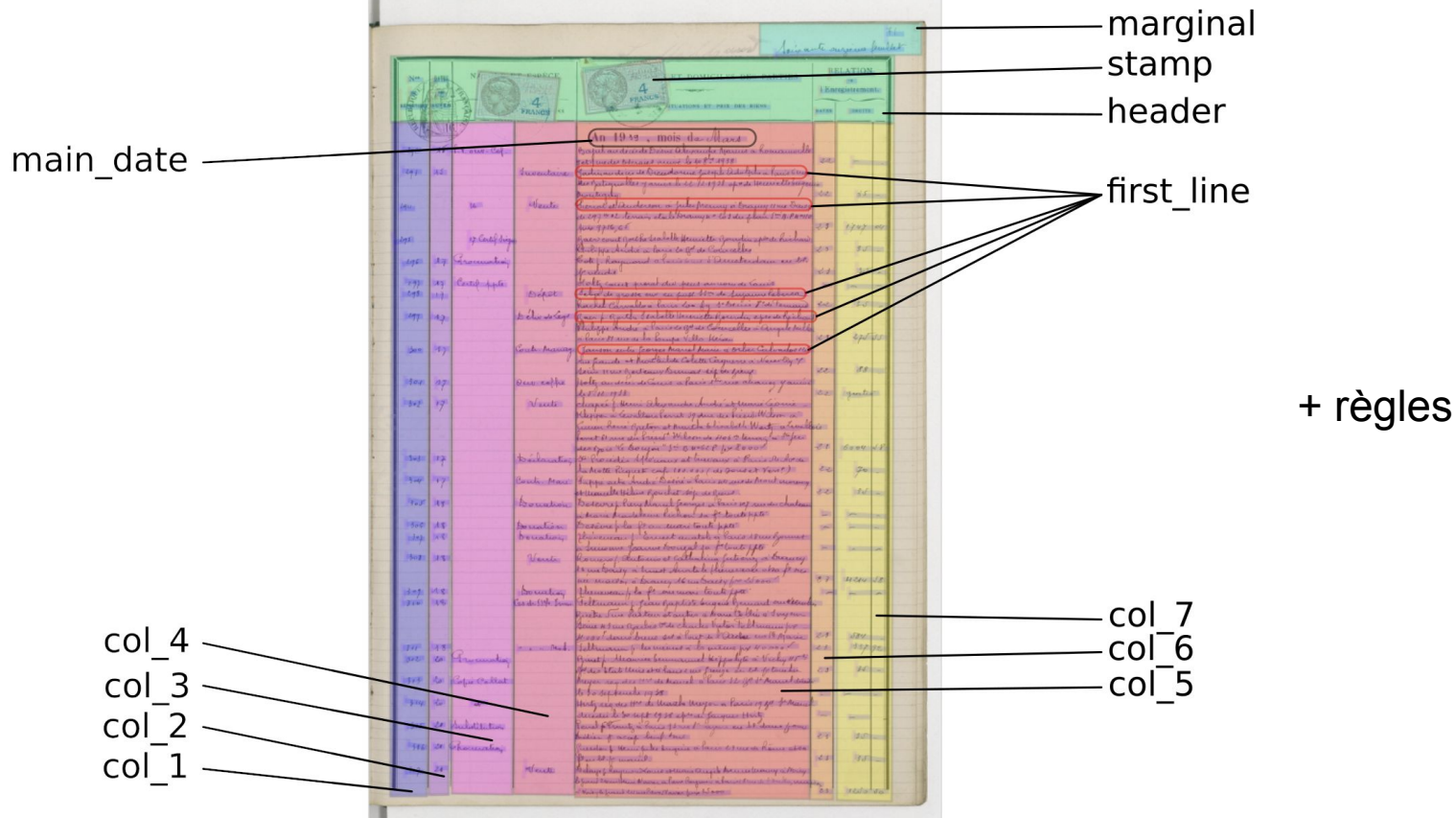
GROBID (GeneRation Of Bibliographic Data)

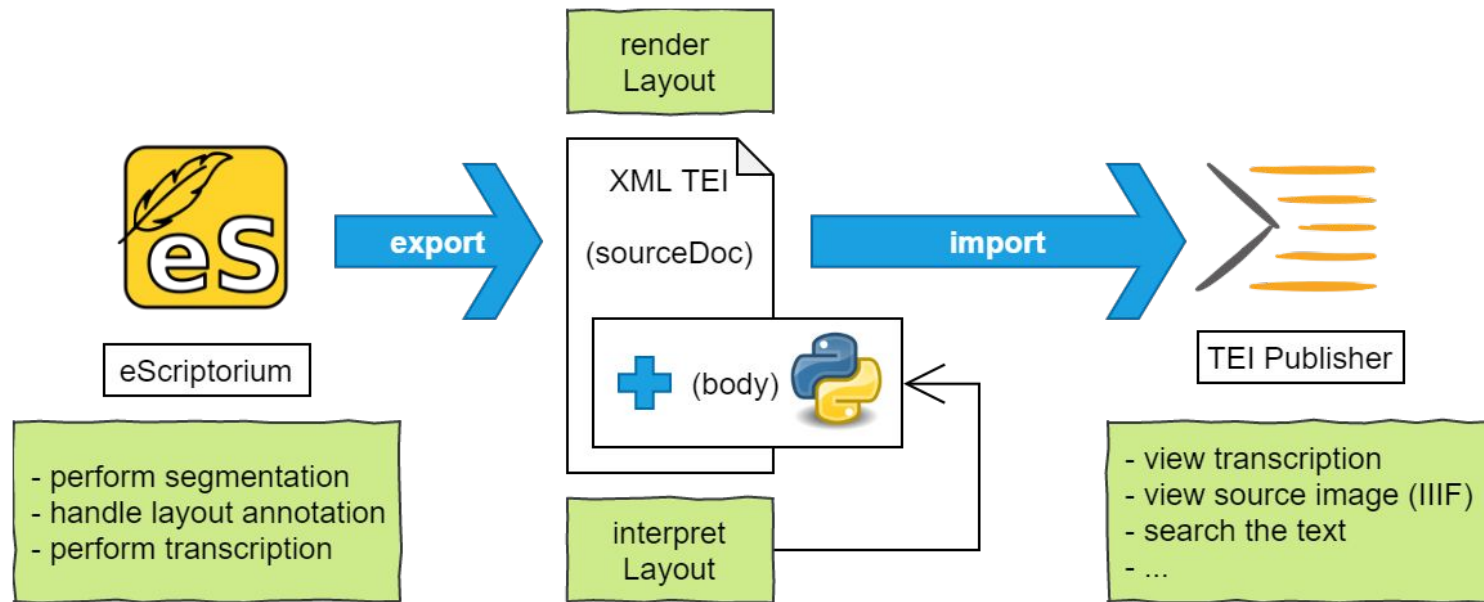
<https://github.com/kermitt2/grobid>

- 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 ^{os} DU RÉPERTOIRE	DATES	NATURE ET ESPÈCE DES ACTES		INDICATIONS, SITUATIONS ET PRIX DES BIENS	RELATION DE l'Enregistrement.	
		DES ACTES	DES BIENS		DATES	DROITS
		 		  		
				An 1927, mois d'Avril		
1358	8	Certif. de pp ^{te}		Baladilhe (concern ^t 5 extraits d'inscrip ^{on} au total de 1419 ^x de rente f ^{ce} au même nom que ci-dessus	11	22 50
1359	8	d ^o		Baladilhe (concern ^t divers extraits d'inscrip ^{on} de rente f ^{ce} au nom de Séphore Desvallières Jeanne Elisabeth Georgina, f ^e de Emile / dit nomm ^e	11	22 50
1360	8	d ^o		Baladilhe (concern ^t divers extraits d'inscrip ^{on} de rente f ^{ce} au même nom que ci-dessus	11	22 50
1361	8	(son N ^o Delarue)	Mainlevée	Blonde et Boigaud (par la N ^{te} / de Paris 14 rue de Valenciennes d'inscrip ^{on} et Raphaël Boigaud et N ^{te} de Paris 24 f ^{ce} du Marché St Honoré		
1362	8	Procuration	(son N ^o Delarue)	Wathoz (par Fernand / de Paris 113 rue du Chemin Vert en blanc pour vendre		
1363	8	Mainlevée	d ^o	Tricotel (par Georges / de Paris 8 bd Poissonnière d'inscrip ^{on} et André Lécuyer et N ^{te} de Paris 207 rue de Clichy		
1364	9	Certif. de vie	d ^o	Bourcier (concern ^t Louis / f ^e Paris 43 imp ^{te} du Puitsseau		

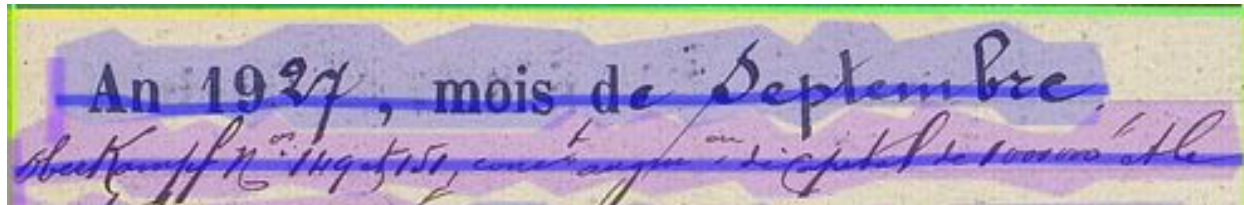
Reconstituer la structure des répertoires pour cibler l'extraction d'information

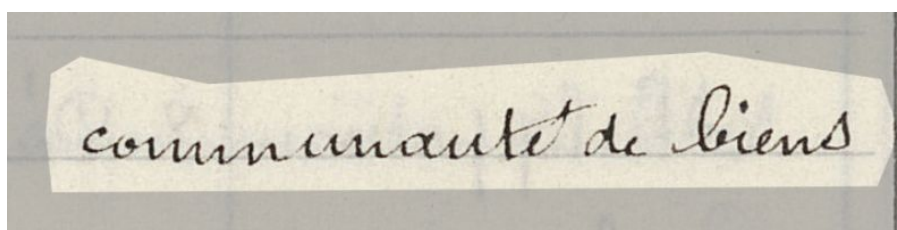
En conclusion



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