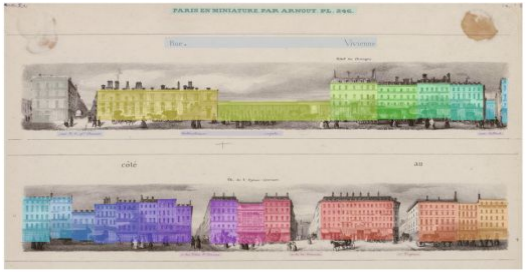
LectAuRep (2018-2021) :Projet de lecture automatique de répertoires de notaires

Aurélia Rostaing, Hugo Scheithauer



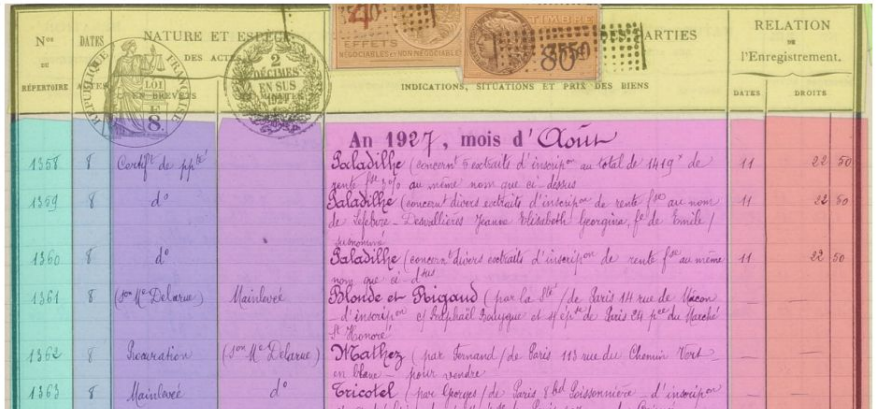
Jules Arnout, *Paris en miniature*, "Rue Vivienne, côté au levant, Rue Vivienne, côté du couchant", pl. 246, musée Carnavalet, G. 22998, segmentation des façades - Projet Richelieu. Histoire du quartier (détail).

Segmenter et annoter les images

déconstruire pour reconstruire

LectAuRep (2018-2021)

Projet de lecture automatique de répertoires de notaires

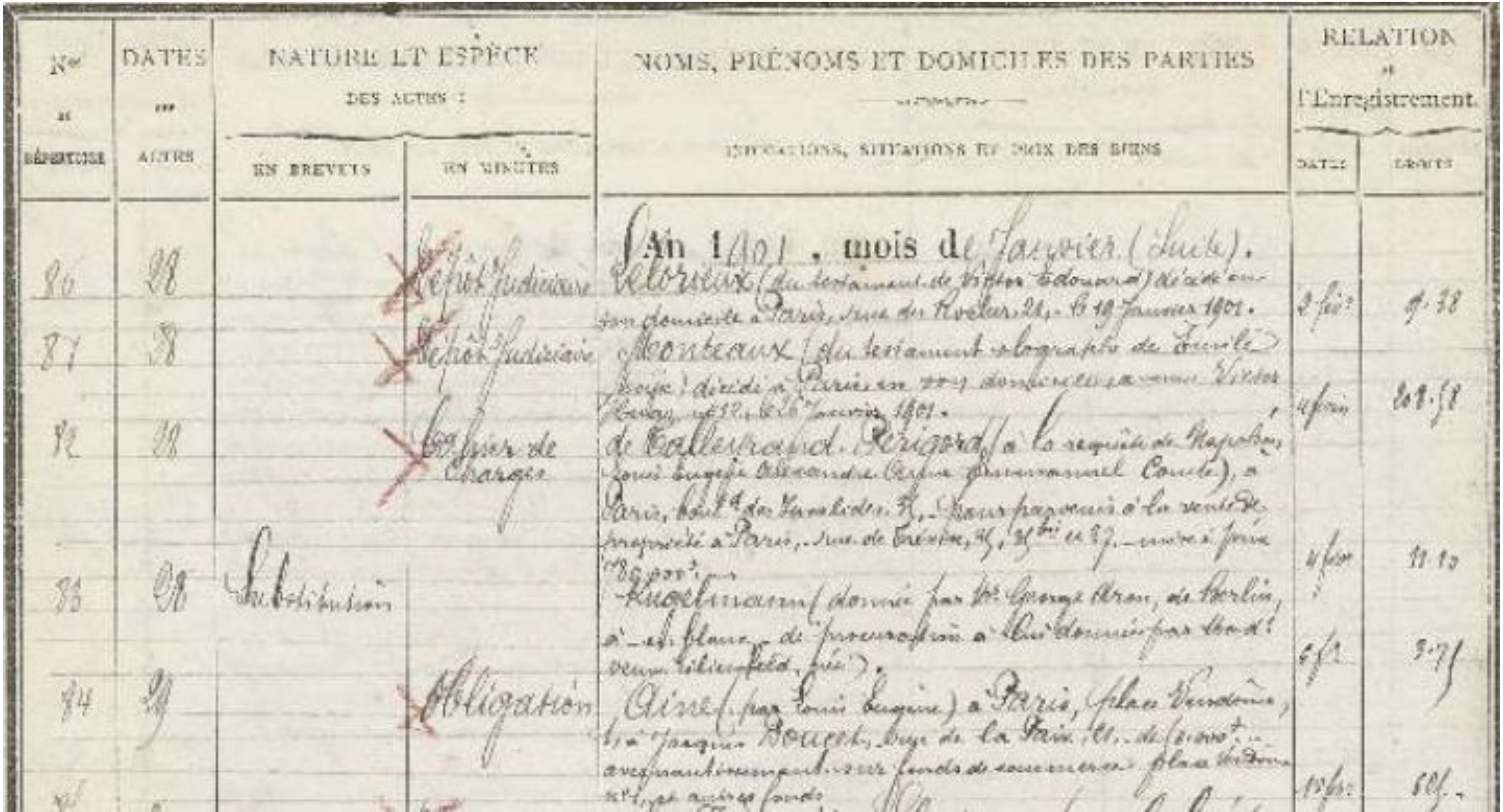


Aurélia Rostaing (Archives nationales)

Hugo Scheithauer (ALMAnaCH, Inria)

INHA, salle Vasari, 15 novembre 2022

1803-années 1940 : une mise en page structurée



Développer l’exploration des données textuelles contenues dans des images de répertoires

HTR (**segmentation linéaire**)

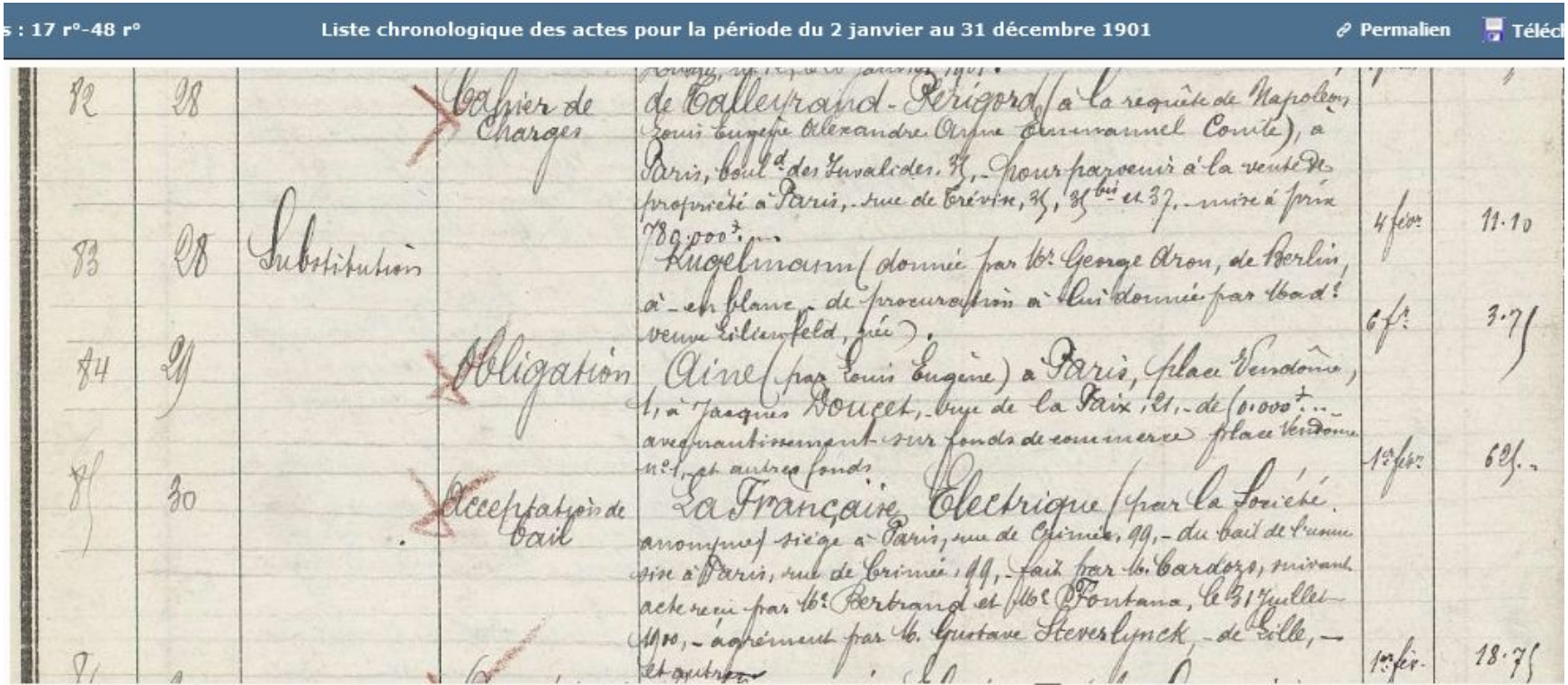
> faciliter la recherche en texte intégral (*ex. Jacques Doucet*)

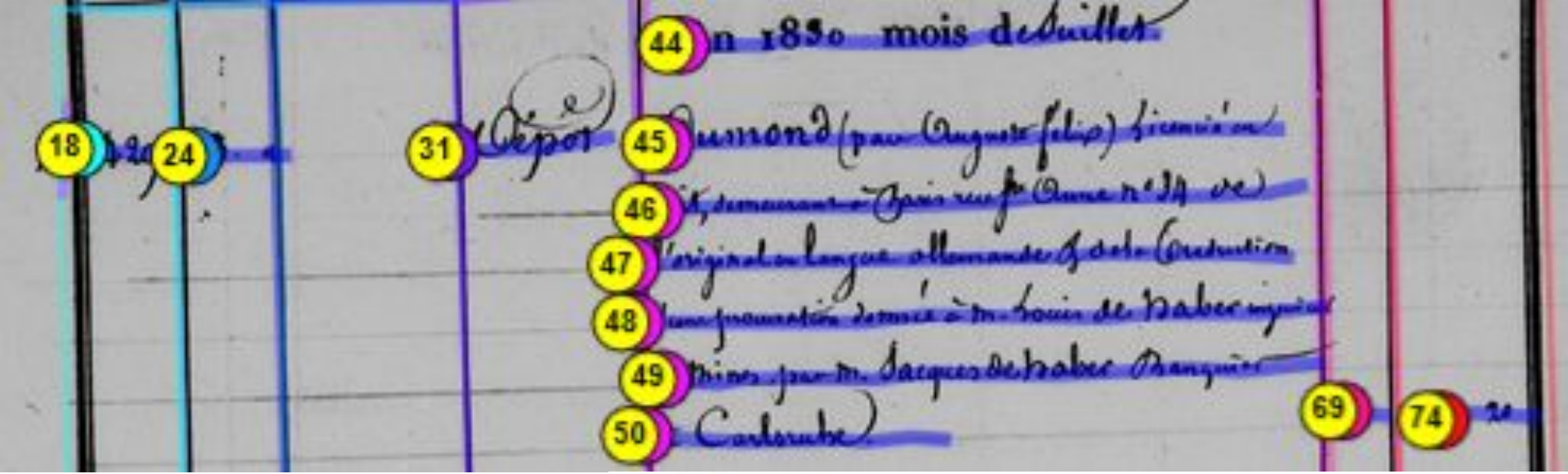
HTR + KWS/REN (**segmentation linéaire discontinue et segmentation par zones**)

> 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).*

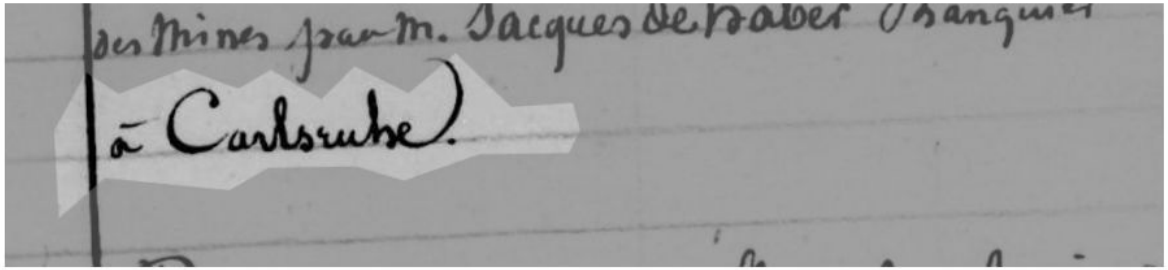
Segmenter les images numériques

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.



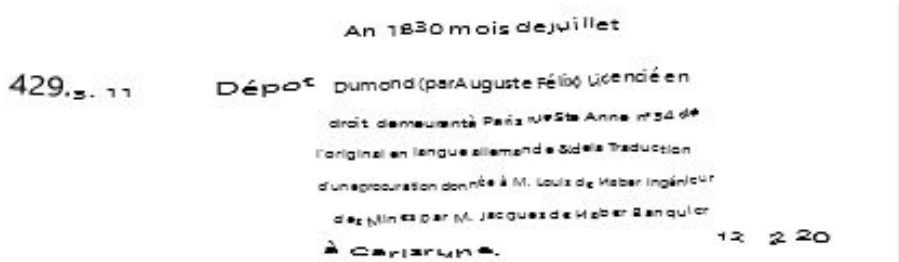


Line #50



à Carlsruhe.

Conserver la structure logique des répertoires de notaires après la transcription


An 1830 mois dejuillet

429.5. 11 Dépot Dumond (parAuguste Félix) Licencié en

droit demeurantà Paris rueSte Anne n°34 de

l'original en langue allemande &dela Traduction

d'uneprocuration donnée à M. Louis de Haber ingénieur

des Mines par M. Jacques de Haber Banquier

à Carlsruhe.

12 2 20

46 droit, demeurantà Paris rueSte Anne n°34 de
47 l'original en langue allemande &dela Traduction
48 d'uneprocuration donnée à M. Louis de Haber ingénieur
49 des Mines par M. Jacques de Haber Banquier
50 à Carlsruhe.

44 An 1830 mois deJuillet
45 Dumond (parAuguste Félix) Licencié en
46 droit, demeurantà Paris rueSte Anne n°34 de
47 l'original en langue allemande &dela Traduction
48 d'uneprocuration donnée à M. Louis de Haber ingénieur
49 des Mines par M. Jacques de Haber Banquier
50 à Carlsruhe.

```
Procuration
Audtorisation
Certificat de ppté
Procuration
Mainlevée
de signification
Mainlevée
Mainlevée
Mainlevée
Mainlevée
Obligation
Dépôt
Transport
Vente
Constitution
deSequestre
Décharge
Procuration
Substitution
Quittance
Vente
An 1914, mois d. Fevrier
à Paris 34 rue Sedaine
Lamour (par Jn B^e) employé à la Caserne de la G^de Républicaine
à Paris rue St Etienne du Mont en blanc pour recueillir SS^on
Boettcher (par Henry Emile) Ing^t conseil à Paris 39 Bd St Martin
à Léontine Jeanne Blanchot safe dt avec lui pour gérer Comce
Boileau-Desombres (parlaSté) à Paris 73 rue Lafayette d'inxn c/
Sophie Marie Louise Dessoliers épse de Edouard Lavelaine de Maubeuge
Gront (par Jean Eugène André) direstt d'usme à Bahigny route des
Petits Ponts à André Valern Fortin ind^el à Paris 34 rue Sedaine
```

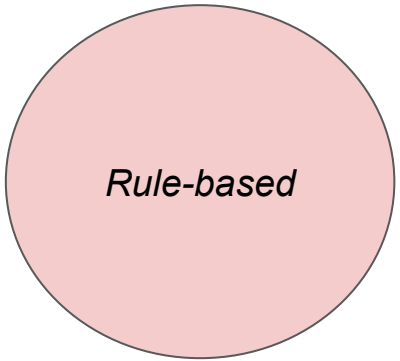
Extrait d’une transcription, exportée en texte brut

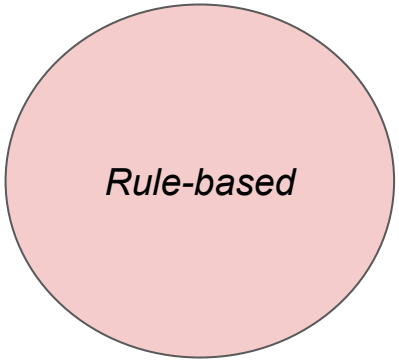
Le texte se retrouve complètement mis à plat, sans information de structure.

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.

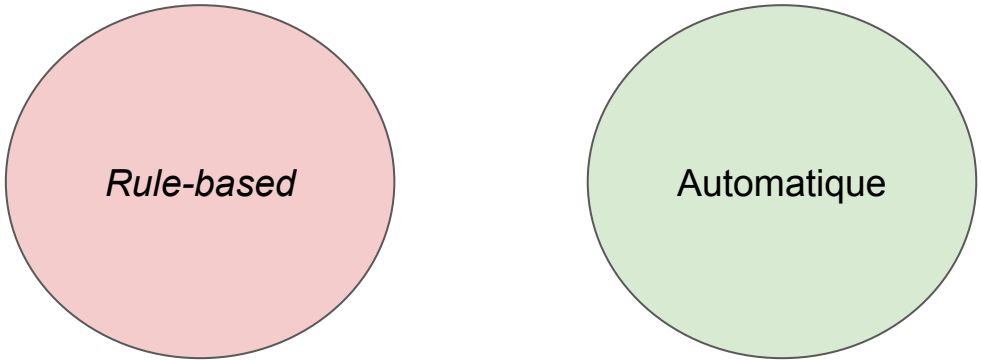
Quelles stratégies pour restructurer les transcriptions ?

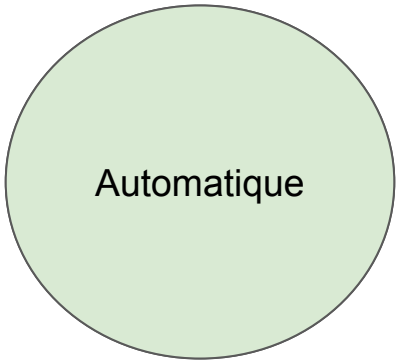




Trouver des marqueurs structurants
et définir un ensemble de règles.

Quelles stratégies pour restructurer les transcriptions ?

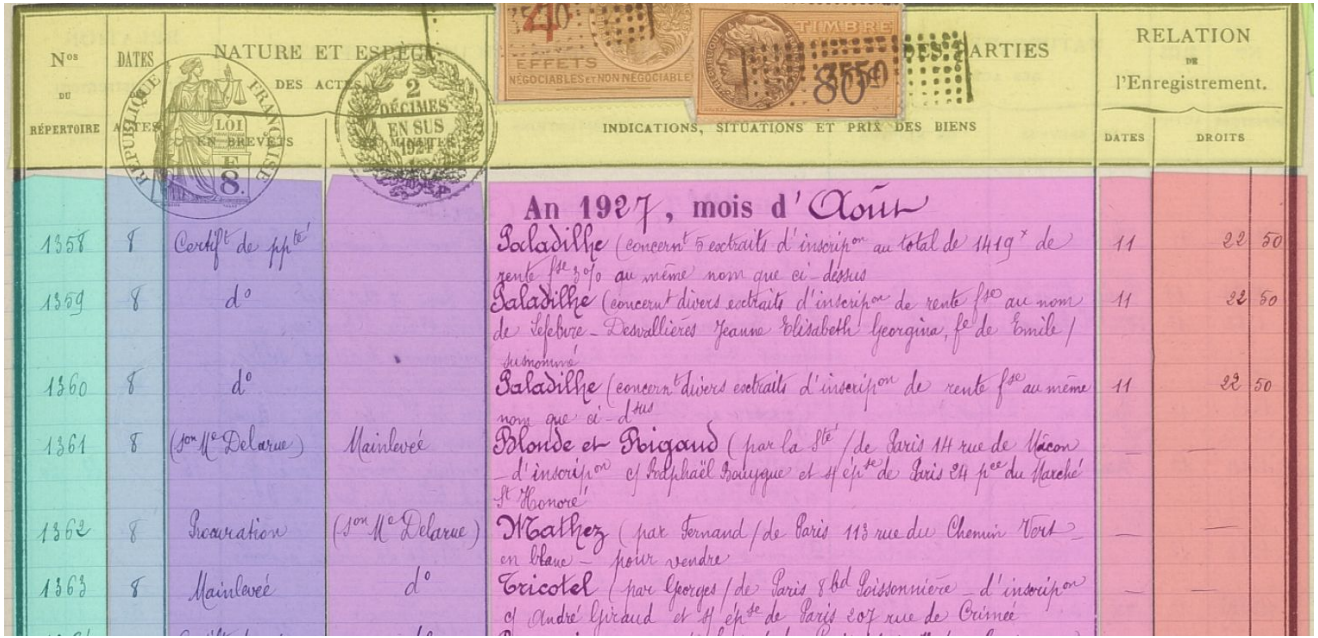




Utiliser des outils d’apprentissage machine :

- 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



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

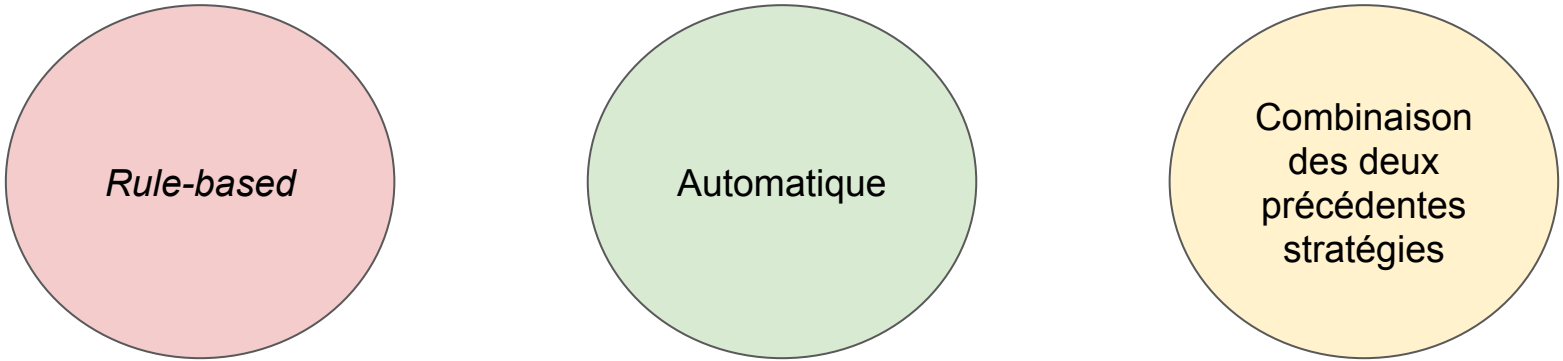
Thibault Clérice. You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine. 2022. ffhal-03723208f

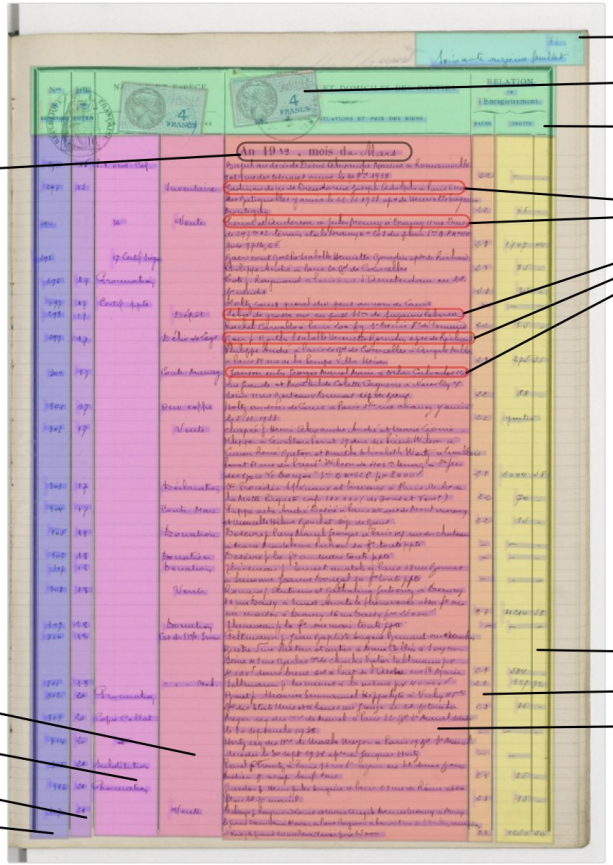
GROBID (GeneRation Of BIbliographic Data)

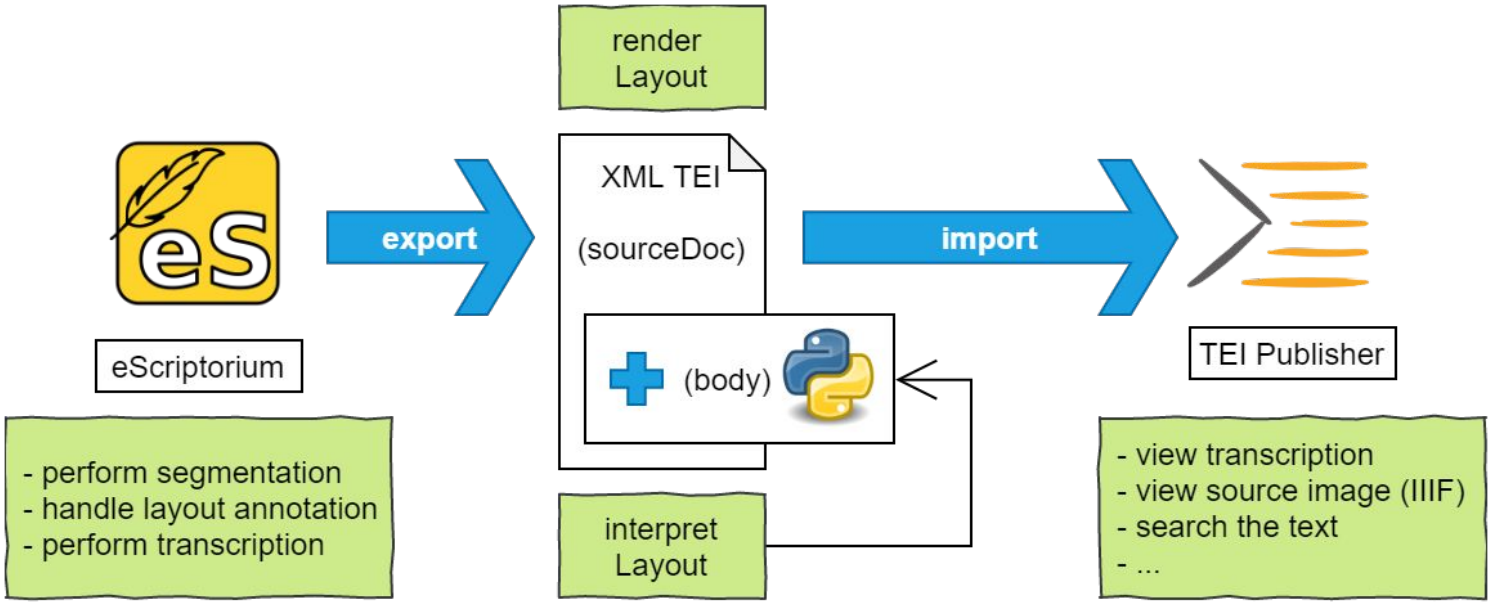
https://github.com/kermitt2/grobid

- 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 ?




marginal
stamp
header
main_date
first_line
+ règles
col_7
col_6
col_5
col_4
col_3
col_2
col_1

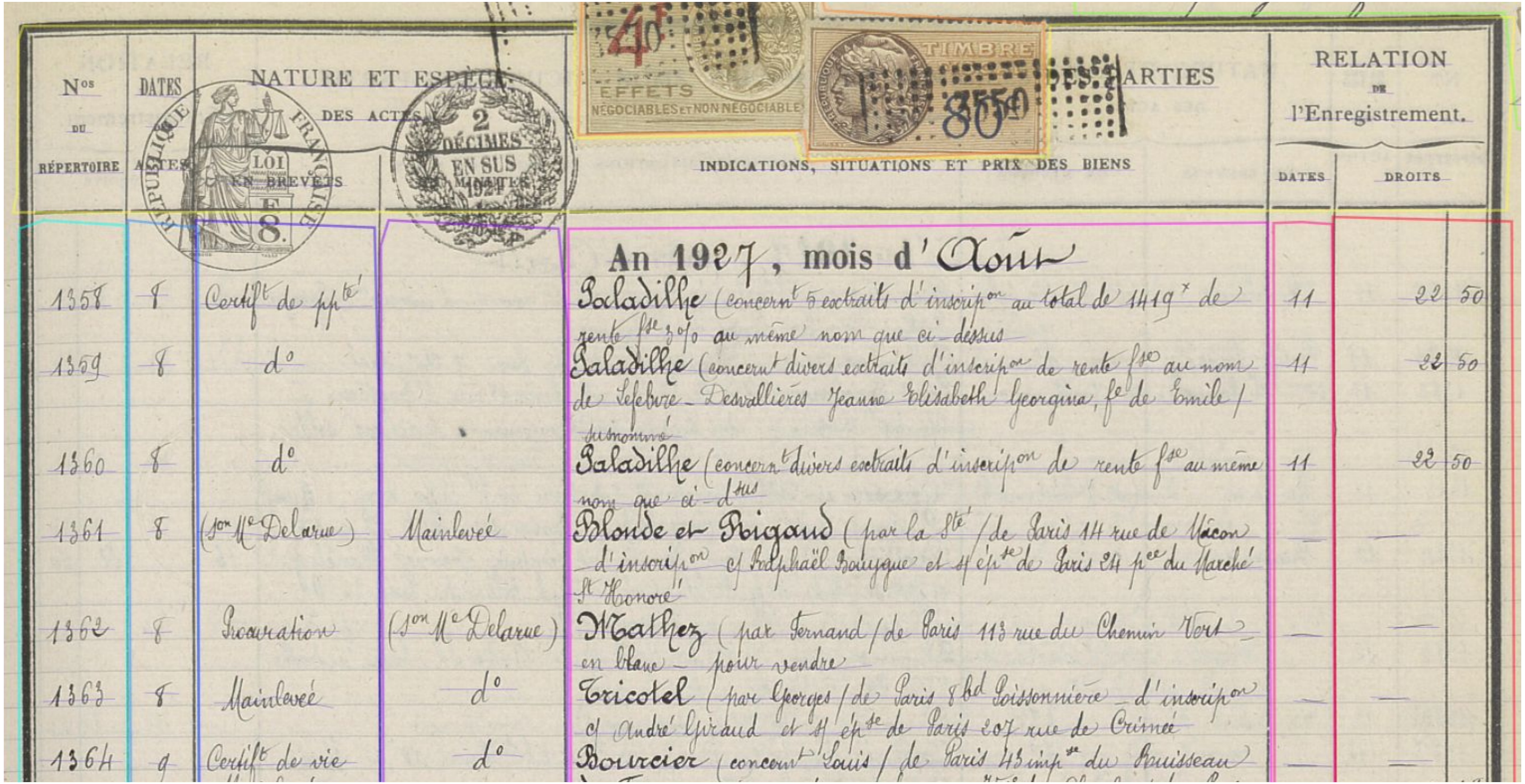


La chaîne de traitement LEPIDEMO, qui vise à restructurer les transcriptions des répertoires de notaires en XML-TEI

```
<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, prénms 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 divorcèe, 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>
     ...
```



Nos du répertoire	Dates des actes	Nature et espèce des actes (en brevets)	(minutes)	Noms des parties — Indications, situations et prix des biens	Relation de l'Enregistrement — Dates	Droits
				An 1927, mois d'Août		
1358	8	Certift de ppté		Paladilhe (concernt 5 extraits d'inscripon au total de 1419x de rente fse 3% au même nom que ci-dessus	11	22 50
1359	8	d°		Paladilhe (concernt divers extraits d'inscripon de rente fse au nom de Lefebvre-Desvallières Jeanne Elisabeth Georgina, fe de Emile / susnommé	11	22 50
1360	8	d°		Paladilhe (concernt divers extraits d'inscripon de rente fse au même nom que ci-dsus	11	22 50
1361	8	(son Me Delarue)	Mainlevée	Blonde et Rigaud (par la Sté / de Paris 14 rue de Mâcon — d'inscripon c/ Raphaël Bouygue et s/ épse de Paris 24 pce du Marché St Honoré	—	—
1362	8	Procuration	(son Me Delarue)	Mathez (par Fernand / de Paris 113 rue du Chemin Vert en blanc — pour vendre	—	—
1363	8	Mainlevée	d°	Tricotel (par Georges / de Paris 8 bd Poissonnière — d'inscripon c/ André Giraud et s/ épse de Paris 207 rue de Crimée	—	—
1364	9	Certift de vie	d°	Bourcier (concernt Louis / de Paris 43 impse du Ruisseau	—	—

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

Visualiser et explorer les répertoires restructurés avec TEI Publisher

tei Publisher

Start Documentation News Download

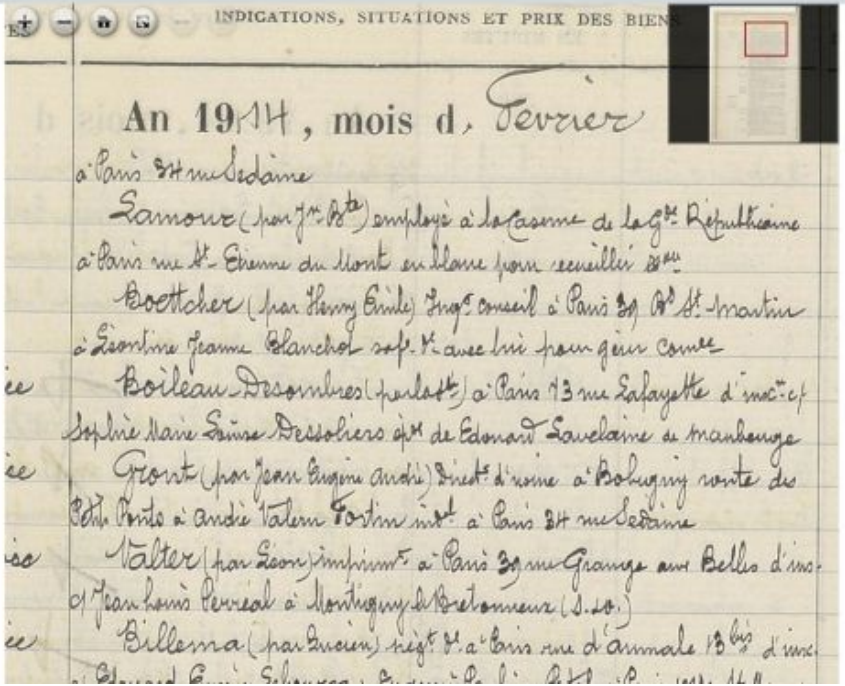
Enter Query

Language English

Login

FRAN_0025_3056_L-0-tei

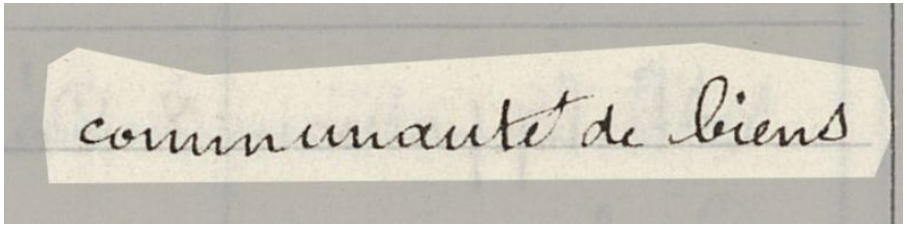
Numéros du répertoire	Dates des actes	Actes en brevets	Actes en minutes	Noms, prénms et domiciles des parties ; indication, situations et prix des biens	Date de l'enregistrement	Droits de l'enregistrement	Autres
198	9	Procuration		Lamour (par Jn Bte) employé à laCaserne de laGde Républicaine à Paris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



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

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

> modèles et méthodes documentés

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

> produits avec un logiciel libre

Merci pour votre attention !

Contacts :

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