LectAuRep : Données d'archives en français des XIXe et XXe siècles

Hugo Scheithauer

▶ **To cite this version:**

Hugo Scheithauer. LectAuRep : Données d'archives en français des XIXe et XXe siècles. Transkribus / eScriptorium : Transcrire, annoter et éditer numériquement des documents d'archives, Bibliothèque nationale de France, May 2022, Paris, France. hal-03666884

HAL Id: hal-03666884
https://inria.hal.science/hal-03666884

Submitted on 12 May 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.

Distributed under a Creative Commons Attribution 4.0 International License

LectAuRep

Données d’archives en français des XIXe et XXe siècles

Transkribus / eScriptorium : Transcrire, annoter et éditer numériquement des documents d’archives (09/05/22)

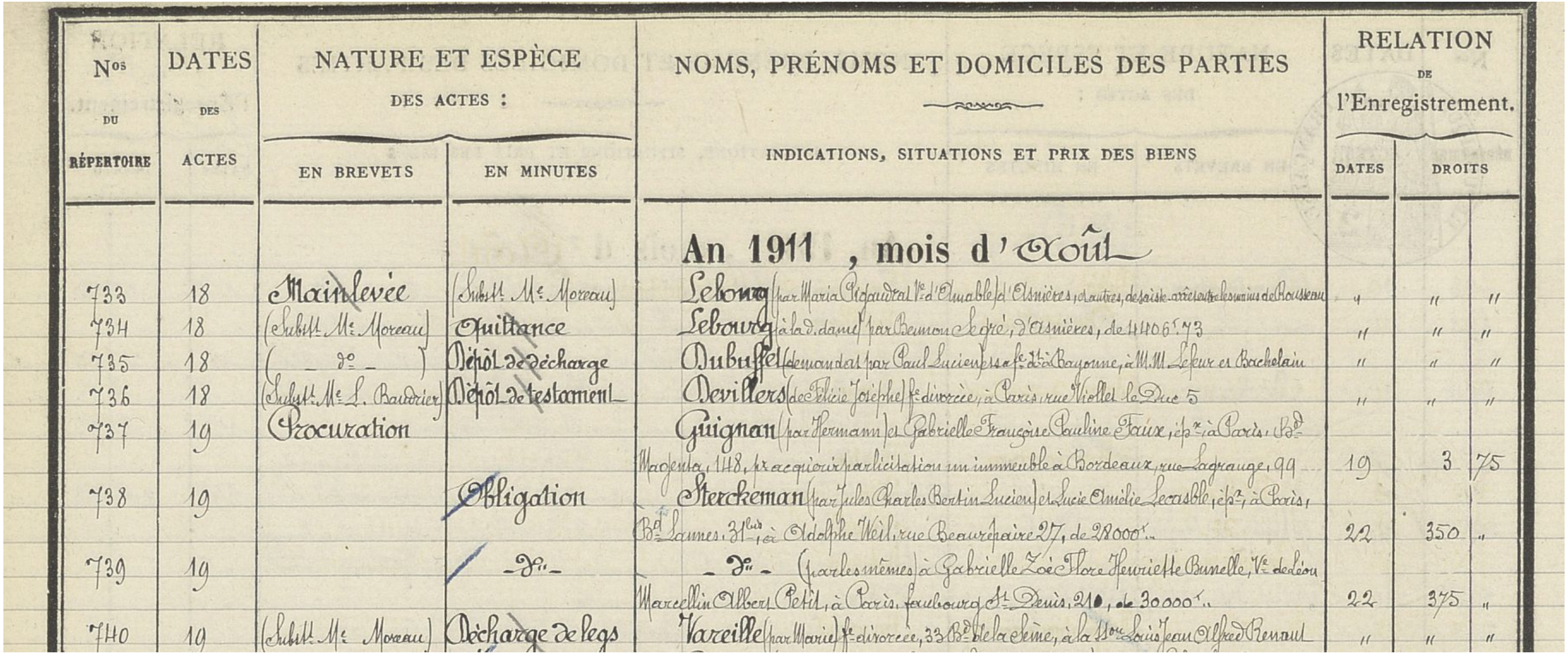
Hugo Scheithauer (Inria)

LectAuRep

Inria - Archives nationales (Département du Minutier Central)

- Faciliter et massifier la consultation des répertoires de notaires conservés au DMC grâce aux technologies de reconnaissance d'écriture manuscrite.
- Offrir aux usagers une lecture enrichie des répertoires et permettre la fouille de données dans le corpus.

Voir Alix Chagué, Rostaing Aurélia. LECTAUREP: Paris Notary Record Books Automated Reading. Fantastic Futures 2021 / Futures Fantastiques 2021, AI4LAM; BnF; Université Paris Saclay, Dec 2021, Paris, France.



Nos du répertoire	Dates des actes	Nature et espèce des actes : en brevets	Nature et espèce des actes : en minutes	Noms, prénoms et domiciles des parties — Indications, situations et prix des biens	Relation de l'Enregistrement : dates	Relation de l'Enregistrement : droits
				An 1911, mois d'Août		
733	18	Mainlevée	(Subst. Me Moreau)	Lebong (par Maria Pigaudral Vve d'Amable) d'Asnières, d'autres, desaisissement ... de Rousseau	"	" "
734	18	(Subst. Me Moreau)	Quittance	Lebourg (à la d. dame) par Bemon Sègre, d'Asnières, de 1406f. 73	"	" "
735	18	(— d° —)	Dépôt de décharge	Dubuffet (demandat par Paul Lucien) ... à Bayonne, à MM. Lefeur et Bachelain	"	" "
736	18	(Subst. Me L. Baudrier)	Dépôt de testament	Devillers (de Félicie Joséphe) fe divorcée, à Paris, rue Viollet le Duc 5	"	" "
737	19	Procuration		Guignan (par Hermann) et Gabrielle Françoise Pauline Faux, épx à Paris, Bd Magenta, 148, pr acquérir par licitation un immeuble à Bordeaux, rue Lagrange, 99	19	3 75
738	19		Obligation	Sterckeman (par Jules Charles Bertin Lucien) et Lucie Amélie Lecasble, épx, à Paris, Bd Lannes, 31bis, à Adolphe Weil, rue Beaurepaire 27, de 28000f.	22	350 "
739	19		d°	d° (par les mêmes) à Gabrielle Zoé Flore Henriette Bunelle, Vve de Léon Marcellin Albert Petit, à Paris, faubourg St Denis, 210, de 30000f.	22	375 "
740	19	(Subst. Me Moreau)	Décharge de legs	Vareille (par Marie) fe divorcée, 33 Bd de la Seine, à la Sr Louis Jean Alfred Renault	"	" "

Extrait d'une page d'un répertoire de notaire.

Transcrire les répertoires de notaires

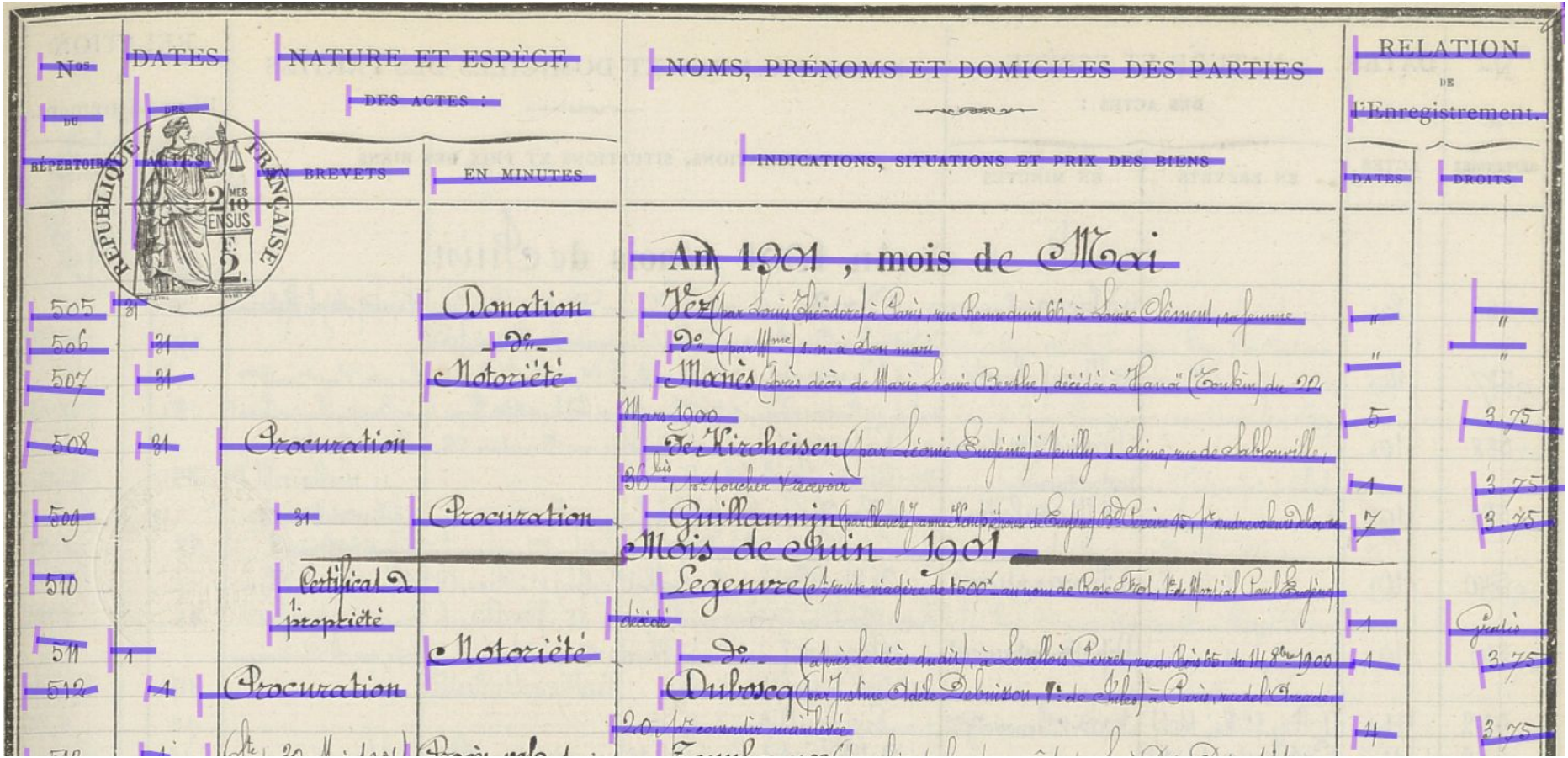
Nos du répertoire	Dates des actes	Nature et espèce des actes : En brevets	Nature et espèce des actes : En minutes	Noms, prénoms et domiciles des parties — Indications, situations et prix des biens	Relation de l'Enregistrement. Dates	Relation de l'Enregistrement. Droits	
				An 1911, mois d'Août			
733	18	Mainlevée	(Subst. Me Moreau)	Lebourg (par Maria Rigaudrat Ve d'Amable) d'Asnières, à autres, de saisie-arrêt entre les mains de Rousseau	"	"	"
734	18	(Subst. Me Moreau)	Quittance	Lebourg à la d. dame (par Bemon Ségré), d'Asnières, de 4406f. 73	"	"	"
735	18	(— d° —)	Dépôt de décharge	Dubuffet (demandat par Paul Lucien) fils et fe d.t à Bayonne, à MM. Lefeur et Bachelain	"	"	"
736	18	(Subst. Me L. Baudrier)	Dépôt de testament	Devillers (de Félicie Joséphe) fe divorcée, à Paris, rue Viollet le Duc 5	"	"	"
737	19	Procuration		Guignan (par Hermann) et Gabrielle Françoise Pauline Faux, ép.x, à Paris, Bd Magenta, 148, p.r acquérir par licitation un immeuble à Bordeaux, rue Lagrange, 99	19	3	75
738	19		Obligation	Sterckeman (par Jules Charles Bertin Lucien) et Lucie Amélie Lecasble, ép.x, à Paris, Bd Lannes, 31 bis, à Adolphe Weil, rue Beaurepaire 27, de 28000f.	22	350	"
739	19		— d° —	— d° — (par les mêmes) à Gabrielle Zoé Flore Henriette Bunelle, Ve de Léon Marcellin Albert Petit, à Paris, faubourg St Denis, 210, de 30000f.	22	375	"
740	19	(Subst. Me Moreau)	Décharge de legs	Vareille (par Marie) fe divorcée, 33 Bd de la Seine, à la Sté Louis Jean Alfred Renault	"	"	"
741	19		Inventaire	Garnier (après décès de M. l'abbé Henri Joseph Jn Bte) curé de l'Église Ste Geneviève des Grandes Carrières, d.t à Paris, rue Championnet 174, décédé à l'hôpital Bichat le 20 juillet 1911	21	7	50

Nos du répertoire	Dates des actes	Nature et espèce des actes … en …	… es	… et domiciles des parties / … situations et prix des biens	Relation de l'Enregistrement. Dates	Droits	
				An 1939, mois de Mars			
292	15	P.V. ouv. Cof.		Bajul au décès de Désiré Alexandre Marius à Romainville 7 et 9 rue des Oseraies arrivé le 10 8bre 1938	22	–	–
293	15		Inventaire	Jadin au décès de Dieudonné Joseph Adolphe à Paris 6 rue des Batignolles y arrivé le 22-12-1938 ép[se] de Henriette Eugénie Montigny	22	35	–
294	16		Vente	Pierrat et Anderson à Jules Merny à Drancy 11 rue Daisy de 297m 02 terrain et asile Drancy n° 203 du plan Son A. P. n° 110 prix 9756,65	23	1747	04
295	17	Certif. Sign.		Baer conct Berthe Isabelle Henriette Bourdin ép[se] de Richard Philippe André à Paris 20 Bd de Courcelles	23	35	–
296	17	Procuration		Cote J. Raymond à Paris 6 rue d'Amsterdam en bl. pr vendre	23	35	–

- Deux modèles génériques avec un taux d’erreur par caractère inférieur à 10%.

- Des modèles spécialisés qui résultent de l’affinage des modèles génériques sur des lots de données plus petits et plus uniformes, atteignant un taux d’erreur par caractère égal voire inférieur à 5%.

Conserver la structure logique des répertoires



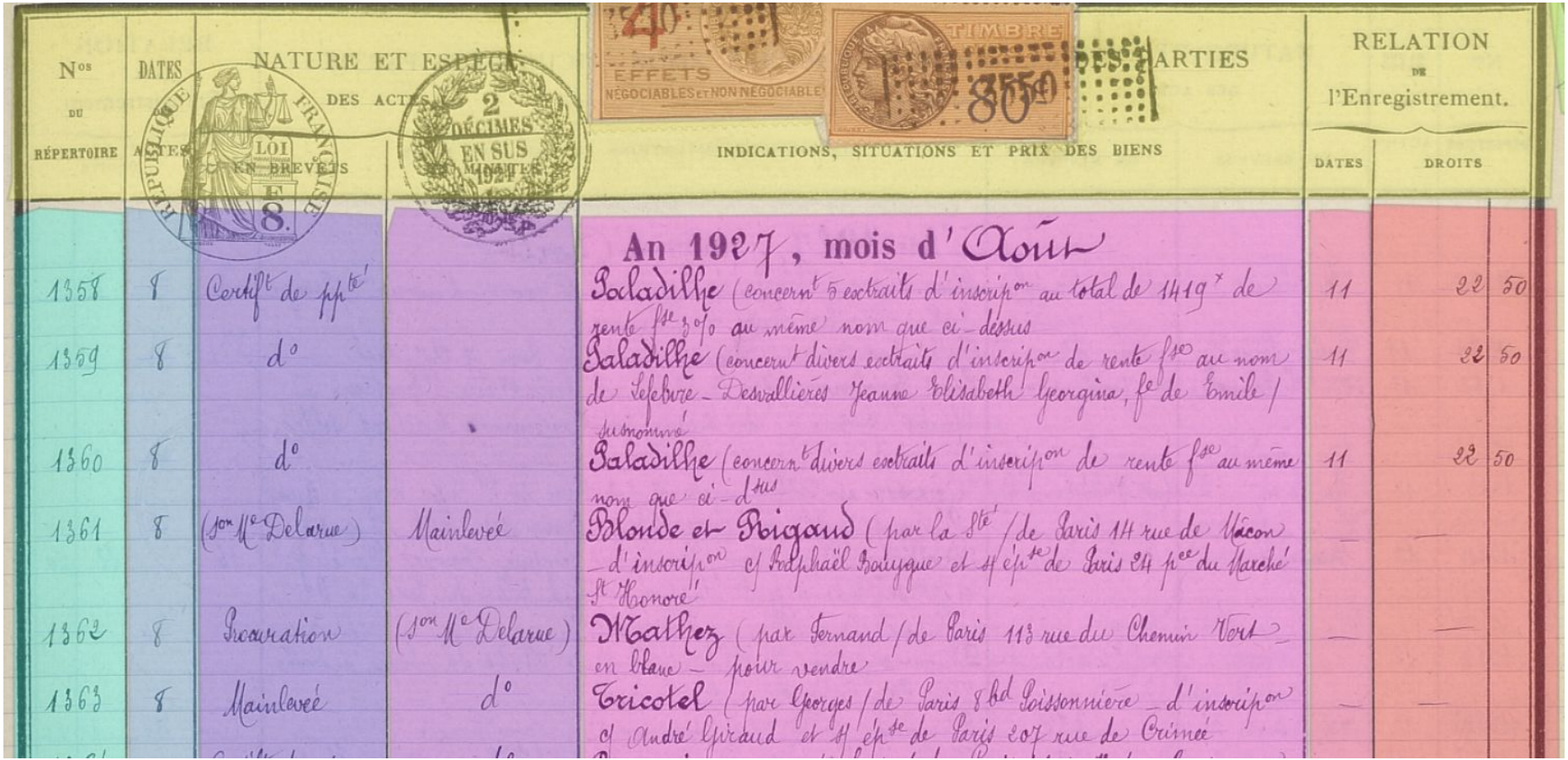
Détection des lignes de texte dans un répertoire de notaire

Entraîner un modèle de segmentation sur des données tabulaires

Conserver la structure logique des répertoires

```
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
```

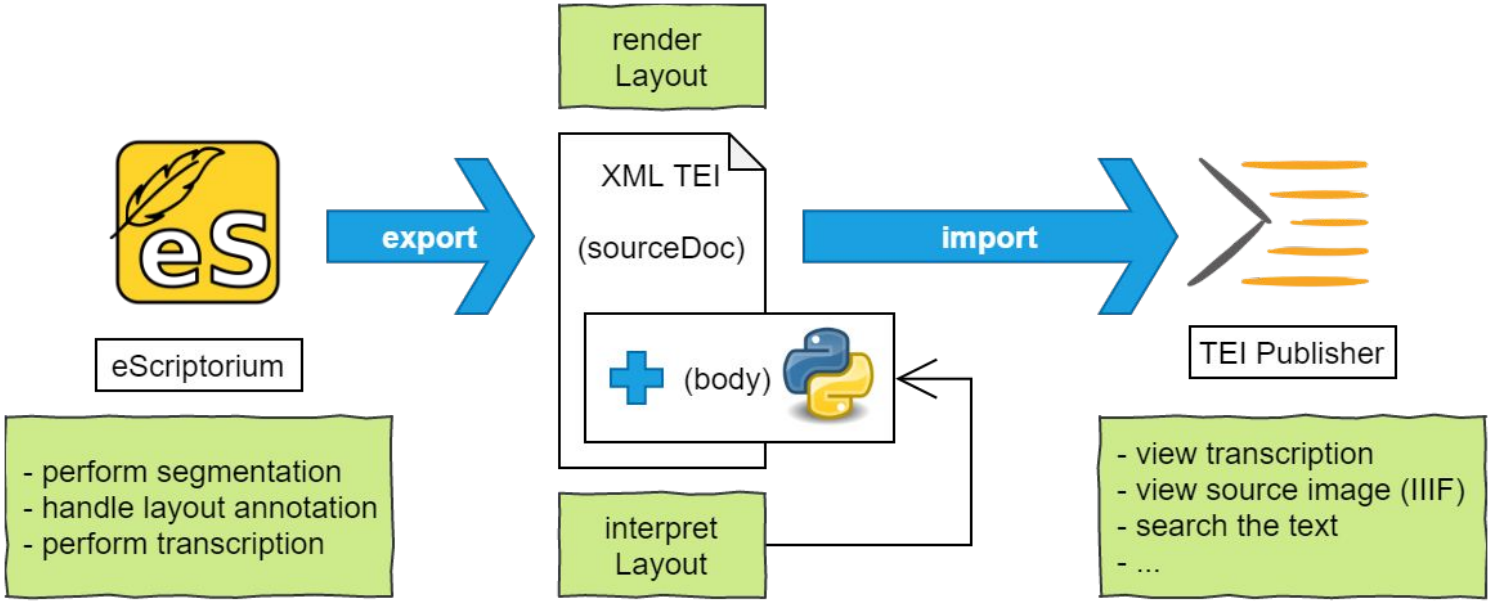
Extrait d’une transcription à plat, exportée en texte brut.



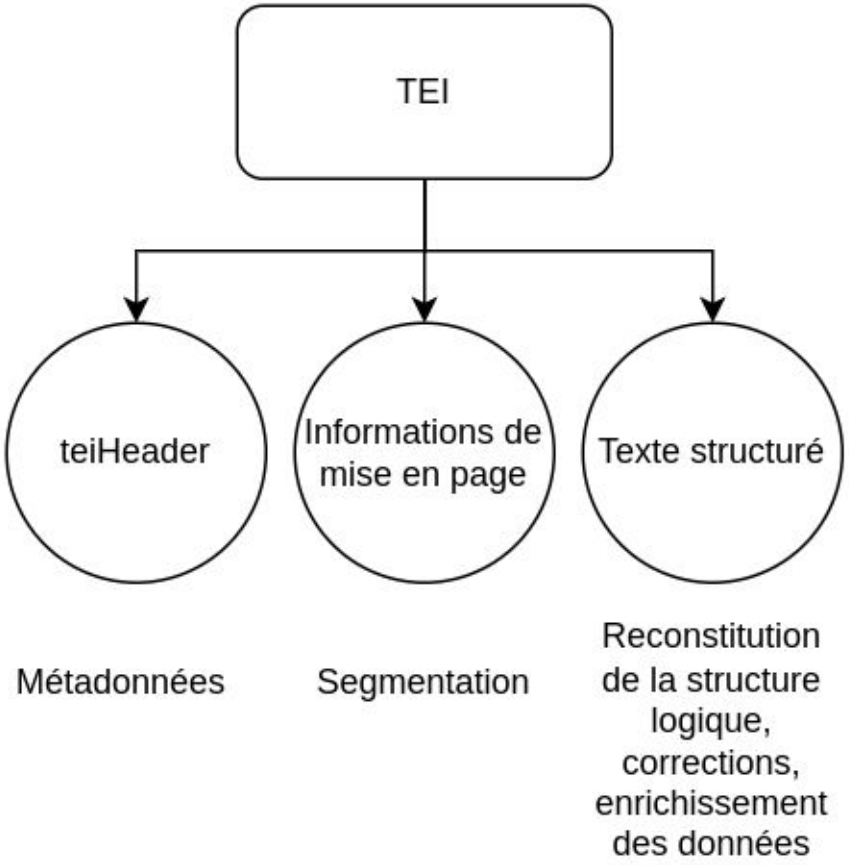
Détection des différentes régions de texte dans un répertoire de notaire

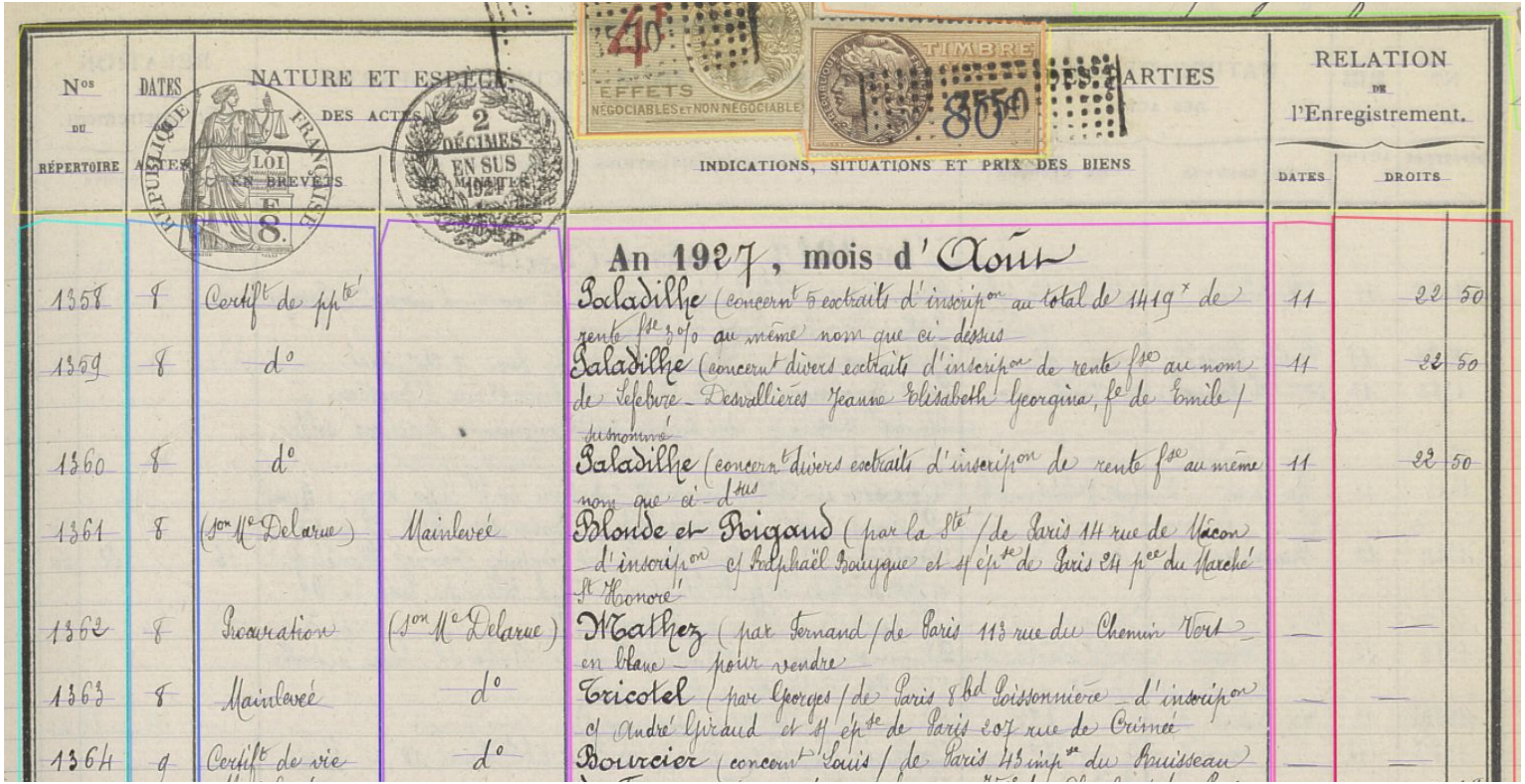
Entraîner un modèle de segmentation sur des données tabulaires

Conserver la structure logique des répertoires



La chaîne de traitement LEPIDEMO


TEI
teiHeader
Informations de mise en page
Texte structuré
Métadonnées
Segmentation
Reconstitution de la structure logique, corrections, enrichissement des données

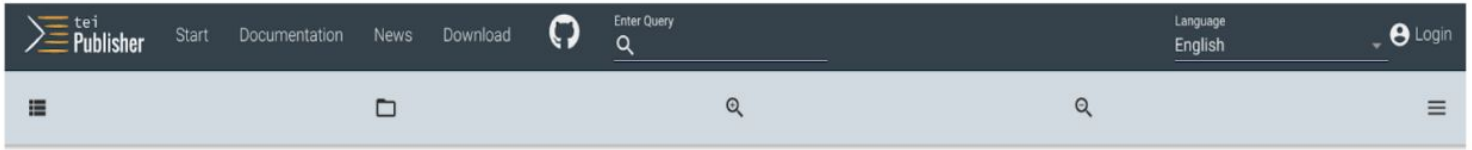


An 1927, mois d'Août

Nos du répertoire	Dates des actes	Nature et espèce des actes: en brevets	Minutes	Noms, prénoms, domiciles des parties; indications, situations et prix des biens	Relation de l'Enregistrement: Dates	Droits
1358	8	Certift de ppté		Saladilhe (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°		Saladilhe (concernt divers extraits d'inscripon de rente fse au nom de Lefebvre-Desvallières Jeanne Élisabeth Georgina, fe de Émile / susnommé	11	22 50
1360	8	d°		Saladilhe (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 Poigaud (par la Sté / de Paris 14 rue de Mâcon d'inscripon c/ Raphaël Bouygue et s/ épse de Paris 24 fce 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 des unités logiques pour signaler l'information avec des traitements par règle ou avec de la reconnaissance d'entités nommées.

tei Publisher



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
292	15	P. V. ouv. Cof.		Bajul au décès de Désiré Alexandre Marius à Romainville 7 et 9 rue des Oseraies arrivé le 10 8^bre 1938	22	-|-	
293	15		Inventaire	Jadin au décès de Dieudonné Joseph Adolphe à Paris 6 rue des Batigolles y arrivé le 22-12-1938 ep^x de Henriette Eugénie Montigny	22	35|-	
294		16	Vente	Pierrat et Anderson à Jules Menny à Drancy 11 rue Daisy de 297^f 02 terrain etoile Drancy n°203 du plan XXX A. P n°110 prix 9756,65	23	1747|04	
295		17 Certif Sign.		Baer conct Berthe Isabelle Henriette Bourdin ep^se de Richard Philippe André à Paris 20 B^d de Courcelles	23	35|-	

Visualiser les répertoires après transcription

Conserver la structure logique des répertoires

Conclusion

- La segmentation est essentielle pour bien délimiter des unités de texte, mais également pour comprendre ces unités une fois transcrites au sein de la mise en page.

- La transcription permet d’obtenir la donnée, et plus un modèle sera robuste, plus la donnée sera de qualité.

Déployer eScriptorium à Inria

- 2019 : machine virtuelle, sans carte graphique.
- 2020 : serveur traces6, 2 cartes graphiques, plus de stockage.
- 2021-2022 : CREMMA, 2 cartes graphiques supplémentaires, plus de stockage, plus de mémoire vive.

Merci pour votre attention !