

LECTAUREP

Lecture Automatique de Répertoires

Archives nationales (DMG/DMOASI)

Equipe ALMAAnaCH

Atelier Culture - Inria, 2 décembre 2019

The background image shows a page from a handwritten archival document, likely a table of contents or index. The text is written in cursive and includes various entries with numbers and dates. The table structure is partially visible, with columns for numbers and text.

36	6	Verific' de present d'...	8	5600
		Cent vingt huit actes, - d.c.c.	8	3.75
31	7	Est devic	8	1.88

Un partenariat pluridisciplinaire

Archives nationales

Département du Minutier Central (DMC)
Département de la maîtrise d'ouvrage du système d'information (DMOASI)

- ★ Virginie Grégoire (DMC)
- ★ Danis Habib (DMC)
- ★ Marie-Françoise Limon-Bonnet (DMC)
- ★ Gaetano Piraino (DMOASI)
- ★ Aurélia Rostaing (DMC)
- ★ Frédéric Zamarreno (DMOASI)

SCRIPTA (EPHE)

- ★ Marc Bui
- ★ Daniel Stökl Ben Ezra

- ★ ++ El Hassane Gargem
- ★ Benjamin Kiessling
- ★ ++ Robin Tissot

ALMAnaCH (Equipe Inria *Automatic Language Modelling and Analysis & Computational Humanities*)

- ★ -- Marie-Laurence Bonhomme
- ★ ++ Alix Chagué
- ★ Eric de la Clergerie
- ★ -- Marie Puren
- ★ -- Charles Riondet
- ★ Laurent Romary
- ★ ++ Lionel Tadjou

Le corpus d'archives



Les répertoires de notaires de Paris (1803-1944)

Plus de 900 notaires

Environ 2000 registres (300 à 500 pages chacun)

Plusieurs milliers de mains de clerks différentes

Le corpus d'archives

93 28. 10. 19

Domicile faculté

N°	DATES	NATURE ET ESPÈCE		NOMS, PRÉNOMS ET DOMICILES DES PARTIES		RELATION	
		EN BREVETS	EN MINUTES	INDICATIONS, SITUATIONS ET VALEURS DES BIENS	INDICATIONS, SITUATIONS ET VALEURS DES BIENS	EN BREVETS	EN MINUTES
An 1920, mois de Janvier							
<p>Ce présent répertoire contenant tous les faits, a été préparé par Monsieur le Greffier du Tribunal Civil de la Seine pour servir à inscrire les notes en minutes et en brevets qui seront reçus par M. L. Lindet, notaire à Paris.</p> <p>Tous les renseignements sont fournis par les Titulaires successifs</p>							
Janvier 1920 (cont.)							
17	5	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
18	6	Chemin		Maucou, Levallois-Perret, Seine	6	1.58	
19	6	Chemin		Maucou, Levallois-Perret, Seine	6	1.58	
20	6	Commune		Rouyer, Levallois-Perret, Seine	6	1.58	
21	6	Chemin		Maucou, Levallois-Perret, Seine	6	1.58	
22	6	Chemin		Maucou, Levallois-Perret, Seine	6	1.58	
23	6	Chemin		Maucou, Levallois-Perret, Seine	6	1.58	
24	6	Chemin		Maucou, Levallois-Perret, Seine	6	1.58	
25	6	Chemin		Maucou, Levallois-Perret, Seine	6	1.58	
26	6	Chemin		Maucou, Levallois-Perret, Seine	6	1.58	
27	6	Chemin		Maucou, Levallois-Perret, Seine	6	1.58	
28	6	Chemin		Maucou, Levallois-Perret, Seine	6	1.58	
29	6	Chemin		Maucou, Levallois-Perret, Seine	6	1.58	
30	6	Chemin		Maucou, Levallois-Perret, Seine	6	1.58	

Domicile faculté

N°	DATES	NATURE ET ESPÈCE		NOMS, PRÉNOMS ET DOMICILES DES PARTIES		RELATION	
		EN BREVETS	EN MINUTES	INDICATIONS, SITUATIONS ET VALEURS DES BIENS	INDICATIONS, SITUATIONS ET VALEURS DES BIENS	EN BREVETS	EN MINUTES
An 1920, mois de Janvier							
<p>Ce présent répertoire contenant tous les faits, a été préparé par Monsieur le Greffier du Tribunal Civil de la Seine pour servir à inscrire les notes en minutes et en brevets qui seront reçus par M. L. Lindet, notaire à Paris.</p> <p>Tous les renseignements sont fournis par les Titulaires successifs</p>							
Janvier 1920 (cont.)							
31	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
32	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
33	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
34	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
35	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
36	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
37	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
38	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
39	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
40	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
41	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
42	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
43	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
44	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
45	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
46	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
47	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
48	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
49	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
50	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	

Domicile faculté

N°	DATES	NATURE ET ESPÈCE		NOMS, PRÉNOMS ET DOMICILES DES PARTIES		RELATION	
		EN BREVETS	EN MINUTES	INDICATIONS, SITUATIONS ET VALEURS DES BIENS	INDICATIONS, SITUATIONS ET VALEURS DES BIENS	EN BREVETS	EN MINUTES
An 1920, mois de Janvier							
<p>Ce présent répertoire contenant tous les faits, a été préparé par Monsieur le Greffier du Tribunal Civil de la Seine pour servir à inscrire les notes en minutes et en brevets qui seront reçus par M. L. Lindet, notaire à Paris.</p> <p>Tous les renseignements sont fournis par les Titulaires successifs</p>							
Janvier 1920 (cont.)							
51	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
52	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
53	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
54	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
55	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
56	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
57	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
58	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
59	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
60	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
61	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
62	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
63	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
64	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
65	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
66	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
67	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
68	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
69	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	
70	6	Chemin		Pastreau, Levallois-Perret, Seine	6	1.58	

La reconstitution des actes

Vide (sauf accident dans colonnes adjacentes)

1654	27	Procuration	An 1919, mois de Novembre		
1655	27	Collep ^{te}	Stoël, (par Louis) St. Guine de la montagne St. Genevieve 47, en blanc, pour recueillir son Bas (cont. armoiries, comm. au duc de Jh. Anne Guillaume) décès en son domicile à Paris, Lasserre le 27 avril 1919 sur rente viagère pour Livilleme n° 1325777 et 1636220 de 100 ⁺ francs	29	3.75

N° de l'acte
Nombre entre 1
et 3000
(estimation)

Typologie de l'acte
Chaîne de caractères
Vocabulaire contrôlé

Date de l'acte (jour)
Nombre entre 1 et 31

Date de l'acte (année et mois)
Écritures mixtes (imprimées et manuscrites)

Description de l'acte
Nom et adresse des signataires, prix de vente d'un
bien, date d'un décès, etc.

Date d'enregistrement (jour)
Nombre entre 1 et 31

Taxes acquittées
Chiffres, chaînes de
caractères (gratis,
etc.)

Détection des unités d'informations

The image shows a handwritten document with several lines of text. A red horizontal line is drawn above the first two lines. A red horizontal line is drawn under the date '29'. Another red horizontal line is drawn under the number '3-75'. Arrows point from these red lines to boxes containing lists of information units.

1654	27	Procuration	An 1919, mois de Novembre	29	3-75
1655	27	Cat lepp ^{te}	Noël, (par l'intermédiaire) de l'ancien de la montagne		
			et Genevieve 47, en blanc, pour reconnaître son		
			Bas, (cont. arroyo, courus au duc de jh Edin		
			Guillaume) décès en son domicile à Paris, les ingénieurs		
			n: 9 le 27 avril 1919 au restaurant pour l'arrêter		
			n: 1325777 et 1636220 de 100 ⁺ chaux		

Limite supérieure de l'acte :

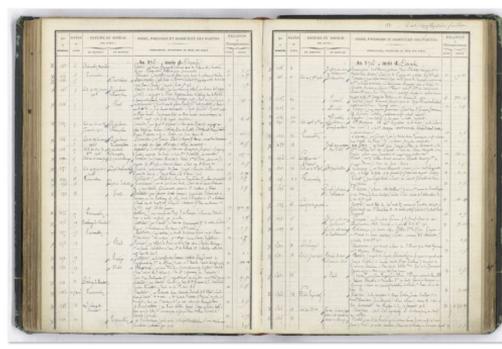
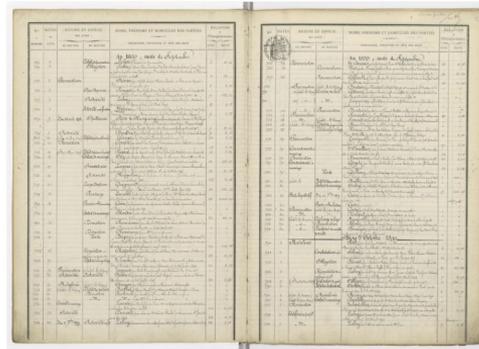
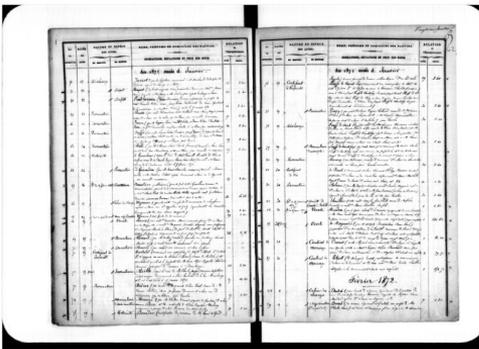
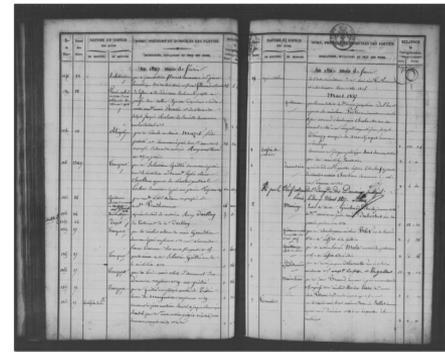
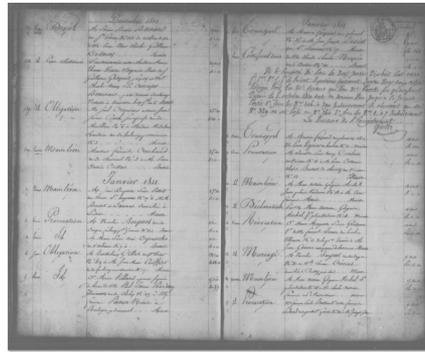
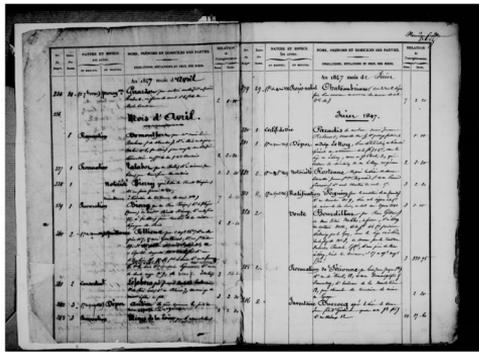
- N° d'ordre
- Date de l'acte
- Typologie
- Nom du signataire en gras

Limite inférieure de l'acte :

- Description de l'acte : dernière ligne parfois incomplète
- Date de l'enregistrement de l'acte
- Taxes acquittées

Le corpus d'images numériques

Près d'un demi-siècle de campagnes de reprographie analogique et numérique

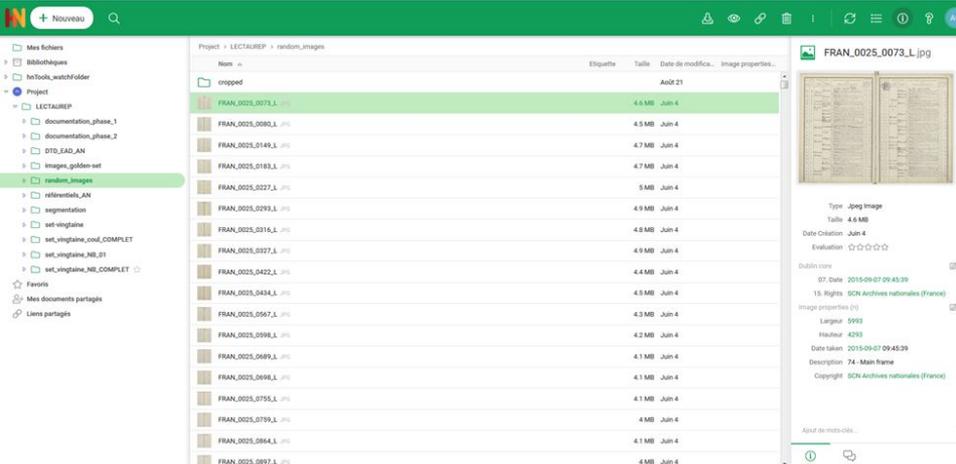


Les sous-corpus d'images numériques

Des dizaines de mains différentes

Un *golden set* : 10000 doubles pages de 41 registres (1789-1875) numérisés en noir et blanc et en couleur, référence pour les phases 2 et 3

Un *random set* : 1000 doubles pages aléatoires de quatre campagnes de numérisation récentes en couleur (années 1880-années 1930)



Basculement du stockage sur Sharedocs (Huma-Num)

Objectifs du projet

Objectifs du projet

Enjeux :

- pour le public des archives : *service de recherche en texte intégral*
- pour les chercheur·ses : *exploitations statistiques sur l'ensemble du corpus*
- pour les services publics d'archives et les institutions patrimoniales : *outil mutualisé, en réseau*

Approche :

- Reconnaissance automatique de structure et d'écriture manuscrite (HTR)
- Indexation et publication sur une plateforme dotée d'une recherche avancée

Chaîne de traitements

1. Cadrage et redressement des tableaux
2. Analyse de la mise en page des répertoires
3. Détection des lignes de texte et segmentation (polygones)
4. Reconnaissance d'écriture manuscrite (HTR + *Word Spotting*)
5. Extraction d'informations (personnes, adresses, lieux, professions, types d'actes, mots clés)
6. Mise à disposition du public et de la communauté des chercheur·ses
7. Brique participative (correction de la segmentation, des transcriptions et de l'indexation)

Objectifs des phases 1 et 2

Phase 1 (2018) :

- état de l'art, banc d'essai et premiers développements à partir de l'exemple de l'étude Marotte
- établissement de préconisations pour le déploiement de la chaîne de traitements

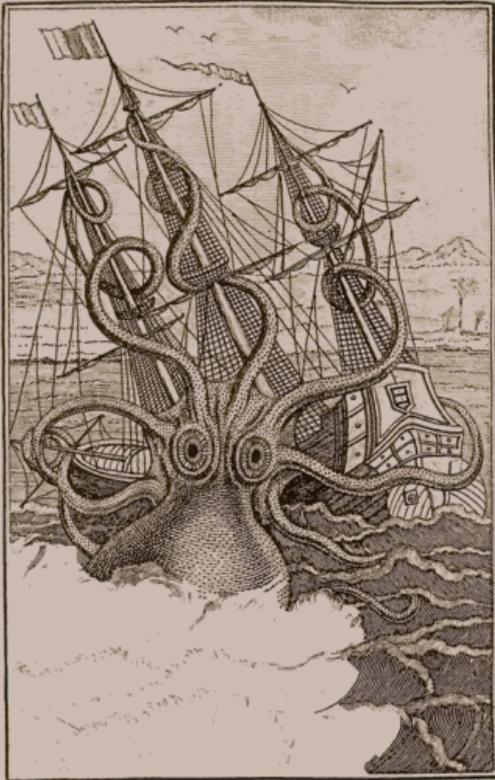
Phase 2 (mai 2019 - novembre 2019) :

4 volets :

1. optimisation des outils, méthodes et flux de travaux pour l'analyse de la structure des documents et leur transcription
2. indexation de la transcription et alignement sur des référentiels
3. étude des scénarios et cas d'usage dans la perspective de développer une plateforme de transcription
4. élaboration de recommandations pour la mise en place d'une plateforme technique et d'une infrastructure d'hébergement

Kraken et eScriptorium

Kraken



En quelques mots :

- un logiciel en ligne de commande (CLI) pour la détection de lignes de texte (segmentation) et la reconnaissance automatique de texte manuscrit (transcription)

Avantages :

- développé par Benjamin Kiessling (Scripta, EPHE-Inria)

Documentation :

- <http://kraken.re>

eScriptorium

En quelques mots :

- une plateforme web dotant **Kraken** d'une interface graphique
- destinée à gérer l'essentiel de la chaîne de traitements des documents

Avantages :

- proximité avec l'équipe de développement
- entièrement gratuit, *open source* et agnostique
- modèles entraînés maîtrisés, accessibles et partageables

Freins :

- développement en cours
- peu documenté

eScriptorium a été choisie dès le début de la phase 2 comme cadre pour l'ensemble des développements produits dans le cadre du projet LECTAUREP.

Changement d'outils

Transkribus

- facile d'utilisation
- travail collaboratif
- formats d'export variés
- customisation faible

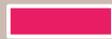


Kraken + eScriptorium

- unité logicielle
- facile d'utilisation
- customisation forte (*open source*)
- maîtrise du stockage des données et des modèles
- développements prometteurs

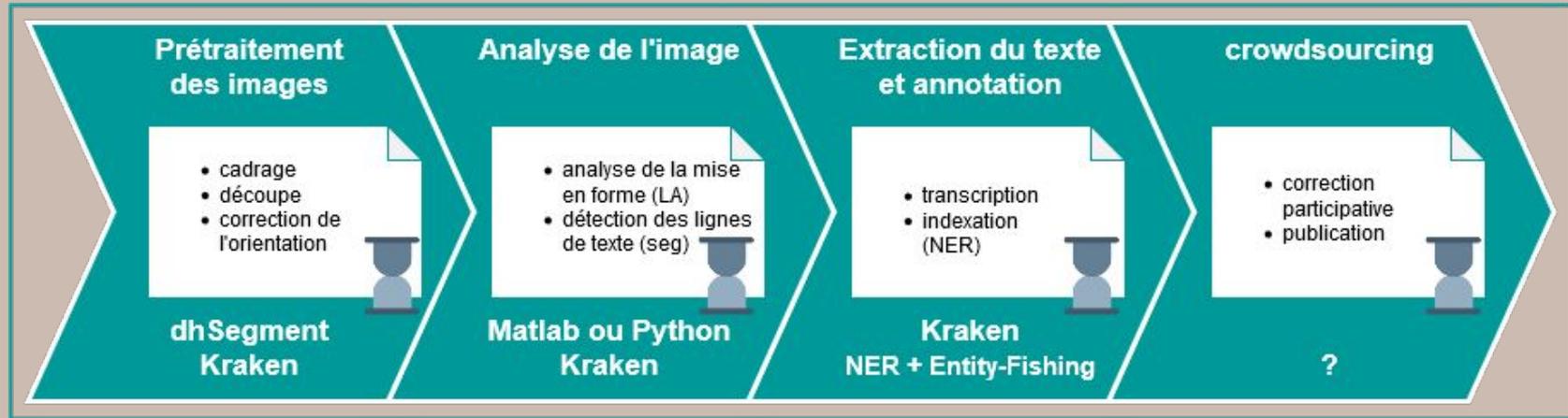


-
- pas de maîtrise du stockage des données et des modèles entraînés
 - modèles non *open source*
 - avenir incertain



- travail collaboratif limité à ce stade
- pas d'instance à jour disponible

Une chaîne complète dans eScriptorium



A terme, eScriptorium intégrera les modules de prétraitement et d'analyse d'images, de gestion des métadonnées, de segmentation, de transcription, d'indexation, accompagnés de fonctionnalités (*features*) permettant d'en faire une plateforme ouverte pour le *crowdsourcing*.

Analyse de la mise en page

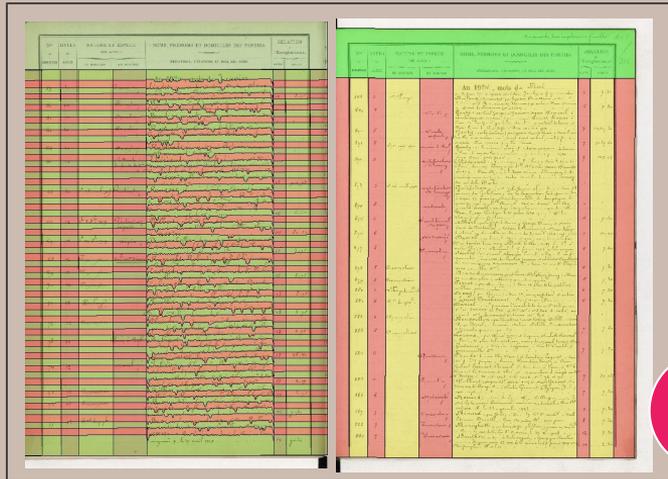
Analyse de la mise en page

Conclusion de la phase 1 :

- algorithme de recadrage (*seam carving*) pour détecter la structure du tableau puis transcrire les cellules

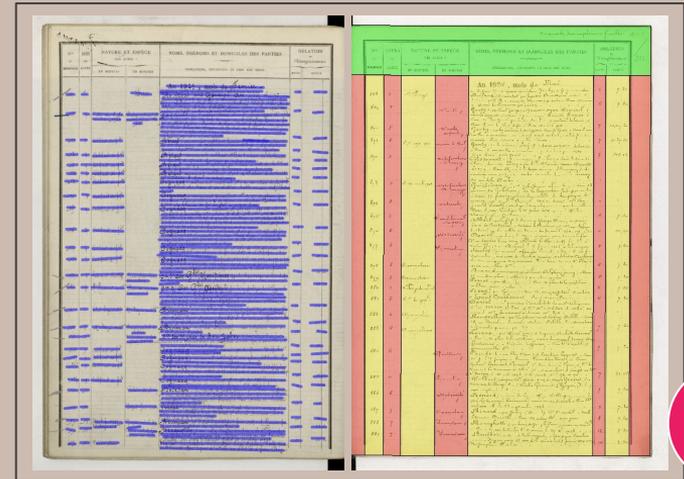
Exploration de la phase 2 :

- détecter les colonnes, détecter les segments de texte sur la page, déduire les lignes à partir des coordonnées des segments



A scanned document page with a table structure. The table has multiple columns and rows. The first column is highlighted in red, and the second column is highlighted in yellow. The rest of the table is in white. The text is in French, and the table appears to be a ledger or account book.

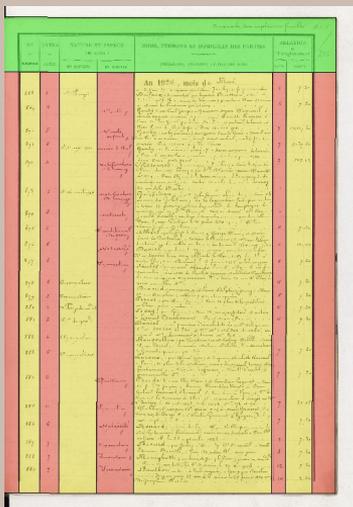
1



A scanned document page with a table structure. The table has multiple columns and rows. The first column is highlighted in blue, the second column in red, and the third column in yellow. The rest of the table is in white. The text is in French, and the table appears to be a ledger or account book.

2

Détection automatique des colonnes



CENTRE DE RECHERCHE		BUREAU NATIONAL DE STATISTIQUE DES PAYS-BAS		PROBABILITE	
1970	1	1	1	1	1
1971	2	2	2	2	2
1972	3	3	3	3	3
1973	4	4	4	4	4
1974	5	5	5	5	5
1975	6	6	6	6	6
1976	7	7	7	7	7
1977	8	8	8	8	8
1978	9	9	9	9	9
1979	10	10	10	10	10
1980	11	11	11	11	11
1981	12	12	12	12	12
1982	13	13	13	13	13
1983	14	14	14	14	14
1984	15	15	15	15	15
1985	16	16	16	16	16
1986	17	17	17	17	17
1987	18	18	18	18	18
1988	19	19	19	19	19
1989	20	20	20	20	20
1990	21	21	21	21	21
1991	22	22	22	22	22
1992	23	23	23	23	23
1993	24	24	24	24	24
1994	25	25	25	25	25
1995	26	26	26	26	26
1996	27	27	27	27	27
1997	28	28	28	28	28
1998	29	29	29	29	29
1999	30	30	30	30	30
2000	31	31	31	31	31
2001	32	32	32	32	32
2002	33	33	33	33	33
2003	34	34	34	34	34
2004	35	35	35	35	35
2005	36	36	36	36	36
2006	37	37	37	37	37
2007	38	38	38	38	38
2008	39	39	39	39	39
2009	40	40	40	40	40
2010	41	41	41	41	41
2011	42	42	42	42	42
2012	43	43	43	43	43
2013	44	44	44	44	44
2014	45	45	45	45	45
2015	46	46	46	46	46
2016	47	47	47	47	47
2017	48	48	48	48	48
2018	49	49	49	49	49
2019	50	50	50	50	50
2020	51	51	51	51	51
2021	52	52	52	52	52
2022	53	53	53	53	53
2023	54	54	54	54	54
2024	55	55	55	55	55
2025	56	56	56	56	56
2026	57	57	57	57	57
2027	58	58	58	58	58
2028	59	59	59	59	59
2029	60	60	60	60	60
2030	61	61	61	61	61
2031	62	62	62	62	62
2032	63	63	63	63	63
2033	64	64	64	64	64
2034	65	65	65	65	65
2035	66	66	66	66	66
2036	67	67	67	67	67
2037	68	68	68	68	68
2038	69	69	69	69	69
2039	70	70	70	70	70
2040	71	71	71	71	71
2041	72	72	72	72	72
2042	73	73	73	73	73
2043	74	74	74	74	74
2044	75	75	75	75	75
2045	76	76	76	76	76
2046	77	77	77	77	77
2047	78	78	78	78	78
2048	79	79	79	79	79
2049	80	80	80	80	80
2050	81	81	81	81	81
2051	82	82	82	82	82
2052	83	83	83	83	83
2053	84	84	84	84	84
2054	85	85	85	85	85
2055	86	86	86	86	86
2056	87	87	87	87	87
2057	88	88	88	88	88
2058	89	89	89	89	89
2059	90	90	90	90	90
2060	91	91	91	91	91
2061	92	92	92	92	92
2062	93	93	93	93	93
2063	94	94	94	94	94
2064	95	95	95	95	95
2065	96	96	96	96	96
2066	97	97	97	97	97
2067	98	98	98	98	98
2068	99	99	99	99	99
2069	100	100	100	100	100
2070	101	101	101	101	101
2071	102	102	102	102	102
2072	103	103	103	103	103
2073	104	104	104	104	104
2074	105	105	105	105	105
2075	106	106	106	106	106
2076	107	107	107	107	107
2077	108	108	108	108	108
2078	109	109	109	109	109
2079	110	110	110	110	110
2080	111	111	111	111	111
2081	112	112	112	112	112
2082	113	113	113	113	113
2083	114	114	114	114	114
2084	115	115	115	115	115
2085	116	116	116	116	116
2086	117	117	117	117	117
2087	118	118	118	118	118
2088	119	119	119	119	119
2089	120	120	120	120	120
2090	121	121	121	121	121
2091	122	122	122	122	122
2092	123	123	123	123	123
2093	124	124	124	124	124
2094	125	125	125	125	125
2095	126	126	126	126	126
2096	127	127	127	127	127
2097	128	128	128	128	128
2098	129	129	129	129	129
2099	130	130	130	130	130
2100	131	131	131	131	131
2101	132	132	132	132	132
2102	133	133	133	133	133
2103	134	134	134	134	134
2104	135	135	135	135	135
2105	136	136	136	136	136
2106	137	137	137	137	137
2107	138	138	138	138	138
2108	139	139	139	139	139
2109	140	140	140	140	140
2110	141	141	141	141	141
2111	142	142	142	142	142
2112	143	143	143	143	143
2113	144	144	144	144	144
2114	145	145	145	145	145
2115	146	146	146	146	146
2116	147	147	147	147	147
2117	148	148	148	148	148
2118	149	149	149	149	149
2119	150	150	150	150	150
2120	151	151	151	151	151
2121	152	152	152	152	152
2122	153	153	153	153	153
2123	154	154	154	154	154
2124	155	155	155	155	155
2125	156	156	156	156	156
2126	157	157	157	157	157
2127	158	158	158	158	158
2128	159	159	159	159	159
2129	160	160	160	160	160
2130	161	161	161	161	161
2131	162	162	162	162	162
2132	163	163	163	163	163
2133	164	164	164	164	164
2134	165	165	165	165	165
2135	166	166	166	166	166
2136	167	167	167	167	167
2137	168	168	168	168	168
2138	169	169	169	169	169
2139	170	170	170	170	170
2140	171	171	171	171	171
2141	172	172	172	172	172
2142	173	173	173	173	173
2143	174	174	174	174	174
2144	175	175	175	175	175
2145	176	176	176	176	176
2146	177	177	177	177	177
2147	178	178	178	178	178
2148	179	179	179	179	179
2149	180	180	180	180	180
2150	181	181	181	181	181
2151	182	182	182	182	182
2152	183	183	183	183	183
2153	184	184	184	184	184
2154	185	185	185	185	185
2155	186	186	186	186	186
2156	187	187	187	187	187
2157	188	188	188	188	188
2158	189	189	189	189	189
2159	190	190	190	190	190
2160	191	191	191	191	191
2161	192	192	192	192	192
2162	193	193	193	193	193
2163	194	194	194	194	194
2164	195	195	195	195	195
2165	196	196	196	196	196
2166	197	197	197	197	197
2167	198	198	198	198	198
2168	199	199	199	199	199
2169	200	200	200	200	200
2170	201	201	201	201	201
2171	202	202	202	202	202
2172	203	203	203	203	203
2173	204	204	204	204	204
2174	205	205	205	205	205
2175	206	206	206	206	206
2176	207	207	207	207	207
2177	208	208	208	208	208
2178	209	209	209	209	209
2179	210	210	210	210	210
2180	211	211	211	211	211
2181	212	212	212	212	212
2182	213	213	213	213	213
2183	214	214	214	214	214
2184	215	215	215	215	215
2185	216	216	216	216	216
2186	217	217	217	217	217
2187	218	218	218	218	218
2188	219	219	219	219	219
2189	220	220	220	220	220
2190	221	221	221	221	221
2191	222	222	222	222	222
2192	223	223	223	223	223
2193	224	224	224	224	224
2194	225	225	225	225	225
2195	226	226	226	226	226
2196	227	227	227	227	227
2197	228	228	228	228	228
2198	229	229	229	229	229
2199	230	230	230	230	230
2200	231	231	231	231	231
2201	232	232	232	232	232
2202	233	233	233	233	233
2203	234	234	234	234	234
2204	235	235	235	235	235
2205	236	236	236	236	236
2206	237	237	237	237	237
2207	238	238	238	238	238
2208	239	239	239	239	239
2209	240	240	240	240	240
2210					

Détection des unités d'informations

The image shows a snippet of a handwritten document with several columns. A red dashed line highlights a specific row. A blue dashed line highlights the first column of this row. A blue dashed box highlights the last two columns of this row. Arrows point from text boxes below to these highlighted areas.

1657	27	Proclamation	Doire (par Frédéricque Elise Françoise Mathilde Mathieu à Saint B ^t de la Madeleine 9, veuve de Marie Valere Gault pour remplir 3 ^m à et Marcel Sachet à Saint B ^t Michel 9	1-	3.75
1658	27	Proclamation	Montoux (par Frédéricque Jules) 5 ^m à Saint Louis		

Début de l'acte : écriture à la première ligne de l'acte

Colonne centrale : Description de l'acte

Fin de l'acte : écriture à la dernière ligne de l'acte

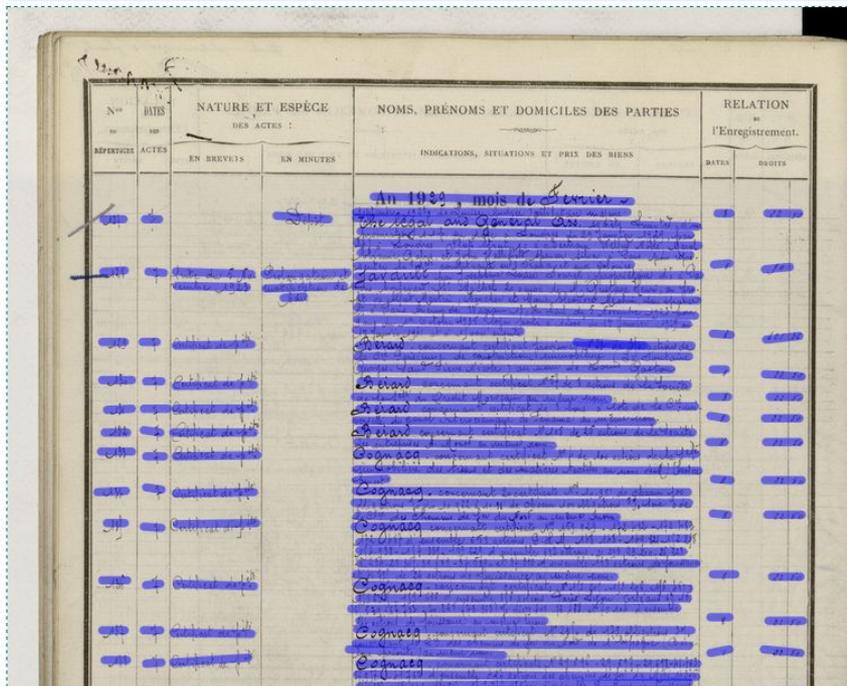
Les résultats d'analyse de la phase 1 sur les répertoires de l'études Marotte ne sont pas généralisables.

Segmentation

Edition des lignes de base (*baselines*) avec eScriptorium

eScriptorium Home About Contact

My Documents Hello testlectaurep ▾



Usage

Drop a picture or click on the dashed rectangle to initialise the baseline editor.

Left click on the image to **create** new line, **Right click** to **add points** and **Left click** to finish it.

You can keep the mouse button pressed for free drawing.
Hitting escape while drawing a line cancels it. It also clears the selection.

Left click on a line to select it, then you can drag it's closest control point.
If you click exactly on a control point, a yellow trash button  appears allowing to delete it.

Double click on the line will create a new control point at the mouse location.
You can go through your changes history back and forth with **Ctrl+Z**(undo) and **Ctrl+Y**(redo) or by using the corresponding buttons.

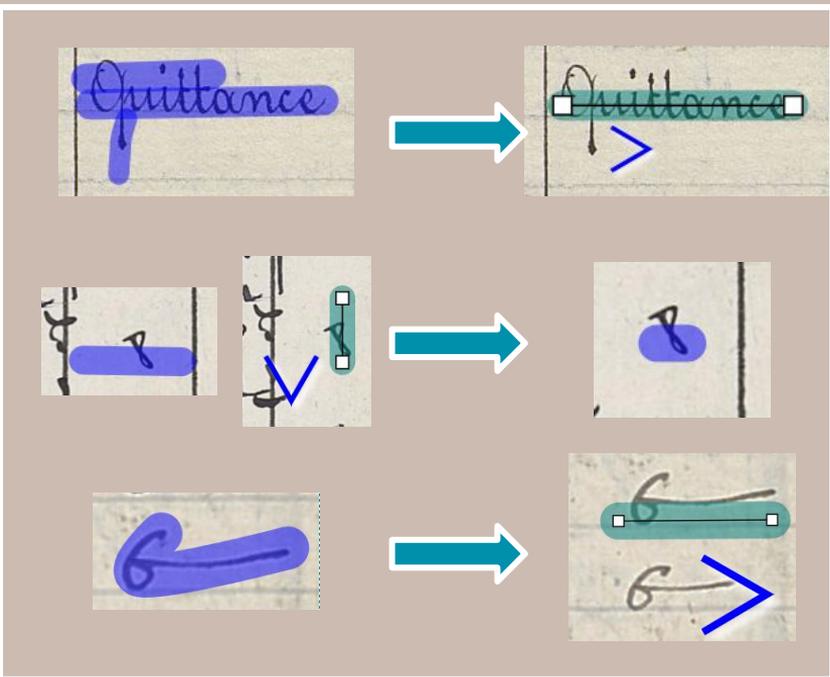
Shift+click allows to add or remove a line from the selection, shift and dragging creates a lasso selection tool.

Ctrl+drag allows to move the entire selection at once.

You can zoom with the mouse wheel or the device equivalent. Then pan the image with **Right click dragging**.

By switching to region mode, you can create and edit regions like you would with lines.

Acquisition des données d'entraînement



exemples d'erreurs de segmentation

Plusieurs temps de formation et de calage :

- pour créer les données d'entraînement
- pour entraîner les modèles

2 campagnes de segmentation :

- 3 versements
- 20 n&b + 22 couleur
- 42 n&b et couleur

6 modèles entraînés :

- variation des paramètres et des corpus de données d'entraînement
- résultats entre 50 % et 60 % d'exactitude
- cible : 70 %

Transcription

HTR : résultats de la phase 1

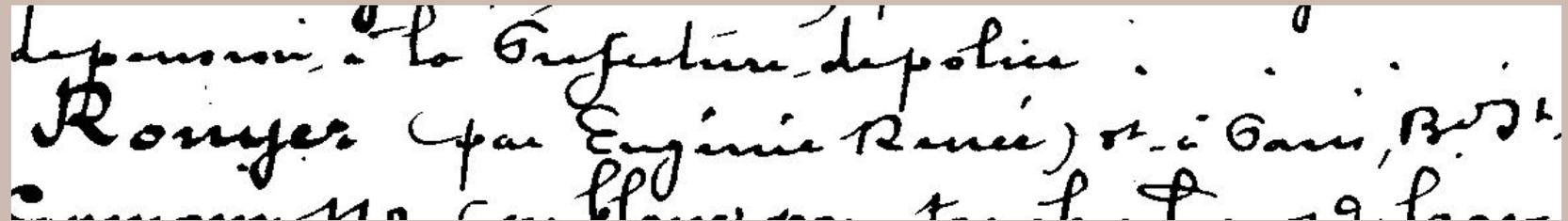
- Entraînement de 2 modèles d'HTR à partir de pages transcrites manuellement (un seul scribe, étude Marotte) :
 - M1 : 40 pages
 - M2 : 50 pages (1 million de mots)
- Taux d'erreur par caractères (CER) sur un échantillon test du même répertoire:
 - M1 : **13,5 %**
 - M2 : **10,4 %**
- Test du modèle M2 sur d'autres registres (écriture et qualité différentes) :
 - taux d'erreur autour de 40 %
 - nécessité de données d'entraînement plus hétérogènes.

Avancées de la phase 2

- 1 modèle entraîné avec Kraken à partir des données d'entraînement produites sur Transkribus :
 - pipeline : Transkribus → XML PAGE → HTML → Kraken
 - 19,36% CER
 - imprécision des segments récupérés avec la pipeline :



Jastean (de et. tit. de mare) dem^t à Paris

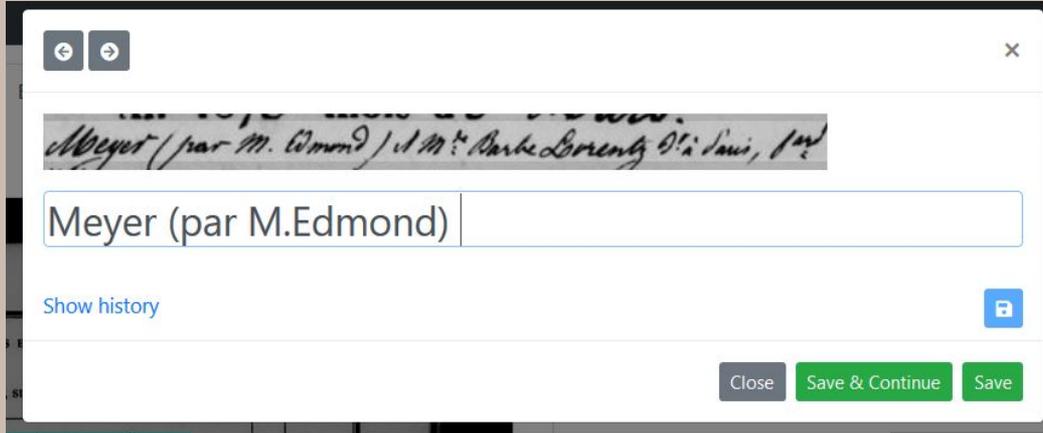


Rouyer (par Engéline Rouée) et à Paris, B. D. H.

- pas davantage d'entraînement en l'attente d'une segmentation opérationnelle et d'une mise à jour de l'interface d'eScriptorium

Transcription dans eScriptorium

- interface ergonomique pour la transcription et l'édition du texte
- eScriptorium et Kraken gèrent désormais les polygones, ce qui améliore la qualité des données d'entraînement



Perspectives

Développements à venir

À approfondir (Kraken / eScriptorium) :

- ☛ Découpage et redressement éventuel des doubles pages après détection des zones des tableaux (module basé sur dhSegment ou Kraken) ;
- ☛ Structuration des analyses à partir des indices de mise en page et de mise en forme ;
- ☛ Détection automatique des mains d'écriture pour adapter le modèle de transcription ;
- ☛ Entraînement de modèles de transcription spécifiques à certaines mains d'écriture ;
- ☛ Mise en production de l'interface de traitement des images.

Développements à venir

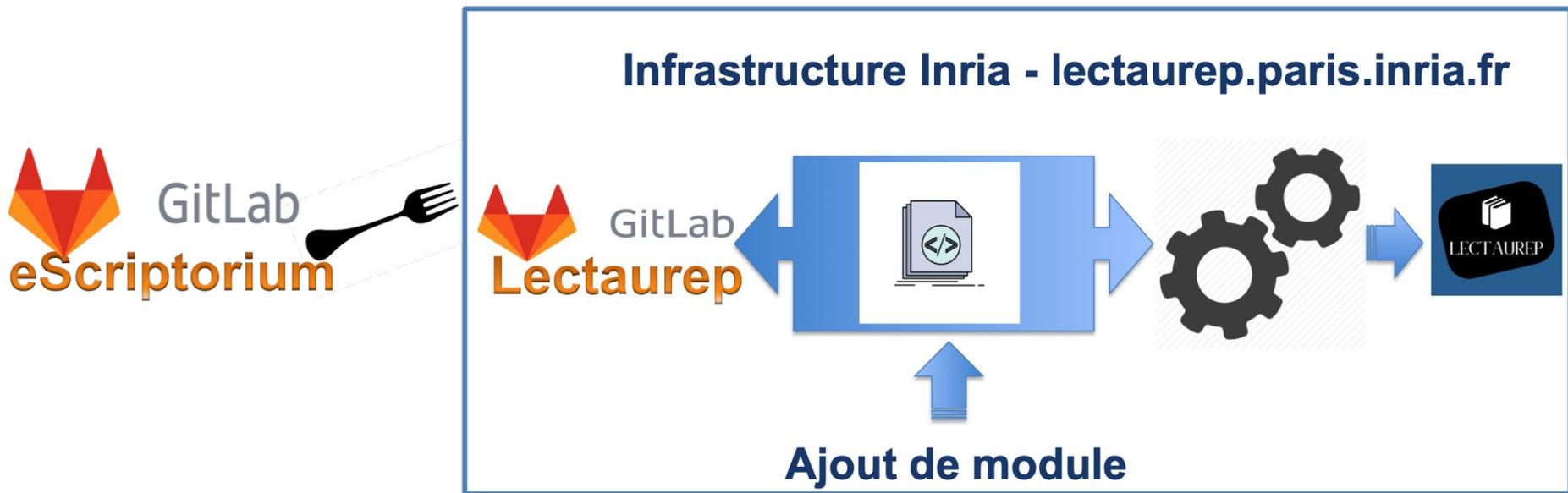
Plusieurs mains de scribes par répertoire, plusieurs pratiques de mise en page

Plus de 1800 répertoires

Un modèle à entraîner pour chaque main...

- ☛ Une interface collaborative adossée à eScriptorium pour entraîner puis corriger les données de segmentation, de transcription, voire d'indexation obtenues par automatisation (plateforme de myriadisation) ;
- ☛ Si possible, des outils de visualisation et de traitement des données ;
- ☛ Si possible, des fonctionnalités de reconnaissance d'entités nommées et de liage de ces entités à des référentiels internes ou externes aux Archives nationales.

Une plateforme collaborative adossée à une instance eScriptorium pour LECTAUREP



Merci !

des questions ?