

Interprétation sémantique de tables historiques : des sources anciennes aux graphes de connaissances

Solenn Tual¹, Nathalie Abadie¹, Joseph Chazalon², Bertrand Duménieu³, Julien Perret¹

¹ LASTIG, Université Gustave Eiffel, IGN-ENSG

² LRE, EPITA

³ CRH, EHESS-CNRS

Journée d'étude HTR/OCR ANR Oceanlinks - "Reconnaissance automatique caractères et extraction d'information spatiale à partir de sources anciennes numérisées"

10 décembre 2025, Paris

(1)



(2)



(3)



Contexte

- **Quantités significatives de documents tabulaires** dans les services d'archives, les bibliothèques, ...
- **Informations intéressantes pour différentes disciplines** (histoire, démographie, économie historique, géographie, etc.)
- **Numérisation** progressive de ces documents... les **informations restent complexes à retrouver**

Exemple : documents tabulaires numérisés et diffusés par les Archives Départementales du Val-de-Marne (94)

5 866 images

Registres détaillés des élèves inscrits à l'école
vétérinaire de Maisons-Alfort (1760-1817)[illegible]

24 012 images

Registres d'états de sections* du cadastre napoléonien (1809-)

CANTON	TOWNSHIP	TAX PAYER	VOTE		TOWNSHIP		TOWNSHIP	TOWNSHIP
			For	Against	For	Against		
1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27
28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45
46	47	48	49	50	51	52	53	54
55	56	57	58	59	60	61	62	63
64	65	66	67	68	69	70	71	72
73	74	75	76	77	78	79	80	81
82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99
100	101	102	103	104	105	106	107	108
109	110	111	112	113	114	115	116	117
118	119	120	121	122	123	124	125	126
127	128	129	130	131	132	133	134	135
136	137	138	139	140	141	142	143	144
145	146	147	148	149	150	151	152	153
154	155	156	157	158	159	160	161	162
163	164	165	166	167	168	169	170	171
172	173	174	175	176	177	178	179	180
181	182	183	184	185	186	187	188	189
190	191	192	193	194	195	196	197	198
199	200	201	202	203	204	205	206	207
208	209	210	211	212	213	214	215	216
217	218	219	220	221	222	223	224	225
226	227	228	229	230	231	232	233	234
235	236	237	238	239	240	241	242	243
244	245	246	247	248	249	250	251	252
253	254	255	256	257	258	259	260	261
262	263	264	265	266	267	268	269	270
271	272	273	274	275	276	277	278	279
280	281	282	283	284	285	286	287	288
289	290	291	292	293	294	295	296	297
298	299	300	301	302	303	304	305	306
307	308	309	310	311	312	313	314	315
316	317	318	319	320	321	322	323	324
325	326	327	328	329	330	331	332	333
334	335	336	337	338	339	340	341	342
343	344	345	346	347	348	349	350	351
352	353	354	355	356	357	358	359	360
361	362	363	364	365	366	367	368	369
370	371	372	373	374	375	376	377	378
379	380	381	382	383	384	385	386	387
388	389	390	391	392	393	394	395	396
397	398	399	400	401	402	403	404	405
406	407	408	409	410	411	412	413	414
415	416	417	418	419	420	421	422	423
424	425	426	427	428	429	430	431	432
433	434	435	436	437	438	439	440	441
442	443	444	445	446	447	448	449	450
451	452	453	454	455	456	457	458	459
460	461	462	463	464	465	466	467	468
469	470	471	472	473	474	475	476	477
478	479	480	481	482	483	484	485	486
487	488	489	490	491	492	493	494	495
496	497	498	499	500	501	502	503	504
505	506	507	508	509	510	511	512	513
514	515	516	517	518	519	520	521	522
523	524	525	526	527	528	529	530	531
532	533	534	535	536	537	538	539	540
541	542	543	544	545	546	547	548	549
550	551	552	553	554	555	556	557	558
559	560	561	562	563	564	565	566	567
568	569	570	571	572	573	574	575	576
577	578	579	580	581	582	583	584	585
586	587	588	589	590	591	592	593	594
595	596	597	598	599	600	601	602	603
604	605	606	607	608	609	610	611	612
613	614	615	616	617	618	619	620	621
622	623	624	625	626	627	628	629	630
631	632	633	634	635	636	637	638	639
640	641	642	643	644	645	646	647	648
649	650	651	652	653	654	655	656	657
658	659	660	661	662	663	664	665	666
667	668	669	670	671	672	673	674	675
676	677	678	679	680	681	682	683	684
685	686	687	688	689	690	691	692	693
694	695	696	697	698	699	700	701	702
703	704	705	706	707	708	709	710	711
712	713	714	715	716	717	718	719	720
721	722	723	724	725	726	727	728	729
730	731	732	733	734	735	736	737	738
739	740	741	742	743	744	745	746	747
748	749	750	751	752	753	754	755	756
757	758	759	760	761	762	763	764	765
766	767	768	769	770	771	772	773	774
775	776	777	778	779	780	781	782	783
784	785	786	787	788	789	790	791	792
793	794	795	796	797	798	799	800	801
802	803	804	805	806	807	808	809	810
811	812	813	814	815	816	817	818	819
820	821	822	823	824	825	826	827	828
829	830	831	832	833	834	835	836	837
838	839	840	841	842	843	844	845	846
847	848	849	850	851	852	853	854	855
856	857	858	859	860	861	862	863	864
865	866	867	868	869	870	871	872	873
874	875	876	877	878	879	880	881	882
883	884	885	886	887	888	889	890	891
892	893	894	895	896	897	898	899	900
901	902	903	904	905	906	907	908	909
910	911	912	913	914	915	916	917	918
919	920	921	922	923	924	925	926	927
928	929	930	931	932	933	934	935	936
937	938	939	940	941	942	943	944	945
946	947	948	949	950	951	952	953	954
955	956	957	958	959	960	961	962	963
964	965	966	967	968	969	970	971	972
973	974	975	976	977	978	979	980	981
982	983	984	985	986	987	988	989	990
991	992	993	994	995	996	997	998	999
1000	1001	1002	1003	1004	1005	1006	1007	1008
1009	1010	1011	1012	1013	1014	1015	1016	1017
1018	1019	1020	1021	1022	1023	1024	1025	1026
1027	1028	1029	1030	1031	1032	1033	1034	1035
1036	1037	1038	1039	1040	1041	1042	1043	1044
1045	1046	1047	1048	1049	1050	1051	1052	1053
1054	1055	1056	1057	1058	1059	1060	1061	1062
1063	1064	1065	1066	1067	1068	1069	1070	1071
1072	1073	1074	1075	1076	1077	1078	1079	1080
1081	1082	1083	1084	1085	1086	1087	1088	1089
1090	1091	1092	1093	1094	1095	1096	1097	1098
1099	1100	1101	1102	1103	1104	1105	1106	1107
1108	1109	1110	1111	1112	1113	1114	1115	1116
1117	1118	1119	1120	1121	1122	1123	1124	1125
1126	1127	1128	1129	1130	1131	1132	1133	1134
1135	1136	1137	1138	1139	1140	1141	1142	1143
1144	1145	1146	1147	1148	1149	1150	1151	1152
1153	1154	1155	1156	1157	1158	1159	1160	1161
1162	1163	1164	1165	1166	1167	1168	1169	1170
1171	1172	1173	1174	1175	1176	1177	1178	1179
1180	1181	1182	1183	1184	1185	1186	1187	1188
1189	1190	1191	1192	1193	1194	1195	1196	1197
1198	1199	1200	1201	1202	1203	1204	1205	1206
1207	1208	1209	1210	1211	1212	1213	1214	1215
1216	1217	1218	1219	1220	1221	1222	1223	1224
1225	1226	1227	1228	1229	1230	1231	1232	1233
1234	1235	1236	1237	1238	1239	1240	1241	1242
1243	1244	1245	1246	1247	1248	1249	1250	1251
1252	1253	1254	1255	1256	1257	1258	1259	1260
1261	1262	1263	1264	1265	1266	1267	1268	1269
1270	1271	1272	1273	1274	1275	1276	1277	1278
1279	1280	1281	1282	1283	1284	1285	1286	1287
1288	1289	1290	1291	1292	1293	1294	1295	1296
1297	1298	1299	1300	1301	1302	1303	1304	1305
1306	1307	1308	1309	1310	1311	1312	1313	1314
1315	1316	1317	1318	1319	1320	1321	1322	1323
1324	1325	1326	1327	1328	1329	1330	1331	1332
1333	1334	1335	1336	1337	1338	1339	1340	1341
1342	1343	1344	1345	1346	1347	1348	1349	1350
1351	1352	1353	1354	1355	1356	1357	1358	1359
1360	1361	1362	1363	1364	1365	1366	1367	1368
1369	1370	1371	1372	1373	1374	1375	1376	1377
1378	1379	1380	1381	1382	1383	1384	1385	1386
1387	1388	1389	1390	1391	1392	1393	1394	1395
1396	1397	1398	1399	1400	1401	1402	1403	1404
1405	1406	1407	1408	1409	1410	1411	1412	1413
1414	1415	1416						

* Registres d'initialisation du cadastre (T0)

163 371 images

*Tables alphabétiques de décès, successions
et d'absences (1759-1974)*

List of Persons named, & of their names, surnames, & ages, after the date of their settlement in the colony, according to the Act in that behalf made, 1704, c. 11.						1714		Inhabitants of the said colony, who were not in the said list, in the year 1714.	Inhabitants of the said colony, who were not in the said list, in the year 1714.
First name.	Surname.	Age.	Sex.	1714	1714				
Bartholomew	Smith	40	M	1714	1714				
Charles	Smith	40	M	1714	1714				
George	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James	Smith	40	M	1714	1714				
Robert	Smith	40	M	1714	1714				
Richard	Smith	40	M	1714	1714				
John	Smith	40	M	1714	1714				
Thomas	Smith	40	M	1714	1714				
William	Smith	40	M	1714	1714				
James									

Question de recherche

Comment produire automatiquement des données annotées sémantiquement à partir de tables historiques numérisées afin de faciliter l'accès aux connaissances qu'elles contiennent ?

Question de recherche

Comment produire automatiquement des données annotées sémantiquement à partir de tables historiques numérisées afin de faciliter l'accès aux connaissances qu'elles contiennent ?

- ➡ Travaux existants sur la représentation d'informations issues de documents historiques sous la forme de graphes de connaissances : [1] [2]
- Ne considèrent pas les tables (textes en prose majoritairement)
 - Les tables ont des caractéristiques différentes des textes en prose.

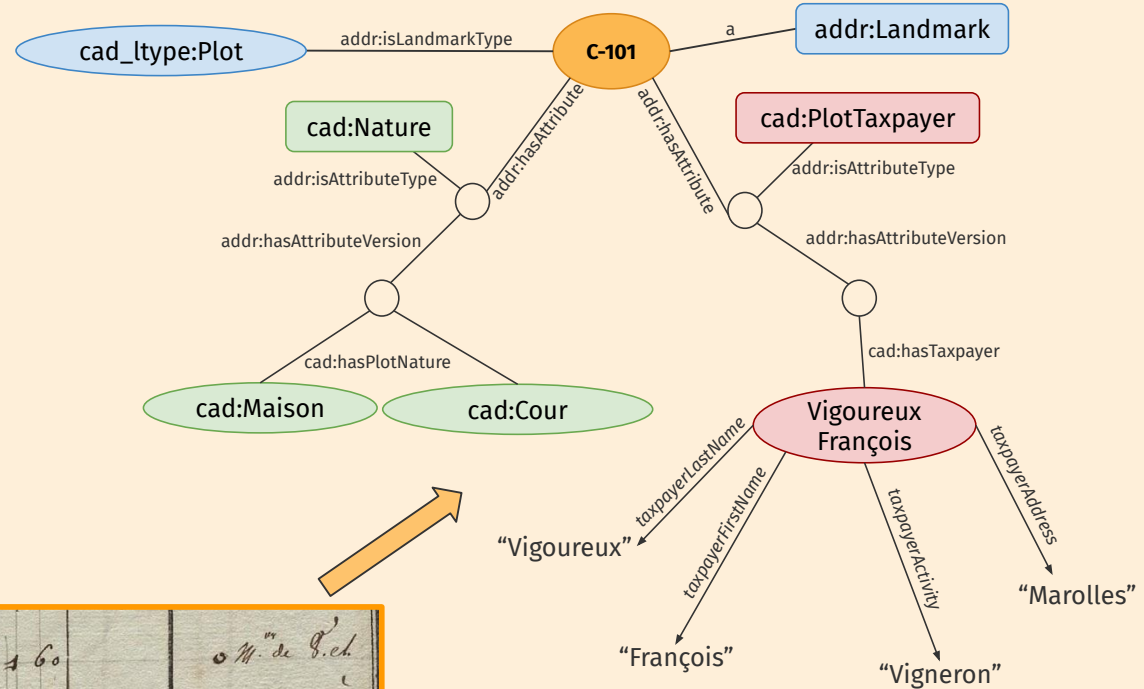
[1] V. Nundloll, et al. Automating the extraction of information from a historical text and building a linked data model for the domain of ecology and conservation science. Heliyon, 8(10) :e10710, 2022.

[2] N. Jain et al. Generating Domain-Specific Knowledge Graphs : Challenges with Open Information Extraction. In Proc. 1st Int. Workshop on Knowledge Graph Generation From Text (Text2KG 2022)

Objectif

CANTON CANTON de Lyon	NUMERO DE FOLIO	SOMME PAR FOLIO	NATURE DE L'IMMEUBLE	CONTENU DE L'IMMEUBLE	CONTENU DE L'IMMEUBLE
Du Village	101	Vigoureux François	Maison	1 60	0 2 1/2 1/2

Du Village	101	Vigoureux François	Maison	1 60	0 2 1/2 1/2
------------	-----	--------------------	--------	------	-------------



Cadastre ancien du XIXème siècle

Registres étudiés décrivent l'état initial des parcelles, produit à la création du cadastre

Tables proches de tables relationnelles, avec des spécificités, telles que:

- Cellules fusionnées
- Valeurs distribuées (*idem*) prenant différentes formes (mot *idem*, traits, cellules vides)

Variabilité de la précision de certaines informations (parfois importante)

Nombreuses **abréviations** ▲

Valeurs manquantes, oublis, etc.

Cellules fusionnées Idem (exemple)

NOMS, PRÉNOMS, PROFESSIONS ET DEMEURES DES PROPRIÉTAIRES.	N° du PLAN rectifié.	CANTONS, TRIAGES ou LIEUX DITS.	NATURE de CULTURE.
197 E. Hilaire Louis f	43	Saône Des	Cer.
201 Mainfroy pr ^e ch ^e lli	44	Postes	C
202 Barre J ^{me} Victor Jhs	45		C

N° Index du
contribuable

Contribuable

Num. Parcelle

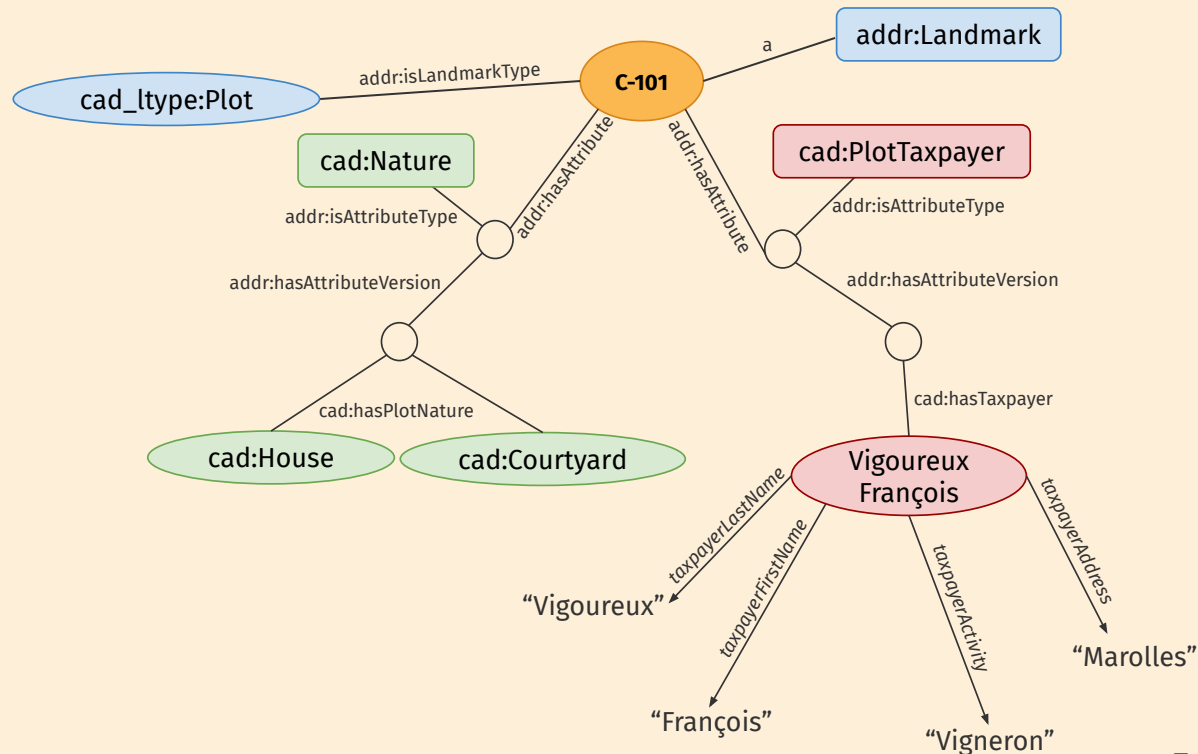
Addr.

Nature

Ontologie de domaine : PeGazUs

PeGaZus: Perpetual Gazetteer of approach-address UtteranceS [1]

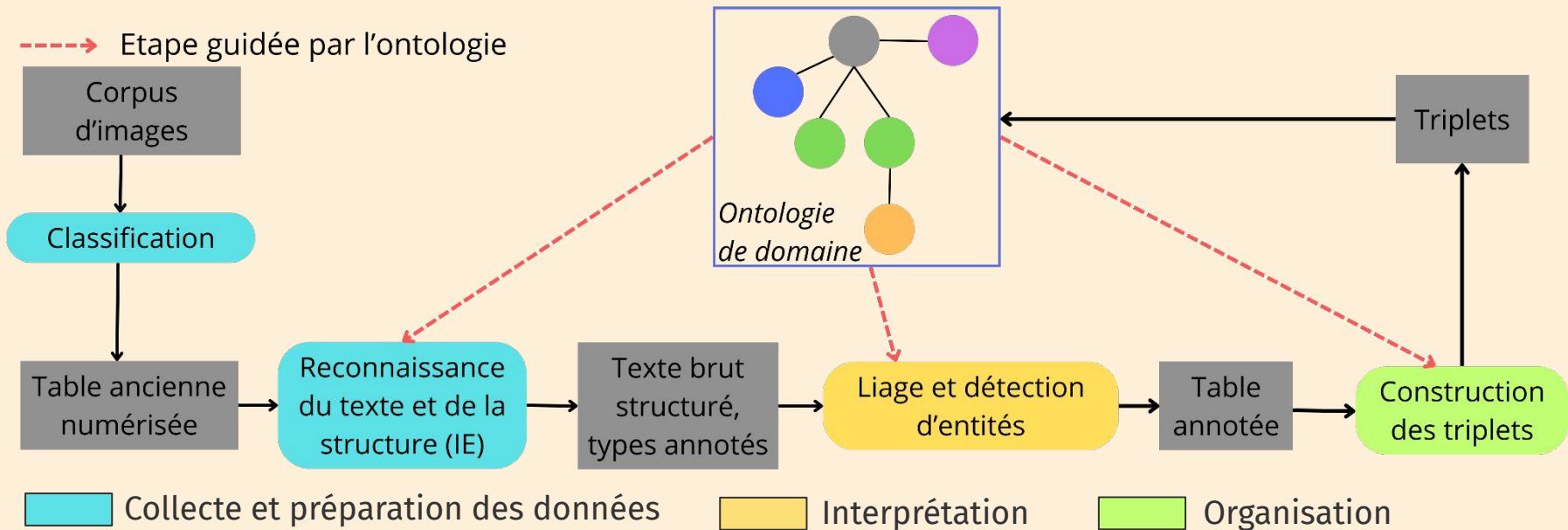
- Ontologie utilisée pour représenter :
 - des entités géographiques et des adresses
 - avec les attributs et leur évolution
 - extension dédiée à la représentation des parcelles



Question de recherche

Comment faciliter l'accès aux connaissances présentées sous la forme de tables contenues dans des documents historiques ?

Application : registres d'états de sections du cadastre ancien (19ème et 20ème siècle)

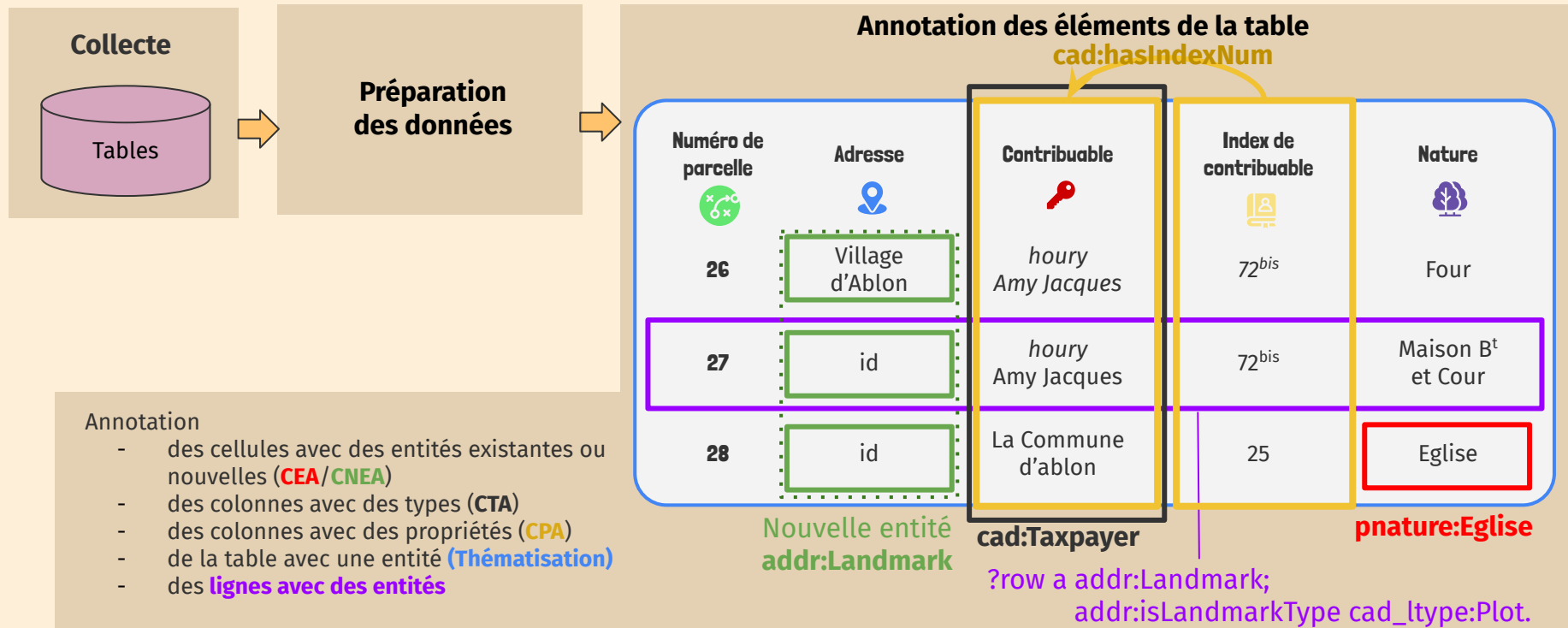


Plan

1. **Etat de l'art**
2. **Notre approche : interprétation sémantique de tables anciennes**
3. **Evaluation**
4. **Conclusion**

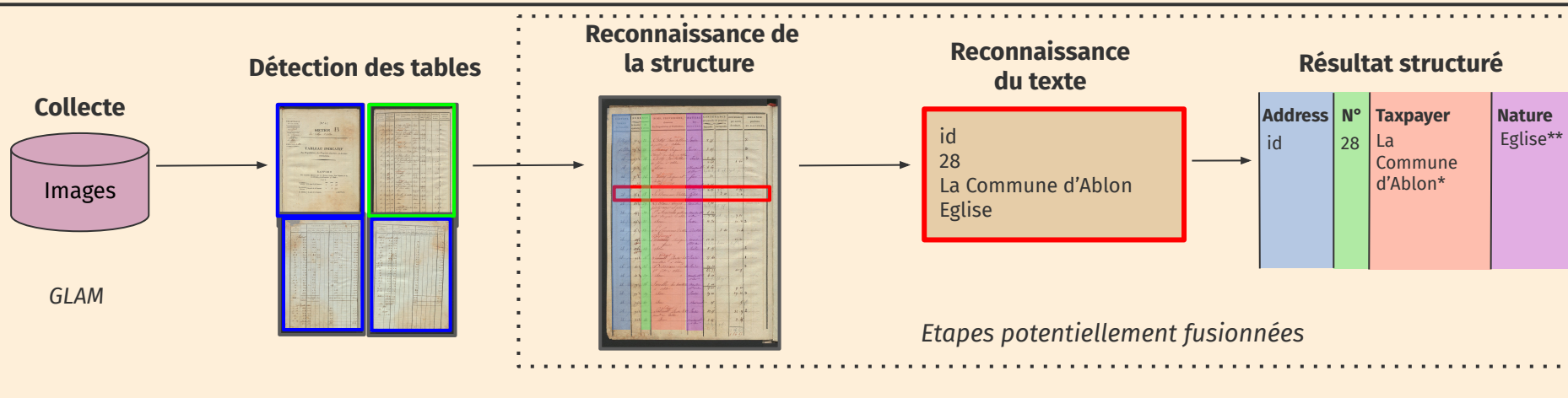
Interprétation sémantique de tables

Annoter les différents éléments qui composent un tableau et leurs relations à l'aide des ressources d'un graphe de connaissances (KG). [11, 12]



Extraction d'informations dans des tables anciennes

- Localiser, transcrire et organiser le texte contenu dans des images de tables vers une représentation informatique structurée exploitable automatiquement
- 2 types d'approches : séquentielles et intégrées
- Tâches (selon les approches, certaines étapes pourront être fusionnées) :



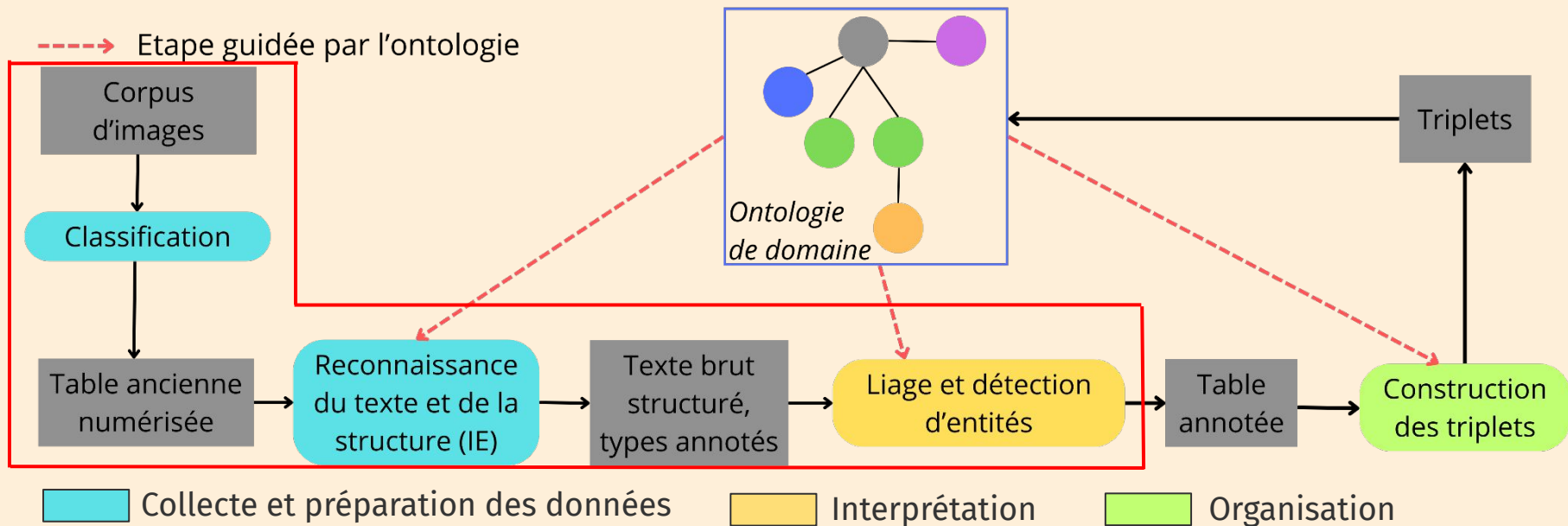
Plan

1. **Etat de l'art**
2. **Notre approche**
3. **Evaluation**
4. **Conclusion**

Interprétation Sémantique de Tables Historiques

Associer les chaînes de traitement d'**Extraction d'Information** et d'**Interprétation Sémantique de Tables** pour structurer et découvrir des connaissances contenues dans des tables historiques.

➡ Toutes les étapes mobilisent une ontologie de domaine



Détection des tables

1. Collecte des images et métadonnées (ici : collecte manuelle)

2. Classification des pages (ici : Yolo-v11 affiné).

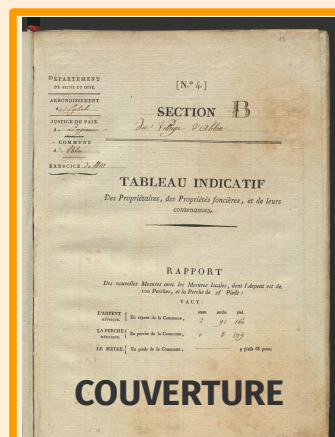


TABLE PRINCIPALE

TABLE SECONDAIRE

RESUME INTERMEDIAIRE

Etats de sections,
Marolles-en-Brie,
1810.

RELEVÉ des Nombres de valeurs

TABLE PRINCIPALE

TABLE PRINCIPALE

PROPRIÉTÉS RATES

SYNTHESE

IE et interprétation sémantique de tables

- La définition du modèle d'annotation des images a été faite en s'appuyant sur l'ontologie : **certaines tâches de STI sont traitées par ce biais.**

Avant 1822

CANTONS, TRIAGES ou lieux-dits.	NUMÉROS de la table alphabétique section.	NOMS, PROFESSIONS, demeures des Propriétaires et Unfruitiers.	NATURE des PROPRIÉTÉS.	CONTENANCE par parcelle de propriété impouable, ou impouable.	par nature de culture.	provisoire DU CLASSEMENT.
Chammi de Brie	95-166	Lefevre	terres	2750	1	1
deuxième	25-167	Guichon, fermier	terres	1135	1	1
	29-168	Caron.	terres	9610	1	1
	95-169	Lefevre	terres	2590	1	1
	25-170	Guichon, fermier	terres	0720	1	1
	25-171	Guichon	terres	0540	1	1
	81-172	Guichon	terres	0615	1	1
	81-173	Guichon	terres	0610	1	1

Après 1822

NOMS, PRÉNOMS, PROFESSIONS ET DEMIÈRES DES PROPRIÉTAIRES.	N° PLAN.	CANTONS ou lieux-dits.	NATURE des PROPRIÉTÉS.	CONTENANCE.	CLASSE.	REVENU.
Guichon, fermier	95-166	25-167	terres	2750	1	1
Guichon, fermier	95-167	25-168	terres	1135	1	1
Guichon, fermier	95-168	25-169	terres	9610	1	1
Guichon, fermier	95-169	25-170	terres	2590	1	1
Guichon, fermier	95-170	25-171	terres	0720	1	1
Guichon, fermier	95-171	25-172	terres	0540	1	1
Guichon, fermier	95-172	25-173	terres	0615	1	1
Guichon, fermier	95-173	25-174	terres	0610	1	1

Second cadastre de la Seine (1835-1845)

NOMS, PRÉNOMS, PROFESSIONS ET DEMIÈRES DES PROPRIÉTAIRES.	N° PLAN.	CANTONS, TRIAGES ou lieux-dits.	NATURE des PROPRIÉTÉS.	CONTENANCE en ares.	CLASSE.	REVENU.
Guichon, fermier	95-166	25-167	terres	2750	1	1
Guichon, fermier	95-167	25-168	terres	1135	1	1
Guichon, fermier	95-168	25-169	terres	9610	1	1
Guichon, fermier	95-169	25-170	terres	2590	1	1
Guichon, fermier	95-170	25-171	terres	0720	1	1
Guichon, fermier	95-171	25-172	terres	0540	1	1
Guichon, fermier	95-172	25-173	terres	0615	1	1
Guichon, fermier	95-173	25-174	terres	0610	1	1

Ici: ontologie “PPerpetual GAZeteer of approach-address Utterances” (PeGazUs) [1]

Annotation des types de colonnes

■ <u>addr:Landmark</u>	■ <u>cad:Taxpayer</u>
■ <u>xsd:string</u>	■ <u>cad:Nature</u>
■ <u>xsd:string</u>	

Annotation des propriétés qui associent les valeurs de deux colonnes

Type des lignes

Here : ?row a add:Landmark;
add:isLandmarkType cad:Plot

Extraction d'informations

- Approche intégrée “au niveau page”
 - ✓ production d'un seul dataset (image → table textuelle) : **Document Attention Network (DAN)** [1]
 - ✓ permet de traiter des pages avec des variations de structure (3 types)
 - ✓ contexte de chaque ligne accessible lors de l'extraction
 - ✓ pas de propagation d'erreurs entre les étapes de reconnaissance de la (a) structure et du (b) texte
 - !! “hallucinations”

CANTON	NOM	PROFESSION	NATURE	CONTENANCE	CONTRIBUTION	COLONNE
Village	26	id	id	id	id	id
Village	27	id	id	id	id	id
Village	28	id	id	id	id	id

Reconnaissance
du texte et de
la mise en page



Numéro de parcelle	Adresse	Contribuable	Index de contribuable	Nature
26	Village d'Ablon	houry Amy Jacques	72 ^{bis}	Four
27	id	houry Amy Jacques	72 ^{bis}	Maison B ^t et Cour
28	id	La Commune d'ablon	25	Eglise

Evaluation : IE dans les tables historiques

- Architecture : Document Attention Network (DAN) [1]
 - Jeu de données : 139 pages annotées manuellement (97/22/28) avec Callico [2]
- CER : 35,3% (min : 7.4%, max : 178.15% = “hallucinations”)

Typography and layout explanation

 Merged Cell
 Ditto
 ↑ Exponent

Additional typography and layout tokens

§ ditto
 ↑ exponent ↓
 go to→the line absent here
to→←merge
×crossed-out± absent here

Ground-truth text

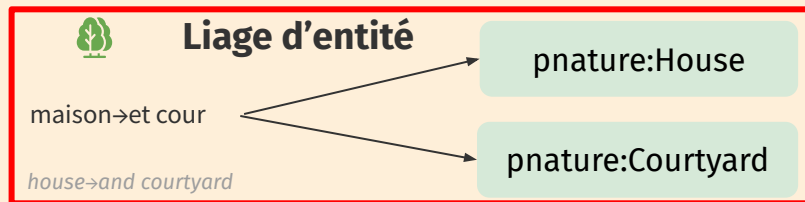
297	Thibault louis f ↑ ois ↓	43	La Voie des→	Terre
201	Mainfray p ↑ re ↓ ch ↑ es ↓	44	←Postes	T
20	Barrié j ↑ ques ↓ victor	45	§	T

Predicted text

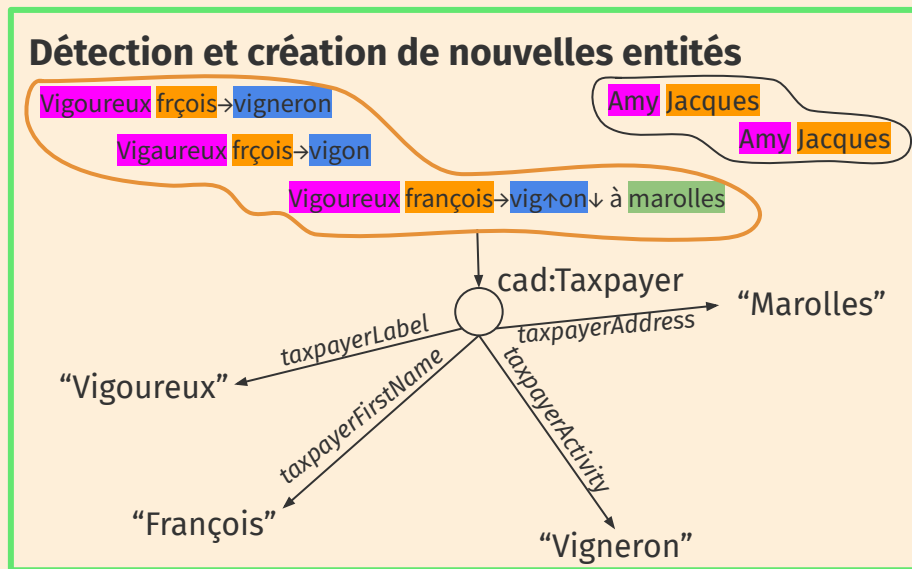
297	Thibault louis f ↑ ois ↓	43	La Voie des→	Terre
201	Mainfray p ↑ re ↓ ch ↑ es ↓	44	§	T
20	Bar m é j ↑ ques ↓ victor	45	§	T

Interprétation sémantique de tables historiques

- Un traitement dédié pour chaque type de colonnes contenant des entités : **annotation des cellules**
 - **CEA**: liage d'entité réalisé en mesurant la similarité des valeurs des cellules et les labels d'entités de l'ontologie.
 - **CNEA**: [NER +] processus de *clustering* utilisé pour créer des **nouvelles entités**
 - Clustering réalisé en comparant les valeurs des cellules, éventuellement structurées avec du NER.



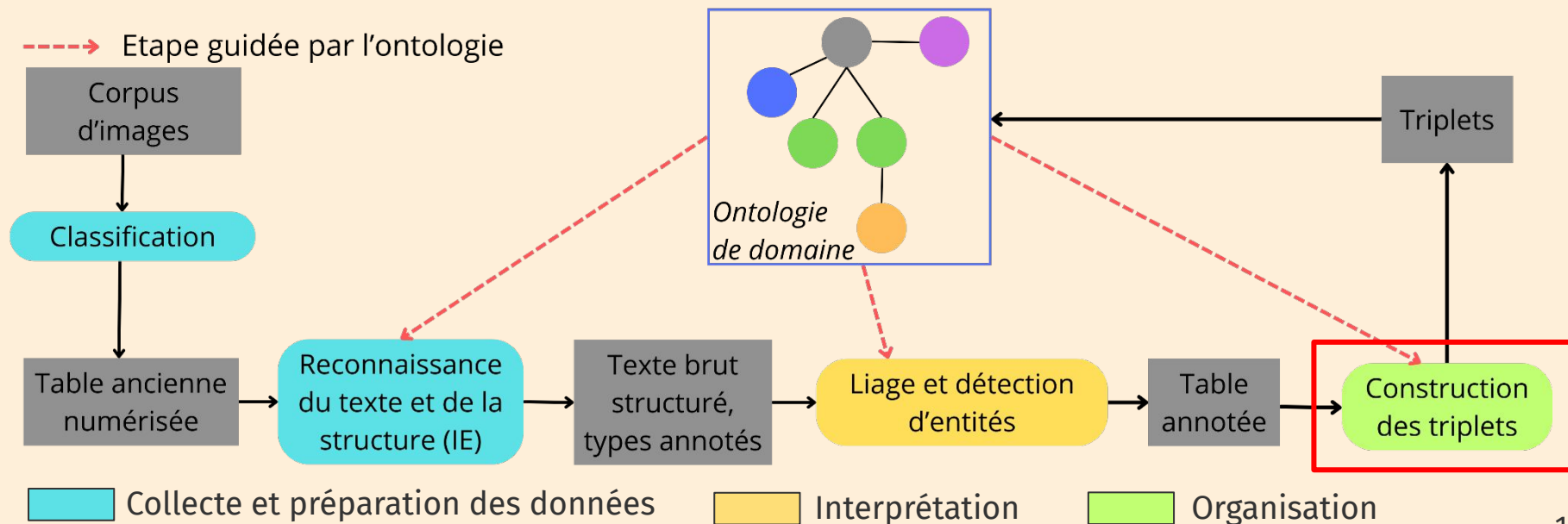
Exemple des "Nature" parcelles



Interprétation sémantique de tables historiques

Associer les chaînes de traitement d'**Extraction d'Information** et d'**Interprétation Sémantique de Tables** pour structurer et découvrir des connaissances contenues dans des tables historiques.

➡ Toutes les étapes mobilisent une ontologie de domaine



Peuplement de l'ontologie

- Peuplement de l'ontologie PeGazUs avec
 - **Les entités et propriétés découvertes dans la table**
 - **Les entités et propriétés créées à partir des métadonnées d'archives**
 - Permet de décrire les documents
 - Conserver le lien entre les connaissances produites et le document initial

Plan

1. **Etat de l'art**
2. **Notre approche**
3. **Evaluation**
4. **Conclusion**

Evaluation du graphe de connaissances

Avec des résultats d'IE obtenus en 1ère intention, est-il possible de créer un graphe de connaissances utilisable pour décrire les parcelles ?

- Jeu de données : 28 pages de L'Hay aux Roses
- Métriques : utilisées pour évaluer des référentiels de données géographiques

	Accord	Déficit	Surplus
Instances	87.3%	12.7%	11.9%
Relations (object type properties)	78.5%	21.5%	19.9%
Datatype properties (sans leur valeur)	87.4%	12.7%	16.7%

Evaluation du graphe de connaissances

Evaluation quantitative

Evaluation des instances créées (par classe)

Type	Agreement	Deficit	Surplus	Actual	Predicted
addr:Event	100	0	0	1	1
rico:RecordSet	100	0	0	2	2
rico:Record	100	0	0	30	30
rico:Instantiation	95.44	4.56	3.33	395	390
addr:Attribute	95.20	4.80	3.43	1084	1050
addr:AttributeVersion	95.20	4.80	3.43	1084	1050
addr:Landmark	94.28	5.72	7.24	367	373
addr:Change	93.91	6.09	3.48	1446	1407
cad:Taxpayer	60.53	39.47	50.27	152	185
addr:LandmarkRelation	47.39	52.61	51.54	728	712

Issues des métadonnées

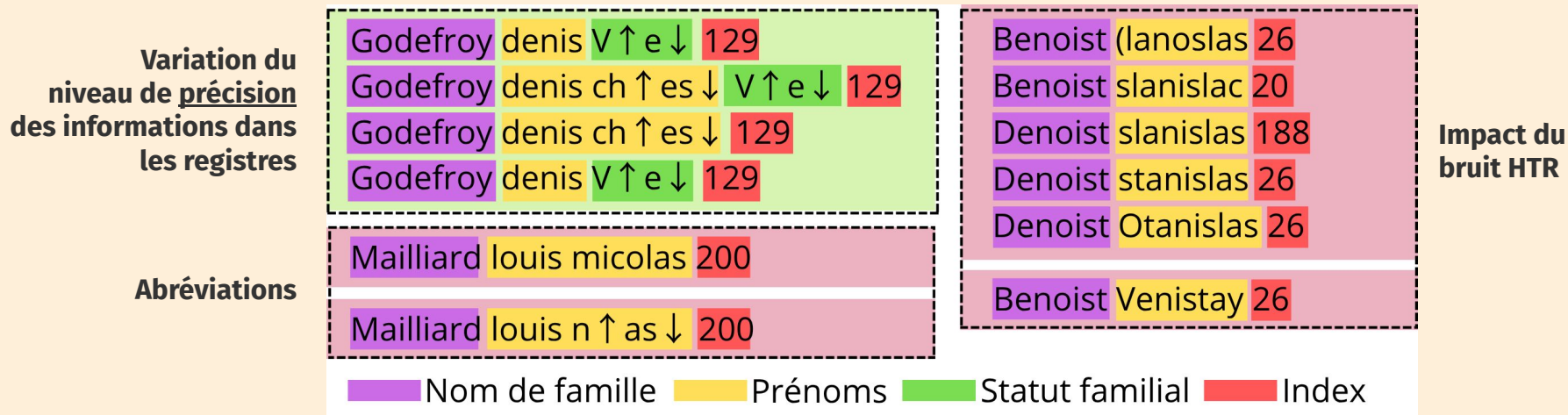
Créées à partir des lignes
de la table (✓)

Créées

- à partir de cellules complexes OU
- d'autres instances

Evaluation du graphe de connaissances

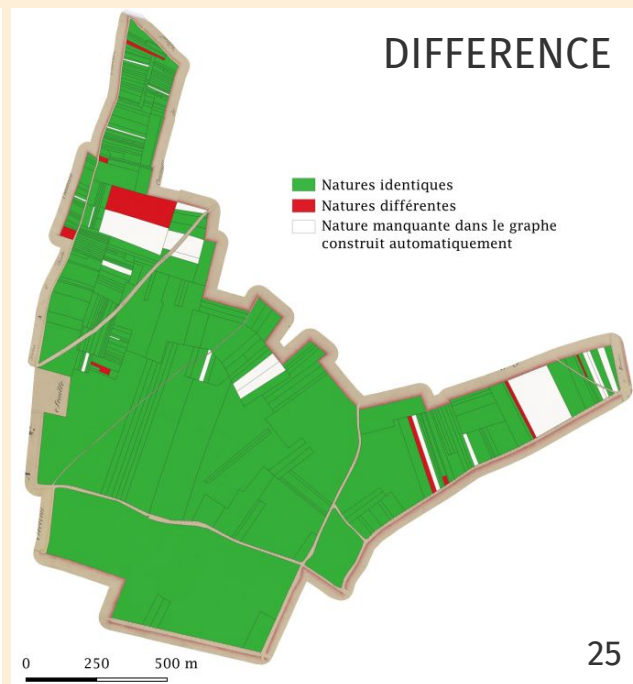
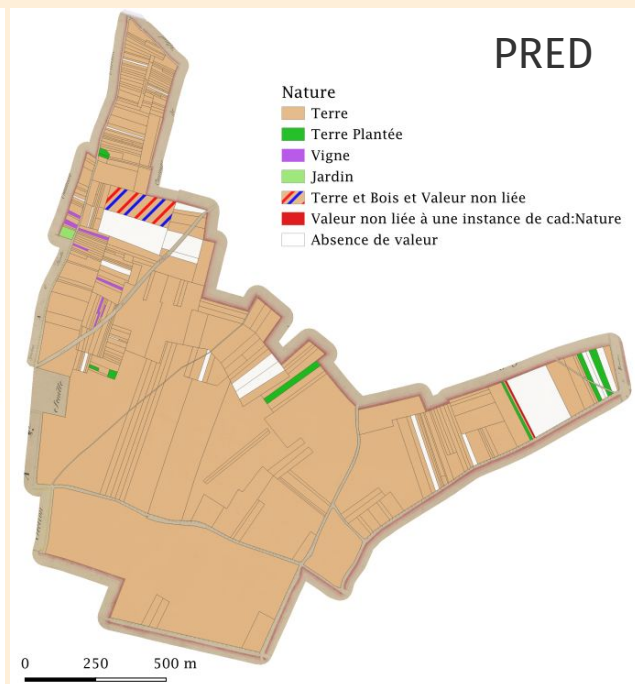
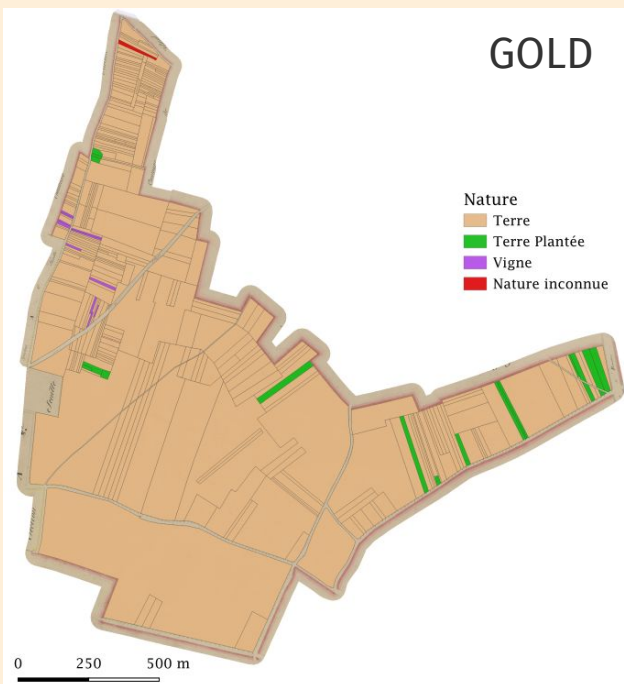
Exemple des création d'entités de type Contribuables



Evaluation du graphe de connaissances

Questions de compétences

Quelles sont les natures des parcelles de la section B de L'Hay en 1841 ?



Conclusion

Preuve de concept d'une chaîne de traitement : des tables anciennes numérisées à un graphe de connaissances

- Utilisation d'outils d'extraction d'informations existants affinés avec des données du domaine
- Processus d'interprétation sémantique intégré à toutes les étapes
- Création de triplets selon une ontologie de domaine
- Préserve le lien à la source (traçabilité des données)

Perspectives

- Étendre la zone d'étude (et donc le graphe de référence)
- Développer les jeux de données d'entraînement et d'évaluation
- Appliquer la méthode à des tables plus complexes : les matrices des contribuables (registres de mutations du cadastre)

Article avec annexes, datasets, modèles, programmes :
https://github.com/solenn-tl/land_registry_tables_processing_TPD_L_2025



Références

M. **Boillet**, S. Tarride, Y. Schneider, B. Abadie, L. Kesztenbaum, and C. Kermorvant. The Socface Project: Large-Scale Collection, Processing, and Analysis of a Century of French Censuses. In ICDAR, pages 57–73, 2024.

M. **Cremaschi**, B. Spahiu, M. Palmonari, E. Jimenez-Ruiz. “Survey on Semantic Interpretation of Tabular Data: Challenges and Directions”, 2024. arXiv :2411.11891.

T. **Constum**, N. Kempf, T. Paquet, P. Tranouez, C. Chatelain, S. Brée, and F. Merveille. Recognition and Information Extraction in Historical Handwritten Tables: Toward Understanding Early 20th Century Paris Census. In Document Analysis Systems, pages 143–157, 2022.

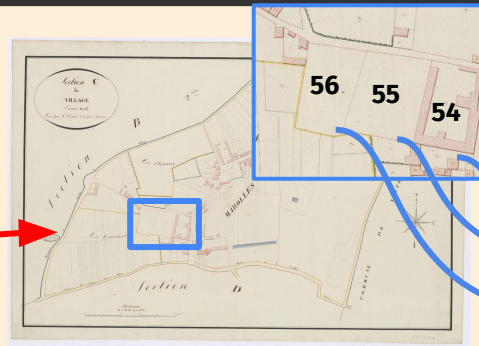
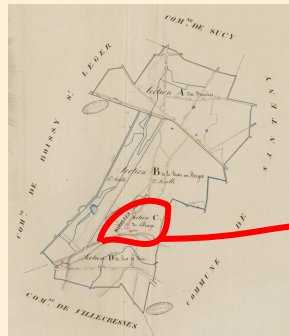
T. **Constum**. DANIEL: a fast document attention network for information extraction and labelling of handwritten documents. IJDAR, pages 1–23, 2025.

D. **Coquenot**, C. Chatelain, T. Paquet. DAN: a Segmentation-free Document Attention Network for Handwritten Document Recognition”. IEEE Transactions on Pattern Analysis and Machine Intelligence, inPress, <10.1109/tpami.2023.3235826>.

J. **Liu**, Y. Chabot, R. Troncy, V.-P. Huynh, T. Labbé, and P. Monnin. “From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods.” J. Web Semant., 76: 100761, 2023.

Présentation du corpus

Ex : Marolle-en-Brie



CANTON, TRIAGES, ou Lieux dits.	NUMÉROS DU PLAN provisoire ou définitif.	NOMS, PROFESSIONS, DEMEURES des Propriétaires ou Usufruitiers.	NATURES des Propriétés.	CONTENANCE par Propriété.	CONTENANCE par nature de culture.	
+	56	Berthier, alar. M. Prins de Napoléon	Maison et Cour	30 45		0. M. 2. 1/2
+	55	Jour	Jour	37 60	37 60	2 ^e
+	54	Jour	Vigne, etc.	29 00	29 00	1 ^e

Tableau d'assemblage

Plan parcellaire (1810)

Etat de sections (1812)

Etat
initial

Mutations

Matrices cadastrales

1812-1822

ARTICLES.	NOMS, PRÉNOMS, PROFESSIONS ET DEMEURES des Propriétaires et Usufruitiers.	INDICATION				NOMBRE D'ARPENS, PERCHES ET MÈTRES,		CLASSES.	REVENU		RENVIS pour les Mutations.	
		de la section.	du N° de la section.	DU CANTON, triage ou lieu dit.	DE LA NATURE de la propriété.	par article DE SECTION.	par article DE MATRICE.		de chaque article DE SECTION.	de chaque article DE MATRICE.	tiré de l'article.	porté à l'article.
3	Berthier, Alexandre Clerc de Napoléon à Gros Boule	B	90	heffard	Gene	arp. p. m. 2 06 99	arp. p. m. 2 06 99		fr. c. 98 99	fr. e. 98 99		74
						0 14 77	0 14 77	1	12 77			45
						0 30 16	0 30 16	2	19 60			
						0 37 60	0 37 60	2	20 70			
						0 29 00	0 29 00	1	24 67			

1822-1914

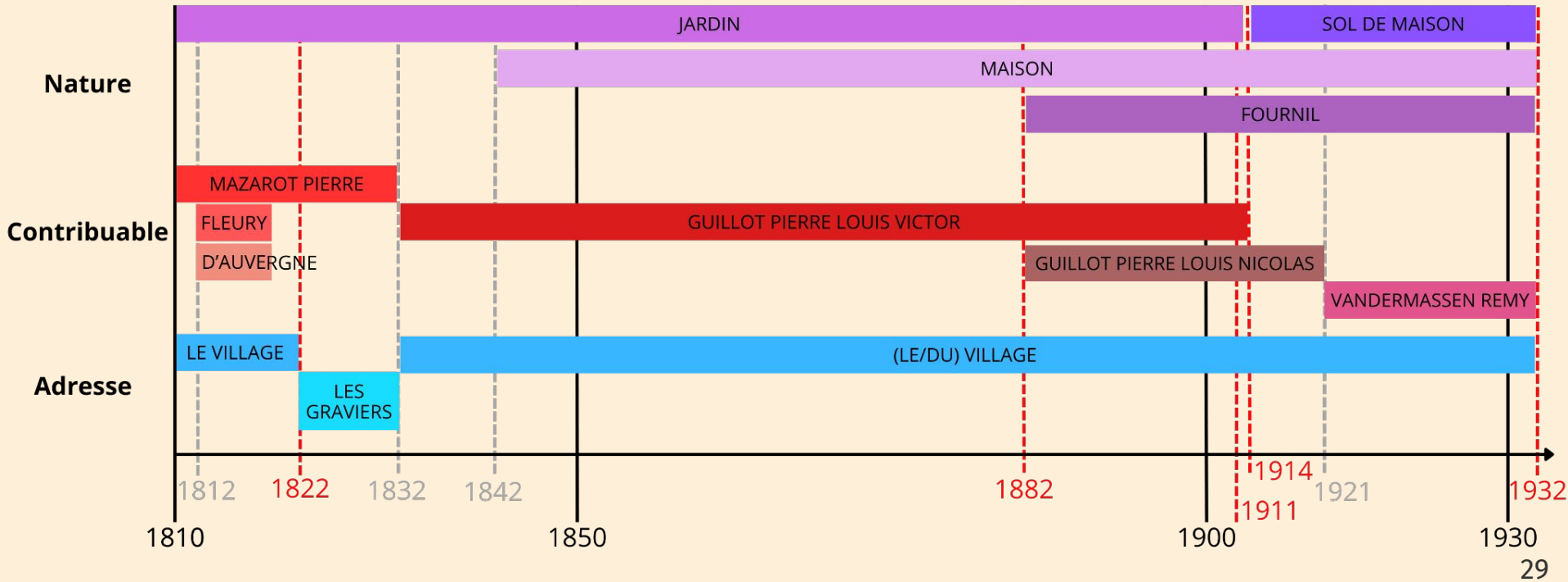
1882-1911

1914-1932

1911-1932

Reconstitution de l'historique d'une parcelle

Parcelle C-191, Marolles-en-Brie



Caractéristiques des tables cadastrales

Variations de structures dans un même corpus : matrices

Contribuable.s

Section

Parcelle

Adresse

Nature

Dates de validité

Tableau des matrices cadastrales (1810-1932) de Marolles-en-Brie. Le tableau est divisé en colonnes : NOMS, PRÉNOMS, PROFESSIONS ET DIMENSIONS; INDICATION; CONTRIBUTION PROPRETÉ; REVENU; and FOLIO. Les données sont inscrites à la main.

Tableau des matrices cadastrales (1810-1932) de Marolles-en-Brie. Le tableau est divisé en colonnes : NOMS, PRÉNOMS, PROFESSIONS ET DIMENSIONS; INDICATION; CONTRIBUTION PROPRETÉ; REVENU; and FOLIO. Les données sont inscrites à la main.

Tableau des matrices cadastrales (1810-1932) de Marolles-en-Brie. Le tableau est divisé en colonnes : NOMS, PRÉNOMS, PROFESSIONS ET DIMENSIONS; INDICATION; CONTRIBUTION PROPRETÉ; REVENU; and FOLIO. Les données sont inscrites à la main.

Interprétation sémantique de tables historiques

Evaluation intermédiaire (1) : extraction d'informations

- Reconnaissance des cellules

Reconnaissance des cellules du sous-ensemble de test

Colonne	Prédictions	Attendus	Matches	Précision	Rappel	F1
Numéro de parcelle	487	467	341	70.02	73.02	71.49
Nature	490	468	335	68.37	71.58	69.94
Numéro d'index du contribuable	384	339	251	65.36	74.04	69.43
Contribuable	485	466	301	62.06	64.59	63.30
Adresse	328	310	178	54.27	57.42	55.80
Ancien.s numéro.s	95	132	49	51.58	37.12	43.17
Ancienne.s nature.s	93	109	41	44.09	37.61	40.59
Total	2362	2291	1496	63.34	65.30	64.30

Cellules régulières

Cellules complexes

Cellules sous
représentées +
complexes

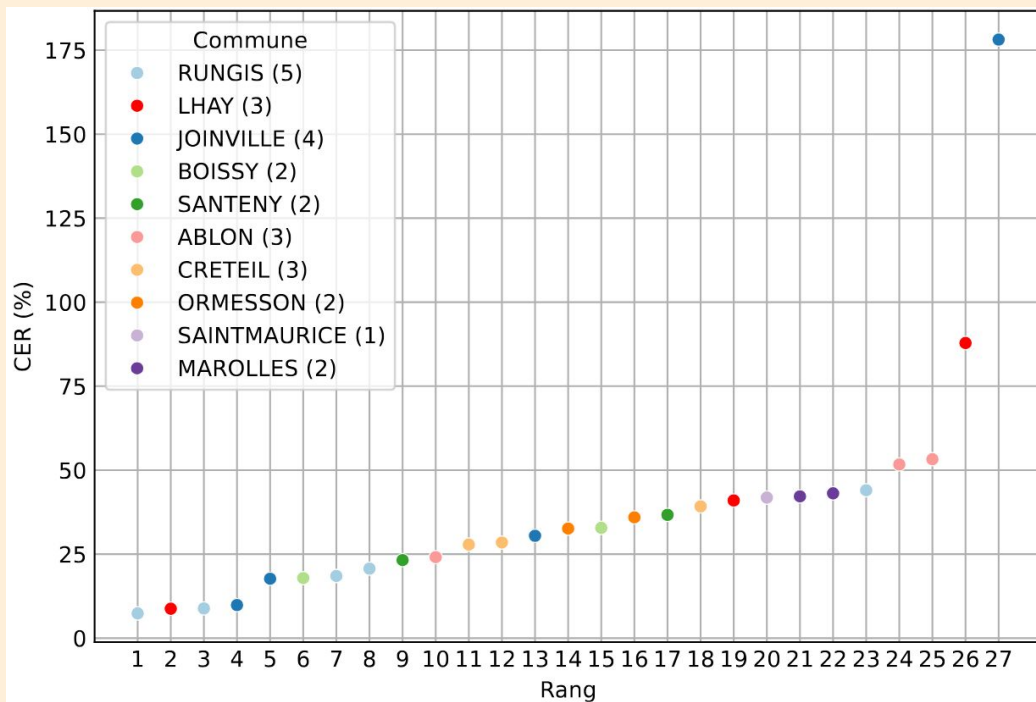
Matching des cellules (prédites et attendues) à l'aide de la bibliothèque Nerval []

Interprétation sémantique de tables historiques

Evaluation intermédiaire (1) : extraction d'informations

- CER moyen : 35.3% | min : 7.4% | max : 178.15% (= contenus indésirables)

Distribution du CER des pages du sous-ensemble de test



Détection des lignes

[illegible]

2ème meilleure page (CER), testset

[illegible]

Pire page (CER), testset