

Extraction automatique d'informations dans les annuaires commerciaux parisiens

Nathalie Abadie ⁽¹⁾, Edwin Carlinet ⁽²⁾, Joseph Chazalon ⁽²⁾, Bertrand Duménieu ⁽³⁾.

Séminaire SoDUCo - BnF, 10 novembre 2022



(1)



(2)



(3)



SODUCO ANR-18-CE38-0013

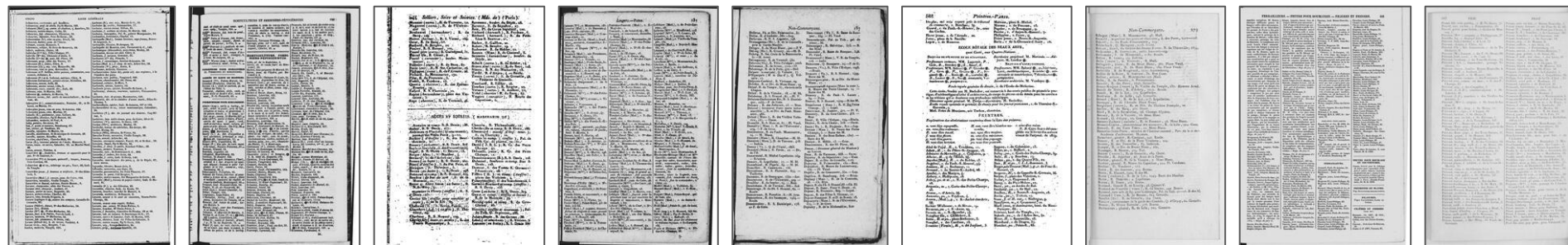
Introduction

Un corpus de **141** annuaires numérisés :

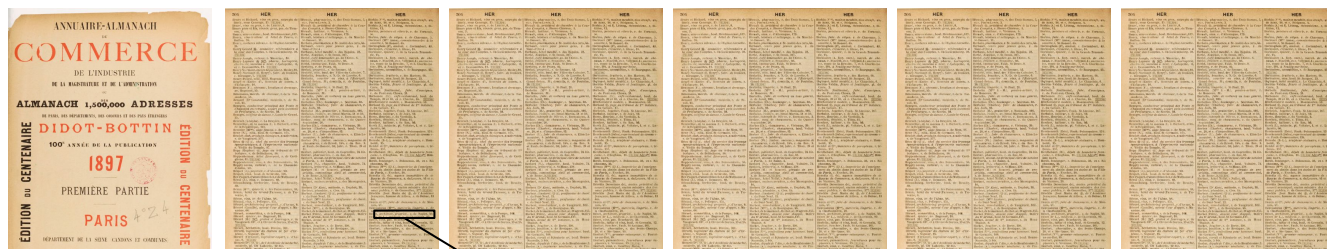
- Plusieurs centaines de milliers de pages,
- Plusieurs millions d'**entrées**, comportant le nom des commerces, leur type d'activité et leur localisation.

⇒ Une immense **source de données** pour le suivi individuel des commerces au cours du 19^e siècle...

... qu'il faut **structurer** pour permettre d'automatiser les analyses et les traitements !



Des documents numérisés aux données structurées



Pages d'annuaires numérisées

Identification des entrées

Entrée d'annuaire

Lemonnyer, plomberie, r. de Bondy, 86, et
r. Bouchardon, 1.

Transcription du texte

Texte numérique de l'entrée

Lemonnyer, plomberie, r. de Bondy, 86, et r. Bouchardon, 1.

Annotation du texte

Entités nommées

PERSONNE ACTIVITÉ RUE NUMÉRO RUE NUMÉRO
Lemonnyer, plomberie, r. de Bondy, 86, et r. Bouchardon, 1.

Structuration des entrées

Base de données intégrée



Construire un outil pour les historiens

Explorer



Visualiser



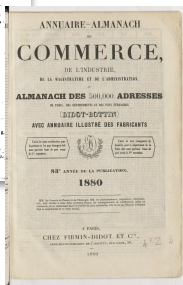
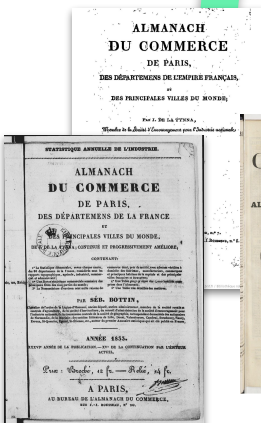
Annoter



Analyser



API IIIF
E.g {BNF Gallica}



Directory Viewer | DocID: 18114.pdf | 14 791 | Export

Entée | 🔍 | 📄 | 📊 | 📌

Éditer en ligne

MÉCANICIENS.

broches de cuisinières, et tout ce qui concerne cette partie ACT | [Temple-loc](#) | [09 comen](#)

Caron (P.), res ingénieur-mécanicien ACT, ci-devant **Faub-St-Martin loc**, [147 comen](#), actuellement, m éme rue loc, [188 comen](#) : fabrique hydro-extracteur, dit séchoir mécanique, sans feu ni pression, avec mouvement progressif; métiers à couvrir la guizune, carissane, laine, gomme, trait d'or et d'argent, milanaise, etc. Ces métiers remplaçant les roulets de Lyon, auxquels on peut adapter ce nouveau système. Métiers continus, à doubler et à retordre, à l'usage des passe-mentiers, retordeurs, fabricants de fils retors et autres ACT : engrenages élastiques ACT : résistants à une vitesse de cinq mille tours par minute, n'ayant pas besoin d'ACT étre grasse ACT, et ne faisant pas de bruit; **SENETIERE** d'ACT: invention sans garantie du gouvernement, en— répond à tout tous travaux, sur idée, en mécaniques, (1849.

Carbonnier res, **Piètre-St-Jacques loc**

O

Carbottaux res, **Neuve-Copeland loc**, [0 comen](#)

Caillon res, appareils et machines à vapeur ACT, 1844, O 1849 ; **Nve: Poinpoint loc**, [0 comen](#)

Caillon res, fait les outils à découper et tout ce qui concerne son état ACT | **Cra-Villiers loc**, [28 comen](#), passage de Rome loc, [02 comen](#)

Cart (A.), res boul. de la Conscience-pe loc, [02 comen](#)

Cave: Charles res, serrurier-mécan, possesseur de sonnettes, cordons de portes-cochères, se charge des tra-vaux de la campagne et fait les réparation ACT | **Michodette loc**, [0 comen](#), près la place Gallien et le **bouc loc** loc

Cave res, **0 comen**, **M.H. comen** 1827, O 1834 et 1844, O S.E. 1834, établis-sements à Paris ACT : **Fb-St-Denis loc**, [02 comen](#) 2, à la **Chapelle loc** : **St-Denis**, rues Cave et d'ACT: Alger, et à Clitichy, en amont du pont d'ACT: **Mezlog** rivière, pour la fabrication des machines à va-peur de toutes forces, destinées aux manufactures, aux usines à fer, aux mines, aux bateaux et aux navires, etc, et pour la construction des locomotives, bateaux en fer, grosse chaudière en cu, sites, pièces de forges de toutes di-mensions, tuyaux étirés au banc et bagues en acier sans soudure ACT :

Conseaux res, **Monimontant loc**, [02 comen](#)

Export structuré
{JSON}
TEI <XML />

Assister le travail humain avec des outils automatiques

Explorer



Visualiser



Annoter



Analyser



API IIIF

E.g {BNF Gallica

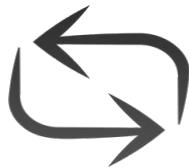
Export structuré

JSON / TEI XML

Interopérabilité

Des micro-services intégrés à l'annotation semi-automatique

Traitement
automatique



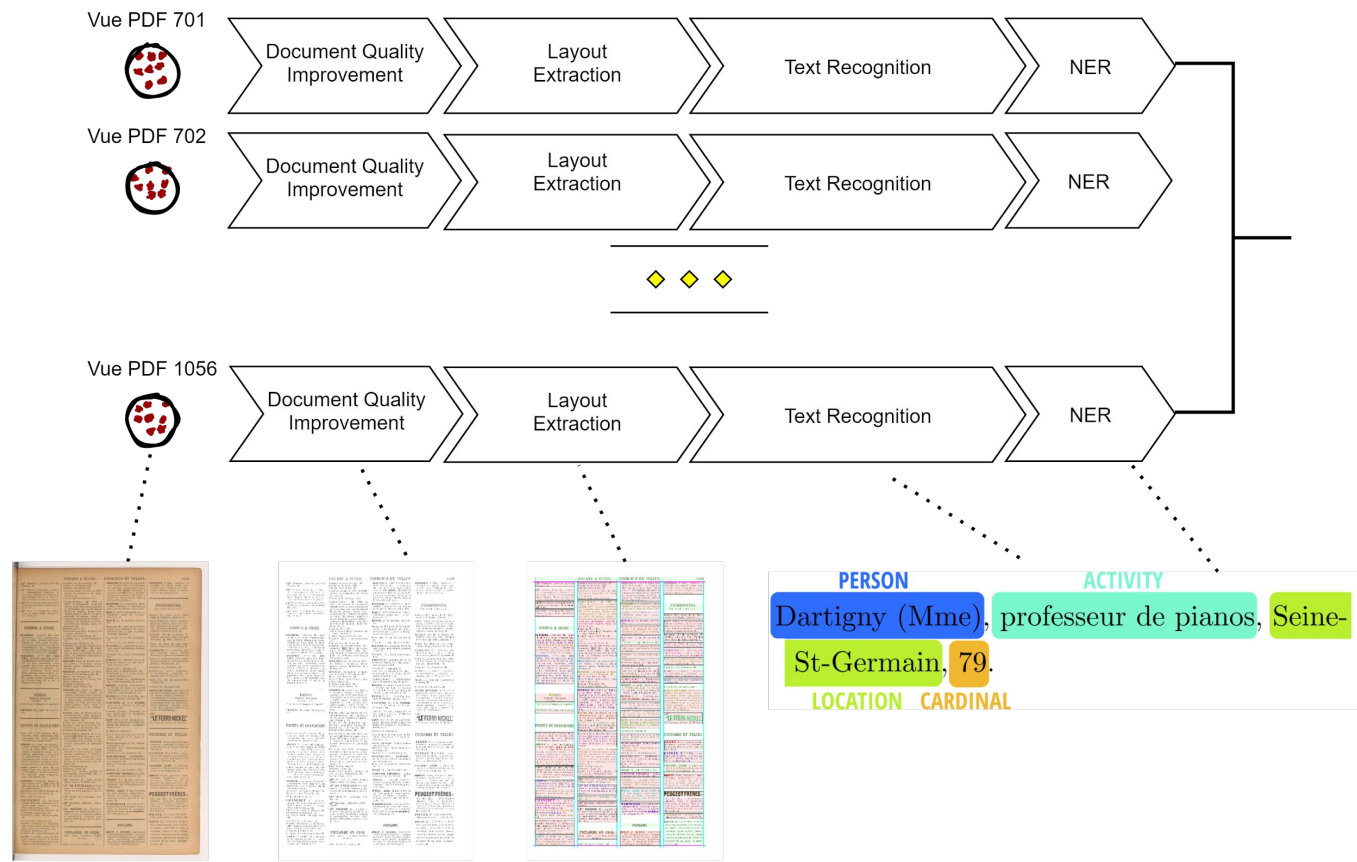
Correction & annotation
manuelle



D'autres avantages:

- App web (distante ou locale) facile à déployer
- Des services de traitements externalisés (on peut changer et combiner des méthodes)

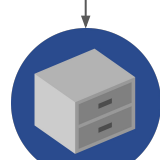
Chaîne de traitement automatique



Une reconnaissance perturbée par différentes imperfections



Archivage



Traces &
Traversées d'encre

Numérisation

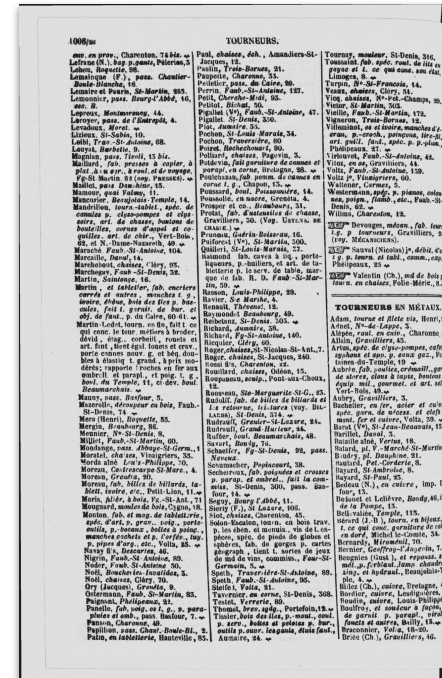
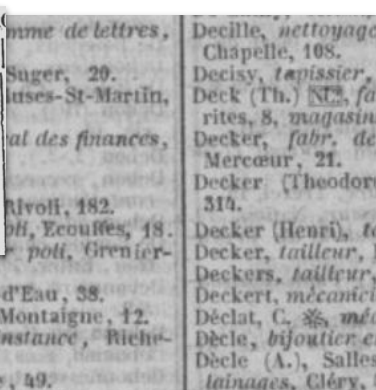


Compression



JPEG

JPEG



Inclinaison &
Rotation & Courbure

Perte des niveaux de
gris &
Artefacts de
compression



Amélioration de la qualité des images

Les OCRs fonctionnent mieux sur des images "propres" :

- ✓ Amélioration de contraste / suppression de fond
- ✓ Atténuation du bruit
- ✓ Correction géométrique (Rotation / Shear / Courbure)
- ✓ (Rapide)

☞ Bronner, ferreur, rue des Carmes, 15.
☞ Jolivet (François), formes et embouteillages chimiques et touristes, Italie et magnésie, boul. Italie, 59 et 61, Neuville-aux-Plâtres (Aisne).
☞ Trévo Gesteauille, machines à faire les formes, r. Popincourt, 11 et 13.
☞ Hoer P., négal bis, sucres, limes et cages pour formiers, rue Valla, 7.

FORMES A SUCRE.

GEIGNERY (Adrien) fils aîné 104 rue des, sucres, maison fondée en 1810, ateliers spéciaux pour la fabrication de formes à sucre, cristaillères, etc., réparation et entretien de formes en service, quai de Seine, 37. [52] 127
Gauldon Collette, ☞ 1853 et 1867, 2 MÉDAILLES D'ARGENT EXPOS. CIVIL 1875, formes à sucre et cristaillères sur cartes dessinées jusqu'à 150 fr. la pièce (fabrication, place du Théâtre français, 5).
Bougeois (B.), méd. ☞ Paris 1889, Gley, A. et Aboult, 87, Lyon, Hôtel de Ville, 5; fab. à Sully Saint-Jean (Somme).
Burkhardt (J.) et C^o, Erbkoupt, 22.
Coutier (J.), rue Parthenon, 11.
DELVIRE (P.), fondais en gros, spec. pour marchands forains, Udry, 25.
Fontaine (E.), boul. Sébastopol, 157.
MILLON (J.), méd. et neveu, fab. de fautes, cristaillères, moules, gâteaux, moules, montres, etc., Rouilly, 30.
MINGET, peripatant et montures, N.-D.-de-Nazareth, 61.
Moix, cité Brevin, 18.
Moutet, boul. de la Ville, 218.
Nache, Aboult, 19.
Papillon, Legrand, 50.
Pechier, quai Vainy, 29.
PIERSON (G.), fab. de fautes, cristaillères, etc., fondeur, Entrepôt d'articles anglais pour

FORMES A SUCRE.

Gosselin, rue des Arènes, 30.
Glatry, r. St-Denis, 222.
Godard, Gravilliers, 18.
Greffelin (Léon) 1878 ☞ 1889, 1890, 1891, 1892, 1893, 1894, 1895, 1896, 1897, 1898, 1899, 1900, 1901, 1902, 1903, 1904, 1905, 1906, 1907, 1908, 1909, 1910, 1911, 1912, 1913, 1914, 1915, 1916, 1917, 1918, 1919, 1920, 1921, 1922, 1923, 1924, 1925, 1926, 1927, 1928, 1929, 1930, 1931, 1932, 1933, 1934, 1935, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 2680, 2681, 2682, 2683, 2684, 2685, 2686, 2687, 2688, 2689, 2690, 2691, 2692, 2693, 2694, 2695, 2696, 2697, 2698, 2699, 2700, 2701, 2702, 2703, 2704, 2705, 2706, 2707, 2708, 2709, 2710, 2711, 2712, 2713, 2714, 2715, 2716, 2717, 2718, 2719, 2720, 2721, 2722, 2723, 2724, 2725, 2726, 2727, 2728, 2729, 2730, 2731, 2732, 2733, 2734, 2735, 2736, 2737, 2738, 2739, 2740, 2741, 2742, 2743, 2744, 2745, 2746, 2747, 2748, 2749, 2750, 2751, 2752, 2753, 2754, 2755, 2756, 2757, 2758, 2759, 2760, 2761, 2762, 2763, 2764, 2765, 2766, 2767, 2768, 2769, 2770, 2771, 2772, 2773, 2774, 2775, 2776, 2777, 2778, 2779, 2780, 2781, 2782, 2783, 2784, 2785, 2786, 2787, 2788, 2789, 2790, 2791, 2792, 2793, 2794, 2795, 2796, 2797, 2798, 2799, 2800, 2801, 2802, 2803, 2804, 2805, 2806, 2807, 2808, 2809, 2810, 2811, 2812, 2813, 2814, 2815, 2816, 2817, 2818, 2819, 2820, 2821, 2822, 2823, 2824, 2825, 2826, 2827, 2828, 2829, 2830, 2831, 2832, 2833, 2834, 2835, 2836, 2837, 2838, 2839, 2840, 2841, 2842, 2843, 2844, 2845, 2846, 2847, 2848, 2849, 2850, 2851, 2852, 2853, 2854, 2855, 2856, 2857, 2858, 2859, 2860, 2861, 2862, 2863, 2864, 2865, 2866, 2867, 2868, 2869, 2870, 2871, 2872, 2873, 2874, 2875, 2876, 2877, 2878, 2879, 2880, 2881, 2882, 2883, 2884, 2885, 2886, 2887, 2888, 2889, 2890, 2891, 2892, 2893, 2894, 2895, 2896, 2897, 2898, 2899, 2900, 2901, 2902, 2903, 2904, 2905, 2906, 2907, 2908, 2909, 2910, 2911, 2912, 2913, 2914, 2915, 2916, 2917, 2918, 2919, 2920, 2921, 2922, 2923, 2924, 2925, 2926, 2927, 2928, 2929, 2930, 2931, 2932, 2933, 2934, 2935, 2936, 2937, 2938, 2939, 2940, 2941, 2942, 2943, 2944, 2945, 2946, 2947, 2948, 2949, 2950, 2951, 2952, 2953, 2954, 2955, 2956, 2957, 2958, 2959, 2960, 2961, 2962, 2963, 2964, 2965, 2966, 2967, 2968, 2969, 2970, 2971, 2972, 2973, 2974, 2975, 2976, 2977, 2978, 2979, 2980, 2981, 2982, 2983, 2984, 2985, 2986, 2987, 2988, 2989, 2990, 2991, 2992, 2993, 2994, 2995, 2996, 2997, 2998, 2999, 3000, 3001, 3002, 3003, 3004, 3005, 3006, 3007, 3008, 3009, 3010, 3011, 3012, 3013, 3014, 3015, 3016, 3017, 3018, 3019, 3020, 3021, 3022, 3023, 3024, 3025, 3026, 3027, 3028, 3029, 3030, 3031, 3032, 3033, 3034, 3035, 3036, 3037, 3038, 3039, 3040, 3041, 3042, 3043, 3044, 3045, 3046, 3047, 3048, 3049, 3050, 3051, 3052, 3053, 3054, 3055, 3056, 3057, 3058, 3059, 3060, 3061, 3062, 3063, 3064, 3065, 3066, 3067, 3068, 3069, 3070, 3071, 3072, 3073, 3074, 3075, 3076, 3077, 3078, 3079, 3080, 3081, 3082, 3083, 3084, 3085, 3086, 3087, 3088, 3089, 3090, 3091, 3092, 3093, 3094, 3095, 3096, 3097, 3098, 3099, 3100, 3101, 3102, 3103, 3104, 3105, 3106, 3107, 3108, 3109, 3110, 3111, 3112, 3113, 3114, 3115, 3116, 3117, 3118, 3119, 3120, 3121, 3122, 3123, 3124, 3125, 3126, 3127, 3128, 3129, 3130, 3131, 3132, 3133, 3134, 3135, 3136, 3137, 3138, 3139, 3140, 3141, 3142, 3143, 3144, 3145, 3146, 3147, 3148, 3149, 3150, 3151, 3152, 3153, 3154, 3155, 3156, 3157, 3158, 3159, 3160, 3161, 3162, 3163, 3164, 3165, 3166, 3167, 3168, 3169, 3170, 3171, 3172, 3173, 3174, 3175, 3176, 3177, 3178, 3179, 3180, 3181, 3182, 3183, 3184, 3185, 3186, 3187, 3188, 3189, 3190, 3191, 3192, 3193, 3194, 3195, 3196, 3197, 3198, 3199, 3200, 3201, 3202, 3203, 3204, 3205, 3206, 3207, 3208, 3209, 3210, 3211, 3212, 3213, 3214, 3215, 3216, 3217, 3218, 3219, 3220, 3221, 3222, 3223, 3224, 3225, 3226, 3227, 3228, 3229, 3230, 3231, 3232, 3233, 3234, 3235, 3236, 3237, 3238, 3239, 3240, 3241, 3242, 3243, 3244, 3245, 3246, 3247, 3248, 3249, 3250, 3251, 3252, 3253, 3254, 3255, 3256, 3257, 3258, 3259, 3260, 3261, 3262, 3263, 3264, 3265, 3266, 3267, 3268, 3269, 3270, 3271, 3272, 3273, 3274, 3275, 3276, 3277, 3278, 3279, 3280, 3281, 3282, 3283, 3284, 3285, 3286, 3287, 3288, 3289, 3290, 3291, 3292, 3293, 3294, 3295, 3296, 3297, 3298, 3299, 3300, 3301, 3302, 3303, 3304, 3305, 3306, 3307, 3308, 3309, 3310, 3311, 3312, 3313, 3314, 3315, 3316, 3317, 3318, 3319, 3320, 3321, 3322, 3323, 3324, 3325, 3326, 3327, 3328, 3329, 3330, 3331, 3332, 3333, 3334, 3335, 3336, 3337, 3338, 3339, 3340, 3341, 3342, 3343, 3344, 3345, 3346, 3347, 3348, 3349, 3350, 3351, 3352, 3353, 3354, 3355, 3356, 3357, 3358, 3359, 3360, 3361, 3362, 3363, 3364, 3365, 3366, 3367, 3368, 3369, 3370, 3371, 3372, 3373, 3374, 3375, 3376, 3377, 3378, 3379, 3380, 3381, 3382, 3383, 3384, 3385, 3386, 3387, 3388, 3389, 3390, 3391, 3392, 3393, 3394, 3395, 3396, 3397, 3398, 3399, 3400, 3401, 3402, 3403, 3404, 3405, 3406, 3407, 3408, 3409, 3410, 3411, 3412, 3413, 3414, 3415, 3416, 3417, 3418, 3419, 3420, 3421, 3422, 3423, 3424, 3425, 3426, 3427, 3428, 3429, 3430, 3431, 3432, 3433, 3434, 3435, 3436, 3437, 3438, 3439, 3440, 3441, 3442, 3443, 3444, 3445, 3446, 3447, 3448, 3449, 3450, 3451, 3452, 3453, 3454, 3455, 3456, 3457, 3458, 3459, 3460, 3461, 3462, 3463, 3464, 3465, 3466, 3467, 3468, 3469, 3470, 3471, 3472, 3473, 3474, 3475, 3476, 3477, 3478, 3479, 3480, 3481, 3482, 3483, 3484, 3485, 3486, 3487, 3488, 3489, 3490, 3491, 3492, 3493, 3494, 3495, 3496, 3497, 3498, 3499, 3500, 3501, 3502, 3503, 3504, 3505, 3506, 3507, 3508, 3509, 3510, 3511, 3512, 3513, 3514, 3515, 3516, 3517, 3518, 3519, 3520, 3521, 3522, 3523, 3524, 3525, 3526, 3527, 3528, 3529, 3530, 3531, 3532, 3533, 3534, 3535, 3536, 3537, 3538, 3539, 3540, 3541, 3542, 3543, 3544, 3545, 3546, 3547, 3548, 3549, 3550, 3551, 3552, 3553, 3554, 3555, 3556, 3557, 3558, 3559, 3560, 3561, 3562, 3563, 3564, 3565, 3566, 3567, 3568, 3569, 3570, 3571, 3572, 3573, 3574, 3575, 3576, 3577, 3578, 3579, 3580, 3581, 3582, 3583, 3584, 3585, 3586, 3587, 3588, 3589, 3590, 3591, 3592, 3593, 3594, 3595, 3596, 3597, 3598, 3599, 3600, 3601, 3602, 3603, 3604, 3605, 3606, 3607, 3608, 3609, 3610, 3611, 3612, 3613, 3614, 3615, 3616, 3617, 3618, 3619, 3620, 3621, 3622, 3623, 3624, 3625, 3626, 3627, 3628, 3629, 3630, 3631, 3632, 3633, 3634, 3635, 3636, 3637, 3638, 3639, 3640, 3641, 3642, 3643, 3644, 3645, 3646, 3647, 3648, 3649, 3650, 3651, 3652, 3653, 3654, 3655, 3656, 3657, 3658, 3659, 3660, 3661, 3662, 3663, 3664, 3665, 3666, 3667, 3668, 3669, 3670, 3671, 3672, 3673, 3674, 3675, 3676, 3677, 3678, 3679, 3680, 3681, 3682, 3683, 3684, 3685, 3686, 3687, 3688, 3689, 3690, 3691, 3692, 3693, 3694, 3695, 3696, 3697, 3698, 3699, 3700, 3701, 3702, 3703, 3704, 3705, 3706, 3707, 3708, 3709, 3710, 3711, 3712, 3713, 3714, 3715, 3716, 3717, 3718, 3719, 3720, 3721, 3722, 3723, 3724, 3725, 3726, 3727, 3728, 3729, 3730, 3731, 3732, 3733, 3734, 3735, 3736, 3737, 3738, 3739, 3740, 3741, 3742, 3743, 3744, 3745, 3746, 3747, 3748, 3749, 3750, 3751, 3752, 3753, 3754, 3755, 3756, 3757, 3758, 3759, 3760, 3761, 3762, 3763, 3764, 3765, 3766, 3767, 3768, 3769, 3770, 3771, 3772, 3773, 3774, 3775, 3

Reconnaissance du texte (OCR)

Mettereau, prop., quai d'Anjou, 7.	Meurgey, épicier-herboriste, Dragon, 33.
Mettemberg, élig., méd., St-Thomas-d'Enf., 5.	Meurice, Chaussée-d'Antin, 3.
Metz (de), rentier, St-Guillaume, 30.	Meurice (Eug.), tapissier, Vivienne, 12.
Metzinger, avocat, Rameau, 6.	Meurillon, marbrier-sculpteur, butte Mont-
Metzmacher, peint. sur émaux, St-Martin, 124.	Parnasse, 15.

Analyse de la mise en page

Mettereau, prop., quai d'Anjou, 7.	Meurgey, épicier-herboriste, Dragon, 33.
Mettemberg, élig., méd., St-Thomas-d'Enf., 5.	Meurice, Chaussée-d'Antin, 3.
Metz (de), rentier, St-Guillaume, 30.	Meurice (Eug.), tapissier, Vivienne, 12.
Metzinger, avocat, Rameau, 6.	Meurillon, marbrier-sculpteur, butte Mont-
Metzmacher, peint. sur émaux, St-Martin, 124.	Parnasse, 15.

OCR

Mettereau, prop., quai d'Anjou, 7.\nMettemberg, élig., méd., St-Thomas-d'Enf., 5.\nMetz (de), rentier, St-Guillaume, 30.\nMetzinger, avocat, Rameau, 6.\nMetzmacher, peint. sur émaux, St-Martin, 124.\nMeurgey, épicier-herboriste, Dragon, 33.\nMeurice, Chaussée-d'Antin, 3.\nMeurice (Eug.), tapissier, Vivienne, 12.\nMeurillon, marbrier-sculpteur, butte Mont-\nParnasse, 15.\n

La modularité de l'app nous a permis de tester plusieurs OCR rapidement (3 sont déjà API-fiés) :

- Kraken
- Tesseract ★
- Pero OCR ★★

Reconnaître et structurer les informations des entrées

Composition des entrées:

- **Nom de la (des) personne(s) qui exerce(nt) une activité ou tien(nen)t un commerce,**
- Titre honorifique ou professionnel,
- Type d'activité ou de commerce,
- Type de local *,
- **Nom de voie ***,
- **Numéro de voie ***,
- Nom de section *.

L'ordre des informations peut varier selon le type d'index ou l'éditeur.

Seules les informations en gras sont systématiquement présentes.

Les informations suivies d'une * peuvent apparaître plusieurs fois par entrée.

Les étapes précédentes peuvent avoir produit un texte bruité : entrées mal délimitées, caractères mal reconnus, manquants, etc.



Une approche de reconnaissance des différents types d'informations (Named Entity Recognition) à base de règles serait coûteuse à développer et aurait peu de chances de traiter tous les cas de figure avec succès.

Reconnaissance d'entités nommées à base de réseaux de neurones profonds

1. Quels sont les modèles **récents disponibles** pour réaliser cette tâche de NER ?
2. Ces modèles de NER ont-ils **besoin de beaucoup de données d'entraînement** pour bien fonctionner sur les annuaires ?
3. Ces modèles de NER peuvent-ils produire de **bons résultats malgré le bruit OCR** ?
4. Peut-on **améliorer la résistance de ces modèles de NER au bruit OCR** ?

N. Abadie, E. Carlinet, Joseph Chazalon, B. Duménieu. A Benchmark of Named Entity Recognition Approaches in Historical Documents: Application to 19th Century French Directories. Document Analysis Systems. DAS 2022., May 2022, La Rochelle, France.

<https://github.com/soduco/paper-ner-bench-das22>

Modèles NER évalués

Nécessaire pour traiter des entités non standards

	Entraînement original <i>modèle sur étagère</i>		Adaptation de domaine <i>entraînement spécifique au corpus</i>	
SpaCy CNN:	Pré-entraînement non supervisé: <u>deep-sequoia</u>	Entraînement supervisé pour le NER: <u>wikiner-fr</u>	Pré-entraînement	Entraînement supervisé pour le NER: <u>FTD labelled</u>
CamemBERT:	Pré-entraînement non supervisé: <u>OSCAR</u>	Entraînement supervisé pour le NER: <u>wikiner-fr</u>	Pré-entraînement	Entraînement supervisé pour le NER: <u>FTD labelled</u>
CamemBERT <i>pré-entraîné</i> :	Pré-entraînement non supervisé: <u>OSCAR</u>	Entraînement supervisé pour le NER: <u>wikiner-fr</u>	Pré-entraînement non supervisé: <u>FTD unlabelled</u>	Entraînement supervisé pour le NER: <u>FTD labelled</u>

Les modèles testés sont disponibles sur nos [dépôts HuggingFace](#) et Zenodo ([10.5281/zenodo.6576008](#)).

Transcriptions Pero OCR non corrigées pour 845,000 entrées brutes (≈7000 p.)

Transcriptions et annotations manuelles pour 8765 entrées (78 p.)

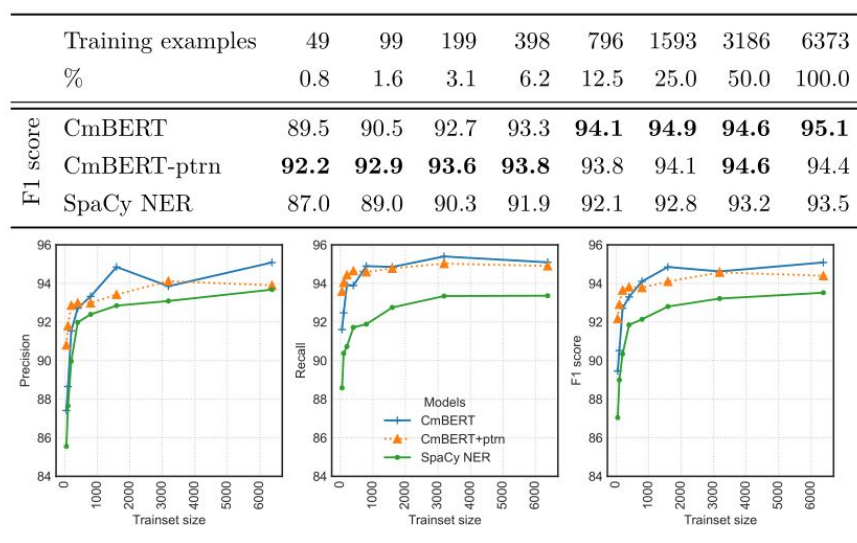
Quantité d'annotations et qualité des résultats

2. A-t-on besoin de beaucoup de données d'entraînement ?

Non ! On obtient de **bon résultats avec peu de données d'entraînement**, surtout si on réalise aussi un **pré-entraînement non supervisé**.

3. Les modèles de NER à base de réseaux de neurones profonds sont-ils adaptés pour traiter des annuaires hétérogènes ?

Oui ! Et les **modèles de type Transformer** obtiennent de très bon résultats.



F1-scores moyens obtenus pour 5 entraînements + tests

Stratégie d'entraînement pour réduire l'influence du bruit OCR sur les résultats

Modèles testés :

- CamemBERT
- CamemBERT pré-entraîné

Entrées :

- Extraits d'annuaires corrigés et annotés manuellement
- Extraits d'annuaires directement produits par OCR (Tesseract V4 et Pero)

Sorties — Prédictions :

- prédictions NER de camemBERT
Avec et sans pré-entraînement

Sorties — Référence :

- Extraits d'annuaires corrigés et annotés manuellement
- Ou annotations manuelles projetées sur les extraits d'annuaires directement produits par OCR (bruités).

Mesures : score F1

Jeux de données utilisés pour entraîner et tester CamemBERT (× 12 variantes)

Jeux de pré-entraînement	Jeux d'entraînement	Jeux de test
• <i>Aucun</i> • PERO (brut)	• Référence • PERO* (* annotations projetées)	• Référence • PERO* • Tesseract* (*annotations projetées)

Extrait du jeu de test de référence:

Annotations manuelles :

Dulay **PER**, chaudronnier **ACT**, r. du Pont- aux Choux **LOC**, 15 **CARDINAL**, 314.

Extrait du jeu de test avec annotations projetées sur du texte bruité :

Dulay **PER**, chandronnier **ACT**, +. du Pont-anx-Cars **LOC** Ge 7 **CARDINAL** Fe ÊR one

Stratégie d'entraînement pour réduire l'influence du bruit OCR sur les résultats

3 bis. Les modèles de NER à base de réseaux de neurones profonds sont-ils adaptés pour traiter des **transcriptions OCR bruitées** ?

Oui ! Mais les résultats sont moins bons que quand on a des textes propres.

4. Peut-on rendre les modèles de NER à base de réseaux de neurones profonds **plus robustes au bruit OCR** ?

Oui ! Il faut les pré-entraîner et les entraîner sur des textes bruités.

⇒ N'entraînez pas votre modèle sur du texte propre si vous voulez traiter des textes OCRisés !



F1-scores moyens obtenus pour 5 entraînements + tests

Démonstration

MÉCANICIENS.

697

Tous ronds et demi-ronds et carvés, et de toute sorte de moulure ; spécialité de fabrication de broches en acier et cuivre pour les peignes à tisser draps, voiles et ouvertures, toile métallique pour tout ce qui concerne le tissage, M. H. 1849, Amandiers-Popincourt, 19.

Auger (Yve), mécanicien, fabr. d'emporte-pièces pour festons de garnitures de robes et de mantelets, emporte-pièces pour cartonnage, semelles de souliers, etc., tient un assortiment de maillets, billets et plombs, pass. de la Trinité, 77.

Avoine-Bainée, serrurier-mécan., fab. de lits en f. r., 1839-14, 1819, Boulangers-St-Victor, 22.

Bijard, quai Jemmapes, 248.

Baptarosse père, tout ce qui a rapport à la fabrique d'indennes, Charenton, 58.

Barbot, Popincourt, 58.

Bardies, Trois-Bornes, 21.

Bardou, Descartes, 38.

Barquand, horloger-mécanicien, comprend toute espèce de fabrication nouvelle, quelle que soit la précision demandée, et construit toutes sortes de machines et outils pour Messieurs les inventeurs; vis cylindriques et coniques, découpage à façon, fendage de roues; pièces détachées fabriquées par procédés mécaniques pour toute espèce d'industrie, 1849, Ecoiffes, 20. au Mar., près le marché des Blancs-Manteaux. *

Baudat, constructeur de mécaniques à scier le bois. Le bois on

Becker, mécanicien-graveur, breveté d'invention (sans garantie du gouvernement), fabrique spéciale de presses à copier, presses à timbre sec et bumide, presses à cacheter, presses à percussion pour l'extraction des matières végétales et pour le satinage des papiers, composteurs pour papiers à lettres, pour raisons de commerce, griffes de toutes sortes, et généralement toute la gravure sur métaux, St-Denis, 380, passage Lemoine. *

Beerstecher (A.), Loureine, 88.
Berendoff (J.), M. H. 1844, Moulletard, 294, atelier de construction de machines à vapeur et autres, 1844, 1844, 1849.

Bergeron, 1849, Marais-du-Temple, 73.

Bernier aîné, outils établis, presses, 1844, Faubourg-St-Antoine, 89.

Bernier, Menilmontant, 90.

Berthet, Simon-le-Franc, 13.

Berton, Neuve-St-Denis, 12.

Bertrand, mécanicien, breveté (sans garantie du gouvernement), construit toutes espèces de machines et outils, spécialité de machines pour les fabricants de matières premières de chapellerie, machines à vapeur et tours pour les fabricants de chapeaux; tondeuses et souffleurs de poils de lièvres et de lapins, réparation et entretien de machines à vapeur; on trouve toujours chez lui des machines à vapeur toutes prêtes à être mises en place, Vieille-du-Temple, 58, anc. 72.

Boslav (Ch.) 1839-1849

province, à des prix très-moderés, St-Jacques, 261. *

Bosche aîné, ingénieur-mécanicien, inventeur de divers perfectionnements au métier Jacquard, 1850, Amandiers-Popincourt, 22, ci-devant St-Maur, 14. *

Boscher, Bondy, 70.

Bosquillon, 1823-27-34, constructeur de perçages accélérés et de mécaniques Jacquard parisiennes, Paradis-Poissonnière, 20.

Boucher, Saint-Pierre-Popincourt, 18.

Bouchon, dépôt de moulins à bras portatifs, Nve-St-Nicolas, 16.

Boulley (Ene), mécanicien, fab. de machines à couper les peaux de lapins, souffleurs de toutes dimensions marchant à bras et à la vapeur, tours de chapeliers, 1849, Francs-Bourgeois-Marais, 3, ci-devant Beaubourg, 59. *

Bouhon, appareil dit cale à léau, breveté sans garantie du gouvernement, destiné au soulagement des chevaux dans les montées, 1849, Société d'encouragement 1850, pl. Dauphine, 71. *

Bourdier, bandagiste-herniaire, mécanicien-orthopédiste, fournisseur de la Société protestante de prévoyance et de secours mutuels de Paris, fabrique spéciale de toutes espèces de ressorts et de bandages, béquilles, appareils orthopédiques, ceintures hypogastriques, bas lacés en tous genres, suspensoirs, objets d'allaitement, et tous articles en gomme élastique, Cadran 5.

gnier, mécanique nouvelle, presse à copier les lettres, perfectionnée 1849, Saint-Jacques, 169, près le Panthéon. *

Brocard (Léon), machines à papier, St-Gilles, 12.

Burckel, fab. d'écritures, fuseaux et broches, Nve-de-Lappe, 3.

Busser, fab. de presses lithographiques, machines à repaier, fait les réparations à des prix modérés, Lamartine, 18-20, ci-devant Coquenard. *

Butt, Buisson-St-Louis, 16.

Cabanis et Cie, association pour la construction des machines à élever les eaux, pompes domestiques, moteurs hydrauliques de toute nature, machines soufflantes pour les forges, machines à eaux de Sellz, presses à découper, outillage et réparation de machines, Vinai-griers, 32.

Zahouet, breveté (sans garantie du gouvern.) pour les porte-moules à bougies emboutis d'un seul morceau, et les moules montés sans aucune soudure, et pouvant les monter et les démonter soi-même; fabricant de mandrins, fondeur de moules de chandelles, bougies, cierges, spécialité, 1849, Marché-aux-Vaux, 4.

Calamel, Dupetit-Thouars, 16.

Calard (T.-F.), fab. de feuilles métalliques percées en fer, cuivre et zinc; tôle en râpe pour le nettoyage des blés, cribles à main pour remplacer la peau. — Atelier de construction de machines à percer, à cintrer les fers à froid et à ajuster les ressorts de voitures. Leclerc. 8. Faub.-St-Jac-

Cabanis et Cie PER, association pour la construction des machines à élever les eaux, pompes domestiques, moteurs hydrauliques de toute nature, machines soufflantes pour les forges, machines à eaux de Sellz, presses à découper, outillage et réparation de machines ACT, Vinai-griers LOC, 32 CARDINAL.

Comment

Tags

☐ Checked

SAVE

Quelques résultats qualitatifs

1. Cas positifs (ancienne interface)

soduco-annot-gui

View 700 Didot 1856a

Local Server

Force compute

Handles size

Bouquet, M. H. Exposition universelle de 1855, breveté s. g. d. g., spécialité de tables de salles à manger, nouveau système mécanique donnant la facilité à une seule personne de pouvoir l'ouvrir, n'importe sa grandeur, Faub.-St-Antoine, 99, passage du Bras-d'Or. *

Bourdier, fab. de divans en tous genres, fait tout ce qui concerne sa partie, Charenton, 16.

Bourdier aîné, Nve-de-Lappe, 16.

Bourgade. Rumford. 14.

(0/159 entries checked)

ENTRY

OCR NER Form JSON

PER ACT LOC CARDINAL FT TITRE

CLEAR

Bouquet, M. H. Exposition universelle de 1855, breveté s. g. d. g., spécialité de tables de salles à manger, nouveau système mécanique donnant la facilité à une seule personne de pouvoir l'ouvrir, n'importe sa grandeur, Faub.-St-Antoine, 99, passage du Bras-d'Or. *

Checked

Comment

EAUX AUROPHILE ET ARGENTO-PHILE Berger, eaux spéciales pour nettoyages, déposées conformément à la loi. L'Aurophile, liquide pour nettoyer avec la plus grande facilité, propreté et sans altération, toutes pièces en bronze dorés, mates ou polies, montées ou non, avec marbre, etc., comme pendules, lustres, candélabres, garnitures de meubles, les statuettes en biscuit, marbre, etc., les fleurs ou autres sujets très-compliqués en porcelaine, les camées, l'ivoire, la nacre, les cristaux polis ou dépolis, et avec beaucoup de précaution, les dorures sur bois. L'Argentophile, liquide pour nettoyer avec la plus grande facilité et sans altération toutes pièces d'orfèvrerie mates ou bruniées, soit argent, argentées ou plaquées : la joaillerie, la bijouterie, ainsi que

20

soduco-annot-gui

View 700 Didot 1856a Local Server Handles size

Force compute

(0/159 entries checked)

ENTRY

OCR NER Form JSON

PER ACT LOC CARDINAL FT TITRE

CLEAR

Bourdier, fab. de divans en tous genres, fait tout ce qui concerne sa partie, Charenton, 16.

Bourdier aîné, Nve-de-Lappe, 16.

Bourgade, Rumford, 14.

Bourgoin, St-Antoine, 176.

Bourlier, Charonne, 99.

Boutard, Traversière, 76.

Boutung (seule maison connue depuis trente ans), fab. et magasin de meubles en tous genres, modernes, antiques et fantaisie; tient aussi le siège; commission et exportation; Faub.-St-Antoine, 97: ci-devant même rue. 23. *

Checked

Comment

EAUX AUROPHILE ET ARGENTO-PHILE Berger, eaux spéciales pour nettoyages, déposées conformément à la loi. L'Aurophile, liquide pour nettoyer avec la plus grande facilité, propreté et sans altération, toutes pièces en bronze dorés, mates ou polies, montées ou non, avec marbre, etc., comme pendules, lustres, candélabres, garnitures de meubles, les statuettes en biscuit, marbre, etc., les fleurs ou autres sujets très-compliqués en porcelaine, les camées, l'ivoire, la nacre, les cristaux polis ou dépolis, et avec beaucoup de précaution, les dorures sur bois. L'Argentophile, liquide pour nettoyer avec la plus grande facilité et sans altération toutes pièces d'orfèvrerie mates ou bruniées, soit argent, argentées ou plaquées: la joaillerie, la bijouterie, ainsi que

View  700 Didot 1856a   ?

Local

Server

          Force compute

Handles size

Bourdier aîné, Nve-de-Lappe, 16.

Bourgade, Rumford, 14.

Bourgoin, St-Antoine, 176.

Bourlier, Charonne, 99.

Boutard, Traversière, 76.

Boutang (seule maison connue depuis trente ans), fab. et magasin de meubles en tous genres, modernes, antiques et fantaisie; tient aussile siège; commission et exportation; Faub.-St-Antoine, 97; ci-devant même rue, 23. *

(0/159 entries checked)

ENTRY

OCR

NER

Form

JSON

PER

ACT

LOC

CARDINAL

FT

TITRE

CLEAR

Bourdier aîné, Nve-de-Lappe, 16.

Checked

Comment

EAUX AUROPHILE ET ARGENTO-PHILE Berger, eaux spéciales pour nettoyages, déposées conformément à la loi. L'Aurophile, liquide pour nettoyer avec la plus grande facilité, propreté et sans altération, toutes pièces en bronze dorés, mates ou polies, montées ou non, avec marbre, etc., comme pendules, lustres, candélabres, garnitures de meubles, les statuettes en biscuit, marbre, etc., les fleurs ou autres sujets très-compliqués en porcelaine, les camées, l'ivoire, la nacre, les cristaux polis ou dépolis, et avec beaucoup de précaution, les dorures sur bois. L'Argentophile, liquide pour nettoyer avec la plus grande facilité et sans altération toutes pièces d'orfèvrerie mates ou bruniées, soit argent, argentées ou plaquées: la joaillerie, la bijouterie, ainsi que

23

soduco-annot-gui

View 700 Didot 1856a Local Server Handles size

Force compute

Bourgade, Rumford, 14.

Bourgoin, St-Antoine, 176.

Bourlier, Charonne, 99.

Boutard, Traversière, 76.

Boutung (seule maison connue depuis trente ans), fab. et magasin de meubles en tous genres, modernes, antiques et fantaisie; tient aussi le siège; commission et exportation; Faub.-St-Antoine, 97; ci-devant même rue, 23. *

(0/159 entries checked)

ENTRY

OCR NER Form JSON

PER ACT LOC CARDINAL FT TITRE

CLEAR

Bourgade, Rumford, 14.

Checked

Comment

EAUX AUROPHILE ET ARGENTO-PHILE Berger, eaux spéciales pour nettoyages, déposées conformément à la loi. L'Aurophile, liquide pour nettoyer avec la plus grande facilité, propreté et sans altération, toutes pièces en bronze dorés, mates ou polies, montées ou non, avec marbre, etc., comme pendules, lustres, candélabres, garnitures de meubles, les statuettes en biscuit, marbre, etc., les fleurs ou autres sujets très-compliqués en porcelaine, les camées, l'ivoire, la nacre, les cristaux polis ou dépolis, et avec beaucoup de précaution, les dorures sur bois. L'Argentophile, liquide pour nettoyer avec la plus grande facilité et sans altération toutes pièces d'orfèvrerie mates ou bruniées, soit argent, argentées ou plaquées: la joaillerie, la bijouterie, ainsi que

soduco-annot-gui

View 700 Didot 1856a Local Server Handles size

Force compute

Bourgoin, St-Antoine, 176.
Bourlier, Charonne, 99.
Boutard, Traversière, 76.
Boutang (seule maison connue depuis trente ans), fab. et magasin de meubles en tous genres, modernes, antiques et fantaisie; tient aussi le siège; commission et exportation; Faub.-St-Antoine, 97; ci-devant même rue, 23. *

(0/159 entries checked)

ENTRY

OCR NER Form JSON

PER ACT LOC CARDINAL FT TITRE

CLEAR

Bourgoin, St-Antoine, 176.

Checked

Comment

• **EAUX AUROPHILE ET ARGENTO-PHILE** Berger, eaux spéciales pour nettoyages, déposées conformément à la loi. L'Aurophile, liquide pour nettoyer avec la plus grande facilité, propreté et sans altération, toutes pièces en bronze dorés, mates ou polies, montées ou non, avec marbre, etc., comme pendules, lustres, candélabres, garnitures de meubles, les statuettes en biscuit, marbre, etc., les fleurs ou autres sujets très-compliqués en porcelaine, les camées, l'ivoire, la nacre, les cristaux polis ou dépolis, et avec beaucoup de précaution, les dorures sur bois. L'Argentophile, liquide pour nettoyer avec la plus grande facilité et sans altération toutes pièces d'orfèvrerie mates ou bruniées, soit argent, argentées ou plaquées: la joaillerie, la bijouterie, ainsi que

Quelques résultats qualitatifs

2. Cas limites (nouvelle interface)

Didot 1851a, vue 800

Figueran, Enghien, 7.

Figueran PER , Enghien LOC , 7 CARDINAL .

Cas typique : entrée courte avec OCR et NER fonctionnels.

Poisat oncle et Cie , médailles
Ⓐ , fabrique d'acide sulfurique,
acide stéarique et oléique, et au-
tres produits, à la Folie-Nan-
terre, Enghien, 19. *

Poisat oncle et Cie PER , médailles , fabrique d'acide sulfurique, acide
stéarique et oléique, et au- tres produits ACT , à la Folie-Nan- LOC
* ACT terre, Enghien LOC , 19 CARDINAL . LOC

Entrée plus longue, avec un échec OCR (position “*”) qui entraîne un échec NER

Didot 1851a, vue 700

Berendorff(J.), M. H. 1844, Mouf-
fetard, 294, *atelier de construction
de machines à vapeur et autres*,
Ⓐ S. E., Ⓐ 1844, Ⓞ 1849.

Berendorff(J.), PER M. H. TITRE 18 44 TITRE , Mouf-
fetard LOC , 294 CARDINAL , atelier de construction de
machines à vapeur et autres ACT , S. TITRE E., O 1844, O
1849.

Cas typique : entrée longue avec entités principales (PER, LOC, CARD.) bien reconnues malgré l'inclusion de (“,”), et des erreurs NER sur les autres entités.

Duverneuill et La Tyna, 1806

Benaed, R. des Bons-Enfants, 19.

Benaed PER , R. des Bons-Enfants LOC , 19 CARDINAL .

Erreur OCR sur le nom ⇒ cible pour la correction par redondance

**Chaussard, R. de Grenelle, Halle-aux
Blés, 33.**

Chaussard PER , R. de Grenelle, Halle-aux Blés LOC , 33 CARDINAL .

Exemple positif, mais fusion de plusieurs niveaux de localisation

Deflandre 1829, vue 500

NOEL aîné et fils. Maison de commerce pour la
lithographie; r. N.-des-Petits-Champs , galerie
Colbert, esc. A.

NOEL aîné et fils PER . Maison de commerce pour la
lithographie ACT ; r. N.-des-Petits-Champs , galerie Colbert LOC . esc. A

Exemple positif, mais fusion de plusieurs niveaux de localisation

Favre et Duchesne, 1897

D'Huez, rue des Poulies, n.º 209, — des Gardes--Françaises.

D'Huez PER, rue des Poulies LOC, n. 209 CARDINAL, — des
Gardes--Françaises LOC.

Micro-variations OCR sur la localisation (“--”), localisations multiples au même plan

Bottin, 1820, vue 52

des Commerçans de Paris.		47
Bourdon, bonnetier, r. S.-Denis, 269.	Bourgeois-Borlot, plaqueur, r. du Bac, 98.	451
(Elij.). 292	Bourgeois-Damoulin, md. de nouveautés, r. Honore, 46.	318
Bourdon, md. de beurre, r. du Faub.-S.-Honore, 46.	Bussi, 2.	410
Bourdon et co., commiss. par eau, q. de la Mégisserie, 38.	Bourgeot, fab. de chocolat, r. S.-Honore, 110.	316
Bourdon, epicier, r. des Quatre-Vents, 11.	Bourgeot, papetier, r. des Fosses-Montmar-tre, 31.	437
Bourdon (V.), epiciere, r. Sainte-Anne, 10.	Bourget et co., commiss. de roulage, r. S.-Denis, 152.	323, 381
Bourdon (N.) aîné et co., miroitiers, r. Bourg-l'Abbe, 48.	Bourget jeune, commis. de roulage, r. Beau-repaire, 3.	323
Bourdon (T.) quincaillier, r. Bourg-l'Abbe.	Boursin, epicier, r. du Faub.-S.-Honore.	

OCR NER

318 Honore, 46. Bourdon et co. PER , commiss. par eau ACT , f. de la
325 Megisserie LOC , 38 CARDINAL . Bussi, 2. 410 Bourge PER ot, fab.
de chocolat ACT , r. S.-Honore LOC , 110 CARDINAL . 316

La détection de la structure reste une difficulté majeure.

Bilan

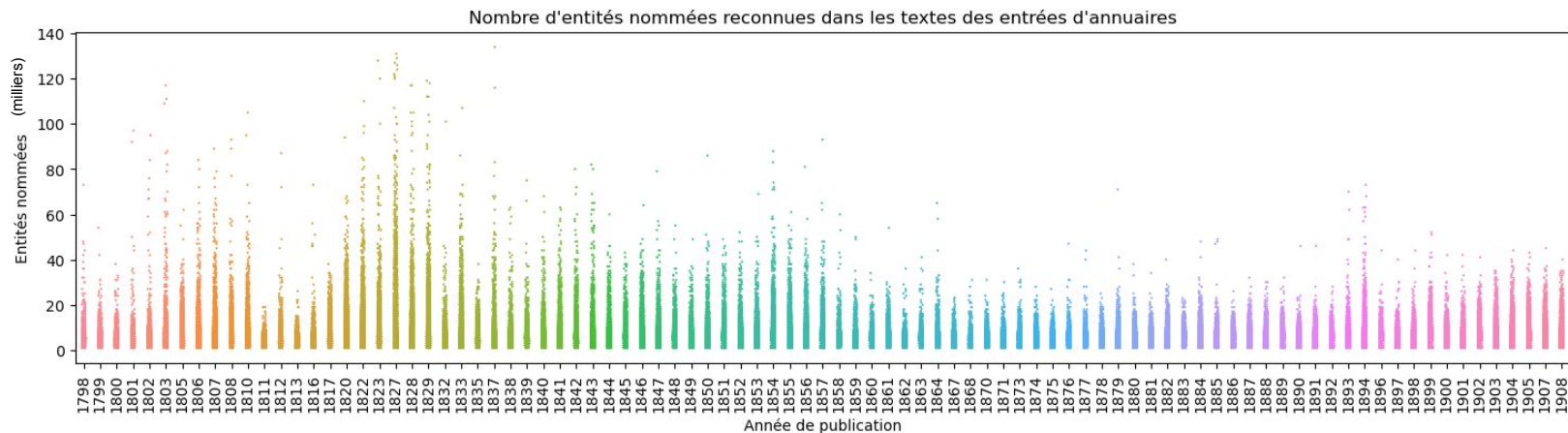
Vue d'ensemble des données extraites

Pour l'ensemble du corpus :

- 9 821 898 entrées extraites
- 7 260 104 (74%) entrées exploitables, de la forme : **PERSON** [**ACT**] **LOC** [**CARDINAL**]

⇒ la forte redondance des entrées est riche en information

Un enjeu restant : identifier, localiser et re-traiter (semi-automatiquement ?) les entrées mal identifiées



“Traîne” des entrées agglomérées : une faible proportion mais des effets potentiellement importants sur les traitements suivants.

Perspectives — Projet Biblissima+

Encore une plateforme d’annotation / reconnaissance de documents imprimés ?

⇒ Non ! Vers une **collection** d’outils **ouverts**, **minimaux**, laissant les historiens **constituer leurs corpus** et y “superposer” **leurs données**.

Nos objectifs à venir :

- **Consolider notre interface** de visualisation et d’annotation (page web unique, sans besoins d’hébergement ni d’installation)
- **Déployer des services** pour une utilisation à la demande sans nécessiter une machine puissante (OCR, NER, etc. — un hébergement HumaNuM ?)
- **Enrichir notre boîte à outils** (correction de courbure des pages, export TEI...)
- **Partager** plus largement ces outils (*open source*, documentation, voire formation)