

# Extraction de données

# *Approche du projet SoDUCo*

De l'image d'annuaire du commerce ancien  
au graphe de données interrogable

N. Abadie, S. Baciocchi, E. Carlinet, J. Chazalon, P. Cristofoli, B. Duménieu, J. Perret, S. Tual  
Lastig (IGN), CRH (EHESS), LRE (EPITA)

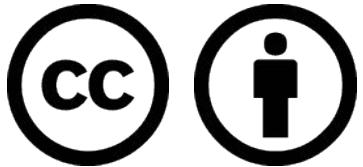


6 nov. 2023, séminaire SoDUCo-BnF



# (Ré)utiliser cette présentation

Licence Creative Common BY 4.0



N'hésitez pas à **réutiliser** ces contenus et nous **citer**.

*Sauf citation ou mention contraire,  
nous sommes les auteurs des contenus.*

Consultez le diaporama pendant et après la présentation

<https://bit.ly/soduco-bnf-20231106-extraction>



# Un corpus hétérogène

Ouvrages produits par des éditeurs variés, numérisés par différents prestataires.

Des différences de

- Longueur, richesse et structuration de l'information
- Mises en page, fontes
- Qualité inégale : papier, résolution des images, présence de marques...

<i>Non-Commerçans. (Paris):</i>		269
<b>Chardin, R. Pavée,</b>	<b>26. — R. G.</b>	<b>Cheviron, R. Chapon,</b>
<b>Chardin, R. Michel Lepelletier,</b>	<b>21.</b>	<b>Chimay, (Mme.) R. de Varennes,</b>
<b>Chardon, (Vve.) R. S. Marc,</b>	<b>15.</b>	<b>Choart-Duplessis, R. de Turenne,</b>
AMADOU ET ALLUMETTES. — Pour les ALLUMETTES OXIGÉNÉES. Voyez BRIQUETS PHYSIQUES.		
DARRAS (Thomas), r. de la Vieille-Monnaie, 10.		GALLIENNE je., r. de la Heaumerie, 3.
Briques et veilleuses, mèches à quinquets, à quinquet, veilleuses mèches, souffrées ; mèches souffrées, pierres, agaric de chêne, pierres, agaric, bouchons, liège.		Brûle-tout, boîtes à briquet, mèches à vin et canicule.
		LEBOY, r. Aubry-le-Boucher, 43.
<b>BAUDOYER (place).</b> <b>IX Arr. Hôtel-de-Ville.)</b> ← Rue Tixeranderie, pourtour St-Gervais, Saint-Antoine et Renaud-Lefèvre,  1 Lissoty (Vve), vins. 2 <sup>e</sup> Privé, distillateur. Lemoine-Cruzet et Leroy, nouveautés. Chantrier, court.-gourn.		
26* Longpré ainé, bijoutier en or et argent. Saint-Omer, émailleur. Cellier (A.), graveur-ciseleur.*	26* Bourguille, fabr. de presses. Vaudain, passementier. Finino Jno, bronze doré. Rabé ainé, fabr. de balcons.*	7 Ecole communale de jeu- nes filles. Berthelot, vins. 6 Verstaen, serrurier-mé- canicien.
Roussau (J.), bijoutier en or.* Benoit, orfèvre-fab. Léresy, doreur.	Gaulin, chapelier. Moisy, tabletier. 30 Bonton, fab. de cuir vernis.* 31 Pardon, vins.	8 Michel, brossier. 9 Labottière, serrurier. 10 Sacrez, vins. 12 Baudoin, épicer. 13 Lejard, clouteries et crê- pins. 14 Baduel (Vve), fab. de
		et tapisseries. 10 Lainé jeune, vins. 11 Jumelius omnibus et en- treprise générale des Omnitibus. 11 Melouzay, vins en gros, et à Bercy, Port, 31. 12 Combaud, coiffeur. Monmain (P.), vins en gros. 13 Dufally, sculpt. fabr. de carton-pierre.



# Un instantané des acteurs commerciaux à un instant

Bibliothèque Ste-Geneviève, Clotilde, 1.  
Bibliothèque de la Ville, quai d'Austerlitz , 33  
(provisoirement).  
Biblique protestante (Société), Moulins, 16.  
Bibron, aide-natural. , au Muséum d'hist. nat.  
Bibus, tailleur, Roule, 21.  
Bibus, tailleur, Richelieu, 31.  
Bical et Dorre, fab. de socques, Vertbois, 14.  
Bicant (Mme), fondeur en cuivre , cour de la  
Corderie-du-Temple, 26.  
Bichard (Mme), Nve-de-Luxembourg, 17.  
Bichard, tabacs et eau-de-vie , Faub.-St-Mar-  
tin, 45.

Didot 1841a - page 95

Bibliothèque Ste-Geneviève, Clotilde, 1.  
Bibliothèque de la Ville, quai d'Austerlitz , 33  
(provisoirement).  
Biblique protestante (Société), Moulins, 16.  
Bibron, aide-natural. , au Muséum d'hist. nat.  
Bibus, tailleur, Roule, 21.  
Bibus, tailleur, Richelieu, 31.  
Bical, fab. de jouets, Montmorency, 33.  
Bical et Dorre, fab. de socques, Vertbois, 14.  
Bican (Vve) et fils, fondeur en cuivre , place  
de la Corderie-du-Temple, 26.  
Bicel, épicier, marché d'Aguesseau, 15.  
Bichard (Mme), Nve-de-Luxembourg, 17.  
Bichard, tabacs et eau-de-vie , Faub.-St-Mar-  
tin, 45.

Didot 1842a - page 117

Bibliothèque Ste-Geneviève, rue des Sept-Voies  
et place du Panthéon.  
Bibliothèque de la Ville , quai d'Austerlitz , 33  
(provisoirement).  
Bibus, tailleur, Richelieu, 31.  
Bical, fab. de jouets, Montmorency, 33.  
Bical et Doire, fab. de socques, Vert-Bois, 14.  
Bican (Vve) et fils, fondeurs en cuivre , place  
de la Corderie-du-Temple, 26.  
Bicel, épicier, Marché-d'Aguesseau, 15.  
Bichard, tabac et eau-de-vie , Faub.-St-Mar-  
tin, 45.

Didot 1843a - page 129

Bibliothèque Ste-Generiève, rue des Sept-Voies  
et place du Panthéon.  
Bibliothèque de la Ville , quai d'Austerlitz , 33  
(provisoirement).  
Bibolet, relieur, passage Sainte-Marie-Saint-  
Germain, 10.  
Bibonne, architecte, Magasins, 12.  
Bibron , aide-naturaliste au Jardin-des-Plan-  
tes, Cuvier, 29.  
Bibus, tailleur, Richelieu, 31.  
Bical, fab. de jouets, Montmorency, 33.  
Bican (Vve) et fils, fondeurs en cuivre , place  
de la Corderie-du-Temple, 26.  
Bicel, épicier, Marché-d'Aguesseau, 15.  
Bichard , tabac et eau-de-vie , Faub.-St-Mar-  
tin, 45.

Didot 1844a - pages 125-126

# Un instantané des acteurs commerciaux à un instant

Grande redondance d'information intra- et inter-annuaires.

Bibliothèque Ste-Geneviève, Clotilde, 1.

Bibliothèque de la Ville, quai d'Austerlitz, 33  
(provisoirement).

Bill que protestante (Société), Moulins, 16.  
Bibron, aide-natural., au Muséum d'hist. nat.  
Bibus, tailleur. Roule, 21.

Bibus, tailleur, Richelieu, 31.  
Bical et Dorre, fab. de socques, Vertbois, 14.  
Bicant (Mme), fondeur en cuivre, cour de la  
Corderie-du-Temple, 26.

Bichard (Mme), Nve-de-Luxembourg, 17.  
Bichard, tabacs et eau-de-vie, Faub.-St-Martin, 45.

Didot 1841a - page 95

Bibliothèque Ste-Geneviève, Clotilde, 1.

Bibliothèque de la Ville, quai d'Austerlitz, 33  
(provisoirement).

Biotique protestante (Société), Moulins, 10.  
Bibron, aide-natural., au Muséum d'hist. nat.  
Bibus, tailleur. Roule, 21.

Bibus, tailleur, Richelieu, 31.  
Bical, fab. de jouets, Montmorency, 33.  
Bical et Dorre, fab. de socques, Vertbois, 14.  
Bican (Vve) et fils, fondeur en cuivre, place  
de la Corderie-du-Temple, 26.

Bicel, épicier, marché d'Aguesseau, 15.  
Bichard (Mme), Nve de Luxembourg, 17.  
Bichard, tabacs et eau-de-vie, Faub.-St-Martin, 45.

Didot 1842a - page 117

Bibliothèque Ste-Geneviève, rue des Sept-Voies  
et place du Panthéon.

Bibliothèque de la Ville, quai d'Austerlitz, 33  
(provisoirement).

Bibus, tailleur, Richelieu, 31.

Bical, fab. de jouets, Montmorency, 33.  
Bical et Doire, fab. de socques, Vert-Bois, 14.  
Bican (Vve) et fils, fondeurs en cuivre, place  
de la Corderie-du-Temple, 26.

Bicel, épicier, Marché-d'Aguesseau, 15.

Bichard, tabac et eau-de-vie, Faub.-St-Martin, 45.

Didot 1843a - page 129

Bibliothèque Ste-Geneviève, rue des Sept-Voies  
et place du Panthéon.

Bibliothèque de la Ville, quai d'Austerlitz, 33  
(provisoirement).

Bibolet, relieur, passage Sainte-Marie-Saint-Germain, 10.

Bibonne, architecte, Magasins, 12.  
Bibron, aide-naturaliste au Jardin-des-Plantes, Cuvier, 20.

Bibus, tailleur, Richelieu, 31.

Bical, fab. de jouets, Montmorency, 33.

Bican (Vve) et fils, fondeurs en cuivre, place  
de la Corderie-du-Temple, 26.

Bicel, épicier, Marché-d'Aguesseau, 15.  
Bichard, tabac et eau-de-vie, Faub.-St-Martin, 45.

Didot 1844a - pages 125-126

# Un instantané des acteurs commerciaux à un instant

Des entrées disparaissent au cours du temps.

Bibliothèque Ste-Geneviève, Clotilde, 1.  
Bibliothèque de la Ville, quai d'Austerlitz , 33  
(provisoirement).

Bibliothèque protestante (Société), Moulins, 16.

Bibron, aide-naturaliste au Muséum d'hist. nat.

Bibus, tailleur, Roule, 21.

Bibus, tailleur, Richelieu, 31.

Bical et Dorre, fab. de socques, Vertbois, 14.

Bicant (Mme), fondeur en cuivre , cour de la Corderie-du-Temple, 26.

Bichard (Mme), Nve-de-Luxembourg, 17.

Bienhard, tabacs et eau-de-vie , Faub.-St-Martin, 45.

Bibliothèque Ste-Geneviève, Clotilde, 1.  
Bibliothèque de la Ville, quai d'Austerlitz , 33  
(provisoirement).

Biblique protestante (Société), Moulins, 16.

Bibus, tailleur, Roule, 21.

Bibus, tailleur, Richelieu, 31.

Bical, fab. de jouets, Montmorency, 33.

Bical et Dorre, fab. de socques, Vertbois, 14.

Bican (Vve) et fils, fondeurs en cuivre , place de la Corderie-du-Temple, 26.

Bicel, épicier, marché d'Aguesseau, 15.

Bichard (Mme), Nve-de-Luxembourg, 17.

Bichard, tabacs et eau-de-vie , Faub.-St-Martin, 45.

Bibliothèque Ste-Geneviève, rue des Sept-Voies et place du Panthéon.

Bibliothèque de la Ville, quai d'Austerlitz , 33  
(provisoirement).

Bibus, tailleur, Richelieu, 31.

Bical, fab. de jouets, Montmorency, 33.

Bical et Doire, fab. de socques, Vert-Bois, 14.

Bican (Vve) et fils, fondeurs en cuivre , place de la Corderie-du-Temple, 26.

Bicel, épicier, Marché-d'Aguesseau, 15.

Bichard, tabac et eau-de-vie , Faub.-St-Martin, 45.

Bibliothèque Ste-Geneviève, rue des Sept-Voies et place du Panthéon.

Bibliothèque de la Ville, quai d'Austerlitz , 33  
(provisoirement).

Bibolet, relieur, passage Sainte-Marie-Saint-Germain, 10.

Bibonne, architecte, Magasins, 12.

Bibron, aide-naturaliste au Jardin-des-Plantes, Cuvier, 29.

Bibus, tailleur, Richelieu, 31.

Bical, fab. de jouets, Montmorency, 33.

Bican (Vve) et fils, fondeurs en cuivre , place de la Corderie-du-Temple, 26.

Bicel, épicier, Marché-d'Aguesseau, 15.

Bichard, tabac et eau-de-vie , Faub.-St-Martin, 45.

Didot 1841a - page 95

Didot 1842a - page 117

Didot 1843a - page 129

Didot 1844a - pages 125-126

# Un instantané des acteurs commerciaux à un instant

D'autres apparaissent.

Bibliothèque Ste-Geneviève, Clotilde, 1.  
Bibliothèque de la Ville, quai d'Austerlitz , 33  
(provisoirement).  
Bibliothèque protestante (Société), Moulins, 16.  
Bibron, aide-natural. , au Muséum d'hist. nat.  
Bibus, tailleur, Roule, 21.  
Bibus, tailleur, Richelieu, 31.  
Bical et Dorre, fab. de socques, Vertbois, 14.  
Bicant (Mme), fondeur en cuivre , cour de la  
Corderie-du-Temple, 26.  
Bichard (Mme), Nve-de-Luxembourg, 17.  
Bichard, tabacs et eau-de-vie , Faub.-St-Mar-  
tin, 45.

Didot 1841a - page 95

Bibliothèque Ste-Geneviève, Clotilde, 1.  
Bibliothèque de la Ville, quai d'Austerlitz , 33  
(provisoirement).  
Bibliothèque protestante (Société), Moulins, 16.  
Bibron, aide-natural. , au Muséum d'hist. nat.  
Bibus, tailleur, Roule, 21.  
Bibus, tailleur, Richelieu, 31.  
Bical, fab. de jouets, Montmorency, 33.  
Bical et Dorre, fab. de socques, Vertbois, 14.  
Bican (Vve) et fils, fondeur en cuivre , place  
de la Corderie-du-Temple, 26.  
Bicel, épicier, marché d'Aguesseau, 15.  
Bichard (Mme), Nve-de-Luxembourg, 17.  
Bichard, tabacs et eau-de-vie , Faub.-St-Mar-  
tin, 45.

Didot 1842a - page 117

Bibliothèque Ste-Geneviève, rue des Sept-Voies  
et place du Panthéon.  
Bibliothèque de la Ville , quai d'Austerlitz , 33  
(provisoirement).  
Bibus, tailleur, Richelieu, 31.  
Bical, fab. de jouets, Montmorency, 33.  
Bical et Doire, fab. de socques, Vert-Bois, 14.  
Bican (Vve) et fils, fondeurs en cuivre , place  
de la Corderie-du-Temple, 26.  
Bicel, épicier, Marché-d'Aguesseau, 15.  
Bichard, tabac et eau-de-vie , Faub.-St-Mar-  
tin, 45.

Didot 1843a - page 129

Bibliothèque Ste-Geneviève, rue des Sept-Voies  
et place du Panthéon.  
Bibliothèque de la Ville , quai d'Austerlitz , 33  
(provisoirement).  
Bibolet, relieur, passage Sainte-Marie-Saint-  
Germain, 10.  
Bibonne, architecte, Magasins, 12.  
Bibron , aide-naturaliste au Jardin-des-Pian-  
tes, Cuvier, 29.  
Bibus, tailleur, Richelieu, 31.  
Bical, fab. de jouets, Montmorency, 33.  
Bican (Vve) et fils, fondeurs en cuivre , place  
de la Corderie du Temple, 26.  
Bicel, épicier, Marché-d'Aguesseau, 15.  
Bichard , tabac et eau-de-vie , Faub.-St-Mar-  
tin, 45.

Didot 1844a - pages 125-126

# Un instantané des acteurs commerciaux à un instant

Et certaines changent.

Bibliothèque Ste-Geneviève, Clotilde, 1.  
Bibliothèque de la Ville, quai d'Austerlitz , 33  
(provisoirement).  
Biblique protestante (Société), Moulins, 16.  
Bibron, aide-natural. , au Muséum d'hist. nat.  
Bibus, tailleur, Roule, 21.  
Bibus, tailleur, Richelieu, 31.  
Bical et Dorre, fab. de socques, Vertbois, 14.  
Bicant (Mme), fondeur en cuivre , cour de la  
Corderie-du-Temple, 26.  
Bichard (Mme), Nve-de-Luxembourg, 17.  
Bichard, tabacs et eau-de-vie , Faub.-St-Mar-  
tin, 45.

Didot 1841a - page 95

Bibliothèque Ste-Geneviève, Clotilde, 1.  
Bibliothèque de la Ville, quai d'Austerlitz , 33  
(provisoirement).  
Biblique protestante (Société), Moulins, 16.  
Bibron, aide-natural. , au Muséum d'hist. nat.  
Bibus, tailleur, Roule, 21.  
Bibus, tailleur, Richelieu, 31.  
Bical, fab. de jouets, Montmorency, 33.  
Bical et Dorre, fab. de socques, Vert-Bois, 14.  
Bican (Vve) et fils, fondeurs en cuivre , place  
de la Corderie-du-Temple, 26.  
Bicel, épicier, marché d'Aguesseau, 15.  
Bichard (Mme), Nve-de-Luxembourg, 17.  
Bichard, tabacs et eau-de-vie , Faub.-St-Mar-  
tin, 45.

Didot 1842a - page 117

Bibliothèque Ste-Geneviève, rue des Sept-Voies  
et place du Panthéon.  
Bibliothèque de la Ville, quai d'Austerlitz , 33  
(provisoirement).  
Bibus, tailleur, Richelieu, 31.  
Bical, fab. de jouets, Montmorency, 33.  
Bical et Dorre, fab. de socques, Vert-Bois, 14.  
Bican (Vve) et fils, fondeurs en cuivre , place  
de la Corderie-du-Temple, 26.  
Bicel, épicier, Marché-d'Aguesseau, 15.  
Bichard, tabac et eau-de-vie , Faub.-St-Mar-  
tin, 45.

Didot 1843a - page 129

Bibliothèque Ste-Geneviève, rue des Sept-Voies  
et place du Panthéon.  
Bibliothèque de la Ville, quai d'Austerlitz , 33  
(provisoirement).  
Bibolet, relieur, passage Sainte-Marie-Saint-  
Germain, 10.  
Bibonne, architecte, Magasins, 12.  
Bibron, aide-naturaliste au Jardin-des-Plan-  
tes, Cuvier, 29.  
Bibus, tailleur, Richelieu, 31.  
Bicel, fab. de jouets, Montmorency, 33.  
Bican (Vve) et fils, fondeurs en cuivre , place  
de la Corderie-du-Temple, 26.  
Bicel, épicier, Marché-d'Aguesseau, 15.  
Bichard, tabac et eau-de-vie , Faub.-St-Mar-  
tin, 45.

Didot 1844a - pages 125-126

# Notre objectif d'extraction de données

Extraire les entrées des annuaires pour les géo-référencer, les dédoublonner, les corriger, les relier...

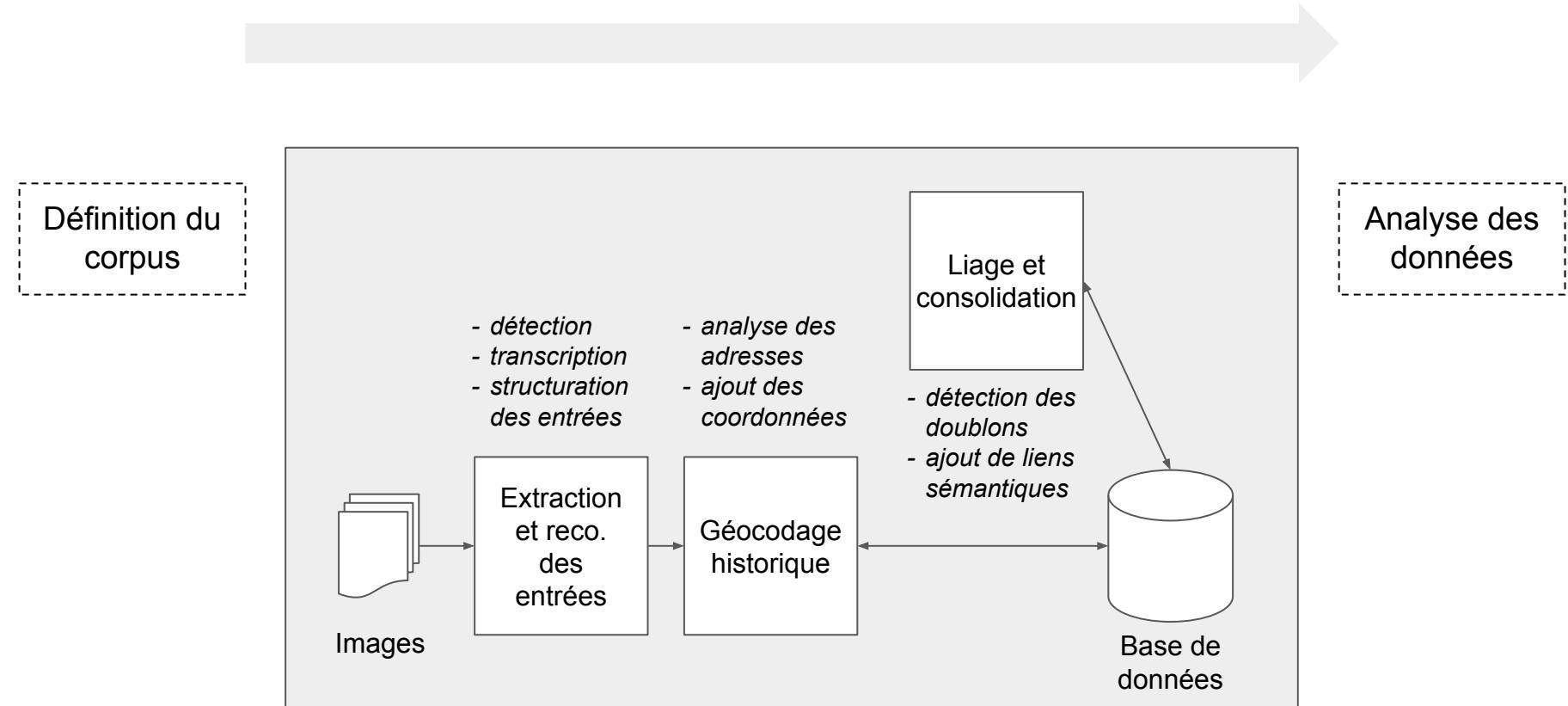
Bibliothèque Ste-Geneviève, Clotilde, 1.  
Bibliothèque de la Ville , quai d'Austerlitz , 33  
(provisoirement).  
Biblique protestante (Société), Moulins, 16.  
Bibron, aide-natural., au Muséum d'hist. nat.  
Bibus, tailleur, Roule, 21.  
Bibus, tailleur Richelieu, 31.  
Bical et Dorre, fab. de socques, Vertbois, 14.  
Bicant (Mme), fondeur en cuivre , cour de la  
Corderie-du-Temple, 26.  
Bichard (Mme), Nve-de-Luxembourg, 17.  
Bichard, tabacs et eau-de-vie , Faub.-St-Martin, 45.

Didot 1841a - page 95

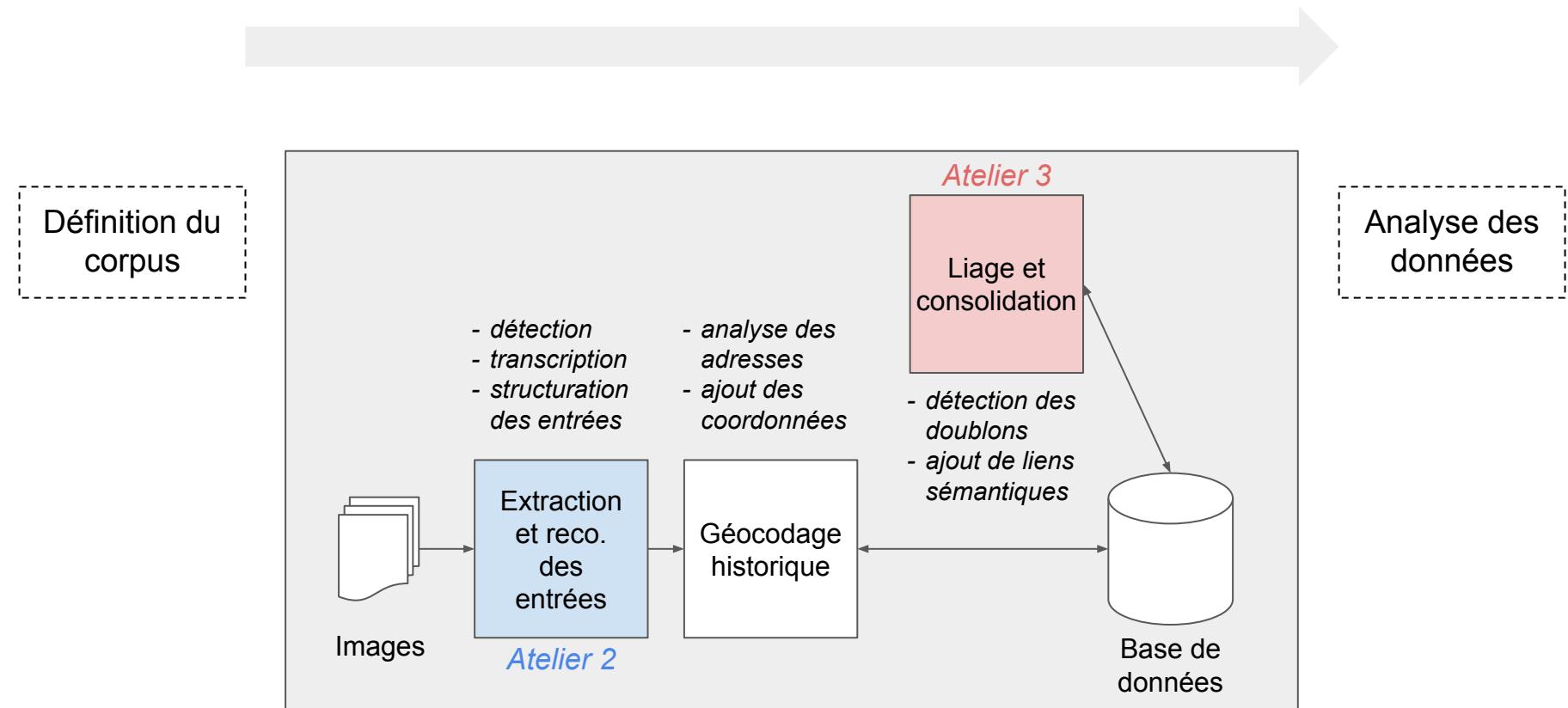
The diagram illustrates the data extraction process. On the left, a rectangular box contains a list of historical business entries from a document. A red dashed arrow points from this box to a large cylinder on the right. The cylinder represents a database structure, specifically a row in a table with six columns. The data from the list is mapped into these columns: 'Bibus', 'tailleur', 'Rue Richelieu', 'no 31, (48,865, 2,337)', 'Didot\_1841a\_p95', and an ellipsis.

Bibus	tailleur	Rue Richelieu	no 31, (48,865, 2,337)	Didot_1841a_p95	...
-------	----------	---------------	------------------------	-----------------	-----

# Vue d'ensemble du processus



# Vue d'ensemble du processus



# Extraction et reconnaissance des entrées



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

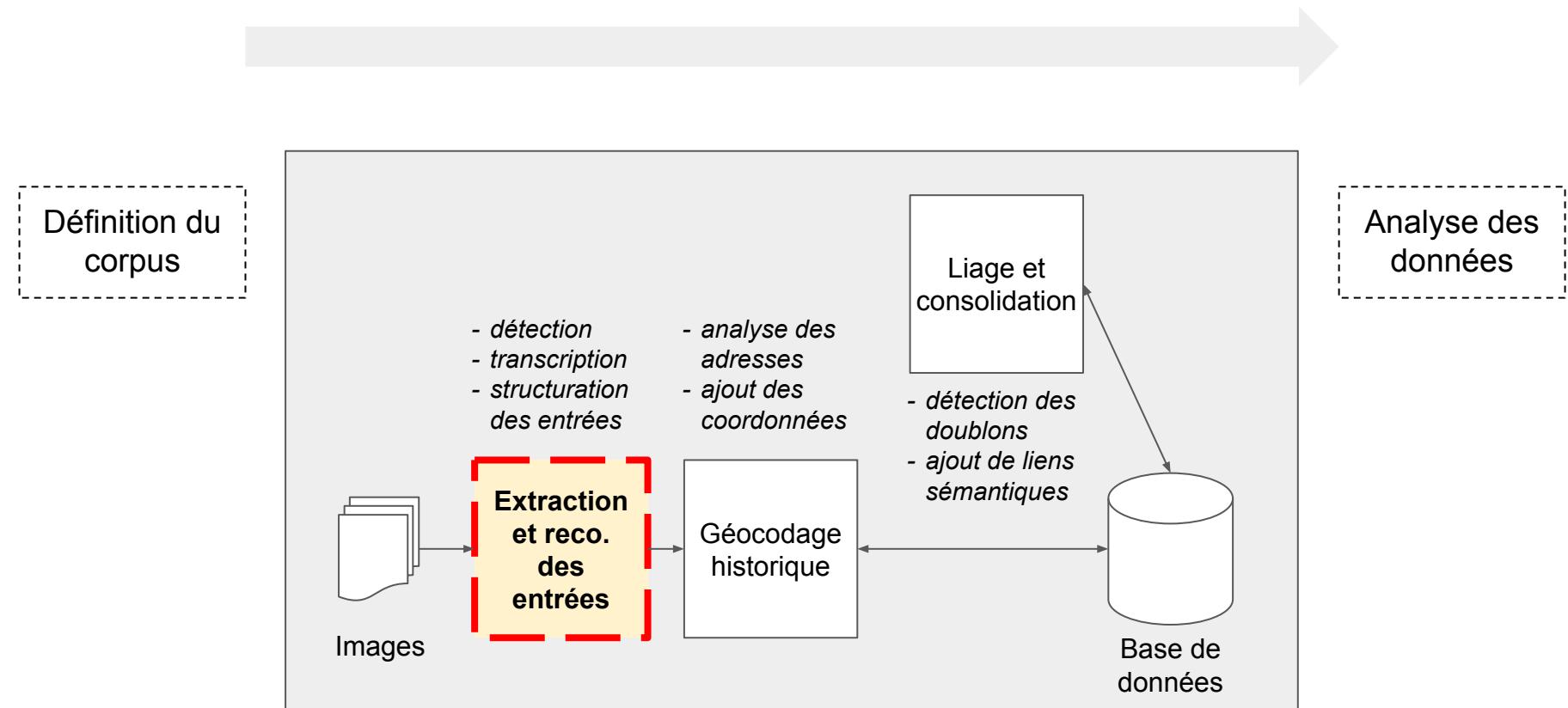


Lemonnyer, plomberie, r. de Bondy, 86, et r. Bouchardon, 1.



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

# Vue d'ensemble du processus



# Chaîne de traitement pour l'extract. et la reco. des entrées

Pré-traitement image  
(correction géométrique,  
réduction de bruit...)

Segmentation du canvas  
Classification des blocs

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

Transcription

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

Extraction des  
informations clés

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



# Chaîne de traitement pour l'extract. et la reco. des entrées

Pré-traitement image  
(correction géométrique,  
réduction de bruit...)

Segmentation du canvas  
Classification des blocs

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

Transcription

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

Extraction des  
informations clés

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



# Amélioration de la qualité d'image

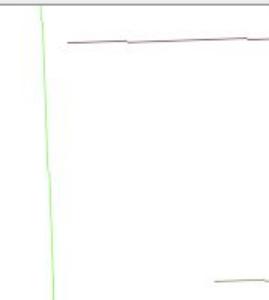
- Réduire la variance des entrées pour la suite des traitements
  - ✓ Amélioration de contraste / suppression de fond
  - ✓ Rotation / Shear

Une contribution notable publiée :  
détection de lignes rapide et fiable

*P. Bernet, J. Chazalon, E. Carlinet, A. Bourquelot et E. Puybareau, “Linear Object Detection in Document Images using Multiple Object Tracking”, in Proc. ICDAR 2023*  
<https://github.com/EPITAResearchLab/bernet.23.icdar>



## *Image originale*



## *Lignes détectées*



## *Image redressée*

1449

**FOURBISSEURS.**  
[Voir aussi Armuriers.]

Bacques (F.), succ. de F. Delageur @ Bacques, Elevir, 7.

Barre (Gustave), armes blanches de baron, sabres, épées de commandant, canons, répartie, armes antiques @ Comm. export., Lions-Saint-Paul, 11.

Debacker (F.) @ et Bacques (F., successeur), fabr. d'armes blanches, ① 1867, Elevir, 7.

Fauve Le Page # (anc. maison Le Page), rue Richelieu, 8.

**GAMBETTE (F.)**, armures, armes blanches, copies d'épées anciennes, Quatre-Septembre, 2.

Glain (L.), armes blanches, spé. de fourreaux et garnitures de sabres, à Beaubourg, 49.

Grégoire (A.), Ecouffes, 20.  
 Lacroix (J.), armes françaises et étrangères, couleaux de chasse, rue de Thorigny, 6.  
 Maria (J.), rue du Quatre-Sep-tembre, 14.  
 Pettilis (V.), Ecouffes, 22.  
**DÉPÔT DE PARIS**, spécialité de sabres et d'épées pour officiers de toutes armes (fantaisie et ordonnance), et pour fonction-

Digitized by srujanika@gmail.com

— 1 —

— 1 —

## BEAUCO

II. Ans

11 АГР.

## *l' Roulc int. B.*

Count-Hone

4744-2

*age*

essée

# Pré-traitements

Très liés à notre approche de segmentation du canevas.

Une contribution notable publiée : détection de lignes rapide et fiable

*P. Bernet, J. Chazalon, E. Carlinet, A. Bourquelot et E. Puybareau, “Linear Object Detection in Document Images using Multiple Object Tracking”, in Proc. ICDAR 2023*

<https://github.com/EPITAResearchLab/bernet.23.icdar>

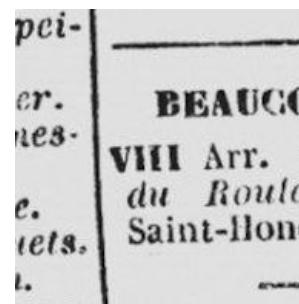
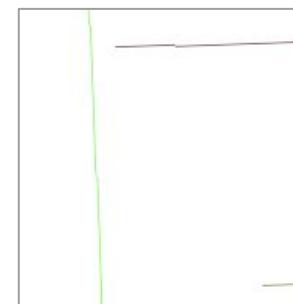


Image  
originale



Lignes  
détectées



Image  
redressée

# Chaîne de traitement pour l'extract. et la reco. des entrées

Pré-traitement image  
(correction géométrique,  
réduction de bruit...)

Segmentation du canvas  
Classification des blocs

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

Transcription

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

Extraction des  
informations clés

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



# Extraction de canvas / Segmentation

Méthode *ad hoc* qui tire profit de la mise en page régulière

1. Séparation en **blocs** (XY cuts, smearing) et Classification des régions (entête, titre...)
2. Détection des **lignes** (watershed)
3. Regroupement des lignes en **entrées** (HMM)

## Avantages

- Très rapide (fraction de seconde par image)
- Aussi efficace que les approches modernes type LayoutParser sur ces données
- Extraction d'entrées intégrée

## Limites

- Extraction d'entrées intégrée
- Pas d'exploitation de l'information textuelle
- Pas de gestion du multi-pages
- Sensible au bruit, d'où le nettoyage en amont

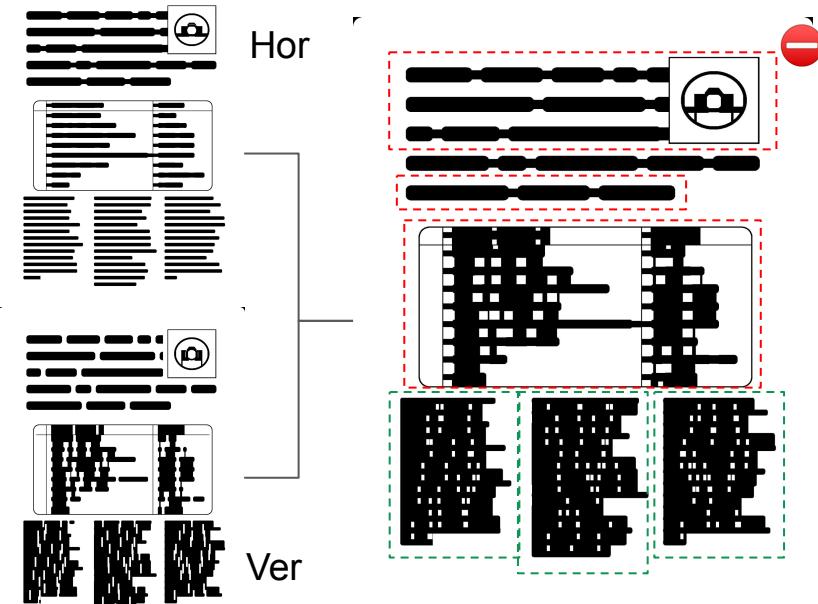
FORMES A SUCRE.	FOURCHES ET PELLES.	1443
<p><b>BROMMER, JEROME</b>, faveur des Clermonts, 1760, 1761, 1762, 1763, 1764, 1765, 1766, 1767, 1768, 1769, 1770, 1771, 1772, 1773, 1774, 1775, 1776, 1777, 1778, 1779, 1780, 1781, 1782, 1783, 1784, 1785, 1786, 1787, 1788, 1789, 1790, 1791, 1792, 1793, 1794, 1795, 1796, 1797, 1798, 1799, 1800, 1801, 1802, 1803, 1804, 1805, 1806, 1807, 1808, 1809, 1810, 1811, 1812, 1813, 1814, 1815, 1816, 1817, 1818, 1819, 1820, 1821, 1822, 1823, 1824, 1825, 1826, 1827, 1828, 1829, 1830, 1831, 1832, 1833, 1834, 1835, 1836, 1837, 1838, 1839, 1840, 1841, 1842, 1843, 1844, 1845, 1846, 1847, 1848, 1849, 1850, 1851, 1852, 1853, 1854, 1855, 1856, 1857, 1858, 1859, 1860, 1861, 1862, 1863, 1864, 1865, 1866, 1867, 1868, 1869, 1870, 1871, 1872, 1873, 1874, 1875, 1876, 1877, 1878, 1879, 1880, 1881, 1882, 1883, 1884, 1885, 1886, 1887, 1888, 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, 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, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2587, 2588, 2589, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2597, 2598, 2599, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 2679, 2680, 2681, 2682, 2683, 2684, 2685, 2686, 2687, 2687, 2688, 2689, 2689, 2690, 2691, 2692, 2693, 2694, 2695, 2696, 2697, 2697, 2698, 2699, 2699, 2700, 2701, 2702, 2703, 2704, 2705, 2706, 2707, 2708, 2709, 2709, 2710, 2711, 2712, 2713, 2714, 2715, 2716, 2717, 2718, 2719, 2719, 2720, 2721, 2722, 2723, 2724, 2725, 2726, 2727, 2728, 2729, 2729, 2730, 2731, 2732, 2733, 2734, 2735, 2736, 2737, 2738, 2739, 2739, 2740, 2741, 2742, 2743, 2744, 2745, 2746, 2747, 2748, 2749, 2749, 2750, 2751, 2752, 2753, 2754, 2755, 2756, 2757, 2758, 2759, 2759, 2760, 2761, 2762, 2763, 2764, 2765, 2766, 2767, 2768, 2769, 2769, 2770, 2771, 2772, 2773, 2774, 2775, 2776, 2777, 2778, 2779, 2779, 2780, 2781, 2782, 2783, 2784, 2785, 2786, 2787, 2787, 2788, 2789, 2789, 2790, 2791, 2792, 2793, 2794, 2795, 2796, 2797, 2797, 2798, 2799, 2799, 2800, 2801, 2802, 2803, 2804, 2805, 2806, 2807, 2808, 2809, 2809, 2810, 2811, 2812, 2813, 2814, 2815, 2816, 2817, 2817, 2818, 2819, 2819, 2820, 2821, 2822, 2823, 2824, 2825, 2826, 2827, 2828, 2829, 2829, 2830, 2831, 2832, 2833, 2834, 2835, 2836, 2837, 2838, 2839, 2839, 2840, 2841, 2842, 2843, 2844, 2845, 2846, 2847, 2848, 2849, 2849, 2850, 2851, 2852, 2853, 2854, 2855, 2856, 2857, 2858, 2859, 2859, 2860, 2861, 2862, 2863, 2864, 2865, 2866, 2867, 2868, 2869, 2869, 2870, 2871, 2872, 2873, 2874, 2875, 2876, 2877, 2878, 2879, 2879, 2880, 2881, 2882, 2883, 2884, 2885, 2886, 2887, 2888, 2889, 2889, 2890, 2891, 2892, 2893, 2894, 2895, 2896, 2897, 2898, 2899, 2899, 2900, 2901, 2902, 2903, 2904, 2905, 2906, 2907, 2908, 2909, 2909, 2910, 2911, 2912, 2913, 2914, 2915, 2916, 2917, 2918, 2919, 2919, 2920, 2921, 2922, 2923, 2924, 2925, 2926, 2927, 2928, 2929, 2929, 2930, 2931, 2932, 2933, 2934, 2935, 2936, 2937, 2938, 2939, 2939, 2940, 2941, 2942, 2943, 2944, 2945, 2946, 2947, 2948, 2949, 2949, 2950, 2951, 2952, 2953, 2954, 2955, 2956, 2957, 2958, 2959, 2959, 2960, 2961, 2962, 2963, 2964, 2965, 2966, 2967, 2968, 2969, 2969, 2970, 2971, 2972, 2973, 2974, 2975, 2976, 2977, 2978, 2979, 2979, 2980, 2981, 2982, 2983, 2984, 2985, 2986, 2987, 2988, 2989, 2989, 2990, 2991, 2992, 2993, 2994, 2995, 2996, 2997, 2998, 2999, 2999, 3000, 3001, 3002, 3003, 3004, 3005, 3006, 3007, 3008, 3009, 3009, 3010, 3011, 3012, 3013, 3014, 3015, 3016, 3017, 3018, 3019, 3019, 3020, 3021, 3022, 3023, 3024, 3025, 3026, 3027, 3028, 3029, 3029, 3030, 3031, 3032, 3033, 3034, 3035, 3036, 3037, 3038, 3039, 3039, 3040, 3041, 3042, 3043, 3044, 3045, 3046, 3047, 3048, 3049, 3049, 3050, 3051, 3052, 3053, 3054, 3055, 3056, 3057, 3058, 3059, 3059, 3060, 3061, 3062, 3063, 3064, 3065, 3066, 3067, 3068, 3069, 3069, 3070, 3071, 3072, 3073, 3074, 3075, 3076, 3077, 3078, 3079, 3079, 3080, 3081, 3082, 3083, 3084, 3085, 3086, 3087, 3088, 3089, 3089, 3090, 3091, 3092, 3093, 3094, 3095, 3096, 3097, 3098, 3099, 3099, 3100, 3101, 3102, 3103, 3104, 3105, 3106, 3107, 3108, 3109, 3109, 3110, 3111, 3112, 3113, 3114, 3115, 3116, 3117, 3118, 3119, 3119, 3120, 3121, 3122, 3123, 3124, 3125, 3126, 3127, 3128, 3129, 3129, 3130, 3131, 3132, 3133, 3134, 3135, 3136, 3137, 3138, 3139, 3139, 3140, 3141, 3142, 3143, 3144, 3145, 3146, 3147, 3148, 3149, 3149, 3150, 3151, 3152, 3153, 3154, 3155, 3156, 3157, 3158, 3159, 3159, 3160, 3161, 3162, 3163, 3164, 3165, 3166, 3167, 3168, 3169, 3169, 3170, 3171, 3172, 3173, 3174, 3175, 3176, 3177, 3178, 3179, 3179, 3180, 3181, 3182, 3183, 3184, 3185, 3186, 3187, 3188, 3189, 3189, 3190, 3191, 3192, 3193, 3194, 3195, 3196, 3197, 3198, 3199, 3199, 3200, 3201, 3202, 3203, 3204, 3205, 3206, 3207, 3208, 3209, 3209, 3210, 3211, 3212, 3213, 3214, 3215, 3216, 3217, 3218, 3219, 3219, 3220, 3221, 3222, 3223, 3224, 3225, 3226, 3227, 3228, 3229, 3229, 3230, 3231, 3232, 3233, 3234, 3235, 3236, 3237, 3238, 3239, 3239, 3240, 3241, 3242, 3243, 3244, 3245, 3246, 3247, 3248, 3249, 3249, 3250, 3251, 3252, 3253, 3254, 3255, 3256, 3257, 3258, 3259, 3259, 3260, 3261, 3262, 3263, 3264, 3265, 3266, 3267, 3268, 3269, 3269, 3270, 3271, 3272, 3273, 3274, 3275, 3276, 3277, 3278, 3279, 3279, 3280, 3281, 3282, 3283, 3284, 3285, 3286, 3287, 3288, 3289, 3289, 3290, 3291, 3292, 3293, 3294, 3295, 3296, 3297, 3298, 3299, 3299, 3300, 3301, 3302, 3303, 3304, 3305, 3306, 3307, 3308, 3309, 3309, 3310, 3311, 3312, 3313, 3314, 3315, 3316, 3317, 3318, 3319, 3319, 3320, 3321, 3322, 3323, 3324, 3325, 3326, 3327, 3328, 3329, 3329, 3330, 3331, 3332, 3333, 3334, 3335, 3336, 3337, 3338, 3339, 3339, 3340, 3341, 3342, 3343, 3344, 3345, 3346, 3347, 3348, 3349, 3349, 3350, 3351, 3352, 3353, 3354, 3355, 3356, 3357, 3358, 3359, 3359, 3360, 3361, 3362, 3363, 3364, 3365, 3366, 3367, 3368, 3369, 3369, 3370, 3371, 3372, 3373, 3374, 3375, 3376, 3377, 3378, 3379, 3379, 3380, 3381, 3382, 3383, 3384, 3385, 3386, 3387, 3388, 3389, 3389, 3390, 3391, 3392, 3393, 3394, 3395, 3396, 3397, 3398, 3399, 3399, 3400, 3401, 3402, 3403, 3404, 3405, 3406, 3407, 3408, 3409, 3409, 3410, 3411, 3412, 3413, 3414, 3415, 3416, 3417, 3418, 3419, 3419, 3420, 3421, 3422, 3423, 3424, 3425, 3426, 3427, 3428, 3429, 3429, 3430, 3431, 3432, 3433, 3434, 3435, 3436, 3437, 3438, 3439, 3439, 3440, 3441, 3442, 3443, 3444, 3445, 3446, 3447, 3448, 3449, 3449, 3450, 3451, 3452, 3453, 3454, 3455, 3456, 3457, 3458, 3459, 3459, 3460, 3461, 3462, 3463, 3464, 3465, 3466, 3467, 3468, 3469, 3469, 3470, 3471, 3472, 3473, 3474, 3475, 3476, 3477, 3478, 3479, 3479, 3480, 3481, 3482, 3483, 3484, 3485, 3486, 3487, 3488, 3489, 3489, 3490, 3491, 3492, 3493, 3494, 3495, 3496, 3497, 3498, 3499, 3499, 3500, 3501, 3502, 3503, 3504, 3505, 3506, 3507, 3508, 3509, 3509, 3510, 3511, 3512, 3513, 3514, 3515, 3516, 3517, 3518, 3519, 3519, 3520, 3521, 3522, 3523, 3524, 3525, 3526, 3527, 3528, 3529, 35</p>		

# Segmentation du canvas (pour les blocs)

XY-Cut (Projection d'histogramme réciproque)

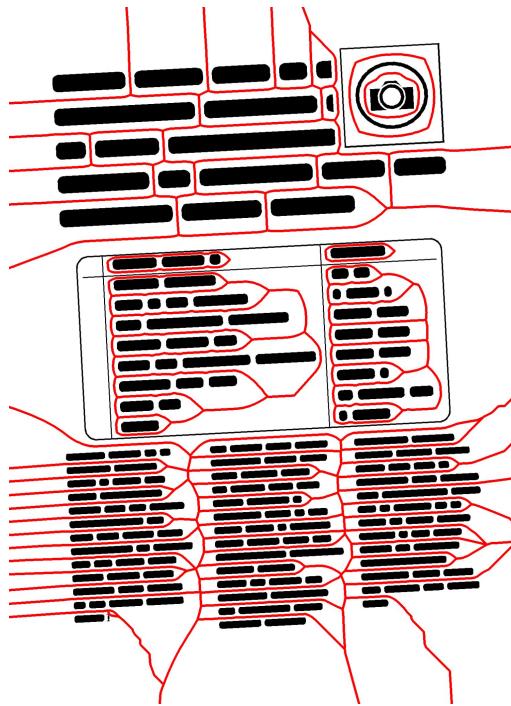


Smearing (ouverture/fermeture morphologique)



# Segmentation du canvas (pour les lignes)

## Ligne de partage des eaux



# Chaîne de traitement pour l'extract. et la reco. des entrées



# Reconnaissance du texte (OCR)

Comparaison de plusieurs systèmes OCR sur notre jeu de données “*A Dataset of French Trade Directories from the 19th Century (FTD)*”

DOI [10.5281/zenodo.6394464](https://doi.org/10.5281/zenodo.6394464)

## Systèmes testés

- [Tesseract v4](#) (v5 pas testée)
- [PERO OCR](#) (code Github + modèle 2020 auteurs)
- [Kraken OCR](#) (modèle générique EN imprimé)

Pas de ré-entraînement.

Bottin 1820  
**Dufort, bottier, Palais-R., gal. vitrée, 215,**  
295

Bottin 1827  
**Baleste, chef aux domaines, S.-Georges, 17.**

Bottin 1837  
**Cattois, pharmac., Bretagne, 46.**

Bottin 1854  
**Fontaine, draperies, Neuve-des-Petits-Champs, 2.**

Cambon Almgene 1841  
**Aron Javal (L.) art. de Paris, r. des Bour-  
donnais, 17.**

Deflandre 1828  
**DEVILLERS, r. Croix-des-Pet.-Champs, 25.  
Cordonn.**

Deflandre 1829  
**Huguenin, épic., r. de Valois, 8, Pal.-Royal.**

Didot 1851  
**Viéville, fab. de boutons, Aumaire, 48, et place  
St-Nicolas-des-Champs, 2.**

# Reconnaissance du texte (OCR)

## Résultats

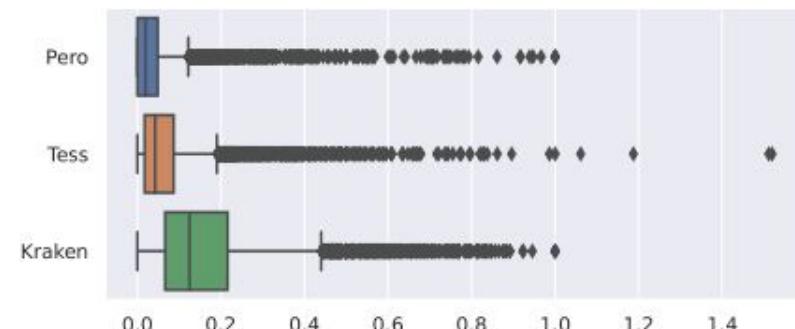
- Très bonne performance de PERO OCR
- Manque de modèles publics pour Kraken (en évolution avec [HTR-United](#))

Autres systèmes libres compétitifs testés depuis

- [Microsoft UniLM/TrOCR](#)
- [Mindee docTR](#)

→ OCR plutôt performant, résultats exploitables

	PERO OCR	Tesseract	Kraken
CER	3.78%	6.56%	15.72%



# Chaîne de traitement pour l'extract. et la reco. des entrées

Pré-traitement image  
(correction géométrique,  
réduction de bruit...)

Segmentation du canvas  
Classification des blocs

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

Transcription

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

Extraction des  
informations clés

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



# Reco. d'entités nommées (*Named Entity Recognition*)

*Identifier et classer des expressions d'un ou plusieurs mots d'un texte en catégories prédéfinies : personne, lieu, organisation, date, prix...*

Entités spécifiques dans le cas des annuaires.

*Exemple:*

Ravrio et comp., fabr. de bronzes et curiosités,  
r. Richelieu, 93; la fabrique rue Montmartre, 161.

PERSON ACTIVITY  
LOCATION CARDINAL FEATURE TYPE LOCATION CARDINAL

# Difficultés : Résultats OCR bruités

Gavarret \*, prof. de physique à la faculté de médecine, Grenelle-St-Germain, 49.

ravarret #, prof. de physique à la faculté de médecine, Grenelle-St-Germain, 49.

Tesseract v4 output

Duffaut, chaudronnier, r. de la Sourdière,  
3<sup>e</sup>

Daflant, c'audronnier, v, de la RARES  
Ge OO a x

Tesseract v4 output

= erreurs

Besoin de robustesse au bruit OCR, voire de post-correction.

# Difficultés : Variation de la structure des entrées

Raison soc.      Activité      Adresse

Durand jeune; pour bas, Charen-  
ton, 12 ancien. 18. \*

No rue

Prévost-Guillaume, f. ta., r. N.-St.-Mart.,  
28.

Lefranc Méquignon et co., satins turcs, prunelle, satins  
et draps de soie, gros de Naples, toiles, coutils, galons,  
rubans, coulisses et lacets pour chaussures de dames,  
sommières, flanelles et molletons de soie pour fourru-  
res, r. des Prouvaires, 32.

Planche , R. de Poitou, 9.-H. Armé.

Baronnat frères, soies teintes et  
écrues, fil-Denis, 257, passage du  
Renard ; maison à Lyon, r. Cen-

Jamain, orangiste, ⓁS. H.1831, Fos-  
sés-St-Marcet, 12.

Appert fils, verres et cristaux,  
21-23 Jour.

Mabire, Lourcine, 124.

Besoin d'un système robuste aux **variations** et **bruit syntaxique**, via un **apprentissage** à base d'**exemples**.

# Approche NER face aux données bruitées

Utilisation d'un modèle NER Transformer **CamemBERT** déjà entraîné, et spécialisation à nos données :

- Pré-entraînement sur données OCR brutes (bruitées)
- Entraînement supervisé sur données OCR bruitées alignées sur la référence

⇒ Gains significatifs en performance, F-score de détection > 94% sur test set

Publication associée :

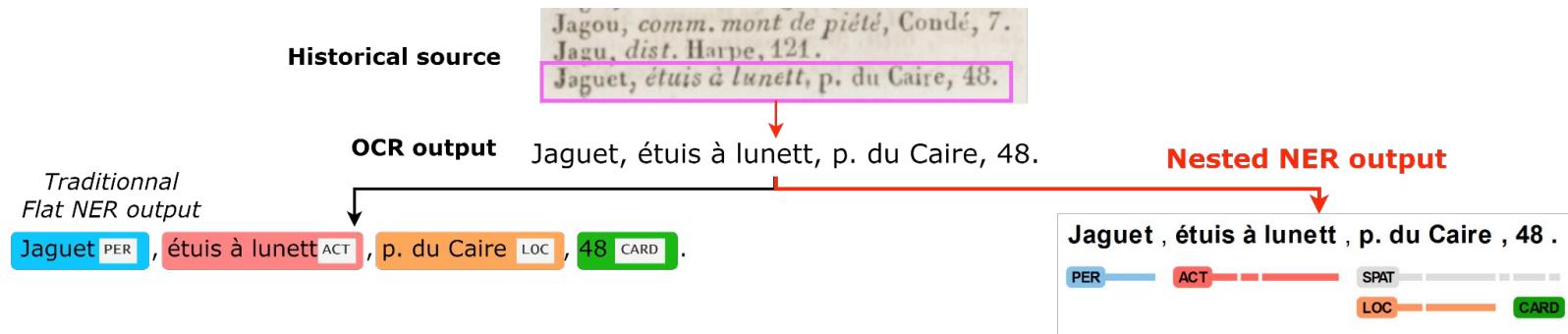
*N. Abadie, E. Carlinet, J. Chazalon et B. Duménieu, “A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories”, in Proc. DAS 2022, <https://github.com/soduco/paper-ner-bench-das22>*

# Expérience NER imbriqué (non déployée)

Publication associée :

S. Tual, N. Abadie, E. Carlinet, J. Chazalon, et B. Duménieu, "A Benchmark of Nested NER Approaches in Historical Structured Documents", in Proc. ICDAR 2023,  
<https://github.com/soduco/paper-nestedner-icdar23-code/>

⇒ Extraction riche possible, performance inchangée



# En pratique...

Extraction “v2”, sur ~110 ouvrages, début 2023

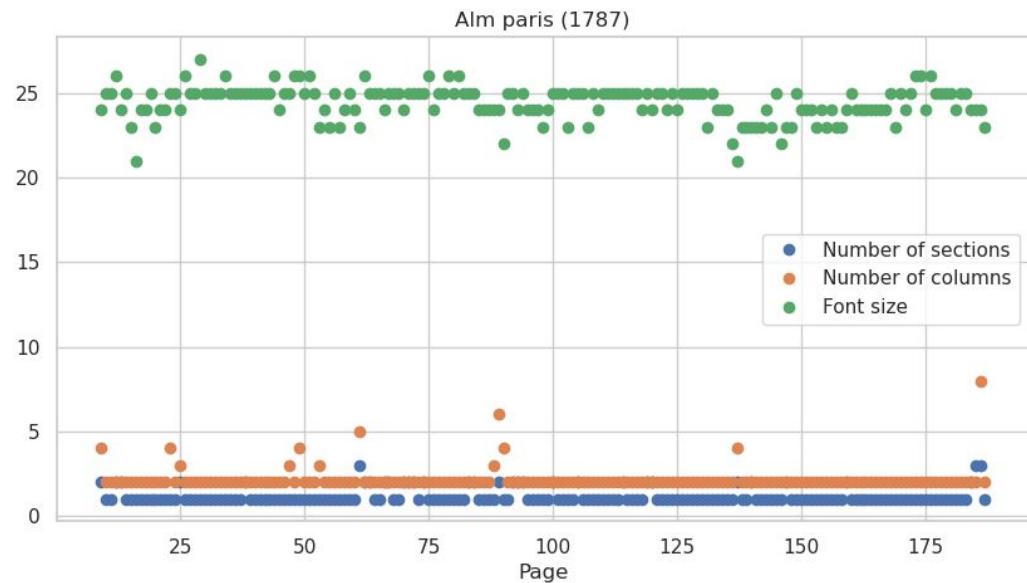
- 9 825 018 entrées extraites
- Environ 74% d'entrées exploitables, de la forme :

**PERSON [ACT] LOC [CARDINAL]**

# Amélioration 1 : Validation de la cohérence inter-pages

Mécanisme en 2 passes :

1. Traiter les images séparément, et calculer certains indicateurs clés : taille des caractères, nombre de colonnes...
2. En déduire les paramètres à appliquer pour les images incohérentes ; relancer les traitements pour les pages / images concernées



# Amélioration 2 : Traitement du texte en flux (en cours)

**Problèmes attaqués :**

1. Comment tirer profit à la fois de l'**information visuelle et textuelle** pour mieux **séparer les entrées** ?
2. Comment s'affranchir des **sauts de colonne** et de **pages** ?
3. Est-il possible de **séparer les entrées** et faire le **NER simultanément** ?

**Opportunité** : documents majoritairement textuels, ordre de lecture simple

**Solution** : un modèle de langage **mixte visuel et textuel**

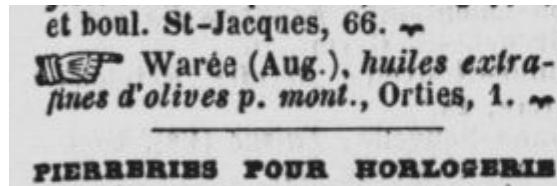


Image originale

```
<page_break><space> et boul. St-Jacques, 66. <space>
<line_break> <space> Warée (Aug.), huiles extra-<space>
<line_break> <space> fines d'olives p. mont., Orties, 1. <space>
<line_break> <space> — <space>
<line_break> <space> PIERRERIES POUR HORLOGERIE <space>
```

Flux d'entrée enrichi pour le système “NER+sép. entrées”

## Amélioration 2 : Traitement du texte en flux (en cours)

Capacité à extraire des entrées à cheval sur plusieurs pages attrante.

Mais une tendance à fusionner les entrées.

⇒ Des résultats intéressants mais à améliorer.

- À combiner avec une approche exploitant la position des éléments type LayoutLM.
- Travaux en cours.



## Au final...

Extraction “v4”, avec séparation simple des entrées fusionnées et réassociation “nom de voie, numéro”, sur 357 listes réparties entre 144 ouvrages et 192 PDFs :

- 22 743 928 entrées
- 23 728 378 entités PER
- 2 106 808 entités TITRE
- 14 663 426 entités ACT
- 27 618 637 addresses (combinaisons de CARDINAL + LOC, ou LOC seules)  
dont 96.3% ont été géocodées

**Soit environ 27 millions de points à placer sur une carte !**

# Géocodeur historique

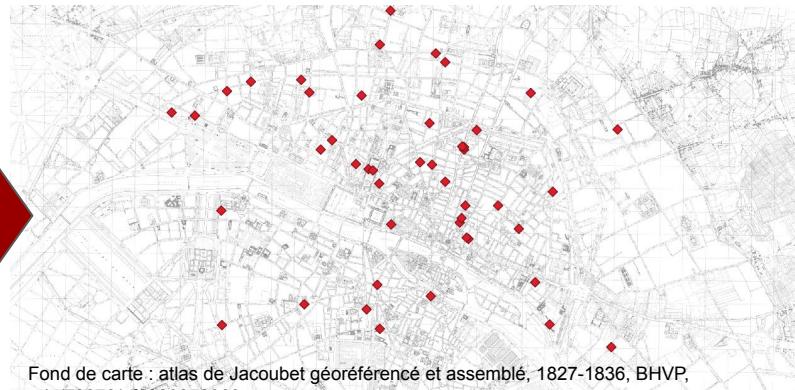
(Lamy\_1840) Annuaire général du commerce, judiciaire et administratif de France et des principales villes du monde, v. 128.

ACA

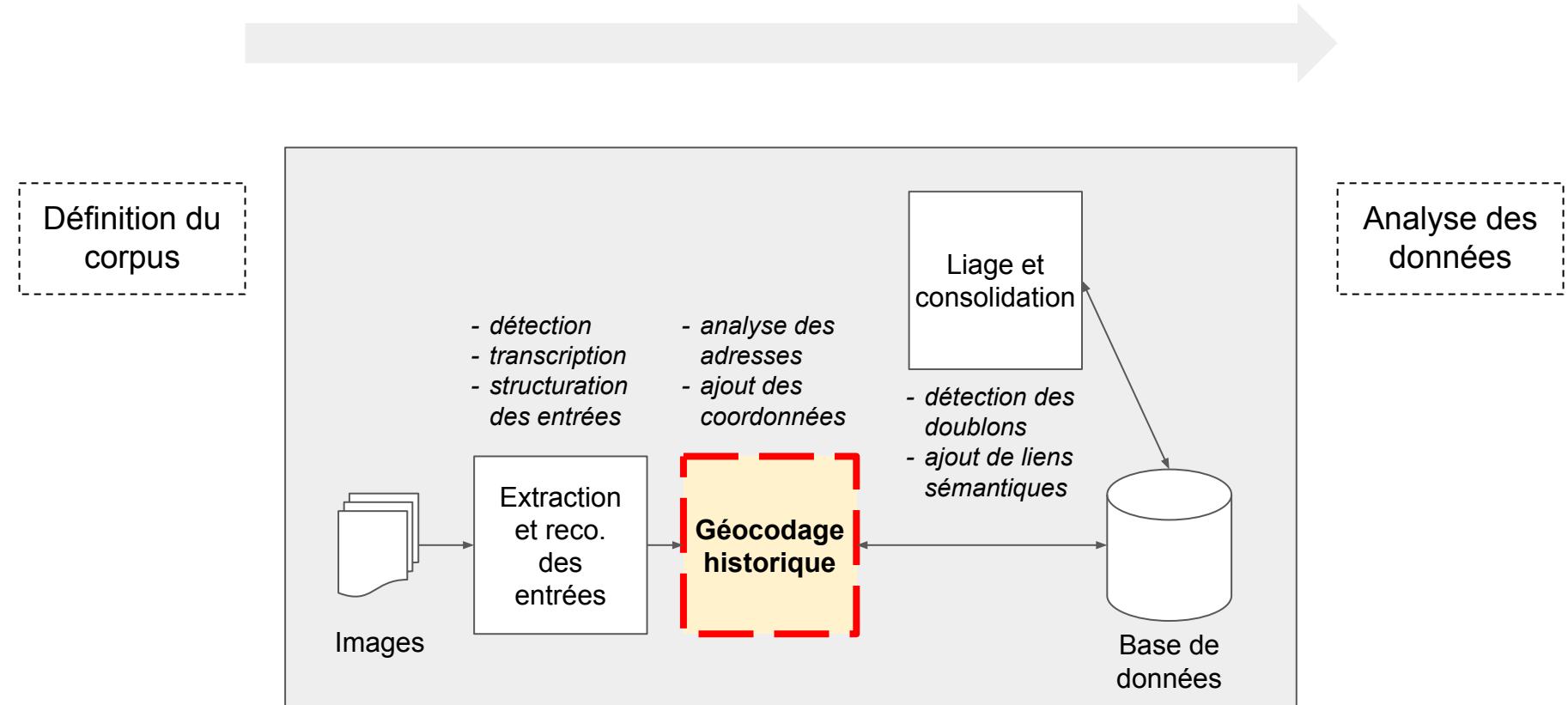
Abadie, perruquier, Miroménil, 21.  
Abadie (A.), él., pharmacien, Ferme, 10.  
Abadie et Degorge, tailleur, Bons-Enfants, 26.  
Abancourt (Vte d') ✽, pair, président à la cour des comptes, Assas, 3 bis.



[https://api.geohistoricaldata.org/directories/export.geojson?source.pdf\\_id=eq.Lamy\\_1840&source.pdf\\_view=eq.0128&](https://api.geohistoricaldata.org/directories/export.geojson?source.pdf_id=eq.Lamy_1840&source.pdf_view=eq.0128&)



# Vue d'ensemble du processus



# Chaîne de traitement

## #1 OCR

> Aaron, bronzes et pendules,  
passage Choiseul, 72 et 74.

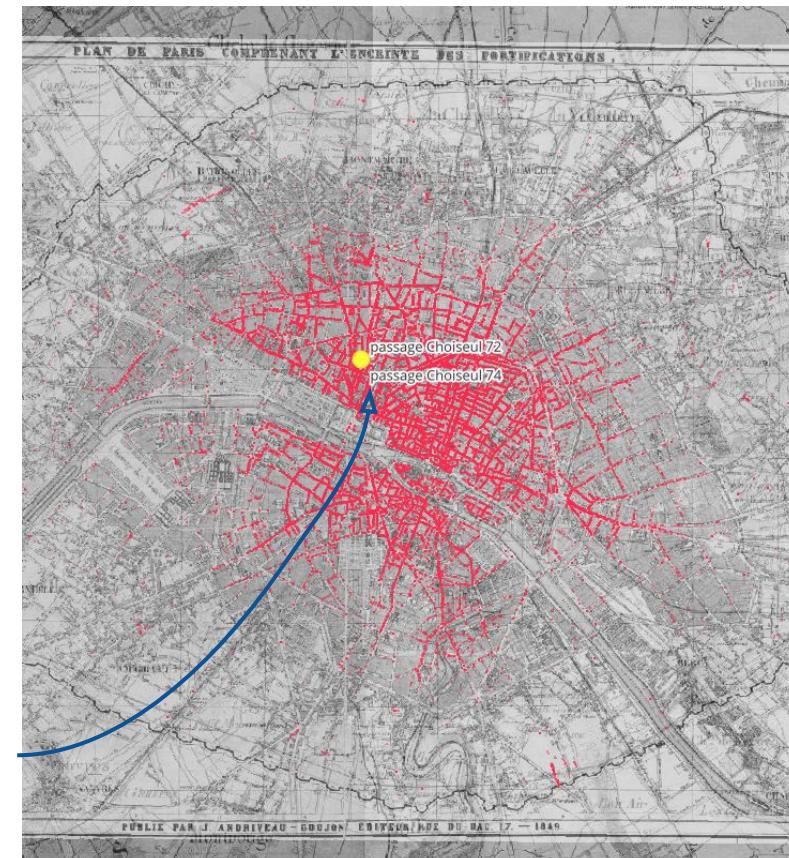
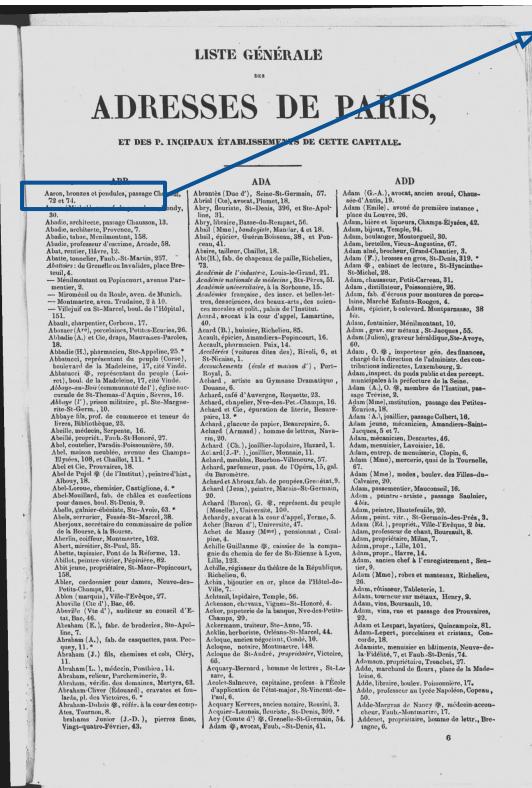
## #2 NER

> Aaron, bronzes et pendules,  
passage Choiseul, 72 et 74.

## #3 Restructuration

Address #1 : passage Choiseul 72

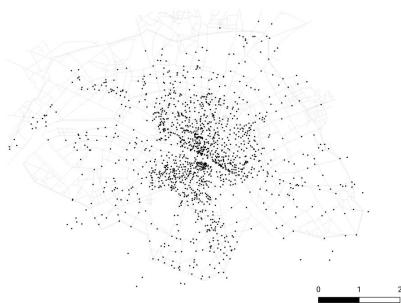
Address #2 : passage Choiseul 74



# Géocodeur historique: un moteur de recherche exploitant les données des atlas de Paris

Recherche multicritères dans les données vectorielles des plans :

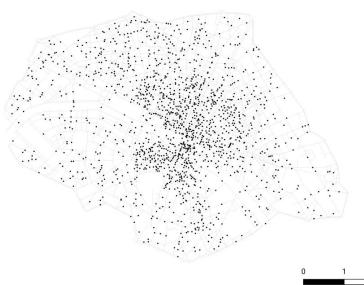
- matching flou du nom + numéro de rue
- distance temporelle



Atlas de Verniquet, 1784-1795  
Filaire des rues



Atlas de Jacoubet, 1827-1836  
Filaire des rues & points adresse

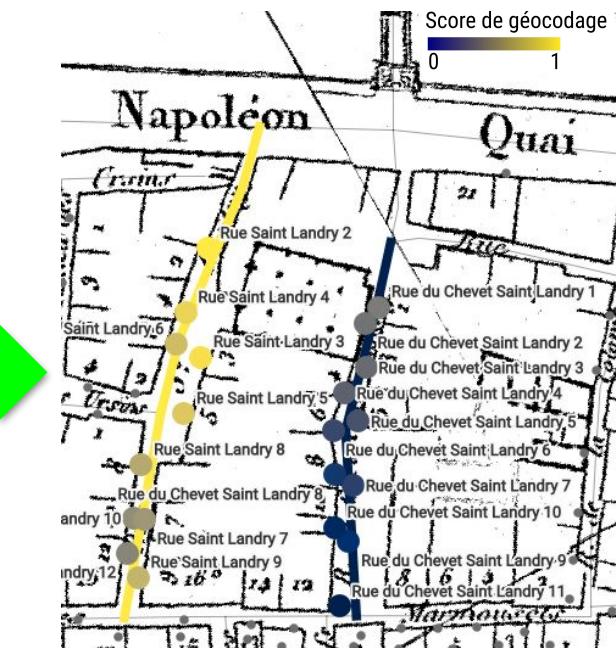
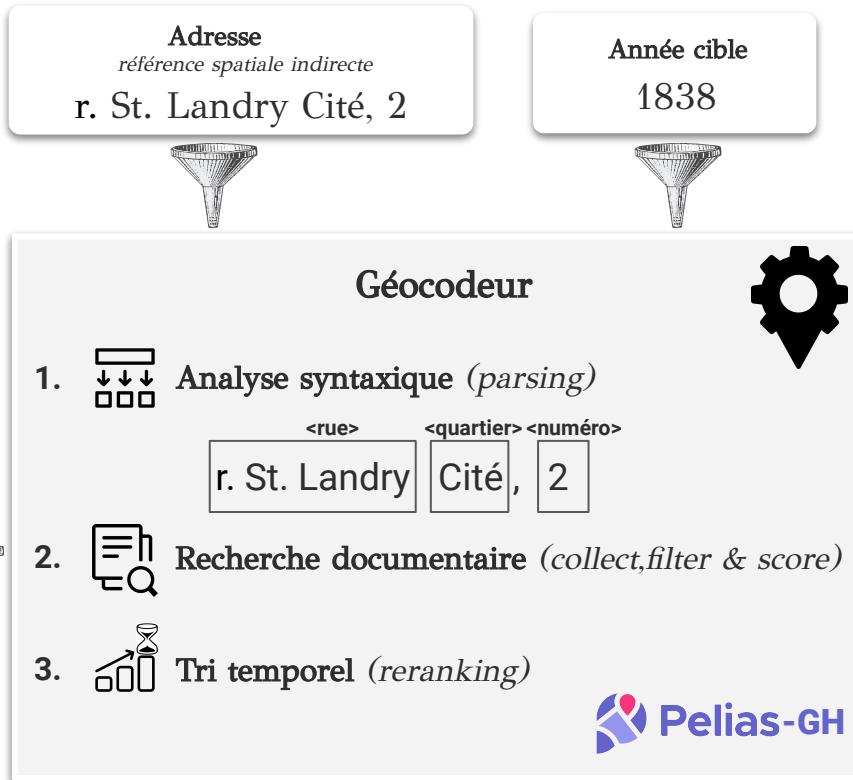


Plan Andriveau-Goujon, 1849  
Filaire des rues



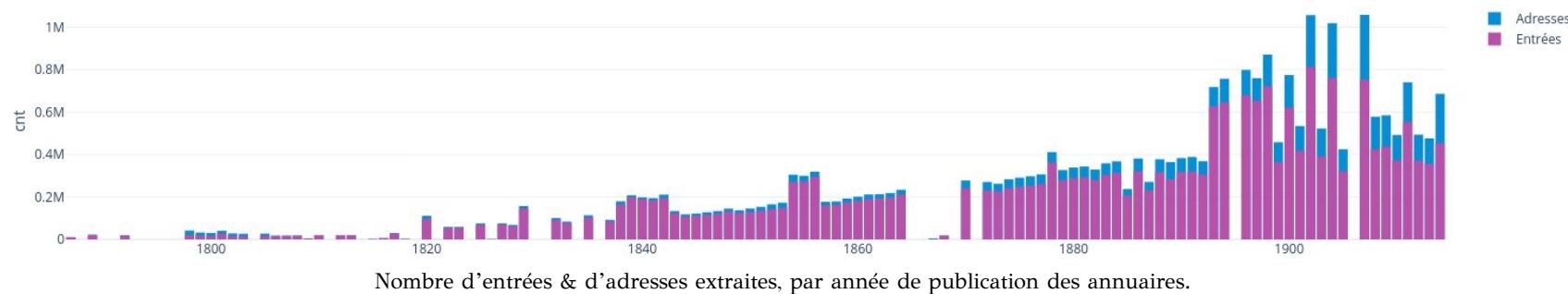
Atlas municipal des 20 arrds., 1888  
Filaire des rues & points adresse

# Un géocodeur historique sensible aux temps valides

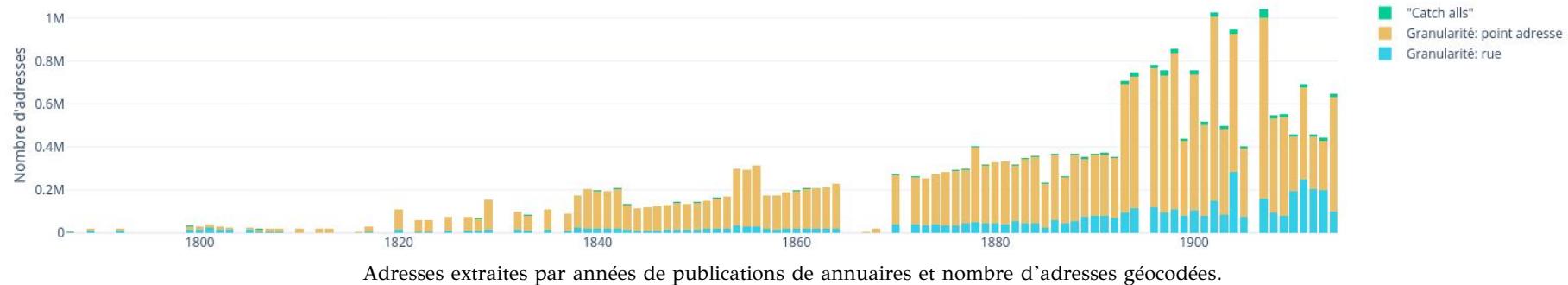


# Géocodage des annuaires : aperçu des résultats

V2, octobre 2023 : 96% d'entrées géocodées (V2 septembre 2022 : 66%)



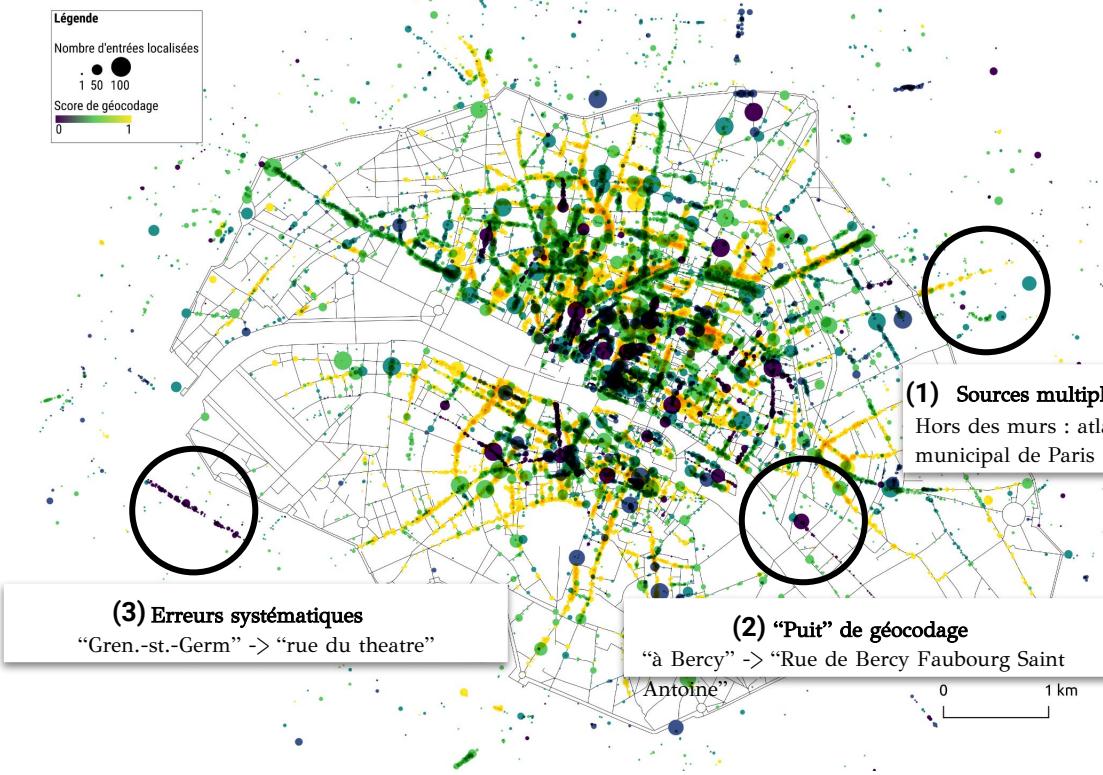
Nombre d'entrées & d'adresses extraites, par année de publication des annuaires.



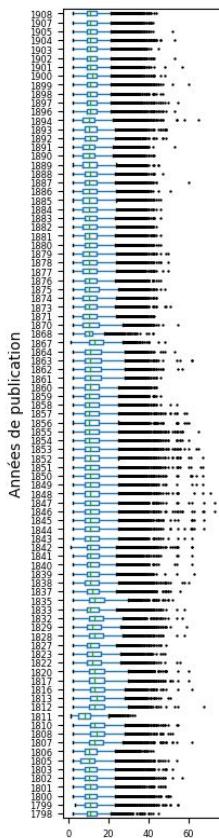
Adresses extraites par années de publications de annuaires et nombre d'adresses géocodées.

# Résultats du géocodage des annuaires

Entrées géocodées pour l'année 1838 : Almanach du commerce de Paris [...] (Bottin) et Annuaire général du commerce, de l'industrie et de l'agriculture de France [...] (Henrichs).



Longueurs des énoncés d'adresses en nombre de caractères et exemples typiques



r. d. des. 123, che 38,  
Bar 63, NAS, r. d., bis,  
Tri, imp, BaC 104, Dou, St-, omb,  
Bac 142., r.,  
e N T, L Ver

Mauvaises-Paroles 15

quai aux Fleurs, au coin de la rue  
de la Cité

avenue de Neuilly, sur la pelouse  
près la rue du chemin de Versailles

Nombre de caractères

# Liage et consolidation

même personne

succession

déménagement

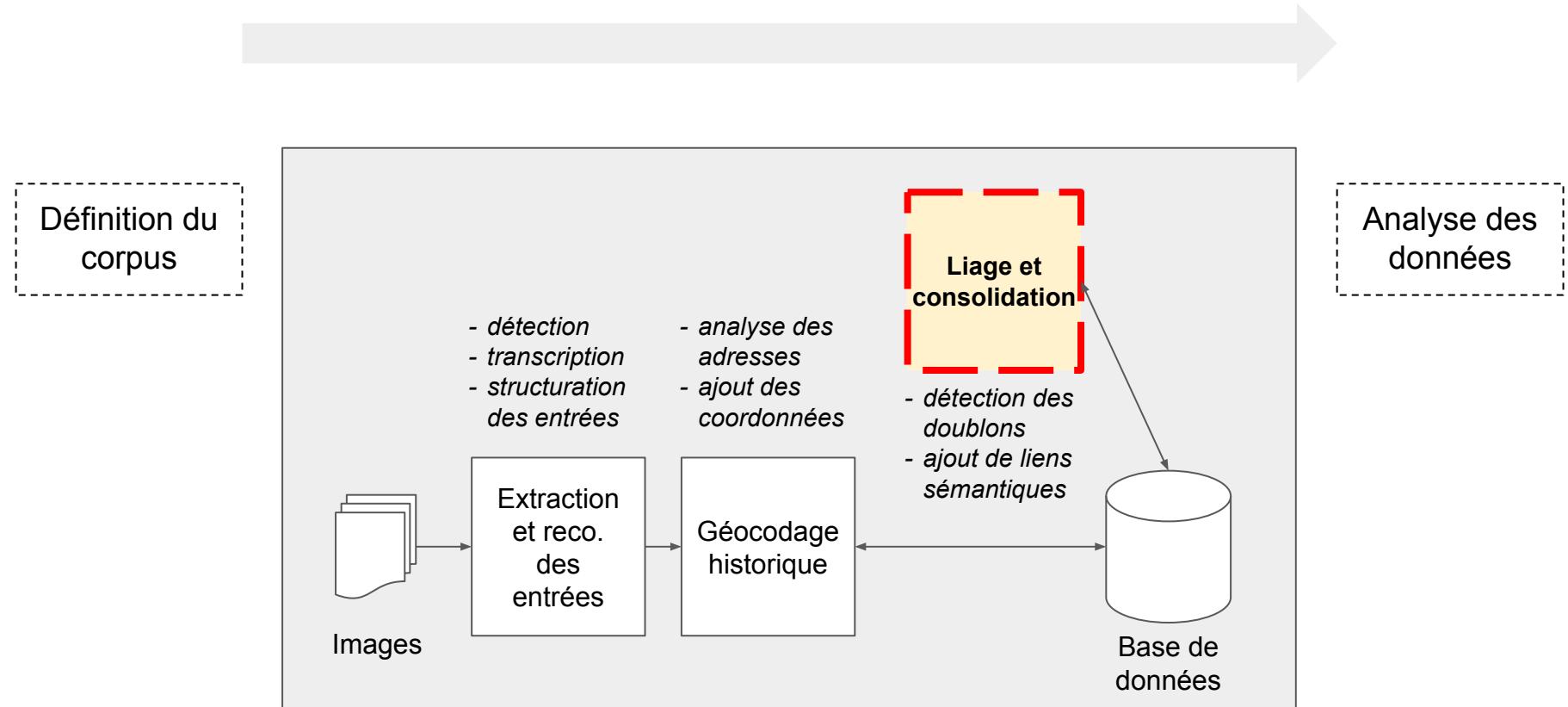
Bibliothèque Ste-Geneviève, Clotilde, 1.
Bibliothèque de la Ville, quai d'Austerlitz , 33 (provisoirement).
Biblique protestante (Société), Moulins, 16.
Bibron, aide-natural., au Muséum d'hist. nat.
Bibus, tailleur. Roule. 21.
Bibus, tailleur, Richelieu, 31.
Bical et Dorre, fab. de socques, Vertbois, 14.
Bicant (Mme), fondeur en cuivre , cour de la Corderie-du-Temple, 26.
Bichard (Mme), Nve-de-Luxembourg, 17.
Bichard, tabacs et eau-de-vie, Faub.-St-Martin, 45.

Bibliothèque Ste-Geneviève, Clotilde, 1.
Bibliothèque de la Ville, quai d'Austerlitz , 33 (provisoirement).
Biblique protestante (Société), Moulins, 16.
Bibron, aide-natural., au Muséum d'hist. nat.
Bibus, tailleur. Roule. 21.
Bibus, tailleur, Richelieu, 31.
Bical, fab. de jouets, Montmorency, 33.
Bical et Dorre, fab. de socques, Vertbois, 14.
Bican (Vve) et fils, fondeur en cuivre , place de la Corderie-du-Temple, 26.
Bicel, épicier, marché d'Aguesseau, 15.
Bichard (Mme), Nve-de-Luxembourg, 17.
Bichard, tabacs et eau-de-vie , Faub.-St-Martin, 45.

Bibliothèque Ste-Geneviève, rue des Sept-Voies et place du Panthéon.
Bibliothèque de la Ville, quai d'Austerlitz , 33 (provisoirement).
Bibus, tailleur, Richelieu, 31.
Bical, fab. de jouets, Montmorency, 33.
Bical et Doire, fab. de socques, Vert-Bois, 14.
Bican (Vve) et fils, fondeurs en cuivre , place de la Corderie-du-Temple, 26.
Bicel, épicier, Marché-d'Aguesseau, 15.
Bichard, tabac et eau-de-vie , Faub.-St-Martin, 45.

Bibliothèque Ste-Geneviève, rue des Sept-Voies et place du Panthéon.
Bibliothèque de la Ville, quai d'Austerlitz , 33 (provisoirement).
Bibolet, relieur, passage Sainte-Marie-Saint-Germain, 10.
Bibonne, architecte, Magasins, 12.
Bibron, aide-naturaliste au Jardin-des-Plantes, Cuvier, 20.
Bibus, tailleur, Richelieu, 31.
Bical, fab. de jouets, Montmorency, 33.
Bican (Vve) et fils, fondeurs en cuivre , place de la Corderie-du-Temple, 26.
Bicel, épicier, Marché-d'Aguesseau, 15.
Bichard, tabac et eau-de-vie , Faub.-St-Martin, 45.

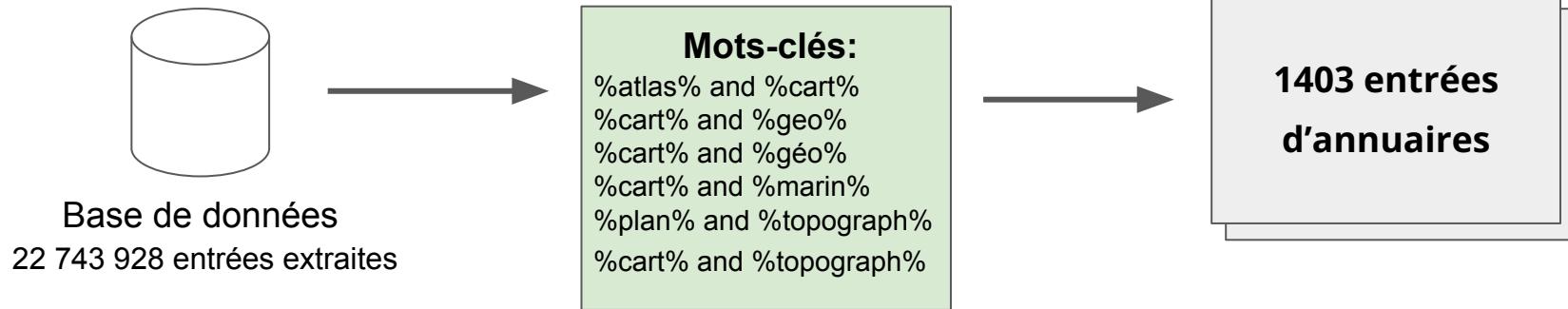
# Vue d'ensemble du processus



# Création de graphes de connaissances géohistoriques professionnels

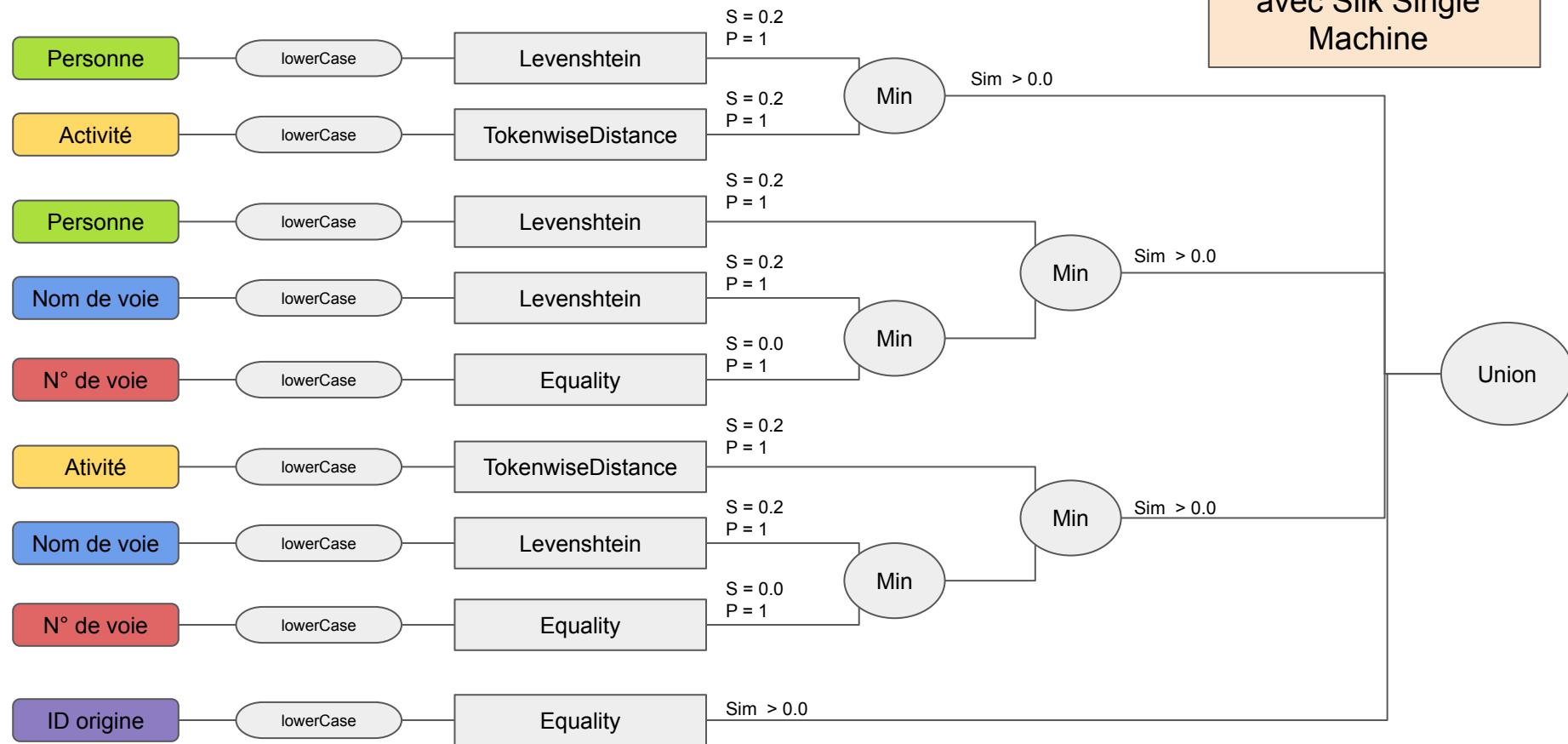
1. Sélection d'un sous-ensemble d'entrées d'annuaires dans la base de données à l'aide d'une liste de mots-clés relatifs à un type d'activité.
2. Création de ressources RDF à partir des enregistrements de la base de données.
3. Utilisation d'une méthode de liage numérique pour comparer les entrées d'annuaires:
  - Noms similaires, activités similaires, adresses similaires → continuation
  - Noms similaires, activités similaires, adresses différentes → déménagement
  - Noms différents, activités similaires, adresses similaires → succession
4. Visualisation du graphe de connaissances géohistorique et évaluation qualitative des liens créés.

## Ex. Sélection des entrées sur les graveurs et marchands de cartes



# Critères de liage

Mise en oeuvre  
avec Silk Single  
Machine



# Résultats du liage des entrées

Exemple des photographes

## Méthode numérique

**20 362 liens**

Paramétrage complexe :  
identifier les seuils de tolérance  
pertinents

## Raisonnement

**16 046 liens**

Propagation des liens  
`owl:sameAs` par transitivité



Comparaison adaptée aux chaînes de  
caractères résultant de l'OCR

## BILAN



**36 408 liens** d'équivalence distincts entre les ressources du graphe

Graphe géohistorique : <https://dir.geohistoricaldata.org/sparql>

# Visualisation et interprétation des résultats



Visualisation des graphes géohistoriques construits à partir des entrées d'annuaires du commerce de Paris (XIXème siècle)

Aide [?](#)

## Dataset

Graveurs et marchands de cartes et plans

[Statistiques du dataset](#)

## Filtres

### Propriétés

#### Raison sociale

Ex : nadar

#### Description

Ex : photo

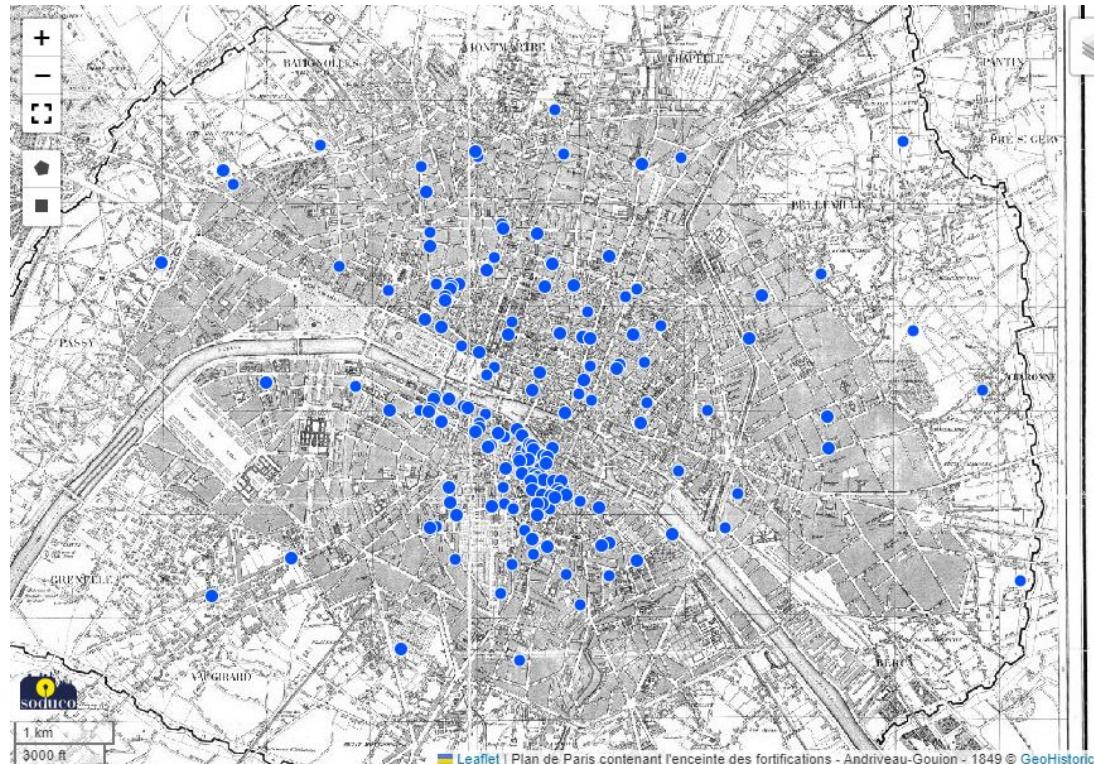
#### Adresse

Ex : rivoli

#### Période

[https://soduco.geohistoricaldata.org/atelier\\_graphes\\_geohistoriques\\_annuaires/](https://soduco.geohistoricaldata.org/atelier_graphes_geohistoriques_annuaires/)

# Visualisation et interprétation des résultats



A historical map of Paris from 1849, showing the city's layout and surrounding fortifications. Numerous blue dots are scattered across the map, primarily concentrated in the central and northern parts of the city, representing locations of interest or search results. The map includes street names like BASTILLE, BONNEUIL, CHAMPEAUX, PANTIN, PRE-Saint-GERVAS, VERNON, GRENELLE, VAUGIRARD, and PASSY. A scale bar at the bottom left indicates 1 km and 3000 ft. A legend in the top right corner provides examples for various search terms: Ex : nadar, Description, Ex : photo, Adresse, and Ex : rivoli. Below these is a timeline labeled "Période" with a green bar highlighting the years 1840 to 1890. A note explains that the temporal filter allows for dynamic updates of the displayed points without a new search. A "Lancer la recherche" button is located at the bottom right.

Ex : nadar  
Description  
Ex : photo  
Adresse  
Ex : rivoli

Période

1840 1890

1790 1821 1853 1884 1915

Le filtre temporel permet de faire varier l'affichage des points préalablement chargés sur la carte sans lancer une nouvelle recherche.  
Données chargées pour la période 1840-1890.

Localisation

Dessinez l'emprise de votre zone de recherche avec l'outil de dessin disponible sur la carte.

Lancer la recherche

[https://soduco.geohistoricaldata.org/atelier\\_graphes\\_geohistoriques\\_annuaires/](https://soduco.geohistoricaldata.org/atelier_graphes_geohistoriques_annuaires/)

# Statistiques sur le graphe géohistorique

Statistiques du jeu de données "Graveurs et marchands de cartes et plans"

Nombre d'entrées d'annuaires

1403

Nombre de ressources RDF

1732

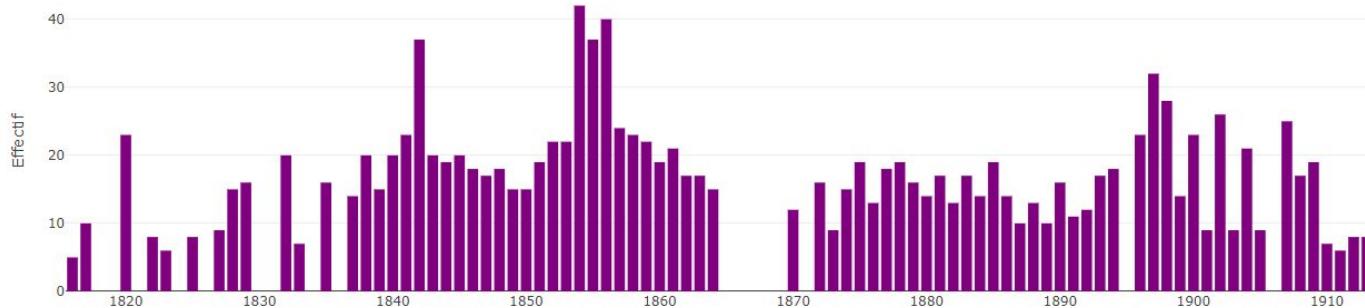
Nombre de triplets RDF

80969

Nombre de liens sameAs entre des ressources différentes

36408

Nombre d'entrées d'annuaires extraites par année



# Suivi individuel des commerces

The screenshot shows a historical map of Paris with various data layers overlaid. A central feature is a modal window for 'Vuillemin' containing detailed metadata about a specific entry.

**Vuillemin**

Adresse (annuaire) : 5 St-Thomas-d'Enfer  
Adresse (géocodage) : 5 Rue Saint Thomas, Paris  
(Source: atlas\_jacoubet\_1836)  
Activité : cartes géographiques  
Année de publication : 1861  
Annuaire : DidotBottin\_1861  
Identifiant de l'entrée : 1d8338bc-8614-54cc-af5-0d412ee29a35

**Frise chronologique**

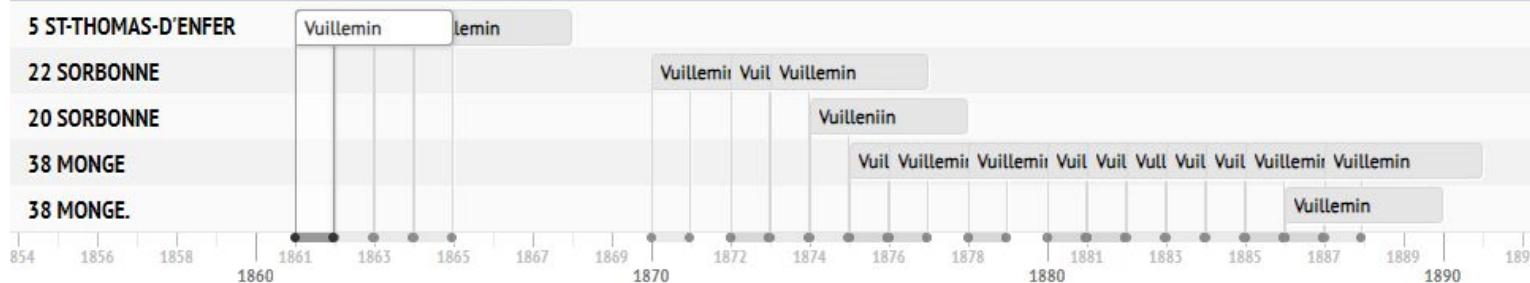
Ex : photo  
Adresse  
Ex : rivoli  
Période  
1790 1915  
1790 1821 1853 1884 1915

Le filtre temporel permet de faire varier l'affichage des points préalablement chargés sur la carte sans lancer une nouvelle recherche.  
Données chargées pour la période 1790-1915.

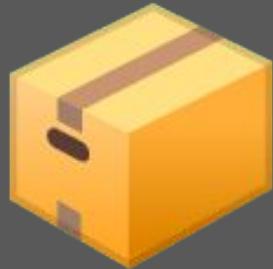
Localisation  
Dessinez l'emprise de votre zone de recherche avec l'outil de dessin disponible sur la carte.

Lancer la recherche

# Suivi individuel des commerces



# Livrables disponibles



## Publications sur le thème “extraction depuis les annuaires”

N. Abadie, E. Carlinet, J. Chazalon et B. Duménieu, “A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories”, in Proc. DAS 2022,  
<https://github.com/soduco/paper-ner-bench-das22>

S. Tual, N. Abadie, E. Carlinet, J. Chazalon, et B. Duménieu, "A Benchmark of Nested NER Approaches in Historical Structured Documents", in Proc. ICDAR 2023, <https://github.com/soduco/paper-nestedner-icdar23-code/>

P. Bernet, J. Chazalon, E. Carlinet, A. Bourquelot et E. Puybareau, “Linear Object Detection in Document Images using Multiple Object Tracking”, in Proc. ICDAR 2023, <https://github.com/EPITAResearchLab/bernet.23.icdar>

S. Tual, N. Abadie, B. Duménieu, J. Chazalon et E. Carlinet. Création d'un graphe de connaissances géohistorique à partir d'annuaires du commerce parisien du 19ème siècle: application aux métiers de la photographie. IC 2023, 34èmes journées francophones d'Ingénierie des connaissances, Strasbourg, France, 3-7 July 2023. <hal-04121643>.  
[https://github.com/soduco/ic\\_2023\\_photographes\\_parisiens](https://github.com/soduco/ic_2023_photographes_parisiens)

## Données sur le thème “extraction depuis les annuaires”

A Dataset of French Trade Directories from the 19th Century (FTD)

DOI [10.5281/zenodo.6394464](https://doi.org/10.5281/zenodo.6394464)

A Dataset of French Trade Directories from the 19th Century for Nested NER task

DOI [10.5281/zenodo.7864174](https://doi.org/10.5281/zenodo.7864174)

Collection SoDUCo sur NAKALA : <https://nakala.fr/collection/10.34847/nkl.abe0gxah>

Annuaires historiques parisiens, 1798-1914. Extraction structurée et géolocalisée à l'adresse des listes nominatives par ordre alphabétique et par activité dans les volumes numérisés  
<https://nakala.fr/10.34847/nkl.98eem49t>

Les datasets par professions incluant les données et les liens sont disponibles sur le SPARQL endpoint du projet : <https://dir.geohistoricaldata.org/>

# Outils et modèles sur le thème “extraction depuis les annuaires”

Pré-traitements et segmentation du canevas : ouverture en cours, détection de segments disponible à  
<https://github.com/EPITAResearchLab/bernet.23.icdar>

OCR : benchmark DAS 2022 : <https://github.com/soduco/paper-ner-bench-das22>

NER : idem + modèles sur HuggingFace Hub, nouveaux prototypes en cours de finalisation

NER imbriqué (code, données et modèles) : <https://github.com/soduco/paper-nestedner-icdar23-code>

Interface Web de visualisation cartographique du graphe des photographes parisiens :  
[https://soduco.geohistoricaldata.org/ic\\_2023\\_photographes\\_pariisiens/](https://soduco.geohistoricaldata.org/ic_2023_photographes_pariisiens/)

Scripts d'extraction d'entrées pour une activité donnée, de liage, de structuration en RDF, de requêtes SPARQL types pour explorer les graphes géohistoriques et interface Web de visualisation cartographique capable d'intégrer plusieurs graphes:

[https://soduco.geohistoricaldata.org/atelier\\_graphes\\_geohistoriques\\_annuaires/](https://soduco.geohistoricaldata.org/atelier_graphes_geohistoriques_annuaires/)

Il reste quelques places pour l'atelier de demain après-midi !

# Travaux en cours & futurs



# Evaluation des étapes de la chaîne de traitement

## Extraction d'entrées :

- échantillonnage → typologie des erreurs
- collection → détection de cas aberrants

**Géocodage** → détection de cas aberrants

**Liage** → évaluation de la cohérence intrinsèque, détection de liens erronés, etc.

**Analyses en aval** → détection de zones d'ombres

Objectif: Aller vers une qualification plus fine des données au regard des problématiques géo-historiques visées.

## Amélioration de la chaîne d'extraction

- Modèles de langue hybrides texte / vision et analyse du texte en flux
- Nouveaux systèmes de transcription (cf présentation C. Kermorvant à suivre)

## Liage en amélioration intensive

- Développement de chaînes de traitement, pour permettre l'extraction de données pour des activités spécifiques (semaine “graphe géo-historique” en oct. 23)
- Liage massif des données qui semble possible à court/moyen terme

## Généralisation de l'approche

- **Ateliers : améliorer la réutilisabilité, la diffusion des données, outils et modèles, pistes de généralisation de la chaîne de traitements...**
- **Mezanno (plan quadriennal BnF) : outils pour constituer un corpus à partir de ressources IIIF et l'annoter de façon semi-automatique, en toute autonomie**
- Plus de tests sur des corpus voisins : annuaires étrangers, de propriétaires, dictionnaires, débats parlementaires...



RÉSERVE

# Distribution des outils

LIB



CLI

API REST

Batch

annotation assistée

## I.2 - Des sources géographiques anciennes à l'analyse géohistorique

### Reconnaissance

Image → Géométries



Scan → Texte

Ravrio et comp., fabr. de bronzes et curiosités, r. Richelieu, 93; la fabrique rue Montmartre, 161.

### Classification

Géométries → Annotations



Texte →  
Entités nommées spatiales

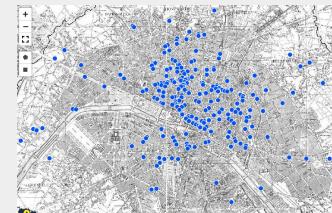
Ravrio et comp. PER , fabr. de bronzes et curiosités ACT ,  
r. Richelieu LOC , 93 CARDINAL ; la fabrique FT  
rue Montmartre LOC , 161 CARDINAL .

### Géo-référencement

Données géométriques →  
Données géographiques



Entités nommées spatiales →  
Données géographiques



### Structuration

Données géographiques →  
Données géohistoriques



Aborder le  
liage ici ?

Données géographiques →  
Données géohistoriques



# Chaîne de traitement automatique : de l'image aux infos sémantiques

Vue PDF 701



Vue PDF 702

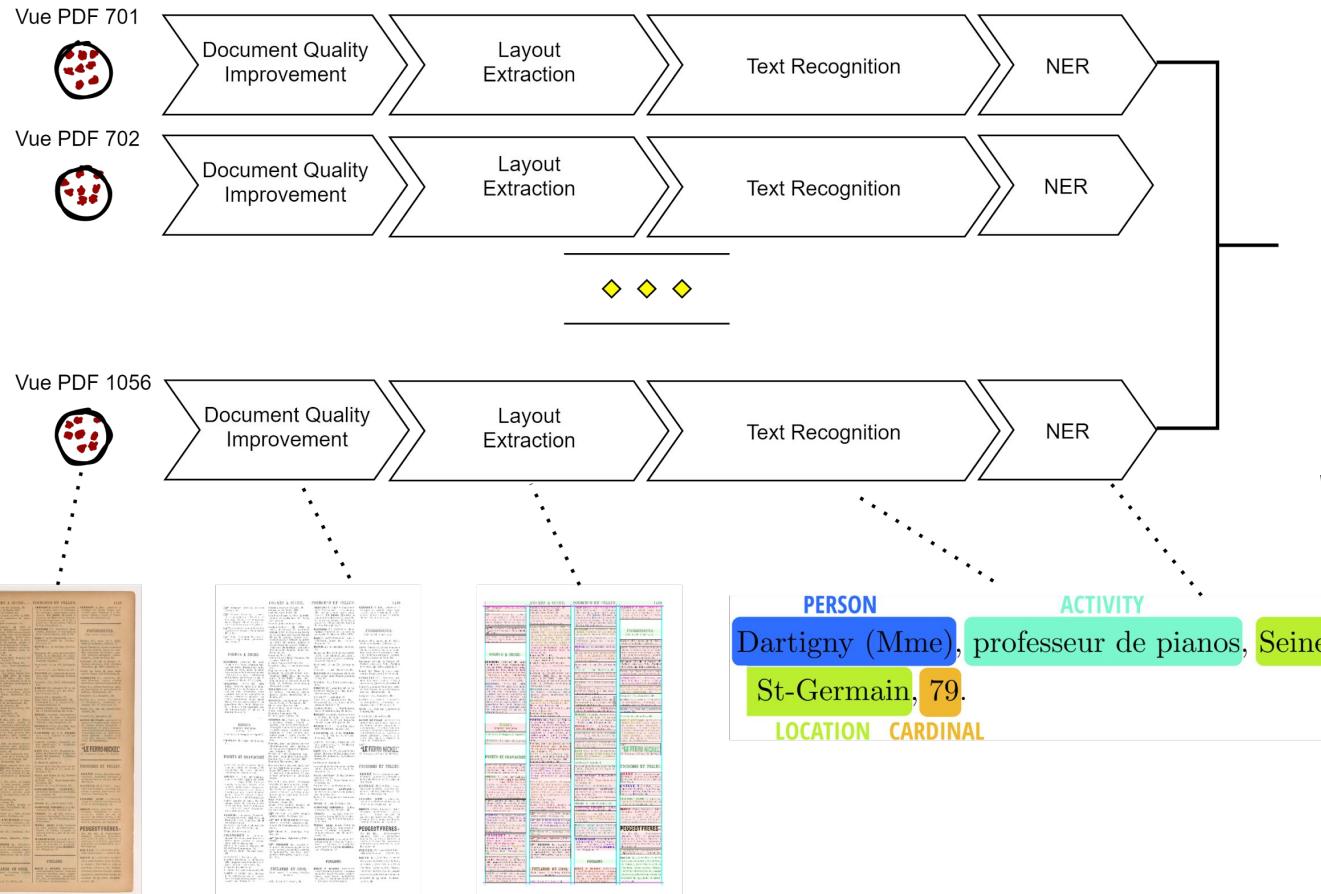


Vue PDF 1056



**PERSON** Dartigny (Mme), professeur de pianos, Seine-  
**ACTIVITY** St-Germain, 79.  
**LOCATION** CARDINAL

# Chaîne de traitement automatique : de l'image aux infos sémantiques



V2 data (2022-06): ~10 M d'entrées

# Chaîne de traitement automatique : ajout des infos spatiales

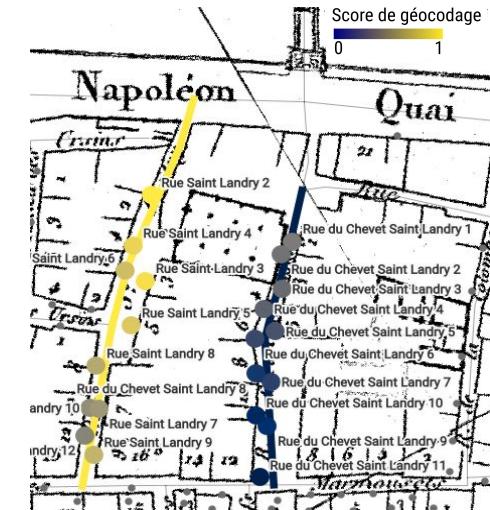
Vue PDF 701



Vue PDF 702



Vue PDF 1056



PERSON

Dartigny (Mme), professeur de pianos, Seine-

St-Germain, 79.

LOCATION CARDINAL

ACTIVITY

# Approche scientifique du projet SoDUCo

**Analyse critique croisée de 2 collections**

*Atlas et annuaires*

**Exploitation des redondances spatiales et temporelles**

*Détection d'erreurs ou de changements, stabilisation de la reconnaissance*

**Extraction d'information pragmatique**

*Assister l'utilisateur expert, viser l'annotation collective dans une certaine mesure*

**Compromis entre généralité et spécificité**

*Viser une solution fonctionnelle sur notre corpus, avec des composants réutilisables*

**Free/open, accessible, indexé, reproductible**

*Dépôts GitHub, <https://geohistoricaldata.org>, jeux de données, publications...*



Directory ViewerError : Network Error [EXPORT](#)

Affichage

Entrée



OCR

NER

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

## MÉCANICIENS.

*lons ronds et demi-ronds et car- rés, et de toute sorte de moulure; spécialité de fabrication de bro- ches en acier et cuivre pour les peignes à tisser draps, voiles et couvertures, toile métallique pour tout ce qui concerne le tissage, M. H. 1849, Amandiers-Popin- court, 19.*

*Auger (Vve), mécanicien, fabr. d'emporte - pièces pour festons de garnitures de robes et de man- telets, emporte - pièces pour car- tonnage, semelles de souliers, etc., tiennent un assortiment de maillets, billets et plombs, pass. de la Tri- nité, 77.*

*Avoine-Bainnée, serrurier-mécan., fab. delets en f.r., ⑩ 1839-14, ⑩ 1819, Boulanger-St-Victor, 22.*

*Biard, quai Jemmapes, 248.*

*Capicrossé pere, tout ce qui a rap- port à la fabrique d'indénnes, Charenton, 58.*

*Barbot, Popincourt, 58.*

*Bardies, Trois-Bornes, 21.*

*Bardon , Deschartes, 38.*

*Bardouïn, horloger-mécanicien, entreprend toute espèce de fa- brication nouvelle, quelle que soit la précision demandée, et cons- truit toutes sortes de machines et outils pour Messieurs les in- venteurs; vis cylindriques et con-iques, découpage à l'écon, fen- dage de roues, pièces détachées fabriquées par procédé mécaniques pour toute espèce d'indus- trie, ⑩ 1849, Ecouffes, 20. au Mar, près le marché des Blances-Manteaux.\**

*Baudat, constructeur de mécani- cunes à scier le placage, le bois ou*

*Becker, mécanicien-graveur, bre- vété d'invention (sans garantie du gouvernement), fabrique spé- ciale de presses à copier, presses à timbre sec et humide, presses à cacherie, presses à percussion pour l'extraction des matières végétales et pour le satinage des papiers, compositeurs pour pa- piers à lettres, pour raisons de commerce, griffes de toutes sortes, et généralement toute la gra- vure sur métal, St-Denis, 380, passage Lemoine.\**

*Boscher, Bondy, 76.*

*Bosquillon \*, ⑩ 1823-27-34,*

*constructeur de perçages accélérés et de mécaniques Jacquard parisienne, Paradis-Poisonnière, 20.*

*Boucher, Saint - Pierre - Popin- court, 18.*

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

*Bouley (Ene), mécanicien, fab. de machines à couper les peaux de lapins, souffleuses de toutes di- mensions marchant à bras et à la vapeur, tours de chapellerie, ⑩ 1849, Francs-Bourgeois-Marais, 3, ci-devant Beaubourg, 59.\**

*Bernier, Ménilmontant, 90.*

*Berthet, Simon-le-Franc, 15.*

*Berton, Neuve-St-Denis, 12.*

*Bertrand, mécanicien, brevete (sans garantie du gouvernement), construit toutes espèces de ma- chines et outils, spécialité de ma- chines pour les fabricants de ma- tières premières de chapellerie, machines à vapeur et tours pour les fabricants de chapeaux; ton- deuses et souffleuses de poils de lièvres ou de lapins, réparation et entretien de machines à va- peur; on trouve toujours chez lui des machines à vapeur toutes prêtes à être mises en place, Vieille-du-Temple, 58, anc. 72.\**

*Beslay (Ch.), ⑩ 1839-1849*

## 97

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

*Bosche aine &c, ingénieur-méca- nicien, inventeur de divers per- fectionnements au métier Jac- quard, ⑩ S.E. 1850, Amandiers- Popincourt, 22, ci-devant St- Maur, 14.\**

*Boscher, Bondy, 76.*

*Bosquillon \*, ⑩ 1823-27-34,*

*constructeur de perçages accélérés et de mécaniques Jacquard parisienne, Paradis-Poisonnière, 20.*

*Butt, Buisson-St-Louis, 16.*

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

*Cabouer, Saint - Pierre - Popin- court, 18.*

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

*Bouley (Ene), mécanicien, fab.*

*de machines à couper les peaux de*

*lapins, souffleuses de toutes di-*

*mensions marchant à bras et à la*

*vapeur, tours de chapellerie, ⑩*

*1849, Francs-Bourgeois-Marais,*

*3, ci-devant Beaubourg, 59.\**

*Bouhon, appareil ditcale à fleau,*

*breveté sans garantie du gou-*

*vernement, destiné au soulagement*

*des chevaux dans les montées,*

*⑩ 1849, ⑩ Société d'encoura-*

*gement 1850, pl. Dauphine, 7.\**

*Bourdier, bandagiste-herrnaria,*

*mécanicien-orthopédiste, fournis-*

*seur de la Société protestante de*

*prévoyance et de secours mutuels*

*de Paris, fabrique spéciale de tou-*

*tes espèces de ressorts et de ban-*

*dages, bêquilles, appareils ortho-*

*pédiques, ceintures hypogastriques,*

*bas lacés en tous genres, suspen-*

*ssoirs, objets d'habillement, et tous*

*articles en gomme élastique, Ca-*

*dran, 5.*

Comment

Tags

 Checked

SAVE ➤

# Phases d'ouverture des outils

Phase 0 : partage en **interne**, validation de l'utilité dans le cadre spécifique de SoDUCo, publication de prototypes isolés (recherche reproductible)

Phase 1 : tests avec **d'autres projets pilotes**, identification des éléments les plus utiles sans adaptation réelle (cas du projet AGODA au BnF Datalab)

Phase 2 : **adaptation à d'autres projets similaires** et choix d'une solution technique pour une évolution à long terme

Phase 3 : **ouverture complète** avec un soin particulier apporté à l'interopérabilité, la documentation, la maintenance, la sécurité, etc. en actant un positionnement et une démarche centrée sur quelques usages

# Perspectives — Phase 2 d'ouverture en cours

*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, intégration IIIF...)
- **Partager** plus largement ces outils (*open source*, documentation, voire formation)

# Liage en amélioration intensive

Développement de chaînes de traitement, pour permettre l'extraction de données pour des activités spécifiques (semaine “graphe géo-historique” en oct. 23)

Croisement massif des données qui semble possible à court/moyen terme

Master puis thèse de Solenn Tual

- Suivi des photographes
- Transposition au cadastre napoléonien

# Amélioration des outils d'annotation et de visualisation

React App

Directory Viewer Didot\_1852a.pdf 14 700 EXPORT SAVE

Affichage Entry

Liguez (P. PER), laminage de tous métaux pour bijouterie, orfèvres ACT , Chapon LOC , 18 CARDINAL ; fabrique de plaqué d'argent ACT , St-Martin LOC , 229 CARDINAL , anoin LOC 175 CARDINAL . —

**LAMINEURS.**

salz , Faub.-St-Martin, 62, Nomaindières, 12, de Cotte, 11, et Penthievre, 36. —

Hébert , en gros , Faub.-St-Denis, 162.

Leforey fils ains Traversière, 34.

Piot et Lefèvre, laiterie en gros , Amsterdam, 39. —

Poinson , nourrisseur-crémier , lait d'ancre et de chèvre à domicile , Chabrol, 32. —

Sarasin , Faub.-St-Martin, 270.

**LAMINEURS.**

Albaret ainé , fondeur et apprêteur de métaux , fab. de mauflechort , prépare pour MM. les orfèvres , lunettiers , couleliers , monteurs de boîtes et garnisseurs , moulures mar. et vif , filets d'ébénistes , entreprend la pièce de fonte , etc. place de l'Ancien-Marché-St-Martin , 7. —

Bachollet et Cuvillier , fab. de plaque or et argent , laminages à fagon de toute sorte de métaux , St-Maur-Popincourt, 134. —

Cailar ( J.-M. ), fondeur , laminage et treflerie de mauflechort pour orfèvres , chirurgie et coutellerie , laminage à fagon de toutes sortes de métaux , exécute

jaune et 1/2 rouge , Montgolfier 6. Marché-St-Martin. —

Liguez (P.) , laminage de tous métaux pour bijouterie, orfèvres , Chapon, 18 ; fabrique de plaqué d'argent , St-Martin LOC , 229 CARDINAL , anoin LOC 175 CARDINAL . —

Naudin (F.) , rue Montmorency , 14.

Oeschger (L.) , Mesdach et Cie , fonderie , affinage de cuivre , de zinc et de plomb , forçô hydrolique de 80 chevaux , quatre laminoirs martinet , à Beaché-St-Vaast , près Arras (Pas-de-Calais) , (P) 1849 ; maison à Paris , r. St-Paul , 28. —

Ponce et Prévost , Baully-St-Martin , 13.

Roussau , ancienne maison Clicquot , (P) 1849 , outils de bijoutiers , rouleaux , acier fondu , unis ou gravés , bâties , fil et plané en tous genres , laminage de métal , r. Beaubourg , 50. —

Wilken , gendre et successeur de Scovena , apprêteur , découpeur et estampeur pour MM. les bijoutiers , achète le vieux cuivre et l'aimaille , Gravilliers , 24 , et pass. de Rome , 36.

**LAMPES (fab. de) . Voyez aussi FERBLANTIERS-LAMPISTES.**

OCR NER

Edit inline

Annotations

Comment

Tags

Checked

SAVE >

soduco.geohistoricaldata.org

Favre\_et\_Duchesne\_1798

Annotations

ANOTTAUX, rue Greneta, n.° 46, —des Amis-de-la-Patrie.

Basset, rue Bailleul, n.º 238, —des Gardes-Françaises.

Bazarme, rue Madeleine, n.º 1429, —du Rôuer. Berçot, rue Guénégaud.

Bourgeois, rue Quincampoix, n.º 58, —des Lombards.

Boussode, rue Eloy, n.º 28, —de la Cité. Bouvrain, rue du Cimet-Je.

Buffard, rue des Moineaux, n.º 413, —de la Butte-des-Moul.

Collin, rue du Bouloy, n.º 21, —de la Halle-au-Bled. V.e Collin, rue Du

Chatelain, rue Antoine, n.º 41, —des Droits-de-l'Homme.

Damour, rue du Rocher, n.º 518, —du Roule.

Delaporte, faub. Laurent, n.º 163, —du Nord. Double, rue de Norma

Ducrot, rue Germain, n.º 87, —du Muséum.

Duherche, rue du faub. Denis, n.º 30, —Poissonnière.

Duprez, rue Beauregard, n.º 224, —de Bonne-Nouvelle.

Durand, Carré-Martin, n.º 9, —des Gravilliers. Durand, rue du Jour, n.

Godet, rue du Temple, n.º 119, —des Gravilliers.

Gouffier, place du Louvre, n.º 55, —du Louvre.

Hulin, rue d'Argenteuil, n.º 7, —de la Bataille-Montmartre.

Jacquier, rue du Temple, n.º 17, —des Gravilliers.

Lebon, rue de l'Assomption, n.º 10, —du Temple.

Marchal, rue de l'Assomption, n.º 10, —du Temple.

Marcou, rue des Vosges, n.º 1218, —de la Porte-de-Grenelle.

Marie, rue Neuve-Petit, n.º 3, —de l'Arrière.

Monceau, rue de l'Assomption, n.º 10, —du Temple.

Villain, rue Phaliusse, n.º 38, —des Gravilliers.

AUBERGISTES.

Auberge à n.º 1 rue Grégoire, n.º 46, —des Amis-de-la-Patrie.

Basset, rue Bailleul, n.º 238, —des Gardes-Françaises.

Bazarme, rue Madeleine, n.º 1429, —du Rôuer. Berçot, rue Guénégaud, n.º 101, —de l'Unité.

Bourgeois, rue Quincampoix, n.º 58, —des Lombards.

Boussode, rue Eloy, n.º 28, —de la Cité. Bouvrain, rue du Cimet-Je.

Buffard, rue des Moineaux, n.º 413, —de la Butte-des-Moul.

Collin, rue du Bouloy, n.º 21, —de la Halle-au-Bled. V.e Collin, rue Du

Chatelain, rue Antoine, n.º 41, —des Droits-de-l'Homme.

Damour, rue du Rocher, n.º 518, —du Roule.

Delaporte, faub. Laurent, n.º 163, —du Nord. Double, rue de Norma

Ducrot, rue Germain, n.º 87, —du Muséum.

Duherche, rue du faub. Denis, n.º 30, —Poissonnière.

Duprez, rue Beauregard, n.º 224, —de Bonne-Nouvelle.

Durand, Carré-Martin, n.º 9, —des Gravilliers. Durand, rue du Jour, n.

Godet, rue du Temple, n.º 119, —des Gravilliers.

Gouffier, place du Louvre, n.º 55, —du Louvre.

Hulin, rue d'Argenteuil, n.º 7, —de la Bataille-Montmartre.

Jacquier, rue du Temple, n.º 17, —des Gravilliers.

Lebon, rue de l'Assomption, n.º 10, —du Temple.

Marchal, rue de l'Assomption, n.º 10, —du Temple.

Marcou, rue des Vosges, n.º 1218, —de la Porte-de-Grenelle.

Marie, rue Neuve-Petit, n.º 3, —de l'Arrière.

Monceau, rue de l'Assomption, n.º 10, —du Temple.

Villain, rue Phaliusse, n.º 38, —des Gravilliers.

Annotations

Page 27 Page 30 Page 31 Page 32 Page 34

# Construire un outil pour les historiens

## Explorer



## Visualise



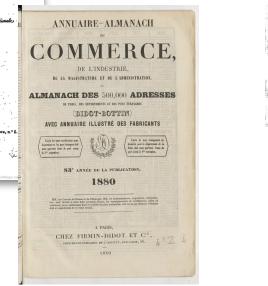
## Annoter



## Analyser



API III



# 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)

# Mezanno – plan quadriennal BnF

Outils de constitution de corpus, annotation, visualisation

Libres et ouverts, minimaux

Fonctionnalités visées : Autonomie des chercheurs SHS

- Pas d'installation requise, pas de serveur requis : une page à ouvrir
- Utilisation de ressources de calcul distantes, à la demande (OCR, NER...)
- Constituer son corpus
- Extraire des données de façon assistée
- Possibilités d'export

Démarrage courant 2024